Anticipating Personalization

Camila Farres*

November 12, 2026

Please click here for the most recent version

Abstract

Institutions often personalize using consumers' revealed preferences. This paper studies how consumers change their behavior when they anticipate personalization. I use a lab experiment that allows me to induce the consumer's information, disclose all consequences of sharing it for personalization, and exogenously manipulate whether personalization is incentive compatible. I find that personalization changes the way people share their information, but not in the way theory predicts; in particular, many share information accurately when they should not. Using a simple task, I can identify the misconceptions that drive suboptimal behavior. Understanding how people respond to anticipated personalization is important for designing it effectively and evaluating consumer welfare.

*I am especially thankful to my advisor, Charles Sprenger, for all his guidance and support. For helpful comments and suggestions, I thank Marina Agranov, Doug Bernheim, PJ Healy, Kirby Nielsen, Axel Niemeyer, Ted O'Donoghue, Tom Palfrey, Antonio Rangel, Jean-Laurent Rosenthal, Kota Saito, Lise Vesterlund, Leeat Yariv, all members of the experimental group at Caltech, and audiences of the Behavioral & Experimental Economics Student Conference at UCSB 2025, World ESA 2026, the ITAM Alumni Conference 2026, and ESA North America 2026. I thank the Linde Institute for Economic and Management Sciences for funding the pilot study and the main experiment. The design was reviewed and granted an exemption by the Institutional Review Board at the California Institute of Technology under protocol number IR25-1538. The study is pre-registered in the AEA RCT Registry under ID AEARCTR-0018307.

1 Introduction

People constantly share their personal information, and institutions increasingly use this information for personalization.[1] Personalization is an interaction where an institution first measures some (normally unobservable) information about the consumer, like preferences or beliefs, and then uses the information to personalize opportunities like prices, alternatives, or recommendations. For example, an online retailer may estimate consumers' preferences using behavior on its platform and then tailor the recommendations and prices it shows each consumer. This means personalization may generate tradeoffs for consumers because the incentives to provide accurate information may conflict with the incentives from future opportunities. In the online retailer example, the incentives to behave naturally conflict with those created by concerns about higher future prices. How consumers respond to this conflict has received little empirical attention despite its importance for consumer protection and policy design, so this paper studies how people change their behavior when they anticipate personalization.[2]

To study the effects of incentives created by personalization on behavior, I design an experiment that allows me to control three essential things. First, I induce consumers' information to measure precisely whether they share it accurately. Second, I disclose all consequences of personalization and test participants' understanding to ensure that they anticipate these consequences correctly. Third, I exogenously manipulate whether personalization creates conflicting incentives by contrasting a control condition in which incentive compatibility is preserved (sharing information accurately is optimal) with a treatment in which personalization compromises it.

I find that participants respond to personalization, but not in the way theory predicts. The control works as expected: when it is incentive compatible to do so, most people share information accurately. However, when personalization compromises incentive compatibility, they share information accurately less often, but still most of the time. Using a simple task at the end of the experiment, I find that sharing information when it is not optimal is related to specific misconceptions of the environment. This provides a new understanding of how people share their information in anticipation of personalization and has important implications for

[1] Acquisti et al. (2016) and Goldfarb and Tucker (2019) survey how firms use consumer data to tailor recommendations and prices.

[2] Personalization is incentive compatible if it is in the consumers' best interest to share their information accurately. Mechanism design has developed theories to design incentive compatible personalization, with the goal of recovering unobservable information accurately (Myerson, 1979). However, personalization is often not incentive compatible in practice (Weitzman, 1980; Pathak and Sönmez, 2008; Björkegren et al., 2020). Thus, in this paper, I study the role of personalization in how people share their information, abstracting away from other important determinants of behavior like privacy, lying aversion, or concerns and beliefs about the institution.

consumer welfare and the design of personalized policy.

Section 2 describes the experiment, which is built around twenty trials of four stages each. In the first stage, I give the participant a token with a number that represents how much money it is worth.[3] In the second stage, participants report their willingness to accept for the token; this is incentivized through a process equivalent to a second-price auction.[4] In the third stage, participants choose between two alternatives. In the last stage, participants get feedback about how much they earn from the trial. The trial earnings are the sum of the outcome from the second-price auction and the payoff from the alternative they choose. In the design, the token represents the unobservable information, the second-price auction is how I measure participants' willingness to accept for the token, and a personalization rule determines the alternatives participants are offered in the third stage.

The only treatment variation is the personalization rule. There is personalization if the rule depends on the willingness to accept that participants report in the second-price auction, because this means that the report in the second stage determines the alternatives offered in the third stage. The main variation across treatments is whether the incentives from the second-price auction conflict with those from personalization. The second-price auction is incentive compatible in isolation because reporting the accurate willingness to accept is a weakly dominant strategy. In this design, the accurate willingness to accept is the value of the token. However, when the report also determines the alternatives, personalization may introduce incentives to misreport[5] and compromise the incentive compatibility of the second-price auction.

I use seven treatments, and the main results focus on two of them, called IC Control and Misreport Up. In the IC Control, there is no personalization, so the second-price auction remains incentive compatible. In Misreport Up, the participants' reported willingness to accept is itself one of the alternatives that they are later offered. A higher report therefore means a better alternative, so the earnings from the third stage increase with the report. This introduces incentives to misreport up, which compromises the incentive compatibility of the second-price auction.

The five extension treatments test robustness on three dimensions of the personalization rule. First, Misreport Down has a personalization rule where alternatives decrease with the willingness to accept. Second, two treatments introduce personalization rules with uncertainty; with some probability, participants face the rule from Misreport Up and with the remaining probability, the one from Misreport Down. Third, two treatments preserve

[3]The token is an induced value (Smith, 1976).

[4]The Becker-DeGroot-Marschak (BDM) mechanism (Becker et al., 1964).

[5]Any reported willingness to accept that is not equal to the token is a misreport.

incentive compatibility but do have personalization, one of them with uncertainty.

After twenty trials, there is a quiz and a survey to assess why participants behave as they do during the trials. The first part of the quiz tests how a participant responds to the different incentives of the trial in isolation and how they integrate them.[6] The second part of the quiz measures factors that could explain why consumers may share information accurately when they should not: narrow framing, failure of backward induction, and lying aversion. The survey measures a fourth, the weight people place on the future consequences of their actions.[7]

Section 3 presents the results, and the main finding is that personalization matters, but participants do not respond to it optimally. The results come in four parts: the first three use the two main treatments, and the last shows the extensions. Personalization matters in the sense that responses to Misreport Up are significantly different from those in the IC Control. This implies that an institution that uses information to personalize in a way that compromises incentive compatibility may not recover information accurately on average. However, participants do not respond optimally to personalization, and they still share accurately most of the time, as if there were no personalization. In the IC Control, 73% of reports are accurate; in Misreport Up, accurate reports fall to 57%, and only 13% of reports are at the optimum.

The aggregate failure to report optimally comes mostly from individual differences, not from different stages of learning. In Misreport Up, 25% of participants report accurately in every trial, behaving as if there were no personalization, and only 3% report optimally throughout. The remaining participants do learn, but not enough: by the last trial, reports cover only 33% of the distance to the optimum, on average.

Thus, while institutions cannot recover information accurately on average, some consumers do share accurately, and the quiz can identify them. This matters because institutions could personalize effectively for the consumers they can identify. The answer to the quiz question about both sets of incentives together is correlated with the reports in the trials. In particular, participants who answer that question with the accurate report also share accurately in the trials, so the answer identifies most of the participants who share accurately in every trial.[8]

[6]Three questions: one about how to best respond to the personalization rule alone, another about how to do so for the second-price auction alone, and one about best responding to both together (asked in the reverse order, the combined question first).

[7]The Consideration of Future Consequences scale, a short psychological questionnaire in which people rate how much statements about weighing immediate against distant outcomes describe them (Strathman et al., 1994; Vilar et al., 2022).

[8]In Misreport Up, the correlation between the answer and how far the participant's reports are from the token in the trials is 0.59. The participants who answer with the token report the token in 71% of their trials, and 90% of the participants who report the token in every trial answer with the token.

The quiz also helps explain why some participants share accurately when it is not optimal. In Misreport Up, most participants answer the quiz question about the incentives of the second-price auction in isolation correctly, about half answer the one about the personalization rule in isolation correctly, and few answer the one about both together correctly.[9] I refer to the failures to best respond in these questions as misconceptions. Since behavior in the trials is heterogeneous, I look at how the answers relate to it. The main finding is that behavior is closely related to best responding to the incentives of personalization in isolation: the participants who report the token in every trial mostly get that question wrong, and those who report closer to the optimum mostly get it right.[10] This is suggestive evidence that the participants who report accurately ignore the incentives of personalization to the point that they cannot best respond to them even when those incentives are isolated.

Narrow framing, a failure of backward induction, and lying aversion are traits that could also explain why participants share accurately when it is not optimal, but I find that they do not. Participants who best respond to both sets of incentives but still share information accurately could be exhibiting narrow framing. Participants who cannot best respond to the personalization rule alone could be failing to do backward induction. And in general, someone who shares accurately could do so due to lying aversion.[11] In the second part of the quiz, I measure these three traits using some unrelated tasks to study whether these are general traits that can explain behavior.[12] In a regression of participant reports on the misconceptions and the measures of the traits, I find that the misconceptions account for a third of the variance while the traits explain almost none.

Results are mostly robust to the personalization rule. In the incentive compatible treatments that add personalization, I detect no response: reports are not distinguished from the IC Control's, consistent with accurate information still being recovered when personalization preserves incentive compatibility. In Misreport Down, I find the analogous result from Misreport Up: reports move in the direction personalization favors (now downward) but fall short of the optimum.

[9] 64%, 46%, and 16% of participants, respectively.

[10] Of the participants who report the token in every trial, 28% answer it correctly, against 94% of those who cover more than half of the distance to the optimum. This is not driven by misunderstanding the personalization rule itself. The rule is tested in a comprehension question right after it is explained, and 100% of the participants who report closer to the optimum answer it correctly on the first attempt, against 66% of those who report the token in every trial and 72% of the rest of the treatment. The last two are not distinguished ($p = 0.632$). So the participants who report closer to the optimum understand the rule better than average, and those who report accurately do not understand it worse.

[11] I tried to frame the environment so reporting accurately would not be understood as truthful. Section 2 discusses the neutral language to do this. And the quiz question about lying aversion gives some evidence that supports this.

[12] It could be narrow framing or a failure of backward induction in the specific environment that drives this, but I wanted to study whether it relates to these traits more generally.

The results from the uncertainty extensions are more nuanced: I detect no response under symmetric uncertainty and only suggestive evidence of one under asymmetric uncertainty. In both treatments, which rule applies, the one from Misreport Up or the one from Misreport Down, is drawn after participants report their willingness to accept for the token. Misreport with Symmetric Uncertainty is a coin flip between Misreport Up and Misreport Down. Misreport with Asymmetric Uncertainty favors Misreport Up over Misreport Down. Neither of the treatments is incentive compatible. However, in Misreport with Symmetric Uncertainty reports are not distinguished from the IC Control's. This suggests that some personalization rules that are not incentive compatible may still largely recover accurate information.[13] In Misreport with Asymmetric Uncertainty fewer reports stay at the token, which suggests that uncertainty does not by itself keep reports accurate.

Section 4 concludes that, when consumers anticipate personalization that compromises incentive compatibility, institutions cannot rely on consumers ignoring the future consequences of personalization to recover their information accurately. However, if institutions can measure consumers' abilities to best respond to the different components, they might be able to personalize effectively for some consumers.[14] Using a task like the quiz, institutions could potentially manipulate the personalization rule (or beliefs about it) to elicit accurate responses without aligning to consumers' incentives.

The results also matter for consumer welfare and the regulation of consumer information. Consumers lose surplus when they share their information accurately under personalization that compromises incentive compatibility. Most regulation requires institutions to disclose the personalization rule to consumers (Ben-Shahar and Schneider, 2014; Loewenstein et al., 2014). But I show in this stylized environment that telling consumers all the consequences of their reports and giving them experience is not enough for all consumers to protect themselves, which might carry over to the field.

I am not the first to use induced-value experiments to study how people behave when they face two sets of conflicting incentives: one to share accurately and one to misreport. However, this has mostly been studied in the context of games, where beliefs and concerns about the other player have often been interpreted as drivers of too much suboptimal accurate sharing.[15] I find this result holds even if those interpretations are shut down, which means

[13] I fail to detect a response, which does not show that there is none. Section 3.4 gives the details.

[14] The quiz predicts what participants do in the trials from their answers to a similar task with a known correct answer. Using such a task outside induced-value experiments is not trivial, because it needs the induced-value analogue of a decision. Bernheim et al. (2026) discuss the caveats of this method, referred to as pattern matching.

[15] The ratchet games (Chaudhuri, 1998; Cooper et al., 1999; Charness et al., 2011), the entry limit pricing games (Cooper et al., 1997a,b; Cooper and Kagel, 2008), the durable-goods monopoly games (Reynolds, 2000; Güth et al., 2004), and the purchase-history pricing games (Brokesova et al., 2014; Tolksdorf, 2022).

some of the results in the literature could be driven by misconceptions rather than by forming beliefs about the other players or some form of social preferences.

More broadly, the results speak to the literature on designing mechanisms that are incentive compatible in practice. On the one hand, induced-value experiments test how much accurate information is recovered when the design is incentive compatible (Kagel and Levin, 1993; Cason and Plott, 2014; Danz et al., 2022). Many find that these designs are not as effective as one would hope, since they recover less than half of the information. However, the way I present and explain the second-price auction recovers most of the information from the start: in the IC Control, 66.3% of participants share accurately in the first trial.[16] On the other hand, there are many experiments in the lab and in the field that are not incentive compatible (Merlo and Schotter, 1999; Toubia et al., 2013; Andreoni et al., 2023; Chapman et al., 2024). Some argue that participants may not understand that the design is not incentive compatible and therefore believe that researchers are still recovering accurate information. I find some evidence that suggests caution. Whenever participants believe there is a clear direction that favors misreporting, some will do it, perhaps enough to bias the conclusions.

2 Experimental design

2.1 Overview

At the core of the design are the trials, which every participant completes twenty times. A trial has four stages. The first three represent the three elements of personalization: the unobservable information, how the institution measures it, and how it personalizes with it. In the fourth, participants get feedback. A trial begins with the *Induced Value*, in which participants receive a token with a known value θ (Smith, 1976); the token stands for the consumer's information, and a report equal to it is the accurate report.[17] In the *Measurement Stage*, they state a willingness to accept for the token, which I call the report and write $\hat{\theta}$. This stage stands for the earlier interactions in which the retailer measures what the consumer is willing to pay. To measure the willingness to accept, I use a mechanism equivalent to a second-price auction. In the *Personalized Menu*, they choose between two monetary alternatives that may be connected to the report, and earn the one they choose. The menu stands for the recommendations or the prices the retailer shows the consumer once it has

[16] Cason and Plott (2014) also use the BDM mechanism with an induced value. In their data, 16.7% of offers are at the value in the first round and 31.1% in the second.

[17] I call this report accurate and not truthful because the design removes the lying context. A quiz item measures lying aversion directly (Section 2.5), which lets me test whether it explains reports at the token.

measured their preferences. A trial ends with the *Interactive Feedback*, in which participants see the outcome of each stage and enter their earnings from the Measurement Stage and from the Personalized Menu, whose sum is their total trial earnings. The shaded box in Figure 1 summarizes the four stages.

The treatment variation is a *personalization rule* T, which maps the induced value θ and the report $\hat{\theta}$ into the two alternatives of the Personalized Menu, $T(\theta, \hat{\theta}) = \{x_1, x_2\}$. Each treatment has its own rule. Nothing else differs across treatments, so comparing reports across treatments isolates the response to the incentives of the Personalized Menu. The design is between subjects: each participant is randomly assigned to one rule for all trials. The personalization rule is explained in the instructions, tested in a comprehension question and in the practice trial, and restated in every trial on the Personalized Menu and Interactive Feedback screens.

Around the trials, participants read instructions, complete a practice trial, and answer a quiz and a survey, as the top row of Figure 1 shows. They first complete a bot check and give consent, then read the instructions one stage at a time (Appendix I shows every page). After each stage's explanation, they answer a short multiple-choice comprehension question that reinforces the main point of the page. They then complete a practice trial with three incentivized practice questions on the elements of a trial. Questions answered incorrectly are retried until correct, and once all are correct, participants move on to the twenty incentivized trials. After the trials, participants answer an incentivized quiz in two parts. The first part asks whether they can best respond to the incentives of the trials, and the second measures three general traits that could explain accurate sharing where accuracy is not optimal. An exit survey, which is not incentivized, measures a fourth.

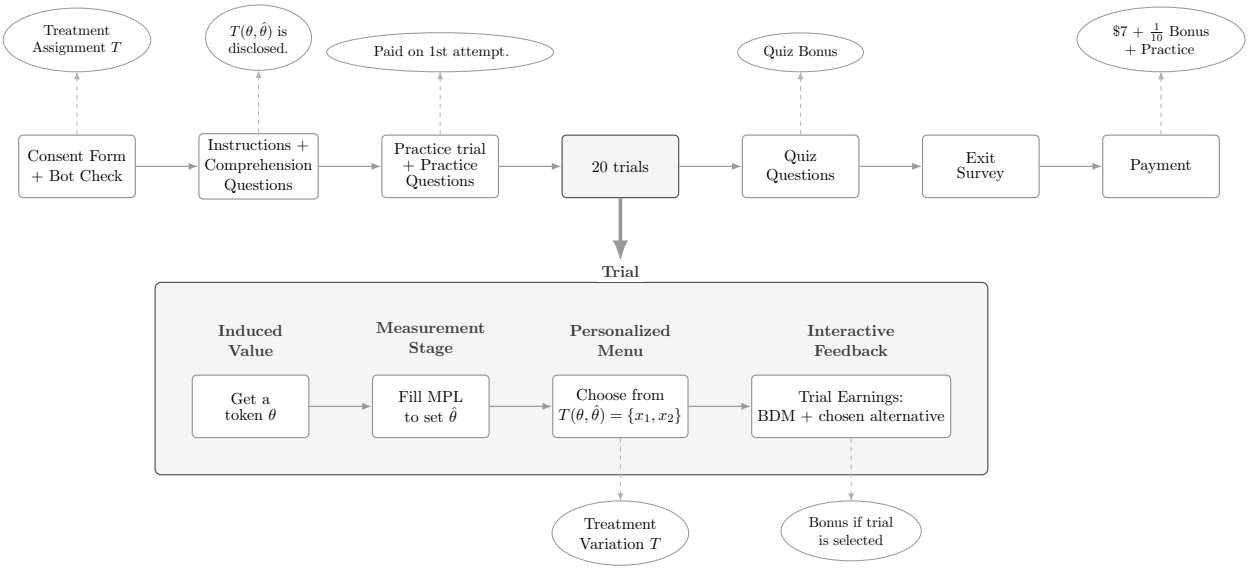
The experiment ran in a single session on Prolific[18] on April 23, 2026. Participants were residents of the United States or Western Europe, fluent in English, with at least a high-school education, a high approval rating, and at least 50 previous submissions, and none had taken part in the pilot study reported in the pre-analysis plan. In total, 702 participants completed the study, each assigned uniformly at random to one of seven treatments, giving roughly 100 participants and 2,000 trial-level observations per treatment. The median completion time was 40.9 minutes.

Payments have three parts. All participants who complete the study receive a \$7 participation fee. The practice questions pay \$0.50 each on the first attempt only, so \$1.50 is the most a participant can earn from them. The average across all treatments was \$1.25, and 63% of participants answered all three correctly the first time.[19] Each participant is selected

[18] Celebi et al. (2026) document high data quality on the platform.

[19] 26% answered two of the three correctly on the first attempt, 9% one, and 2% (13 participants) none.

Figure 1: Experimental sequence.



Notes: The top row shows the overall study timeline. The shaded box expands the four stages within each trial. Dashed callouts indicate treatment assignment, information disclosure, and bonus components. $T(\theta, \hat{\theta}) = \{x_1, x_2\}$ denotes the two alternatives generated by the personalization rule; the Personalized Menu pays the one chosen.

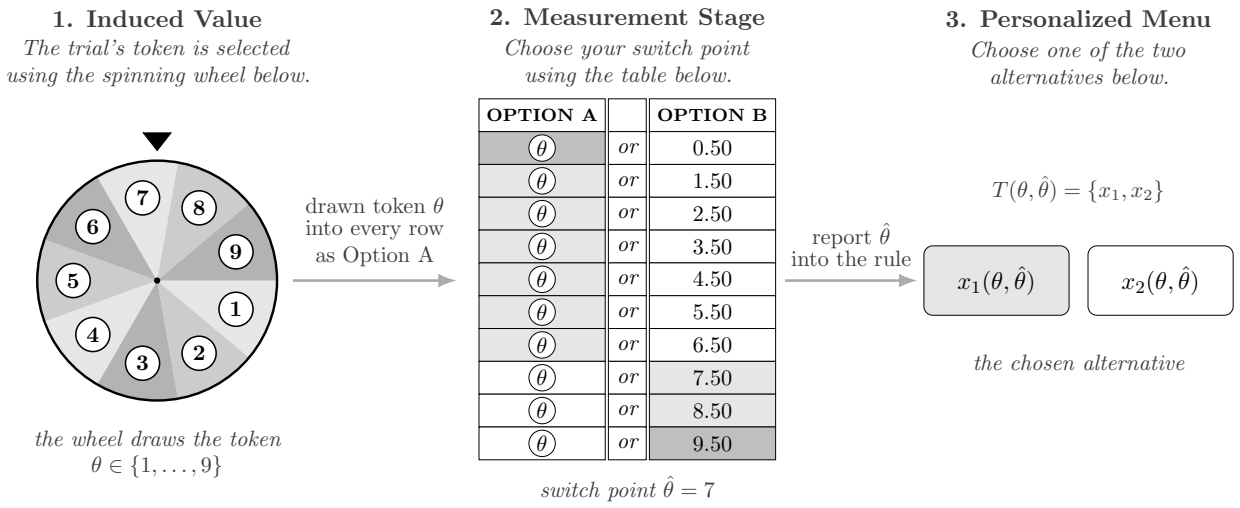
for a bonus payment independently with probability one in ten: for each of those selected, one item is drawn uniformly at random from the pooled set of the twenty trials and the quiz questions, and the bonus is the earnings from that item. Those selected received a bonus of \$12.03 on average.[20]

The design, the sample size, and the primary analysis were pre-registered. The plan sets about 100 participants per treatment from simulation-based power calculations on a pilot. Its main hypothesis is that each treatment changes reports relative to the IC Control. The test for each treatment is the joint test that its effect is zero at every token, in a regression that pools the seven treatments. Table 1 in Section 3.4 reports that regression for every treatment. Appendix C describes the plan and where the paper departs from it.

The rest of this section describes the stages of a trial (Section 2.2), the treatments (Section 2.3), the prediction for each (Section 2.4), and the quiz and the exit survey (Section 2.5).

[20]Appendix B reports demographics, balance, dropout, and comprehension diagnostics per treatment. It also shows where the dropouts stopped relative to the page that discloses the rule, and the main comparison with demographic controls.

Figure 2: Trial screens: the token and the two choices.



Notes: The first three screens a participant sees in each trial, in order. *Left:* the spinning wheel assigns the token value $\theta \in \{1, 2, \ldots, 9\}$, each value equally likely. *Middle:* the multiple price list, in which Option A is always the token drawn by the wheel and Option B rises by \$1 per row; gray cells illustrate an example response, the darker gray in the first and last rows marks the pre-selected endpoints, and the implied switch point is displayed below the table as participants make their selections. *Right:* the two alternatives generated by the personalization rule $T(\theta, \hat{\theta})$; participants choose one, and the chosen alternative is highlighted. The interface also displays the personalization rule explicitly, as a reminder to participants, rather than only the alternatives, so participants see how each alternative is constructed from the token and their own report.

2.2 Trials

Each trial begins with the Induced Value: a spinning wheel like the one in the left panel of Figure 2 draws the token value $\theta \in \{1, \ldots, 9\}$ uniformly at random.[21]

The Measurement Stage is a multiple price list (MPL), shown in the middle panel of Figure 2.[22] Each row is a binary choice between Option A, which is always the token θ, and Option B, an amount b that rises by \$1 per row, with

$$b \in \{0.5, 1.5, \ldots, 9.5\}.$$

The interface avoids any mention of willingness to accept and calls the report a *switch point*. Participants click Option A or Option B in any row. The interface then selects Option A in every row above and Option B in every row below, and participants can revise the choice. Option A is pre-selected in the first row and Option B in the last, so every list has a single

[21] The token is redrawn in every trial so that each participant is observed across the whole range of values.

[22] Cason and Plott (2014) measure an induced value with a direct BDM. I use a price list instead because price lists recover induced values better (Vossler and McKee, 2006; Murphy et al., 2010).

switch point. The report is its midpoint,

$$\hat{\theta} = \frac{b_A + b_B}{2}$$

where b_A is the largest Option B amount turned down and b_B the smallest accepted; the interface displays it below the table as participants make their selections. Reports therefore lie on the same grid as the token, $\hat{\theta} \in \{1, \ldots, 9\}$.

The earnings from the Measurement Stage are the outcome of the BDM mechanism (Becker et al., 1964): one row is drawn uniformly at random, and the participant receives their choice in it, the token if Option A is selected there and that row's b if Option B is. Reporting a switch point equal to the token chooses the larger amount in every row, so it is weakly dominant.[23] These are the incentives for accuracy.

The Personalized Menu offers the two alternatives $\{x_1, x_2\}$ that the treatment's personalization rule $T(\theta, \hat{\theta})$ produces, as in the right panel of Figure 2. Both are monetary amounts shown on the screen, so choosing the larger is a dominant strategy, and the report enters this stage only through the alternatives it produces. A trial therefore has two sets of incentives: the incentives for accuracy of the Measurement Stage and the incentives of the Personalized Menu.

The Interactive Feedback closes the trial. The screen shows the price list with the participant's selections and the drawn row highlighted, and the two alternatives with the chosen one marked. Participants then enter their earnings from each stage themselves, the interface displays the implied total, and both entries must be correct to proceed. Total trial earnings are the sum of the two stages: the choice in the randomly drawn row of the Measurement Stage plus the alternative chosen in the Personalized Menu.

2.3 Treatments

I compare the report $\hat{\theta}$ with two benchmarks, the token and the optimal report. The first is the *token* itself, the accurate report $\hat{\theta} = \theta$, which is optimal whenever the Personalized Menu is ignored, since the Measurement Stage incentivizes accuracy. The second is the *optimal* report $\hat{\theta}^*$, which maximizes expected total trial earnings and therefore takes both stages into account.

Each treatment is a different personalization rule, and the rule determines whether the Measurement Stage is *incentive compatible*. It is incentive compatible when the two

[23] The BDM is equivalent to a second-price auction: the switch point is the participant's bid and the Option B amount of the randomly drawn row is the other bid. When the other bid is higher than the switch point, the participant sells the token for the other bid, and when it is lower, they keep the token. As in a second-price auction, bidding one's value is weakly dominant.

benchmarks coincide, $\hat{\theta}^*(\theta) = \theta$ at every token, so that sharing accurately remains optimal once the Personalized Menu is taken into account. It is not when the optimum is a misreport at some token. The Personalized Menu can compromise the incentive compatibility of the Measurement Stage if it creates a *conflict* with its incentives for accuracy, that is, if it pays more for some misreport than for an accurate report. A conflict need not break incentive compatibility, since the Measurement Stage's cost of a misreport can outweigh the menu's gain, but under every rule in this experiment that creates a conflict it does for some values of the token.

The reports show whether participants share accurately under a given rule. In every treatment, any report other than the token is a *misreport*. When the Measurement Stage is incentive compatible, every misreport is suboptimal. When it is not, the optimum is itself a misreport at some tokens.

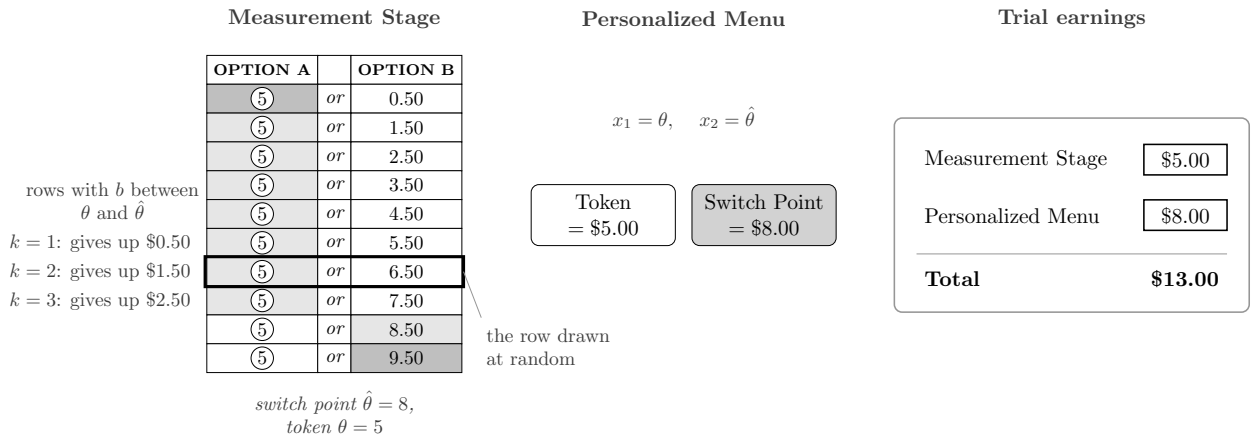
The personalization rules of the two main treatments differ only in the second alternative. Both set $x_1 = \theta$, which represents the consumer's outside option. The second alternative, x_2, is the institution's offer. The *IC Control* sets $x_2 = 5$, a constant at the middle of the token range, which keeps the control's stages and screens the same as in the other treatments. The control is precisely the case without personalization: nothing the participant reports affects the Personalized Menu, so the Measurement Stage is the only place the report matters and the token is the optimal report. There is no conflict, and the Measurement Stage remains incentive compatible. The main treatment, *Misreport Up*, sets $x_2 = \hat{\theta}$. The report is now one of the alternatives: a participant who reports above the token can take the report in place of the token. The menu therefore favors overreporting and creates a conflict with the Measurement Stage, which is no longer incentive compatible. Section 2.4 works out the predictions.

The other five treatments are extensions of the two main ones. Each changes one feature of the menu, and I introduce the details with their results in Section 3.4. I use *the control* for the IC Control, *the two main treatments* for the IC Control and Misreport Up together, and *extension treatments* for the other five. The treatment names say whether the Measurement Stage is incentive compatible: the name has *IC* when it is and *Misreport* when it is not, since the optimum is then a misreport for some tokens.

2.4 Predictions

A misreport has a cost in the Measurement Stage. Figure 3 shows the cost in an example from Misreport Up, in which the token is $\theta = 5$ and the participant reports $\hat{\theta} = 8$. By reporting 8, the participant selects the larger amount in every row except the three rows where Option B

Figure 3: The cost and the gain of a misreport in Misreport Up, shown on the Interactive Feedback screen.



Notes: The figure uses the Interactive Feedback screen of one trial to show the two sides of a misreport under the Misreport Up rule: the token is $\theta = 5$ and the participant reported $\hat{\theta} = 8$. *Left, the cost:* the report turns down Option B on the $|d| = 3$ rows whose amount b lies between the token and the report, labeled $k = 1, 2, 3$; these are the rows that equation (1) sums over, and if row k is drawn the report gives up $k - \frac{1}{2}$ dollars. Here the row drawn at random for the BDM, outlined in bold, is one of them, so the Measurement Stage pays the token, \$5.00, rather than the \$6.50 that row's Option B would have paid. *Middle, the gain:* the report becomes the second alternative, $x_2 = \hat{\theta} = 8$, against $x_1 = \theta = 5$; the chosen one is shaded and the Personalized Menu pays \$8.00, which is \$3.00 more than the token would have paid, and this gain is certain. *Right:* participants type the earnings from each stage into the boxed fields themselves, and the interface displays the implied total as they do so. The largest possible cost, \$2.50 on the row $k = 3$, falls short of the \$3.00 gain.

is \$5.50, \$6.50, and \$7.50. If one of those three rows is drawn, the participant gives up \$0.50, \$1.50, or \$2.50, respectively. In the figure, the row with \$6.50 is drawn, so the participant receives the \$5.00 of the token instead of \$6.50, and the report costs \$1.50. Each row is drawn with probability $\frac{1}{10}$, so the expected cost of this report is $(0.50 + 1.50 + 2.50)/10 = \$0.45$.

The cost of any misreport depends only on its distance from the token. Write $d = \hat{\theta} - \theta$ for the deviation of a report from the token. A report $|d|$ positions away from the token, above or below it, gives up $k - \frac{1}{2}$ on the k-th row from the token, so in expectation it costs

$$\frac{1}{10} \sum_{k=1}^{|d|} \left(k - \tfrac{1}{2}\right) \;=\; \frac{d^2}{20}. \tag{1}$$

Every prediction compares this cost with what the Personalized Menu pays for the same report.

A misreport can have a gain in the Personalized Menu when the personalization rule depends on the report. In the IC Control there is no personalization, so a misreport has a

cost and no gain, and reporting the token is optimal. In Misreport Up the second alternative is the report, so a report above the token creates a better alternative. In the example, the participant reports 8, chooses \$8.00 over \$5.00 in the Personalized Menu, and gains \$3.00 with certainty. In general, a report d positions above the token gains d dollars. A report below the token gains nothing, because the token is then the larger alternative.

In Misreport Up, any report above the token dominates reporting accurately: it earns more whichever row is drawn. In the example, the gain is \$3.00 and the largest possible loss in the Measurement Stage is \$2.50, on the row with \$7.50. The same holds for every report above the token, because the largest possible loss, $d - \frac{1}{2}$, is always half a dollar smaller than the gain, d. This does not depend on the participant's risk preferences.

How far to misreport upward requires weighing the cost against the gain, and in expectation the optimum is the highest report, 9. A report d positions above the token gains d dollars and costs $d^2/20$ in expectation, so it is worth $d - d^2/20$ more than reporting the token. This amount increases with d over the at most eight positions available, so $\hat{\theta}^*(\theta) = 9$ at every token.[24]

Panel (a) of Figure 4 illustrates the calculation of the optimum at one token, $\theta = 5$: it shows the expected earnings of each possible report, by stage and by treatment, and where each is maximized. The Measurement Stage peaks at the token, and the Personalized Menu in Misreport Up rises one dollar per position above the token, so their sum is highest at 9. Panel (b) summarizes the result for every token: the optimal report is the token in the IC Control and 9 in Misreport Up.

The two treatments make the same prediction at $\theta = 9$, where the optimum is the token, and differ everywhere else. I call *conflict tokens* the ones where accuracy is not optimal; in Misreport Up they are $\theta \in \{1, \ldots, 8\}$, and the IC Control has none.[25]

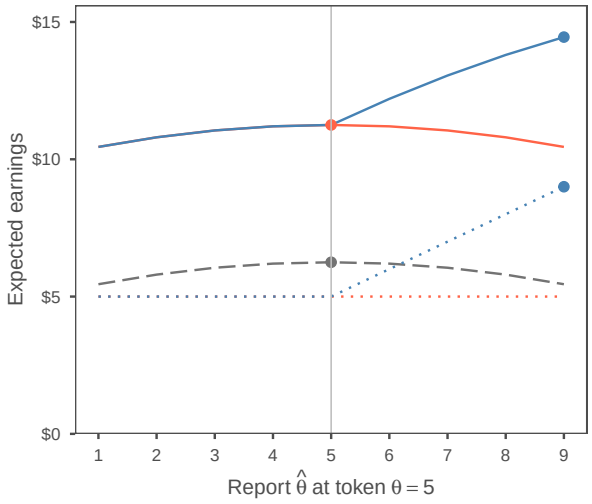
The conflict tokens fix a rule for the rest of the paper: whenever a figure or a test pools across tokens, I restrict to the conflict tokens shared by the treatments in the comparison that are not incentive compatible. For the main comparison that is $\theta \in \{1, \ldots, 8\}$.[26]

[24] This assumes risk neutrality, but the optimum does not depend on it: it is 9 for every risk-averse expected-utility maximizer of total trial earnings and under standard probability weighting, and only strong risk seeking changes it, to 8 at the lowest tokens. Appendix A shows the calculation.

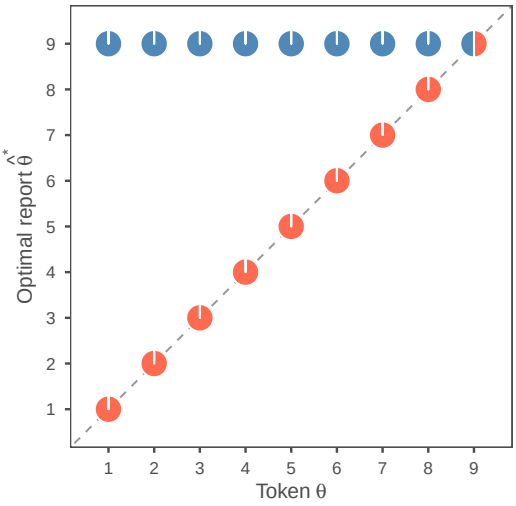
[25] The design itself uses all nine tokens rather than the conflict tokens alone, because I wanted the available reports to coincide with the possible token values. On a common grid the highest token is necessarily one whose optimal report is the token.

[26] Dropping $\theta = 9$ from the IC Control as well keeps the two treatments on the same range of θ. The five extension treatments of Section 3.4 have conflict tokens on different ranges, so when all seven treatments are compared at once the same rule gives the intersection of those ranges, $\theta \in \{2, \ldots, 7\}$. Appendix A shows which treatment contributes which bound. Figures that report by token rather than pooling over it say in their notes which tokens they show.

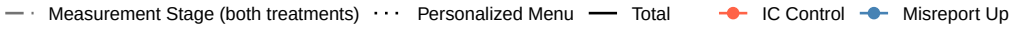
Figure 4: The earnings of each stage at one token, and the optimal report at every token.



(a) Expected earnings by report at $\theta = 5$, by stage



(b) Optimal report, by token



Notes: Panel (a): the calculation at one token, $\theta = 5$, with the report on the horizontal axis and each component of expected earnings drawn on its own, expectations taken over the price-list row that will be drawn. The Measurement Stage is common to both treatments and peaks at the token. The Personalized Menu is flat at 5 in the IC Control, where the second alternative is 5 whatever is reported, and in Misreport Up it is $\max(\theta, \hat{\theta})$, which coincides with the control's at reports below the token, where the token is the better alternative, and rises one dollar per position above it. The total is the sum. A filled point marks a curve's unique maximizer; to the left of the token the two treatments coincide and the blue lines cover the red; the control's Personalized Menu has no unique maximizer, since it is flat. Panel (b): the result of the same calculation at every token, with the token on the horizontal axis and the report maximizing expected total trial earnings on the vertical axis. The dashed line is $\hat{\theta} = \theta$. In the IC Control, the optimum lies on that line at every token; in Misreport Up it is 9 at every token. Appendix A shows the calculation for every treatment.

2.5 The quiz and the exit survey

The quiz that follows the trials has two incentivized parts. The first part turns the incentives of the trials into three questions. Each pays \$10 if correct and has the format of a trial: a token worth $\theta = 5$,[27] the price list of the Measurement Stage, a restatement of the participant's own personalization rule, and a request for the switch point that maximizes one component of trial earnings. The *measurement question* asks for the report that maximizes Measurement Stage earnings alone, the *menu question* for the report that maximizes Personalized Menu earnings alone, and the *combined question* for the report that maximizes total trial earnings.[28] A button below the list, "all switch points give the same earnings", is the correct answer

[27] A token of 5 has the same number of possible reports above and below it.

[28] The combined question is asked first, before the two component questions, so that answering the components does not lead participants to the total step by step.

when the component does not depend on the report. An answer is therefore one of the nine switch points or the button, and random guessing is right one time in ten.[29]

Panel (a) of Figure 4 shows what each question asks for at that token: the three curves are the Measurement Stage earnings, the Personalized Menu earnings, and the total, and the filled points, where each curve is maximized, are the correct answers. In Misreport Up, the correct answer is 5 to the measurement question and 9 to the menu and combined questions. In the IC Control, it is 5 to the measurement and combined questions and the button to the menu question, since the control's Personalized Menu curve is flat and has no unique maximizer, which is what "all switch points give the same earnings" refers to. The component questions ask whether a participant can best respond to each set of incentives on its own. The combined question asks whether they can put the two together, which is the step that determines the optimum when the menu creates a conflict.

Together the three answers define six *misconception groups*, to which I refer throughout.[30] A participant who answers the combined question correctly *integrates both stages*, whatever they answered on the components. Among those who do not, a participant who answers only the measurement question correctly *neglects the Personalized Menu*, and one who answers only the menu question correctly *neglects the Measurement Stage*. A participant who answers both component questions correctly but fails the combined one *understands both* stages and cannot combine them. This group divides according to the wrong combined answer: a misreport (they *misreport partially*), or the token or the button (they *report accurately*). The remainder *understands neither*.

The second part measures three general traits that could explain why a participant reports the token where it is not optimal, each with one incentivized item. The first is narrow framing, evaluating a decision in isolation from the others it is tied to (Tversky and Kahneman, 1981; Read et al., 1999; Barberis et al., 2006; Rabin and Weizsäcker, 2009; Ellis and Freeman, 2024): a participant who frames the two stages separately would answer the Measurement Stage on its own terms. I measure narrow framing with two payoff-equivalent choices between two coins, one with the endowment kept separate and one with it folded into the outcomes. The second is a failure of backward induction (McKelvey and Palfrey, 1992; Binmore et al., 2002; Johnson et al., 2002; Levitt et al., 2011): a participant who does not reason backward from the Personalized Menu would not carry its incentives into the report. I measure the failure of backward induction with a short two-move game against a computer that always takes its best branch. The third is lying aversion (Gneezy, 2005; Fischbacher and Föllmi-Heusi, 2013;

[29] Appendix F describes the three questions, and Appendix I shows them as participants saw them.

[30] The misconception groups are mutually exclusive and exhaustive: every participant answered the three questions, so each belongs to exactly one group. Appendix F gives the counts.

Gneezy et al., 2013, 2018; Abeler et al., 2019): a participant averse to misstating a value is more likely to report the token if they read a switch point other than the token as a lie. The design abstracts from the lying context, and the item tests whether participants bring it in nonetheless. I measure lying aversion with an item that pays participants the number they report and asks for the number a wheel drew, which may also pick up demand effects.[31]

The exit survey measures a fourth trait. It is not incentivized. It uses the ultra-short Consideration of Future Consequences scale of Vilar et al. (2022), a short form of the scale of Strathman et al. (1994): six items on a seven-point scale that measure how much weight a person puts on the later consequences of what they do now, split into a CFC-Future and a CFC-Immediate subscale. It closes with an open box for comments.

3 Results

This section has four parts. Section 3.1 asks whether the Personalized Menu changes what participants report at the Measurement Stage on average, and whether they best respond to it. Section 3.2 asks whether the average response reflects reports that lie between the token and the optimum or a mixture of accurate and optimal reports, and whether the two come from different participants. Section 3.3 asks whether the participants who report accurately can be identified using a task outside the trials, and why participants report accurately, optimally, or in between, including whether general traits account for it. Section 3.4 asks whether the results are robust to the personalization rule.

The first three sections use the two treatments of the main comparison, the IC Control with 98 participants and Misreport Up with 116, and the behavior of interest is the Measurement Stage report.[32]

3.1 Reports move in the direction personalization favors and stop short of the optimum

In Misreport Up, the distance between the token and the optimum is different at each token, so I need a common scale to compare reports across tokens. I measure each report by its deviation from the token, $d = \hat{\theta} - \theta$, as a share of the distance available in that direction.

[31] Appendix F describes the three items, their scoring, and their relation to the reports in both main treatments.

[32] The Personalized Menu is a choice between two known amounts, so the only behavior it can reveal is whether a participant takes the strictly worse one. Dominance violations are 5.7% in the IC Control and 4.7% in Misreport Up. Participants must enter both stages' earnings correctly on the feedback screen before continuing, and mostly do so at once: over all seven treatments the average Interactive Feedback takes 1.24 submissions, and 90% of trials are entered correctly on the first attempt.

The available distance is $9-\theta$ above the token and $\theta-1$ below it, so

$$\tilde{d} \;=\; \begin{cases} d/(9-\theta) & \text{if } d>0, \\ 0 & \text{if } d=0, \\ d/(\theta-1) & \text{if } d<0. \end{cases} \tag{2}$$

I report $\tilde{d}$ as a percentage: it is 100% when the report is 9 and -100% when the report is 1.[33] On this scale, the predictions are the same at every conflict token. In the IC Control the optimal report is the token, so the prediction is $\tilde{d}=0\%$. In Misreport Up the optimal report is 9, so the prediction is $\tilde{d}=100\%$. The variable tells me four things: whether a participant reports the token, in which direction they move away from it, how far they move, and whether they reach the optimum.

Panel (a) of Figure 5 shows the average report at each token in each treatment, together with the predictions: the dotted 45° line is the token, the prediction for the IC Control, and the stars are the prediction for Misreport Up at its conflict tokens.[34] Panel (b) shows the average $\tilde{d}$ at the eight conflict tokens.[35] Both panels show three things: reports in the IC Control are close to the token; reports in Misreport Up are above the control's at every conflict token, in the direction personalization favors; and they are far below the prediction of 9. I test the three observations with the same regression setup. All tests use every decision on the conflict tokens, with token and trial fixed effects and standard errors clustered on the participant.[36]

The first observation is that the control works as expected. The prediction for the IC Control is $\tilde{d}=0\%$ at every token, so I test jointly that its eight token means are zero. Reports in the IC Control cover 7% of the available distance. This is close to the token, but not exactly at it: the joint test rejects ($F(8,213)=2.33$, $p=0.020$).[37] This is a small amount of noise

[33] The exceptions are the two ends of the grid. A report of 9 at $\theta=9$, or of 1 at $\theta=1$, is the token itself, so $\tilde{d}=0\%$ there. The variable is always defined because a denominator is never zero: a report above the token requires $\theta<9$, and a report below the token requires $\theta>1$.

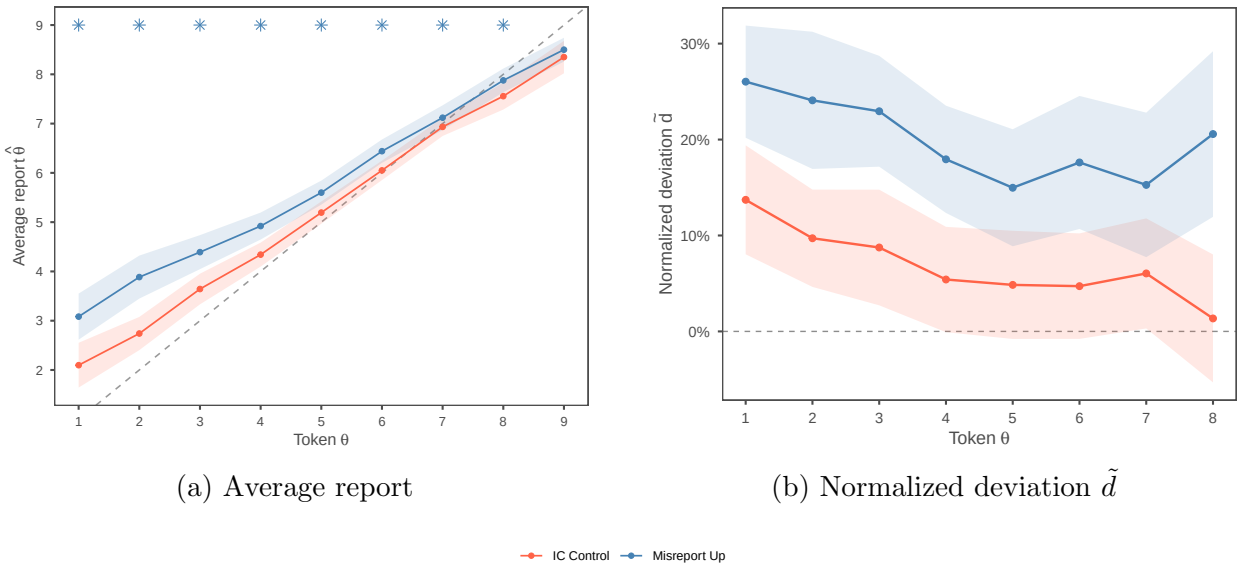
[34] The confidence intervals use the participant means within each token cell as the unit of observation. The tests weight decisions rather than participants, with standard errors clustered on the participant, as pre-registered.

[35] At $\theta=9$ the optimum and the token are the same report, so there is no conflict, and the only deviation available is downward. The panel would show a drop there that is unrelated to the treatment. For the same reason, I drop $\theta=9$ from the rest of the analysis of the two main treatments, following the rule of Section 2.4.

[36] Joint tests report the conventional cluster-robust F, with the number of participants minus one as denominator degrees of freedom. Appendix D repeats them with the CR2 small-sample correction, and the conclusions do not change. I report the pre-registered test of the main comparison as it is.

[37] The noise comes mostly from one group. The 37% of control participants who fail the backward-induction item of the quiz, a task unrelated to the trials, report the token in 48% of their trials, and the rest report it in 87% (Appendix F). Section 3.3 discusses this.

Figure 5: Average reports by token.



(a) Average report

(b) Normalized deviation $\tilde{d}$

Notes: Panel (a) shows the average report $\hat{\theta}$ at each of the nine tokens, in each of the two treatments, over all trials and participants. The dotted 45° line is the token, which is the optimal report in the IC Control, and the stars mark the optimal report in Misreport Up, $\hat{\theta}^*(\theta) = 9$, on its conflict tokens. Panel (b) shows the average normalized deviation $\tilde{d}$ of equation (2) at each of the eight conflict tokens: the optimum is $\tilde{d} = 100\%$ at each of them in Misreport Up, and $\tilde{d} = 0\%$ in the IC Control. In both panels every point is a mean of participant means within the token cell rather than a mean over decisions, since a participant contributes a random number of decisions to each cell, and bands are the 95% confidence intervals that go with it.

compared to what other experiments find when they test incentive compatible mechanisms with induced values.[38]

The second observation is that Misreport Up moves reports in the direction personalization favors. This observation compares the two treatments, so I test it with the coefficient on an indicator for Misreport Up. Reports in Misreport Up cover 21% of the available distance, which is 13 percentage points more than the control ($p < 0.001$). The effect is not driven by particular tokens, and it does not depend on the unit of observation or on which treatments are in the regression.[39]

The third observation is that reports in Misreport Up stop far short of the optimum. The prediction is $\tilde{d} = 100\%$ at every conflict token, so I test jointly that its eight token means are

[38] Over all nine tokens, 74.3% of the control's reports are at the token. A written BDM recovers the value in 16.7% to 31.1% of offers (Cason and Plott, 2014), a sealed-bid second-price auction in about 30% of bids (Kagel and Levin, 1993), and the binarized scoring rule recovers the prior in 58.5% of reports when participants are told its incentives (Danz et al., 2022).

[39] When I let the effect differ by token, the joint test that the treatment indicator and its seven interactions are all zero rejects ($p = 0.007$). With one observation per participant, the participant's mean $\tilde{d}$, the estimate is the same, 14 percentage points ($p < 0.001$). The pre-registered specification, which pools all seven treatments on their common tokens, gives 13 ($p < 0.001$). Table D1 in Appendix D reports the four specifications side by side.

100%. The joint test rejects ($F(8, 213) = 141.81$, $p < 0.001$): at 21%, reports are closer to the token than to the optimum. Participants respond to personalization, but they do not best respond to it.

These three results imply that the Measurement Stage recovers accurate information when it is incentive compatible, and that, on average, an institution cannot recover consumers' information when personalization compromises incentive compatibility. Participants in the IC Control best respond to their incentives, up to a small amount of noise. When personalization adds incentives that compromise the incentive compatibility of the Measurement Stage, responses change, and they move in the direction personalization favors. However, they are far from optimal. Moreover, even though consumers know the personalization rule, the institution cannot invert it to recover the accurate information from the responses, because the inversion assumes that consumers respond optimally, and they do not.[40]

3.2 The average is made of participants who respond differently

An average of 21% is consistent with many patterns of reports. At one extreme, every report moves about 21% of the way to 9. At the other, most reports are at the token and a few are at 9, so that each report is either accurate or optimal. The data could lie anywhere between the two. Figure 5 cannot say where, so this section looks at the distribution of decisions and then of participants.

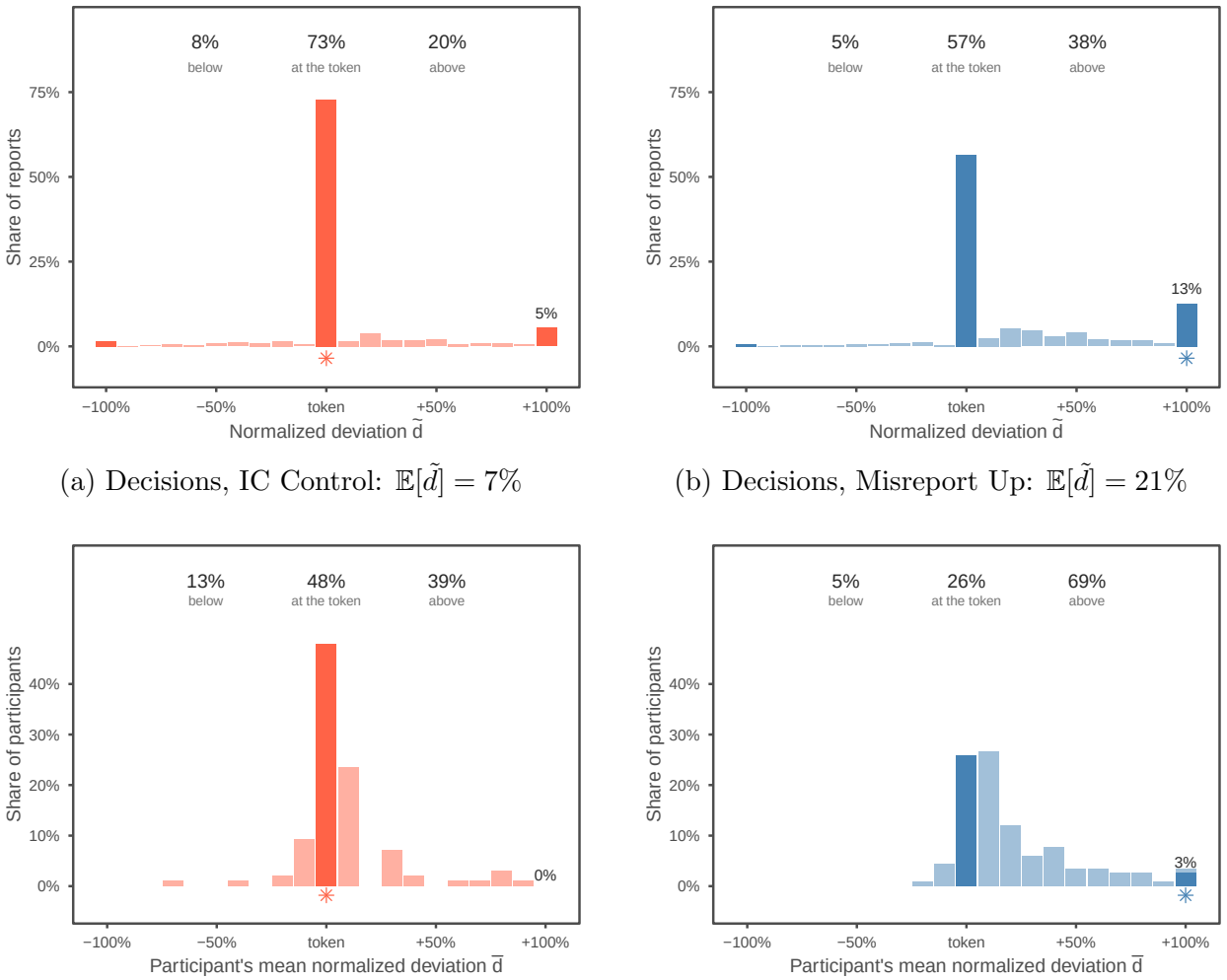
The top row of Figure 6 pools every decision and shows the distribution of $\tilde{d}$ in each treatment. It shows four extensive margins. Three are the shares of decisions at the token ($\tilde{d} = 0$), below it ($\tilde{d} < 0$), and above it ($\tilde{d} > 0$), which add up to one. The fourth is the share that reaches 9, which is part of the share above the token and is marked above its own bar at +100%.

Three of the four extensive margins move, and all in the same direction. Misreport Up has fewer decisions at the token than the IC Control, 57% against 73% ($p = 0.002$), and more above it, 38% against 20% ($p < 0.001$), of which more are at 9, 13% against 5% ($p = 0.010$). The share below the token is the one margin that does not differ (5% against 8%, $p = 0.242$).[41] Across decisions, the Personalized Menu moves some reports and leaves others at the token, and for those 57% of decisions an institution could still read the consumer's information from the report. Of the reports that leave the token, most stop short of the

[40] Inverting the rule would not work in this design even if participants reported optimally, because the optimal report is 9 at every token and so reveals nothing about the token. A rule whose optimal report differed across tokens would allow it. The point here is that participants do not report optimally, so the inversion would fail under such a rule too.

[41] Table D2 in Appendix D reports the regressions behind these tests.

Figure 6: Distribution of $\tilde{d}$ across decisions and across participants.



(a) Decisions, IC Control: $\mathbb{E}[\tilde{d}] = 7\%$

(b) Decisions, Misreport Up: $\mathbb{E}[\tilde{d}] = 21\%$

(c) Participants, IC Control: $\mathbb{E}[\bar{d}] = 7\%$

(d) Participants, Misreport Up: $\mathbb{E}[\bar{d}] = 21\%$

Notes: Tokens 1 to 8 in both treatments. Top row: every decision, with heights the shares of that treatment's decisions. Bottom row: every participant, with $\bar{d}$ the participant's mean $\tilde{d}$ over their trials on those tokens and heights the shares of that treatment's participants; a non-zero mean is never binned into the slot at the token. In every panel the three solid bars are exact values, at the token and at either end of the scale, and the lighter bars in between are 0.1-wide bins. The three shares along the top of a panel are the shares below, at, and above the token; the share above its own bar at +100% is the share at the top of the scale. The star marks the optimum, which is +100% at every conflict token in Misreport Up and the token itself in the IC Control.

optimum.[42]

[42]Conditional on leaving the token, misreports cover similar distances in the two treatments: upward misreports cover 61% of the available distance in Misreport Up and 56% in the IC Control ($p = 0.705$), and downward misreports −47% and −51% ($p = 0.443$). This is not an estimate of the intensive margin, because it conditions on an outcome the treatment changes. About 49% of the upward misreports in Misreport Up are decisions that would have stayed at the token without the treatment, so the two means are taken over different decisions. If those decisions cover shorter distances, the comparison understates the response of the decisions that would have left the token anyway, and the data cannot separate the two.

The bottom row of Figure 6 asks whether the decisions at the token belong to particular participants. For this, I create a participant-level variable, $\bar{d}$, by averaging each participant's $\tilde{d}$ over their trials on the conflict tokens, and the panels plot its distribution in each treatment. Misreport Up has more participants above the token than the control, 69% against 39%, and fewer at it, 26% against 48%. Nearly all of those at zero are at the token in every trial.[43] At the far end, 3% of participants in Misreport Up report 9 in every trial, against 0% in the control. The 83 participants in between are almost all above the token, 93% against 7% below. In the control, the participants in between split 75% against 25%.

The average of 21% is therefore made of three report groups. The first, 25% of participants, report the token in every trial, so from them the Measurement Stage recovers accurate information despite the conflict. The second, 3%, report the optimum in every trial. The rest leave the token without reaching the optimum: 66% misreport in the direction personalization favors but stop short, and 5% misreport in the other direction.

Stopping short of the optimum is costly, and twenty trials with feedback do little to change it. The reports in Misreport Up imply average expected trial earnings of \$11.45, against \$13.90 had every report been the optimum. By report group, the 83 participants in between expect \$11.72 a trial and the 30 whose mean deviation is zero expect \$10.47.[44] Among the participants who move, $\tilde{d}$ rises from 17% in the first trial to 33% in the last, still far from the optimum, and the report groups are the same at the end as at the start.[45]

3.3 The quiz identifies who reports accurately and why

The three report groups of Section 3.2 raise two issues: whether the participants who report accurately can be identified, and why participants report accurately, optimally, or in between. In the trials, I know who reports accurately by design, because I know the token. With unobservable information this is typically not possible, so I ask whether a task outside the trials can identify them. The task is the quiz of Section 2.5. Its combined question is very similar to a trial: it shows a token, the price list, and the participant's own personalization rule, and asks for the report that maximizes trial earnings. I therefore expect the answer to

[43] A mean of zero can come from deviations that offset, and this happens once. In Misreport Up, 30 participants have a mean of zero and 29 of them never leave the token, so 25% of the treatment is at the token in every trial, not 26%. In the IC Control the two counts are the same, 47.

[44] Expectations are taken over the price-list row that will be drawn, so these are the earnings a participant expects when reporting, not the earnings they realized.

[45] These are fitted levels from a linear trend on their decisions. In the IC Control, the participants who move show the same increase, from 8% to 19%, and the two slopes are not distinguished ($p = 0.471$). Appendix E gives the slopes for every participant and for those who move, and plots both treatments by trial. The accurate reports are not a matter of learning the Measurement Stage either: in the three incentive compatible treatments, 68.7% of participants report the token in the first trial and 68.4% in the second, and no effect of a costly draw on the next report is detected once the token and the report are held fixed (Appendix H).

track behavior in the trials.[46]

The second issue is why participants do what they do. For this I use the two component questions, which separate the two sets of incentives, to see whether participants fail to best respond to some of them and whether that relates to their behavior in the trials. The second part of the quiz and the exit survey then test whether general traits explain that behavior.

On the first issue, the answer to the combined question tracks what the participant did in the trials. On the scale of the trials, with the quiz token at 5, an answer of 9 is $\tilde{d} = 100\%$ and an answer of 5 is $\tilde{d} = 0\%$. Its correlation with the participant's mean $\tilde{d}$ over their trials, across the 116 participants in Misreport Up, is 0.59 ($p < 0.001$; Figure G3 in Appendix G). The relation is not particular to the treatment in which the optimum is a misreport: in the IC Control, where the correct answer is the token, the correlation is 0.57 across its 98 participants ($p < 0.001$).

The answer to the combined question identifies who reports accurately and indicates, with an overstatement, what the others do. Panel (a) of Figure 7 shows both by grouping participants on the answer. First, in Misreport Up, the 61% of participants who answer with the token have 71% of their reports at the token. In the IC Control, the 82% who answer with the token have 85%.[47] Second, in Misreport Up, the 16% who answer 9, the correct answer, cover 52% of the available distance in the trials, and those who give any other answer cover 28% against the 34% they state. A single question, answered once after the trials, therefore separates the participants whose reports can be taken as accurate from those whose reports cannot, and bounds how far the rest are from the token.

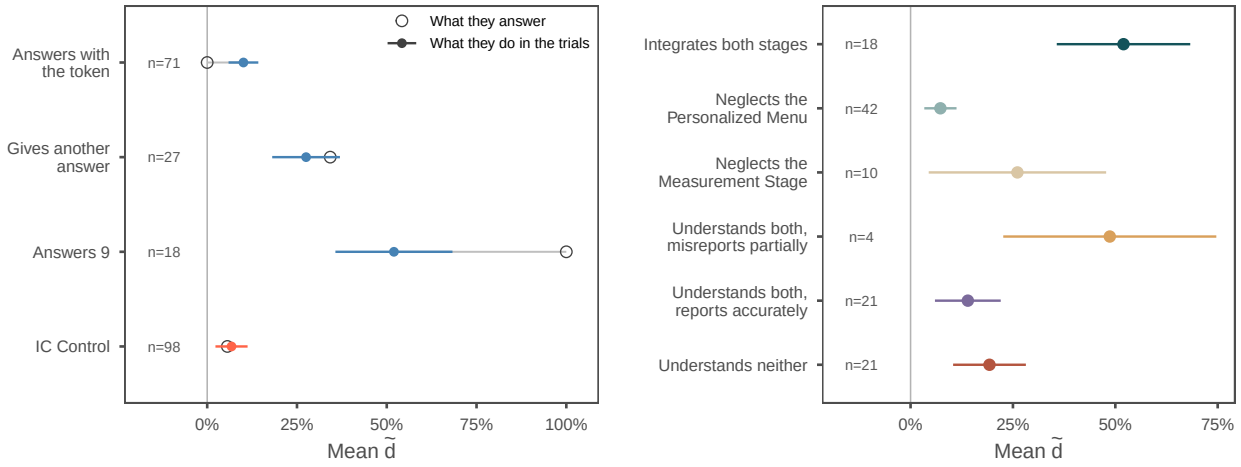
On the second issue, the two component questions show which incentives participants can best respond to. In Misreport Up, 64% of participants answer the measurement question correctly and 46% the menu question, both far above chance, against the 16% who answer the combined one. Best responding to the incentives for accuracy on their own is therefore common. Best responding to the Personalized Menu on its own is less common, and best responding to both together is rare. Sorting the 116 participants in Misreport Up on all three answers gives the six misconception groups of Section 2.5: 16% integrate both stages, 36% neglect the Personalized Menu, 9% neglect the Measurement Stage, 3% understand both and misreport partially, 18% understand both and report accurately, and 18% understand neither.

The participants who stay at the token, and those who get close to the optimum, belong

[46]Building the same kind of task in the field is not trivial, because it needs the induced-value analogue of a decision; Bernheim et al. (2026) discuss the caveats of this method, referred to as pattern matching.

[47]The identification is close but not exact: of the 29 participants in Misreport Up who report the token in every trial, 26 answer with the token (90%), while of the 71 who answer with the token, 37% report it in every trial.

Figure 7: Average $\tilde{d}$ in Misreport Up, by answer to the combined question and by misconception group, with the IC Control.



(a) By answer to the combined question

(b) By misconception group

Notes: Misreport Up, and the IC Control in the last row of panel (a). Each participant's mean $\tilde{d}$ over their trials on the conflict tokens, averaged within group, with 95% confidence intervals and the group's size beside each row. Panel (a) groups participants by their answer to the combined question: the token or the button reading "all switch points give the same earnings", 9, or any other switch point. The hollow marker is the group's answer on the scale of the trials, at the quiz token of 5, so that the token and the button are 0 and an answer of 9 is 100%, and the filled marker is what the group does in the trials. The last row, in red, is the IC Control with all answers together: its mean answer is 6% and its mean $\tilde{d}$ in the trials 7%. Panel (b) groups them by misconception group. The misconception groups use the quiz alone and are fixed before any behavior enters them, and the quiz is answered once, after the trials.

mostly to particular misconception groups. Of the 29 participants who report the token in every trial, 66% neglect the Personalized Menu and a further 21% understand both but report accurately. Of the 16 who cover more than half the available distance, 62% integrate both stages, 12% understand both and misreport partially, and 12% neglect the Measurement Stage (Figure G2 in Appendix G). Panel (b) of Figure 7 summarizes the same relation by misconception group, as the average $\tilde{d}$ within each. How far a group's reports are from the token rises with how much its answers take the Personalized Menu into account. The two misconception groups whose answers stop at the Measurement Stage are closest to the token, at 7% and 14% of the available distance. The groups that best respond to the Personalized Menu on its own, or integrate it with the Measurement Stage, are furthest, at up to 52%.[48]

The two misconception groups closest to the token motivate the traits that the second part of the quiz measures. A participant who neglects the Personalized Menu knows the personalization rule, since the instructions test it, but does not work out the best response to

[48]The other three levels are 26% for the participants who neglect the Measurement Stage, 49% for those who understand both and misreport partially, and 19% for those who understand neither.

it, which could reflect a failure to reason backward from the Personalized Menu. A participant who understands both stages and reports accurately works out each best response but leaves the Personalized Menu out when the two are combined, which could reflect narrow framing. If these are general traits of a participant, and not responses to the trials in particular, measures of them taken in unrelated tasks might predict who reports the token.

General traits measured in unrelated tasks do not account for the report groups. This is the second part of the second issue: whether narrow framing and a failure of backward induction, or lying aversion and the weight placed on future consequences, explain who reports the token. I regress each participant's mean $\tilde{d}$ in Misreport Up on the misconception groups, on the four general traits, and on both. The misconception groups alone account for a third of the variance, with an R^2 of 0.34. The traits alone account for little, with an R^2 of 0.05, and adding them to the misconception groups raises the R^2 only to 0.39. No trait moves reports toward the token. Narrow framing, a failure of backward induction, and the weight placed on future consequences are not distinguishable from zero, and the lying item, the one trait that is distinguishable from zero before adjusting for the four tests ($p = 0.029$, Holm $p = 0.115$), has the sign opposite to its prediction.[49]

Two of the items are nonetheless informative, though not in the way the traits predict. The first is the lying item, whose sign is informative. If participants experienced a misreport as a lie, those averse to lying should stay closer to the token. Instead, 71% of participants report the wheel's number honestly, and honest reporting goes with the share of reports at the token only weakly in the IC Control ($r = 0.176$, $p = 0.084$). In Misreport Up, within misconception groups, honest reporters are 11 percentage points further from the token, a difference that does not survive the adjustment for the four traits. This supports the design choice to remove the lying context from the trials (Section 2.1).

The second is the backward-induction item, which does not explain the lack of response to personalization but measures something else that is common to both treatments. Three facts support this. First, in the IC Control, where the token is the optimal report, failing the item goes with leaving it: the correlation between failing and the share of reports at the token is -0.509, and the 37% of control participants who fail report the token in 48% of their trials, against 87% for the rest. Second, without them the noise shrinks enough that the control is at the token: among the participants who solve the item, the joint test that the

[49] I read the four trait coefficients together to reach this conclusion, so I also report each p-value adjusted for the four tests with Holm's method. Table F3 in Appendix F reports the three regressions with every coefficient. The participants who report honestly are 11 percentage points further from the token in the full specification ($p = 0.029$). The appendix also describes the three items and gives the correlation of each with the share at the token in both main treatments. In the IC Control no trait moves reports toward the token either, with one difference, the backward-induction item, which I discuss below.

eight token means are zero does not reject ($F(8, 61) = 1.32$, $p = 0.251$). Third, in Misreport Up the participants who fail the item behave like those who fail it in the control, with a mean $\tilde{d}$ of 16% against 15% and 53% of reports at the token against 48%, which suggests that their reports are similarly noisy under both rules. As a result, the treatment effect on $\tilde{d}$ in the main specification is 21 percentage points among the 133 participants who solve the item ($p < 0.001$) and 1 among the 81 who fail it ($p = 0.918$; Appendix F).[50] Within Misreport Up, however, the item does not sort those who stay at the token from those who move.

What the backward-induction item measures is not clear. It could be attention, ability, or something else, and the rest of the data fit more than one reading. The participants who fail it make other mistakes: they take the dominated alternative in the Personalized Menu more often, 11% against 3% of control decisions, and need more submissions to pass the three practice questions, 4.8 against 3.6 ($p = 0.003$). They do not go faster: they spend more time on each trial, 37 against 22 seconds ($p < 0.001$), and pass the four comprehension questions in as many submissions as the rest, 5.5 against 5.3 ($p = 0.249$).

The quiz therefore partially identifies how far a participant's reports are from the token, through the combined question, and separates the responses by which incentives a participant can best respond to. Narrow framing, a failure of backward induction, lying aversion, and the weight placed on future consequences, measured as general traits, do not explain the reports in the direction each trait predicts. The lying item indicates instead that participants did not experience a misreport as a lie, and the backward-induction item identifies, in both treatments, the participants whose reports follow neither rule. The participants who accurately report the token mostly best respond to the incentives for accuracy and not to the Personalized Menu, or answer with accuracy alone when asked to combine the two. For an institution, the combined question tells whose reports can be read as the consumer's information, and the component questions tell why the others' cannot.

3.4 Five extension treatments: reports respond to the direction personalization favors

The five extension treatments ask how the results depend on the personalization rule. The main comparison changes one thing between its treatments: whether the report can create a better second alternative. The incentives it adds are of one kind: the menu favors reports above the token, the direction is certain, and the optimum is 9 on every conflict token. The extensions vary that rule to address three issues: whether the response survives a reversal of

[50] Splitting the participants by narrow framing, by the lying item, or by the weight placed on future consequences instead leaves the treatment effect the same in both halves. Figure F2 in Appendix F shows the four splits.

the direction personalization favors; what happens when the direction is uncertain at the time of the report; and whether reports stay accurate when there is personalization that preserves the incentive compatibility of the Measurement Stage.[51] Each treatment keeps the Measurement Stage, and with it the cost of a misreport, and changes one feature of the personalization rule, so I describe each rule next to its results, one block per issue.[52]

The comparisons use tokens 2 to 7, which are conflict tokens in every treatment that is not incentive compatible (Appendix A), so the numbers for Misreport Up here differ from those of Section 3.1 on its wider range.[53] One regression pools all seven treatments, with the control as the omitted category, token and trial fixed effects, and standard errors clustered on the participant. equation (D1) in Appendix D gives the regression. It is the pre-registered test of the treatment effect, for every treatment at once. The joint test of each treatment's effect at every token, from the same regression with treatment-by-token interactions, is in the joint-test column of Table 1, with its Holm-adjusted value in brackets below it.[54] Table 1 reports the regression, and Figure 8 plots each treatment's coefficient on the two outcomes that mean the same thing in every treatment, the deviation and the share at the token. On these tokens the IC Control has mean $\tilde{d}$ of 7% and a share at the token of 73%.

The first block shows that the response survives a reversal of the direction. *Misreport Down* sets $x_2 = 10 - \hat{\theta}$, so a low report creates a high alternative and the menu favors reports below the token. Because a high token is already a good outside option, the optimum is $\hat{\theta}^* = 1$ only on $\theta \in \{2, \ldots, 7\}$, and the token itself at the other tokens. Reversing the direction moves the deviation and not the share at the token. Misreport Down, with 100 participants, has mean $\tilde{d}$ of -3% against 7% in the control ($p = 0.007$, Holm $p = 0.034$), while its share at the token, 64% against 73%, is not distinguished from the control's ($p = 0.116$). Its share at the optimum is 7%, against the 2% of control reports at the same report ($p = 0.003$), and 11% in Misreport Up on the same tokens.[55] The response therefore survives a change of direction, in the deviation if not in the share at the token: reports move in the direction personalization

[51] The IC Control preserves incentive compatibility by having no personalization.

[52] Appendix A derives the optimum in each and Figure A3 plots them token by token.

[53] On these tokens Misreport Up has mean $\tilde{d}$ of 20% ($p < 0.001$ against the control) and a share at the token of 56% ($p = 0.002$).

[54] I read the six comparisons with the control together, so for each outcome I also report the p-values adjusted for the six tests with Holm's method.

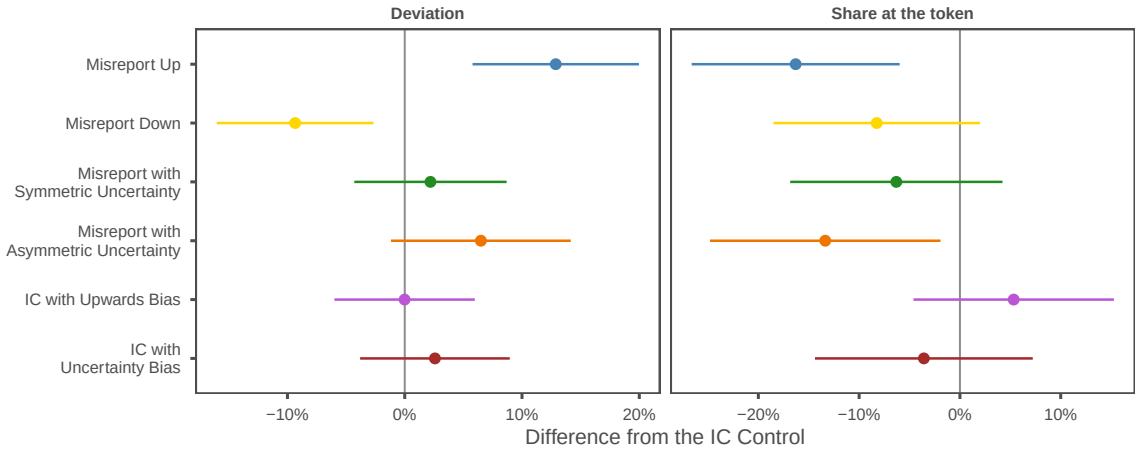
[55] Among reports that are neither the token nor the optimum, only 35% go in the direction personalization favors, against 84% in Misreport Up.

Table 1: Every treatment against the IC Control.

Treatment	N	Mean $\tilde{d}$	Difference	Joint test (p)	Share at token	Difference	Share at optimum
IC Control	98	7%	—	—	73%	—	—
Misreport Up	116	20%	12.9	0.018	56%	-16.3	11%
			(3.6)	[0.106]		(5.3)	
Misreport Down	100	-3%	-9.3	0.050	64%	-8.3	7%
			(3.4)	[0.198]		(5.2)	
Misreport with Symmetric Uncertainty	103	9%	2.2	0.405	66%	-6.3	3%
			(3.3)	[1.000]		(5.4)	
Misreport with Asymmetric Uncertainty	86	13%	6.5	0.039	59%	-13.4	7%
			(3.9)	[0.193]		(5.8)	
IC with Upwards Bias	101	7%	0.0	0.964	78%	5.3	—
			(3.1)	[1.000]		(5.1)	
IC with Uncertainty Bias	98	9%	2.6	0.541	69%	-3.6	—
			(3.3)	[1.000]		(5.5)	

Notes: Decisions on tokens 2 to 7, the conflict tokens common to every treatment that is not incentive compatible. The mean deviation, the share at the token and the share at the optimum are the raw values in each treatment; in the three incentive compatible treatments the optimum is the token, so the share at the optimum is the share at the token and is not repeated. Each Difference is the treatment's coefficient, in percentage points, in a regression of the outcome to its left that pools the seven treatments, with the IC Control omitted and token and trial fixed effects; CR2 standard errors clustered on the participant in parentheses. The joint test is the pre-registered test of the treatment effect on the normalized deviation: from the same regression with the interaction of each treatment with each token, the HTZ Wald test that the treatment's main effect and its interactions are all zero, so that its effect is zero at every token. In brackets, the joint test's p-value adjusted for the six treatments by Holm's step-down method.

Figure 8: Each treatment against the IC Control.



Notes: The treatment coefficients of the pooled regression on tokens 2 to 7, with the IC Control as the omitted category, token and trial fixed effects, and standard errors clustered on the participant, with 95% confidence intervals. *Deviation* is $\tilde{d}$ and *share at the token* an indicator for reporting it, so a point to the right of zero is a treatment whose reports lie further from the token than the control's, or that leaves the token more often. The share at the optimum is not shown, since the optimum is the token in the incentive compatible treatments and an end of the scale in the others, so differencing it against the control would compare hitting one target with hitting a different one; Table 1 reports it per treatment.

favors, down as well as up, and in both directions stop far short of the optimum.[56]

[56]Three points qualify the comparison with Misreport Up. First, the optimum in Misreport Down is not the same report at every token, as it is in Misreport Up, where it is 9 throughout. It is 1 on tokens 2 to 7 and the token itself elsewhere. Second, the IC Control mean is itself above zero, so the two treatments' gaps from the control are not measured from the token. Misreport Down's gap is in the opposite direction to the control's own upward deviation, and Misreport Up's is in the same direction. Third, less is at stake in Misreport Down: on tokens 2 to 7, the optimal report earns on average \$1.74 more per trial in expectation than reporting the token in Misreport Down, against \$3.34 in Misreport Up.

The second block shows that no response is detected when the two directions are equally likely. The direction may be uncertain in practice, when a consumer does not know which direction the institution's menu will favor. *Misreport with Symmetric Uncertainty* applies the rule of Misreport Up with probability one half and the rule of Misreport Down otherwise: the second alternative, drawn before the Personalized Menu (but after the Measurement Stage), is $X = \hat{\theta}$ or $X = 10 - \hat{\theta}$, and the participant chooses between θ and X. The Measurement Stage is not incentive compatible: the two directions are equally likely, but the optimum is still not the token, 1 below $\theta = 5$ and 9 above it. With 103 participants, no response is detected: the treatment is not distinguished from the control on the two outcomes that can register a response in this treatment, the share at the token and the share at the optimum. Its share at the token is 66% against the control's 73% ($p = 0.242$), a difference whose 95% confidence interval runs from -17 to $+4$ percentage points, so the comparison does not exclude a reduction of the size seen in Misreport Up. Its share at the optimum is 3%, no higher than the 4% of control reports at the same reports ($p = 0.922$).[57]

When one direction is more likely, the evidence that reports leave the token is suggestive. A flat response under symmetric uncertainty could come from the uncertainty itself or from the two directions being equally likely. *Misreport with Asymmetric Uncertainty* separates the two: it keeps the draw but applies the rule of Misreport Up with probability 0.9 and the rule of Misreport Down with probability 0.1, so the optimum is 9 at every token, as in Misreport Up. With 86 participants, it has mean $\tilde{d}$ of 13% ($p = 0.099$) and a share at the token of 59%, which differs from the control's before adjusting for the six treatments ($p = 0.023$) but not after ($p = 0.116$). The pre-registered joint test gives the same picture ($p = 0.039$, Holm $p = 0.193$). Its share at the optimum is 7%, against the 5% of control reports at the same report ($p = 0.397$).

The third block shows no response when the report enters the Personalized Menu without changing the optimum. It asks whether the responses above come from the conflict or from the report entering the Personalized Menu at all. Recall that in the IC Control the report never enters the Personalized Menu, so the control differs from the treatments that are not incentive compatible in two ways at once: the report enters the Personalized Menu, and entering changes the optimum. Two treatments separate them by replacing the token with 10 as the first alternative. *IC with Upwards Bias* offers $\{10, \hat{\theta}\}$, the counterpart of Misreport Up, and *IC with Uncertainty Bias* offers $\{10, X\}$, with X being the draw of Symmetric Uncertainty. No report exceeds nine, so 10 is always the better alternative: the report appears

[57]The mean $\tilde{d}$ is not informative in this treatment. Because the optimum is 1 below $\theta = 5$ and 9 above it, a participant who reported the optimum at every token would have deviations of -100% and $+100\%$ that average near zero. The mean is 9%, not distinguished from the control's ($p = 0.508$). Table 1 reports it with the other treatments.

in the Personalized Menu but cannot change what the participant earns there, and the token remains optimal. In both, no response is detected, as expected where the Measurement Stage is incentive compatible: neither treatment is distinguished from the control on either outcome. IC with Upwards Bias, with 101 participants, has mean $\tilde{d}$ of 7% ($p = 0.997$) and a share at the token of 78% ($p = 0.294$); IC with Uncertainty Bias, with 98, has mean $\tilde{d}$ of 9% ($p = 0.430$) and a share at the token of 69% ($p = 0.517$). The intervals are not narrow: the 95% confidence intervals for the effect on the share at the token run from -5 to $+15$ and from -14 to $+7$ percentage points, and for the effect on $\tilde{d}$ from -6 to $+6$ and from -4 to $+9$.

On tokens 2 to 7, the response is robust to the direction, is suggestive when one direction is more likely, is not detected when the two directions are equally likely, and is not detected when the Measurement Stage stays incentive compatible. Reports move down in Misreport Down as they move up in Misreport Up, on the deviation if not on the share at the token. Uncertainty about the direction does not by itself keep reports at the token. When one direction is more likely, the share at the token falls, though neither it nor the deviation is distinguished from the control's once the six treatments are tested together. No response is detected only when the two directions are equally likely, with reports not distinguished from the control's, though the interval leaves room for one. And no response to the report entering the Personalized Menu is detected: neither treatment in which it enters without changing the optimum differs from the control, within intervals that allow moderate differences.

Restricting the comparisons to the participants who solve the backward-induction item sharpens them. Section 3.3 found that the participants who fail the item report alike under both main rules, so Appendix F refits the regression within the 424 participants who solve it and within the 278 who fail it (Table F2 and Figure F1). Among those who solve it, the control is at the token and the treatments order as the direction personalization favors would predict: Misreport Up moves reports the most, Misreport with Asymmetric Uncertainty less, and Misreport with Symmetric Uncertainty least, with a response that is small, suggestive, and upward on both sides of token 5. Among those who fail it, no treatment is distinguished from the control: the noise they add is alike across treatments.

4 Discussion

People's differences can be used to help them or exploit them. With the growing availability of personal data,[58] we can design policies that better target different individuals, and companies can price discriminate or tailor their recommendations. However, none of this can be done

[58] Firms collect personal data and trade it through data brokers. Acquisti et al. (2016) and Bergemann and Bonatti (2019) survey the market.

effectively if we do not understand the effect personalization has on consumers' behavior. If consumers do not share their information accurately, personalization may not work effectively.

In this paper, I studied how behavior changes when consumers anticipate personalization and varied whether it is in their best interest to share their information. I used the induced-value technique to observe whether they share information accurately or not. The main comparison is between an incentive compatible control with no personalization and a treatment that adds personalization that compromises incentive compatibility. I find that, on average, consumers behave differently, but many act as if there were no personalization, and those who respond to personalization struggle to do so optimally. Suboptimal behavior is largely explained by participants' misconceptions, that is, by which sets of incentives they cannot best respond to.

This has implications for regulation on how institutions are allowed to use personal data. The status quo requires firms to ask for permission to collect personal information and disclose what they do with the information.[59] However, I show in this stylized environment that disclosure might not be enough to protect consumers, since many do not know how to best respond to the consequences of personalization and share their information accurately. Moreover, it is possible to identify who shares information accurately using a simple task that measures their ability to best respond, which means institutions could use this to exploit certain consumers or even tune the personalization rule to one that recovers accurate information from most of them. This suggests that personalization should be carefully studied to evaluate how it should be regulated to protect consumers.

A necessary next step is to understand what needs to change in the environment for consumers to best respond to personalization. This could inform policies on data protection and the design of effective personalization. A natural step after that is to move away from induced values and study how the incentives of personalization interact with other important elements I abstracted away from, like privacy concerns, trust in the institution, and lying aversion.

References

Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for truth-telling. *Econometrica*, 87(4):1115–1153.

[59]The General Data Protection Regulation in the European Union (European Parliament and Council, 2016), the California Consumer Privacy Act (California Legislature, 2018), and California's Delete Act of 2023, which created a registry of data brokers (California Legislature, 2023).

Acquisti, A., Taylor, C., and Wagman, L. (2016). The economics of privacy. *Journal of Economic Literature*, 54(2):442–492.

Andreoni, J., Callen, M., Hussain, K., Khan, M. Y., and Sprenger, C. (2023). Using preference estimates to customize incentives: An application to polio vaccination drives in Pakistan. *Journal of the European Economic Association*, 21(4):1428–1477.

Barberis, N., Huang, M., and Thaler, R. H. (2006). Individual preferences, monetary gambles, and stock market participation: A case for narrow framing. *American Economic Review*, 96(4):1069–1090.

Becker, G. M., DeGroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3):226–232.

Ben-Shahar, O. and Schneider, C. E. (2014). *More Than You Wanted to Know: The Failure of Mandated Disclosure*. Princeton University Press, Princeton, NJ.

Bergemann, D. and Bonatti, A. (2019). Markets for information: An introduction. *Annual Review of Economics*, 11(1):85–107.

Bernheim, B. D., Lucia, A., Nielsen, K., and Sprenger, C. D. (2026). When are decisions improvable? An evaluation of diagnostic methods. Working Paper 34979, National Bureau of Economic Research.

Binmore, K., McCarthy, J., Ponti, G., Samuelson, L., and Shaked, A. (2002). A backward induction experiment. *Journal of Economic Theory*, 104(1):48–88.

Björkegren, D., Blumenstock, J. E., and Knight, S. (2020). Manipulation-proof machine learning. arXiv preprint arXiv:2004.03865.

Brokesova, Z., Deck, C., and Peliova, J. (2014). Experimenting with purchase history based price discrimination. *International Journal of Industrial Organization*, 37:229–237.

California Legislature (2018). California Consumer Privacy Act of 2018. Cal. Civ. Code §§1798.100–1798.199.

California Legislature (2023). Delete Act, Senate Bill 362. Cal. Civ. Code §§1798.99.80–1798.99.89.

Cason, T. N. and Plott, C. R. (2014). Misconceptions and game form recognition: Challenges to theories of revealed preference and framing. *Journal of Political Economy*, 122(6):1235–1270.

Celebi, C., Exley, C., Harrs, S., Kivimaki, H., Serra-Garcia, M., and Yusof, J. (2026). The collection of high-quality online data. Working Paper 12535, CESifo, Munich.

Chapman, J., Snowberg, E., Wang, S. W., and Camerer, C. F. (2024). Dynamically optimized sequential experimentation (DOSE) for estimating economic preference parameters. Working Paper 33013, National Bureau of Economic Research.

Charness, G., Kuhn, P., and Villeval, M. C. (2011). Competition and the ratchet effect. *Journal of Labor Economics*, 29(3):513–547.

Chaudhuri, A. (1998). The ratchet principle in a principal agent game with unknown costs: An experimental analysis. *Journal of Economic Behavior & Organization*, 37(3):291–304.

Cooper, D. J., Garvin, S., and Kagel, J. H. (1997a). Adaptive learning vs. equilibrium refinements in an entry limit pricing game. *Economic Journal*, 107(442):553–575.

Cooper, D. J., Garvin, S., and Kagel, J. H. (1997b). Signalling and adaptive learning in an entry limit pricing game. *RAND Journal of Economics*, 28(4):662–683.

Cooper, D. J. and Kagel, J. H. (2008). Learning and transfer in signaling games. *Economic Theory*, 34(3):415–439.

Cooper, D. J., Kagel, J. H., Lo, W., and Gu, Q. L. (1999). Gaming against managers in incentive systems: Experimental results with Chinese students and Chinese managers. *American Economic Review*, 89(4):781–804.

Danz, D., Vesterlund, L., and Wilson, A. J. (2022). Belief elicitation and behavioral incentive compatibility. *American Economic Review*, 112(9):2851–2883.

Ellis, A. and Freeman, D. J. (2024). Revealing choice bracketing. *American Economic Review*, 114(9):2668–2700.

European Parliament and Council (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). Official Journal of the European Union, L 119, 4 May 2016, pp. 1–88.

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

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review*, 95(1):384–394.

Gneezy, U., Kajackaite, A., and Sobel, J. (2018). Lying aversion and the size of the lie. *American Economic Review*, 108(2):419–453.

Gneezy, U., Rockenbach, B., and Serra-Garcia, M. (2013). Measuring lying aversion. *Journal of Economic Behavior & Organization*, 93:293–300.

Goldfarb, A. and Tucker, C. (2019). Digital economics. *Journal of Economic Literature*, 57(1):3–43.

Gonzalez, R. and Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, 38(1):129–166.

Güth, W., Kröger, S., and Normann, H.-T. (2004). Durable-goods monopoly with privately known impatience: A theoretical and experimental study. *Economic Inquiry*, 42(3):413–424.

Johnson, E. J., Camerer, C., Sen, S., and Rymon, T. (2002). Detecting failures of backward induction: Monitoring information search in sequential bargaining. *Journal of Economic Theory*, 104(1):16–47.

Kagel, J. H. and Levin, D. (1993). Independent private value auctions: Bidder behaviour in first-, second- and third-price auctions with varying numbers of bidders. *Economic Journal*, 103(419):868–879.

Levitt, S. D., List, J. A., and Sadoff, S. E. (2011). Checkmate: Exploring backward induction among chess players. *American Economic Review*, 101(2):975–990.

Loewenstein, G., Sunstein, C. R., and Golman, R. (2014). Disclosure: Psychology changes everything. *Annual Review of Economics*, 6(1):391–419.

McKelvey, R. D. and Palfrey, T. R. (1992). An experimental study of the centipede game. *Econometrica*, 60(4):803–836.

Merlo, A. and Schotter, A. (1999). A surprise-quiz view of learning in economic experiments. *Games and Economic Behavior*, 28(1):25–54.

Murphy, J. J., Stevens, T. H., and Yadav, L. (2010). A comparison of induced value and home-grown value experiments to test for hypothetical bias in contingent valuation. *Environmental and Resource Economics*, 47(1):111–123.

Myerson, R. B. (1979). Incentive compatibility and the bargaining problem. *Econometrica*, 47(1):61–73.

Pathak, P. A. and Sönmez, T. (2008). Leveling the playing field: Sincere and sophisticated players in the Boston mechanism. *American Economic Review*, 98(4):1636–1652.

Prelec, D. (1998). The probability weighting function. *Econometrica*, 66(3):497–527.

Pustejovsky, J. E. and Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, 36(4):672–683.

Rabin, M. and Weizsäcker, G. (2009). Narrow bracketing and dominated choices. *American Economic Review*, 99(4):1508–1543.

Read, D., Loewenstein, G., and Rabin, M. (1999). Choice bracketing. *Journal of Risk and Uncertainty*, 19(1–3):171–197.

Reynolds, S. S. (2000). Durable-goods monopoly: Laboratory market and bargaining experiments. *RAND Journal of Economics*, 31(2):375–394.

Smith, V. L. (1976). Experimental economics: Induced value theory. *American Economic Review*, 66(2):274–279.

Strathman, A., Gleicher, F., Boninger, D. S., and Edwards, C. S. (1994). The consideration of future consequences: Weighing immediate and distant outcomes of behavior. *Journal of Personality and Social Psychology*, 66(4):742–752.

Tolksdorf, M. (2022). Experimenting with purchase history based price discrimination: A comment. Discussion Paper 310, CRC TRR 190 Rationality and Competition.

Toubia, O., Johnson, E., Evgeniou, T., and Delquié, P. (2013). Dynamic experiments for estimating preferences: An adaptive method of eliciting time and risk parameters. *Management Science*, 59(3):613–640.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323.

Vilar, R., Milfont, T. L., Araújo, R. d. C. R., Coelho, G. L. d. H., Soares, A. K. S., and Gouveia, V. V. (2022). Consideration of future consequences (CFC): Validation and proposition of an ultra-short scale. *Current Psychology*, 41(6):3428–3442.

Vossler, C. A. and McKee, M. (2006). Induced-value tests of contingent valuation elicitation mechanisms. *Environmental and Resource Economics*, 35(2):137–168.

Weitzman, M. L. (1980). The "Ratchet Principle" and performance incentives. *Bell Journal of Economics*, 11(1):302–308.

Appendix contents

A Optimal reports in every treatment **37**

B Sample and data quality **43**

C Pre-registration **45**

D Estimation **46**

E Learning **48**

F The quiz and the general traits **51**

G The quiz answers in detail **59**

H The role of feedback: costly misreports **61**

I The screens **64**

I.1 Instructions and comprehension questions 66

I.2 The practice trial 90

I.3 A trial 101

I.4 The follow-up questions 107

I.5 The exit survey and the payment 126

A Optimal reports in every treatment

This appendix reads the optimal reports off the expected earnings of the two stages, in the way panel (a) of Figure 4 shows them: the Measurement Stage and the Personalized Menu as functions of the report, and their sum, whose maximum is the optimal report. The shapes are the same at every token, so the optimum at the token of 5 in the figures carries over to the other tokens, with the two exceptions noted below.

The Measurement Stage. Reporting $\hat{\theta}$ selects the token on the rows whose Option B amount $b_k = k - \frac{1}{2}$ lies below $\hat{\theta}$ and the amount b_k on the rest, and one of the ten rows is drawn. Averaging over the rows, the expected earnings of the Measurement Stage are

$$M(\hat{\theta}, \theta) \;=\; 5 + \frac{\theta^2}{20} - \frac{(\hat{\theta} - \theta)^2}{20}, \tag{A1}$$

where each term has a simple meaning. The constant 5 is the average of the ten Option B amounts, which is what a participant earns in expectation by taking Option B in every row. The term $\theta^2/20$ is what reporting the token adds to that. On the θ rows whose Option B amount lies below the token, a participant who reports the token keeps the token instead and gains $\theta - b_k$ on each; these gains, $\theta - \frac{1}{2}, \theta - \frac{3}{2}, \ldots, \frac{1}{2}$, sum to $\theta^2/2$, and dividing by the ten equally likely rows gives $\theta^2/20$. The last term is the cost of a misreport, equation (1) of the main text. On each of the $|\hat{\theta} - \theta|$ rows between the token and the report, the participant takes the worse of the two options and gives up $k - \frac{1}{2}$ on the k-th such row; these losses sum to $(\hat{\theta} - \theta)^2/2$, or $(\hat{\theta} - \theta)^2/20$ once divided by the ten rows. So M is a parabola in the report that peaks at the token and falls by $d^2/20$ at a distance d from it, in either direction. This is the dashed curve of the figures, and it is the same in every treatment. Only its height depends on the token, through $\theta^2/20$; its shape does not.

The Personalized Menu. Participants choose the better of the two alternatives, so the Personalized Menu pays $\max(x_1, x_2)$, averaged over the draw of the second alternative in the two uncertainty treatments. Table A1 gives this payment for each rule, and each expression follows from which two amounts are on the screen. In the IC Control the alternatives are the token and 5, so the menu pays $\max(\theta, 5)$, which does not depend on the report. In Misreport Up they are the token and the report, so the menu pays the report when it is above the token and the token otherwise. In Misreport Down the second alternative is $10 - \hat{\theta}$, so the menu pays more the lower the report, once $10 - \hat{\theta}$ exceeds the token. In the two uncertainty treatments the second alternative is drawn after the report, so the expected payment is the average of the two rules' payments with the probabilities of the draw as weights, one half

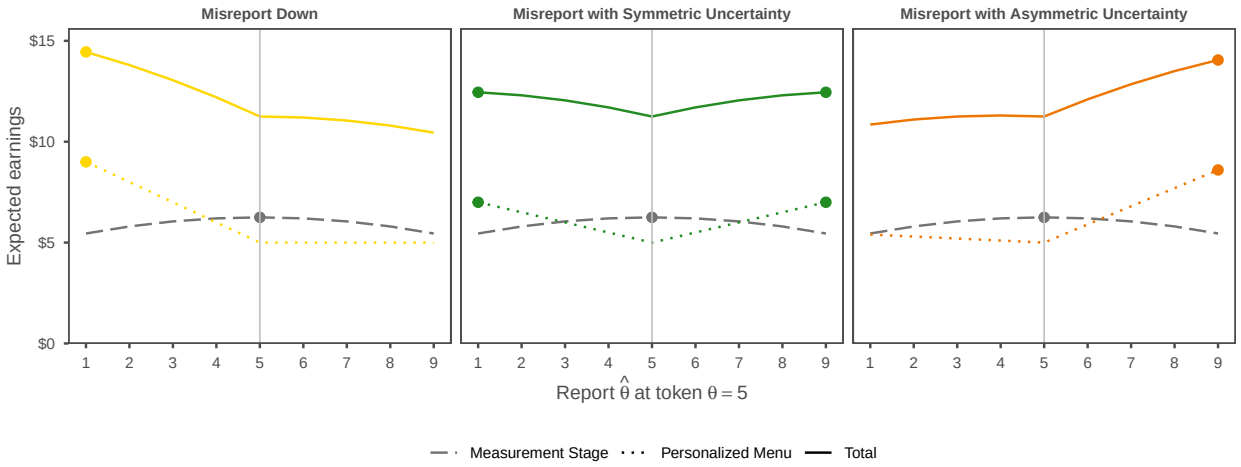
each under Symmetric Uncertainty and 0.9 and 0.1 under Asymmetric Uncertainty. In IC with Upwards Bias and IC with Uncertainty Bias the first alternative is 10, which exceeds any report or draw, so the menu pays 10 whatever the report. The payment is the dotted curve of the figures: flat in the IC Control; flat at the token below it and rising one dollar per position above it in Misreport Up; the mirror image in Misreport Down; and, in the uncertainty treatments, a mix of the last two with the weights of the draw.

The sum and its maximum. Expected trial earnings are the sum $M(\hat{\theta}, \theta) + A_T(\hat{\theta}, \theta)$, the solid curve, and the optimal report is the report at its highest point. Panel (a) of Figure 4 shows the two main treatments at $\theta = 5$ and Figure A1 the three misreport extension treatments at the same token. Reading the maxima:

- In the IC Control the menu adds a constant, so the sum peaks where the Measurement Stage does: at the token.
- In Misreport Up each position above the token adds one dollar in the menu and costs $d^2/20$ in the Measurement Stage, so the sum rises by $1 - (2d - 1)/20 > 0$ at every step up to the top of the scale: the optimum is 9.
- In Misreport Down the menu starts paying once $10 - \hat{\theta}$ exceeds the token, that is, for reports below $10 - \theta$. At $\theta = 5$ that is every report below the token, and each step down adds a dollar against the same cost, so the sum peaks at 1.
- In Misreport with Symmetric Uncertainty the menu pays half a dollar per position in either direction once it pays at all, and at $\theta = 5$ the two ends tie: the sum has two maxima, at 1 and at 9.
- In Misreport with Asymmetric Uncertainty the menu pays 0.9 per position above the token and at most 0.1 per position below it, so the sum peaks at 9, as in Misreport Up.
- In IC with Upwards Bias and IC with Uncertainty Bias the first alternative is 10, which beats any report or draw, so the menu pays 10 whatever the report and the sum peaks at the token.

The other tokens. Moving the token shifts the Measurement Stage parabola along with it and leaves its shape unchanged. In the IC Control and in Misreport Up the menu's kink sits at the token and moves with it, so the picture at $\theta = 5$ repeats at every token. In the rules that pay for low reports the menu also has a kink at $10 - \theta$, which moves the other way, and this produces two exceptions in the optimum and one change of slope that leaves the optimum where it is.

Figure A1: Expected earnings by report at $\theta = 5$, by stage, in the three misreport extension treatments.



Notes: As panel (a) of Figure 4: at the token $\theta = 5$, the expected earnings of each possible report, with the Measurement Stage, which is the same in every treatment, the Personalized Menu, and their total drawn on their own, expectations taken over the price-list row and, in the two uncertainty treatments, the draw of the second alternative. A filled point marks a curve's maximizer; in Misreport with Symmetric Uncertainty the total has two, at 1 and 9. The vertical line is the token.

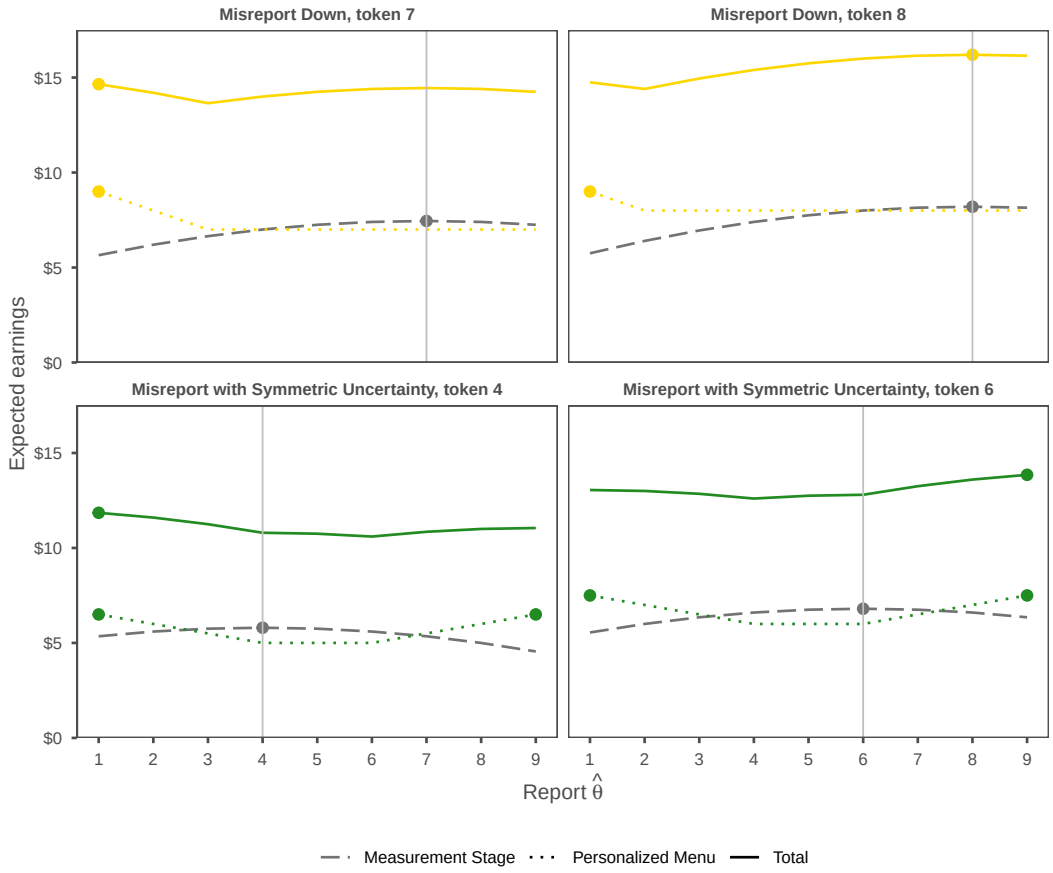
The first exception is Misreport Down. Its menu pays only for reports below $10 - \theta$, which is below the token itself once $\theta > 5$; reaching 1 then means paying for the steps down to $10 - \theta$ before the menu starts to pay, and that is worth doing at $\theta = 6$ and 7 but not at $\theta = 8$ and 9, where the token is optimal.

The second exception is Misreport with Symmetric Uncertainty. The two branches of the draw offset between θ and $10 - \theta$, where the menu is flat, and each pays half a dollar per position beyond that range; the two ends of the scale pay the same, $(\theta + 9)/2$, so the tie at $\theta = 5$ breaks toward the end nearer the token, whose Measurement Stage cost is smaller: 1 at lower tokens and 9 at higher ones.

The change of slope is in Misreport with Asymmetric Uncertainty: below $\theta = 5$ the $10 - \hat{\theta}$ branch of the draw still pays more than the token for the first reports above it, so the menu rises by 0.8 rather than 0.9 per position until the report reaches $10 - \theta$; every step up still beats its cost, the last one at $\theta = 1$ by \$0.05, so the optimum stays at 9.

Figure A2 shows the curves at the tokens where the optimum changes: Misreport Down at $\theta = 7$, where 1 still wins by \$0.20, and at $\theta = 8$, where the token wins; and Misreport with Symmetric Uncertainty at $\theta = 4$, where the tie breaks toward 1, and at $\theta = 6$, where it breaks toward 9. Figure A3 plots every optimum, and the last column of Table A1 lists them.

Figure A2: Expected earnings by report at the tokens where the picture changes.



Notes: As Figure A1, at the tokens named in each panel: Misreport Down at $\theta = 7$ and $\theta = 8$ (top row) and Misreport with Symmetric Uncertainty at $\theta = 4$ and $\theta = 6$ (bottom row). The Measurement Stage is the same curve in every panel up to its height, in Misreport Down the Personalized Menu pays only once $10 - \hat{\theta}$ exceeds the token, in Misreport with Symmetric Uncertainty it is flat between θ and $10 - \theta$ and rises half a dollar per position beyond either end of that range, and a filled point marks a curve's maximizer. The vertical line is the token.

Conflict tokens. A conflict token is one whose optimum is not the token. Misreport Up and Misreport with Asymmetric Uncertainty have $\theta \in \{1, \ldots, 8\}$; Misreport Down has $\{2, \ldots, 7\}$, since its optimum is the token at $\theta = 1$ and from $\theta = 8$; Misreport with Symmetric Uncertainty has $\{2, \ldots, 8\}$, since its optima 1 and 9 are the token at the two ends. The pooled comparisons of Section 3.4 use the intersection, $\theta \in \{2, \ldots, 7\}$, with Misreport Down setting the upper bound and, with Misreport with Symmetric Uncertainty, the lower; the main comparison uses Misreport Up's own range, $\theta \in \{1, \ldots, 8\}$.

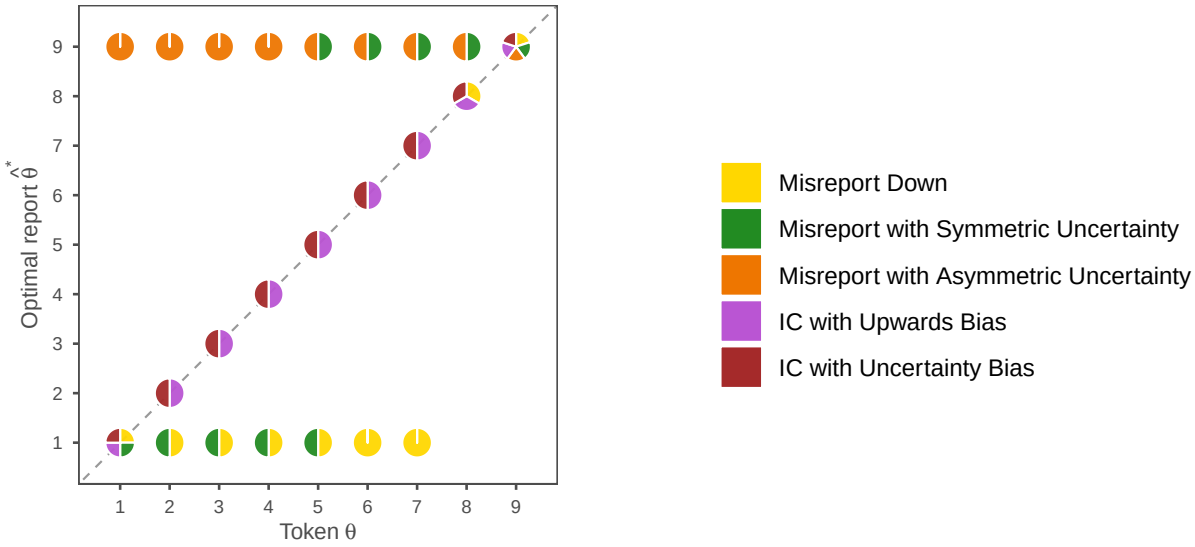
Risk preferences. The curves above are expected values. Recomputing the maxima under CRRA utility over total trial earnings, risk averse ($\rho \in \{0.5, 1, 2, 5, 10, 30\}$) and risk seeking

Table A1: What the Personalized Menu pays, and the optimal report, by treatment.

Treatment	Personalized Menu pays $A_T(\hat{\theta}, \theta)$	Optimal report $\hat{\theta}^*(\theta)$
IC Control	$\max(\theta, 5)$	θ
IC with Upwards Bias	$\max(10, \hat{\theta}) = 10$	θ
IC with Uncertainty Bias	$\max(10, X) = 10$	θ
Misreport Up	$\max(\theta, \hat{\theta})$	9
Misreport with Asymmetric Uncertainty	$0.9 \max(\theta, \hat{\theta}) + 0.1 \max(\theta, 10 - \hat{\theta})$	9
Misreport Down	$\max(\theta, 10 - \hat{\theta})$	1 at $\theta \leq 7$ θ at $\theta \in \{8, 9\}$
Misreport with Symmetric Uncertainty	$\frac{1}{2} \max(\theta, \hat{\theta}) + \frac{1}{2} \max(\theta, 10 - \hat{\theta})$	1 at $\theta < 5$ 9 at $\theta > 5$ 1 or 9 at $\theta = 5$

Notes: $X \in \{\hat{\theta}, 10 - \hat{\theta}\}$ is the draw of the uncertainty treatments. In every treatment the Measurement Stage pays $M(\hat{\theta}, \theta)$ of equation (A1), and the optimal report maximizes $M + A_T$ over $\hat{\theta} \in \{1, \ldots, 9\}$.

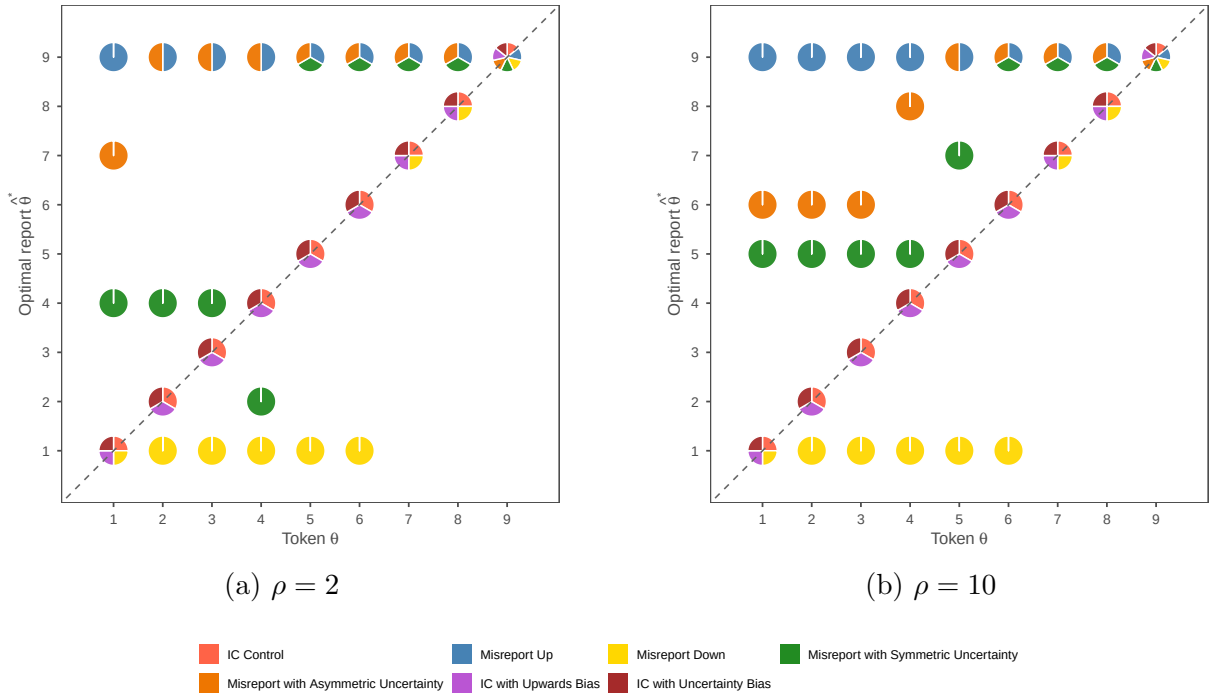
Figure A3: Optimal reports in the five extension treatments, by token.



Notes: The token on the horizontal axis and the report that maximizes expected total trial earnings on the vertical axis, with the dashed line at $\hat{\theta} = \theta$. A marker split between treatments is a cell whose optimum those treatments share. Misreport with Symmetric Uncertainty appears twice at $\theta = 5$, where 1 and 9 tie.

($\rho \in \{-1, -2, -5\}$), and under rank-dependent expected utility with the inverse-S weighting functions of Tversky and Kahneman (1992), Prelec (1998), and Gonzalez and Wu (1999) at their standard estimates, leaves the three incentive compatible treatments and Misreport Up unchanged, except that strong risk seeking ($\rho \leq -2$) moves Misreport Up's optimum to 8 at the lowest tokens. Misreport Down's optimum at $\theta = 7$ rests on an expected margin of \$0.20 against a possible loss of \$5.50, so concave utility from $\rho = 2$, or any of the weighting functions, makes the token optimal there, which removes $\theta = 7$ from its conflict tokens. In the two uncertainty treatments both departures value a narrower spread between the two possible draws, so the optima move toward the middle of the scale at low tokens and the tie at $\theta = 5$ breaks, replacing one misreport optimum with another and leaving the conflict tokens intact. Figure A4 shows the optima at $\rho = 2$ and $\rho = 10$, and Figure A5 under the Tversky and Kahneman (1992) function at two estimates; the other cases lie between them.

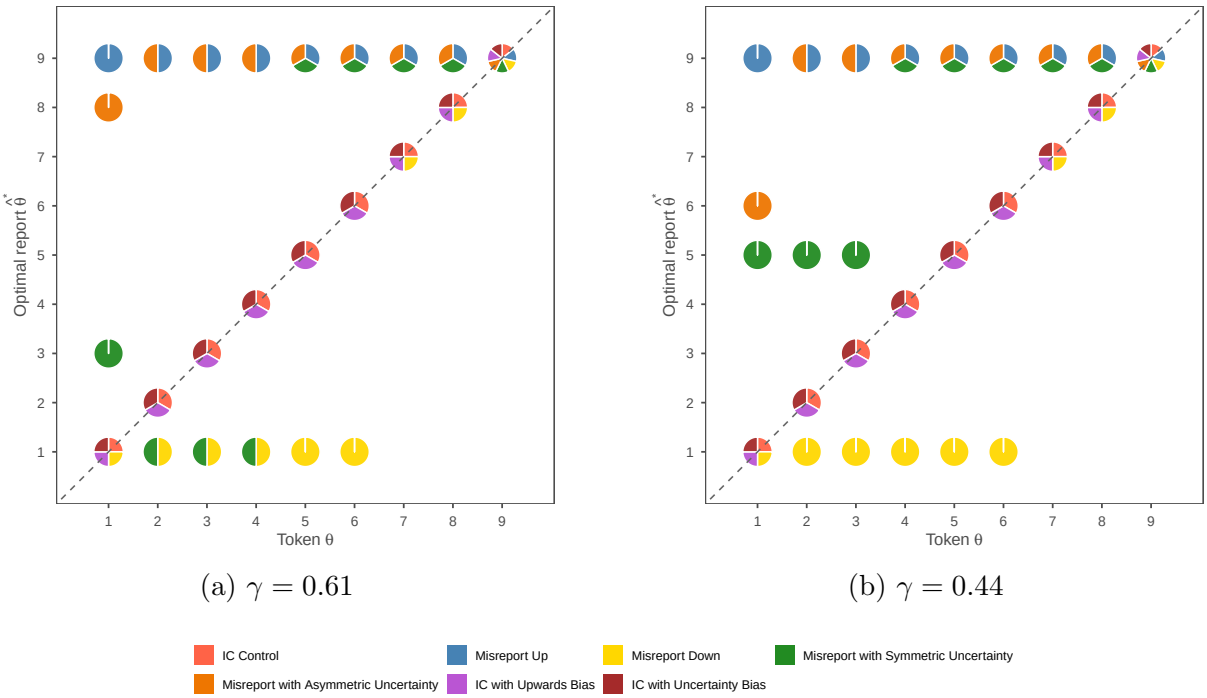
Figure A4: Optimal reports under CRRA utility over total trial earnings.



(a) $\rho = 2$ (b) $\rho = 10$

Notes: The optimal report by token in all seven treatments under $u(x) = x^{1-\rho}/(1-\rho)$, with the dashed line at $\hat{\theta}^* = \theta$. Against the risk-neutral optima of Figure A3 and panel (b) of Figure 4, Misreport Up and the three incentive compatible treatments are unchanged at every ρ on the grid; Misreport Down differs at $\theta = 7$ from $\rho = 2$; and the two uncertainty treatments differ at low tokens and at $\theta = 5$, at more tokens the larger ρ.

Figure A5: Optimal reports under probability weighting.



(a) $\gamma = 0.61$ (b) $\gamma = 0.44$

Notes: The optimal report by token in all seven treatments under rank-dependent expected utility with linear utility and the Tversky and Kahneman (1992) weighting function $w(p) = p^\gamma/(p^\gamma + (1-p)^\gamma)^{1/\gamma}$, at their estimate $\gamma = 0.61$ and at the Gonzalez and Wu (1999) estimate $\gamma = 0.44$, with the dashed line at $\hat{\theta}^* = \theta$. The optima are unchanged in Misreport Up and the three incentive compatible treatments under every weight examined, including Prelec (1998) with $\alpha \in \{0.65, 0.50\}$ and the pure pessimism and optimism weights; they differ in Misreport Down at $\theta = 7$ and in the two uncertainty treatments at low tokens and at $\theta = 5$.

B Sample and data quality

The experiment ran in a single session on Prolific on April 23, 2026. In total, 702 participants completed the study. Table B1 reports, treatment by treatment, the numbers that Section 2.1 quotes pooled: everyone who started and the share who dropped out before finishing; demographics from the Prolific records; completion time and the three parts of the payment, with the number drawn for the bonus in each treatment and their average bonus; the comprehension questions, retried until correct, as the mean number of attempts, and the three practice questions, paid on the first attempt only, as the share answered correctly the first time; and the two data-quality measures of the trials, the share of Personalized Menu decisions that take the strictly worse of the two known amounts and the number of submissions the Interactive Feedback takes before both stages' earnings are entered correctly. The last column tests equality across treatments on one value per participant, Kruskal-Wallis for means and medians and chi-squared for shares.

Attrition. Of the 820 who started, 118 did not finish. The instructions disclose the personalization rule on the Stage 3 page, so a dropout after that page could depend on the treatment. Of the 118, 43 stopped before reaching it, and none stopped on the page itself; the other 75 stopped later, 56 of them on the two screens that cannot be passed without entering the trial earnings correctly, the third practice question and the Interactive Feedback. Counting only the dropouts after the rule is disclosed, as a share of those who started, the IC Control loses 7% and Misreport Up 4%, and the five extension treatments between 7% and 12%. The demographic differences in Table B1, the female share above all, do not change the main comparison: with age, sex, a degree, and residence in the United States as controls, the four demographics with no missing values, the Misreport Up effect on $\tilde{d}$ in the main specification is 12.7 percentage points (standard error 3.3), against 13.3 (3.5) without them on the same 214 participants, and within each treatment women and men report alike, with mean $\tilde{d}$ of 7% and 7% in the control and 18% and 22% in Misreport Up.

Table B1: Sample, participants, payments, comprehension and data quality, by treatment.

	IC Control	Misreport Up	Misreport Down	Symmetric Uncertainty	Asymmetric Uncertainty	IC Upwards Bias	IC Uncertainty Bias	All	p
Sample									
Started	110	124	123	124	111	115	113	820	
Completed	98	116	100	103	86	101	98	702	
Dropout rate	10.9%	6.5%	18.7%	16.9%	22.5%	12.2%	13.3%	14.4%	0.011
Participants									
Age, mean	37.7	37.1	35.6	35.4	36.6	36.6	37.6	36.7	0.736
Female	60%	39%	46%	50%	49%	49%	58%	50%	0.038
Resident in the United States	49%	43%	46%	51%	51%	51%	43%	48%	0.723
Undergraduate degree or higher	77%	71%	58%	63%	69%	58%	62%	65%	0.053
Student	19%	21%	22%	24%	28%	23%	17%	22%	0.685
Employed full time	44%	52%	52%	56%	47%	44%	46%	49%	0.625
Time and payments									
Completion time, median (min)	37.2	41.2	41.3	40.7	43.2	35.4	47.7	40.9	<0.001
Practice bonus, mean ($)	1.27	1.27	1.19	1.25	1.21	1.36	1.20	1.25	0.041
Drawn for the bonus	13	13	10	12	10	15	7	80	0.744
Bonus if drawn, mean ($)	12.23	8.35	13.75	13.71	9.65	12.73	15.00	12.03	0.039
Total payment, mean ($)	9.89	9.20	9.57	9.85	9.33	10.25	9.28	9.62	0.017
Comprehension									
Comprehension question 1, attempts	1.17	1.14	1.19	1.17	1.15	1.20	1.16	1.17	0.949
Comprehension question 2, attempts	1.16	1.22	1.18	1.20	1.28	1.28	1.26	1.22	0.727
Comprehension question 3, attempts	1.85	1.33	1.26	1.15	1.22	1.25	1.17	1.32	<0.001
Comprehension question 4, attempts	1.31	1.59	1.40	1.59	1.41	1.47	1.37	1.45	0.004
Practice question 1, right first time	99%	94%	90%	96%	95%	98%	99%	96%	0.015
Practice question 2, right first time	84%	89%	78%	77%	72%	97%	73%	82%	<0.001
Practice question 3, right first time	71%	71%	70%	78%	74%	76%	68%	73%	0.724
All three right first time	68%	66%	59%	60%	56%	74%	56%	63%	0.063
Trials									
Dominance violations	5.7%	4.7%	7.0%	6.4%	5.5%	4.7%	9.5%	6.2%	0.432
Interactive Feedback, submissions	1.20	1.29	1.24	1.28	1.34	1.14	1.19	1.24	0.028
Interactive Feedback, right first time	91.3%	87.8%	90.7%	87.7%	88.4%	92.7%	91.3%	90.0%	0.016

Notes: Completers only, except the first block, which counts everyone who started. Demographics come from the Prolific export and are shares of the participants whose data were not withdrawn or expired. Comprehension questions follow each stage explanation and are retried until right; practice questions are paid on the first attempt. Dominance violations are shares of Personalized Menu decisions, and the Interactive Feedback counts the successful submission, so one submission means right first time. The p column tests equality across the seven treatments: Kruskal–Wallis for means and medians, chi-squared for shares and for dropout; the decision-level rows are tested on one value per participant.

C Pre-registration

The pre-analysis plan was written on a pilot and posted in the AEA RCT Registry on April 21, 2026, under ID AEARCTR-0018307. It fixes the seven treatments, a sample of about 100 participants per treatment from simulation-based power calculations on the pilot, the recruitment filters of Section 2.1, and the inclusion of every completer, and it names one primary analysis: for each treatment, the joint test that its effect relative to the IC Control is zero at every token, from a regression that pools the seven treatments on their common conflict tokens with token and trial fixed effects, treatment-by-token interactions, and standard errors clustered on the participant, equation (D1) below. Table 1 reports that test for every treatment, beside the average effect from the same equation without the

interactions, which column (4) of Table D1 repeats for Misreport Up. Everything else in the plan is exploratory.

The paper departs from the plan in four ways, none of which changes a conclusion. It measures the deviation as a share of the available distance, $\tilde{d}$, rather than in positions, so that the optimum in Misreport Up is 100% at every conflict token;[60] it fits the main comparison on the two main treatments and Misreport Up's own conflict tokens, the sample its figures show, with the pre-registered fit alongside in Table D1; it treats Misreport Down as an extension treatment rather than as a third main treatment; and it sorts participants into six misconception groups rather than the plan's four, a split the plan flags as one to examine, built from the quiz alone as the four were. The decomposition into margins, the use of the combined question to identify who reports accurately, and the comparison of each treatment's share at its optimum with the control's share on the same reports are not in the plan.

D Estimation

Every pooled test in Section 3 comes from one regression, fitted on the two main treatments over $\theta \in \{1, \ldots, 8\}$ in the main text and on all seven treatments over $\theta \in \{2, \ldots, 7\}$ in the pre-registered version. For participant i in trial r,

$$y_{ir} = \alpha + \sum_{k} \gamma_k \, \mathbb{1}\{\theta_{ir} = k\} + \sum_{\ell} \delta_\ell \, \mathbb{1}\{r = \ell\} + \sum_{j \neq C} \beta_j \, T_{j,i} + \sum_{j \neq C} \sum_{k} \lambda_{jk} \left(T_{j,i} \times \mathbb{1}\{\theta_{ir} = k\} \right) + \varepsilon_{ir}, \quad \text{(D1)}$$

with token fixed effects (the lowest token in the sample omitted), trial fixed effects (trial 1 omitted), treatment indicators $T_{j,i}$ (IC Control omitted), and the interaction of each treatment with each token, so that a treatment's effect may differ from one token to the next. The null that treatment j does nothing is the joint restriction $\beta_j = 0$ and $\lambda_{jk} = 0$ for all k, one restriction per token plus the main effect; testing β_j alone would ask only whether the treatments differ at the omitted token. Where the text reports one difference rather than a joint test, the estimate is β_j from the same equation without the interaction terms, the treatment's average effect over the sample. Standard errors are clustered on the participant throughout. The text reports the conventional cluster-robust F with the number of participants minus one as denominator degrees of freedom; with CR2 standard errors and Satterthwaite degrees of freedom (Pustejovsky and Tipton, 2018) the two joint tests of Section 3.1 are $F(8, 85.8) = 2.14$ ($p = 0.041$) for the control's mean $\tilde{d}$ being zero at every conflict token and $F(8, 102.8) = 132.81$ ($p < 0.001$) for Misreport Up's being at the

[60] In the plan's own specification, with the deviation in positions, the joint test for Misreport Up has $p = 0.008$.

optimum.[61]

Table D1 reports the Misreport Up effect on $\tilde{d}$ under four specifications: the two-treatment fit with and without the interactions, one observation per participant, and the pre-registered seven-treatment fit. Table D2 reports the margins of Section 3.2: the same regression with, as dependent variable, an indicator for a misreport, for an upward one, for a downward one, and for a report of 9, and the distance d and the normalized deviation $\tilde{d}$ on the subsamples of upward and of downward misreports.

Table D1: The treatment effect on the normalized deviation $\tilde{d}$.

	(1)	(2)	(3)	(4)
Misreport Up	13.3	7.9	13.9	12.9
	(3.5)	(4.6)	(3.4)	(3.6)
Treatment × token interactions	No	Yes	No	No
Joint test, treatment effect at every token (p)		0.007		
Joint test, interactions only (p)		0.317		
Token and trial fixed effects	Yes	Yes	No	Yes
Unit of observation	Decision	Decision	Participant	Decision
Treatments	2	2	2	7
Tokens	1–8	1–8	1–8	2–7
Mean $\tilde{d}$ in the IC Control	7.0	7.0	6.8	6.7
Observations	3,826	3,826	214	9,328
Participants	214	214	214	702

Notes: The dependent variable is $\tilde{d}$ in percentage points of the available distance, so a coefficient of 10 is a report a tenth of the way further toward 9 than in the IC Control. Each column reports the treatment indicators of one specification, with the IC Control omitted. Columns (1), (2) and (4) use every decision on the tokens shown, with token and trial fixed effects and CR2 standard errors clustered on the participant in parentheses. Column (2) adds the interaction of the treatment with each token; its Misreport Up coefficient is the effect at $\theta = 1$, and the two joint tests are HTZ Wald tests of the treatment main effect together with its seven interactions, and of the interactions alone. Column (3) uses one observation per participant, the participant's mean $\tilde{d}$ over tokens 1 to 8, with HC2 standard errors. Column (4) is the pre-registered specification, all seven treatments on the conflict tokens common to every treatment that is not incentive compatible; only its Misreport Up coefficient is shown here, and the coefficients of the five extension treatments are reported in Section 3.4.

[61]The tests weight decisions, as pre-registered, while the figures weight participants equally within a token cell. The two estimate the same quantity, because the token is drawn at random and every participant completed all twenty trials.

Table D2: The response decomposed by direction.

Outcome	IC Control	Misreport Up	Difference	p	Joint p
Misreports, $\mathbb{1}\{d \neq 0\}$	0.27	0.43	0.16	0.002	0.027
Upward, $\mathbb{1}\{d > 0\}$	0.20	0.38	0.18	< 0.001	0.002
Downward, $\mathbb{1}\{d < 0\}$	0.08	0.05	-0.02	0.242	0.744
At the optimum, $\mathbb{1}\{\hat{\theta} = 9\}$	0.05	0.13	0.07	0.010	0.004
Distance, d among upward misreports	2.73	2.82	0.09	0.768	0.576
Normalized deviation, $\tilde{d}$ among upward misreports	0.56	0.61	0.02	0.705	0.747
Normalized deviation, $\tilde{d}$ among downward misreports	-0.51	-0.47	0.04	0.443	0.179

E Learning

Within each treatment T and for each outcome Y, the deviation $\tilde{d}$, the share of reports at the token, and the share reporting 9, I estimate by OLS on $\theta \in \{1, \ldots, 8\}$

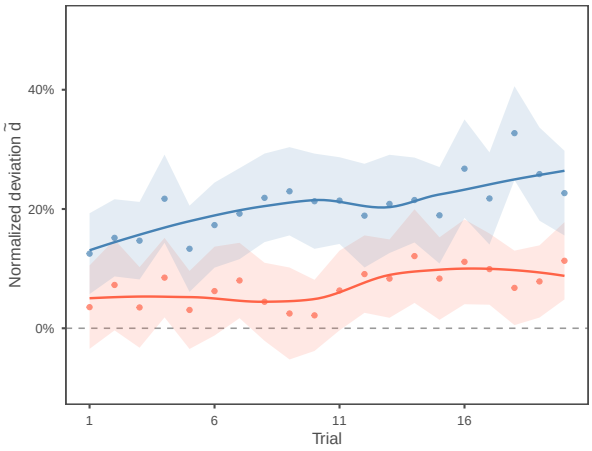
$$Y_{ir} = \alpha_T^Y + \beta_T^Y \cdot (r - 1) + u_{ir}, \tag{E1}$$

with CR2 standard errors clustered on the participant, so α_T^Y is the trial-one level and β_T^Y the per-trial change. Table E1 reports both for every participant (Panel A) and for the participants who report neither the token nor 9 in every trial, whose reports can move (Panel B). In Misreport Up $\tilde{d}$ rises from 15% in trial one to 26% in trial twenty, 0.6% per trial (0.2%, $p < 0.001$), which closes 14% of the gap to the optimum; the IC Control moves the same way from a lower level, and Misreport Up's slope is not distinguished from the control's (0.3% per trial larger, $p = 0.199$). The share of reports at the optimum stays between 8% and 18% throughout. These estimates pool every participant, including those who report the token in every trial and those who report the optimum in every trial, whose reports cannot move; Panel B repeats the regressions for the participants who do neither, 84 in Misreport Up and 51 in the IC Control. Figure E1 shows both samples trial by trial, every participant in the top row and the participants who move in the bottom row: a modest rise over the twenty trials that stops far from the optimum, and shares that move little.

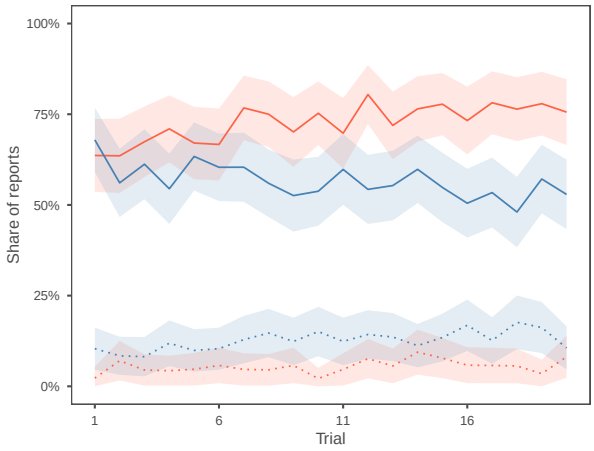
Table E1: Learning slopes: each outcome regressed on the trial, by treatment and sample, with standard errors in parentheses below each slope.

Outcome	Treatment	Trial one	Slope	p	Decisions	Participants
Panel A: all participants						
Deviation $\tilde{d}$	IC Control	4.1%	0.30	0.053	1,757	98
			(0.15)			
	Misreport Up	14.8%	0.61	< 0.001	2,069	116
			(0.18)			
Share at the token	IC Control	66.5%	0.66	0.001	1,757	98
			(0.20)			
	Misreport Up	61.4%	-0.50	0.032	2,069	116
			(0.23)			
Share reporting 9	IC Control	4.3%	0.12	0.186	1,757	98
			(0.09)			
	Misreport Up	10.0%	0.28	0.060	2,069	116
			(0.15)			
Panel B: participants who move						
Deviation $\tilde{d}$	IC Control	7.9%	0.56	0.056	927	51
			(0.29)			
	Misreport Up	16.9%	0.83	< 0.001	1,505	84
			(0.24)			
Share at the token	IC Control	36.2%	1.28	< 0.001	927	51
			(0.34)			
	Misreport Up	50.4%	-0.69	0.030	1,505	84
			(0.31)			
Share reporting 9	IC Control	8.2%	0.23	0.201	927	51
			(0.18)			
	Misreport Up	10.3%	0.38	0.063	1,505	84
			(0.20)			

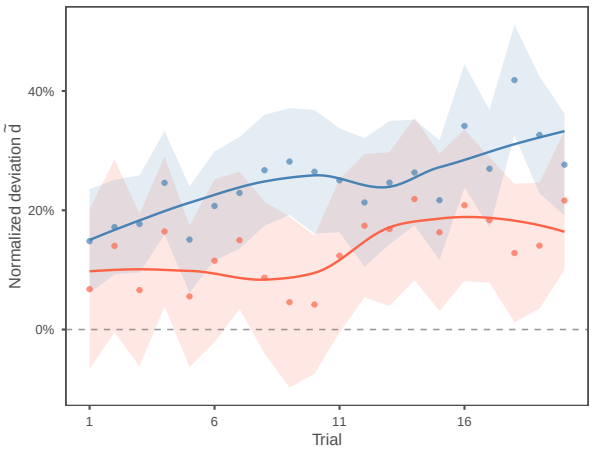
Figure E1: Average $\tilde{d}$ and report shares by trial.



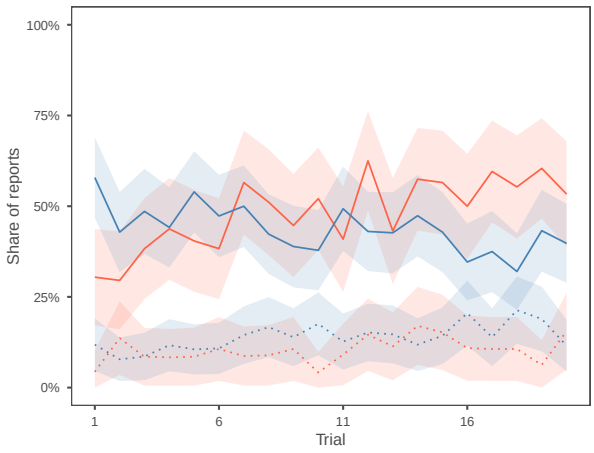
(a) All participants: normalized deviation $\tilde{d}$



(b) All participants: share at the token (solid) and share reporting 9 (dotted)



(c) Participants who move: normalized deviation $\tilde{d}$



(d) Participants who move: share at the token (solid) and share reporting 9 (dotted)

Notes: Both treatments on tokens 1 to 8. Top row: every participant. Bottom row: the participants who report neither the token in every trial nor 9 in every trial, 51 in the IC Control and 84 in Misreport Up. Left panels: mean $\tilde{d}$ by trial, with 95% confidence intervals and a loess fit, on the same axes in both rows. Right panels: the share of reports equal to the token (solid) and the share equal to 9 (dotted), with 95% confidence intervals. On these tokens 9 is the optimum in Misreport Up and neither the optimum nor the token in the IC Control.

F The quiz and the general traits

The three questions on the incentives. Each is put in the format of the trials: it shows a token worth $5.00 and the price list of the Measurement Stage, restates the participant's own personalization rule, and asks for the switch point that maximizes one component of expected trial earnings, with a button below the list, "all switch points give the same earnings", for the case in which that component does not depend on the report. The *measurement question* asks for the switch point that maximizes expected Measurement Stage earnings, ignoring the Personalized Menu; the *menu question* for the one that maximizes expected Personalized Menu earnings, ignoring the Measurement Stage; and the *combined question* for the one that maximizes expected total trial earnings. Each pays $10.00 if correct and is eligible for the bonus draw, and each page carries a button that expands to the definition of expected earnings used in the instructions and to the reminder of how the relevant stages pay. In Misreport Up the correct answers are 5, 9, and 9; in the IC Control they are 5, the button, and 5.

Table F1 defines the six misconception groups from the three answers and gives their sizes in Misreport Up. The groups are mutually exclusive and exhaustive, since all 116 participants answered all three questions.

Table F1: The six misconception groups.

Group	Measurement	Menu	Combined	Misreport Up
Integrates both stages	any	any	correct	18
Neglects the Personalized Menu	correct	wrong	wrong	42
Neglects the Measurement Stage	wrong	correct	wrong	10
Understands both, misreports partially	correct	correct	a misreport	4
Understands both, reports accurately	correct	correct	the token or the button	21
Understands neither	wrong	wrong	wrong	21

Notes: Each row is defined by the answers to the three questions; the last column is the number of participants in Misreport Up. Of those who integrate both stages, 39% also answer both component questions correctly. Of those who understand both and report accurately, 18 named the token and 3 pressed the button.

The three incentivized items on general traits. Each pays its realized outcome if drawn for the bonus and yields one indicator.

Narrow framing: a choice between two 50-50 coins, asked twice in payoff-equivalent versions at the first and last positions of the quiz, in an order randomized across participants.

In the background version the participant is given \$10 and chooses between a coin paying \$3 or \$7 and a coin paying \$1 or \$10; in the integrated version the \$10 is folded into the outcomes, \$13 or \$17 against \$11 or \$20. The indicator equals 1 if the participant chooses differently in the two versions; it is unrelated to reporting the token in either treatment (IC Control $r = -0.052$, $p = 0.614$; Misreport Up $r = -0.147$, $p = 0.116$).

Backward induction: a two-move game against the computer, played once. The participant chooses Left or Right; the computer then picks Up or Down on that branch, always the option that pays the computer more, which is Down on both branches. The four terminal payoffs are hidden behind boxes the participant can click to reveal. Anticipating the computer, the participant should choose Left; the participant's largest possible payoff sits at Right followed by Up. The indicator equals 1 if the participant does anything other than reveal all four boxes and choose Left. In the IC Control it is strongly related to reporting the token ($r = -0.509$, $p < 0.001$): the 36 participants who fail the item, 19 of them without opening all four boxes, report the token in 48% of their trials, against 87% for the 62 who solve it.

The relation is not specific to the report. In the control, leaving the token is a mistake, and the participants who fail the item make more mistakes in the other parts of the trial as well: they take the dominated alternative in 11% of their Personalized Menu decisions, against 3%, and their reports lie 1.5 positions from the token on average, against 0.3. In Misreport Up the participants who fail the item also take the dominated alternative more often, in 9% of their menu decisions against 2%, but there leaving the token is both a mistake and the direction personalization favors, so the participants who solve the item leave it too, and the correlation with reporting the token is near zero ($r = -0.061$, $p = 0.516$). Participants who fail the item report the token less often in every one of the seven treatments, by 6 to 37 percentage points over all nine tokens, the least in Misreport Up and the most in the IC Control.

The treatment effect depends on the item. Fitting the main specification, column (1) of Table D1, within each group: among the 133 participants who solve the item, Misreport Up raises $\tilde{d}$ from 2% in the control to 23%, by 21 percentage points (standard error 3.8, $p < 0.001$), and moves the share at the token from 87% to 59% ($p < 0.001$); among the 81 who fail it, the two treatments are not distinguished, with $\tilde{d}$ at 15% and 16% (a difference of 1 on the same scale, standard error 6.6, $p = 0.918$) and the share at the token at 48% and 53% ($p = 0.510$). Figure F2 shows the two comparisons token by token, in the format of panel (b) of Figure 5, in its top row; the other rows split the participants by the three other general-trait measures, and in none of them do the two halves differ as they do here. The response to the Personalized Menu in Section 3.1 therefore comes from the participants who solve the item; those who fail it report alike under the two rules. Within Misreport Up, however, the item

does not separate the participants who stay at the token from those who move: 13% of those who fail it report the token in every trial, against 32% of those who solve it, and their mean $\tilde{d}$ is 16% against 23%, which is what the coefficient in Table F3 reflects. The item does not explain the failure to best respond in Misreport Up; it identifies, in both treatments, the participants whose reports do not follow the incentives of either rule.

Table F2 extends the split to the seven treatments: the regression of Table 1, on tokens 2 to 7, fitted within the 424 participants who solve the item and within the 278 who fail it. Figure F1 plots their coefficients in the format of Figure 8. As in Table 1, I also report the p-values adjusted for the six treatments with Holm's method, within each group. The panels read as follows, and Section 3.4 summarizes them.

Among the participants who solve the item, the control is at the token, with mean $\tilde{d}$ of 2% and 88% of reports at it, every treatment that is not incentive compatible is distinguished from it before the adjustment and all but Misreport with Symmetric Uncertainty after it, and the treatments order as the direction personalization favors would predict: Misreport Up at 20.8 percentage points of $\tilde{d}$, Misreport with Asymmetric Uncertainty at 13.6, and Misreport with Symmetric Uncertainty at 5.0 ($p = 0.055$).

The response under symmetric uncertainty is therefore smaller than under certainty, and the evidence that there is one at all is suggestive: its share at the token differs from the control's by -12.2 points before adjusting for the six treatments ($p = 0.026$) but not after ($p = 0.078$). It also has a direction: among these participants the reports that leave the token go up on both sides of token 5, 15% against the control's 5% above it, where 9 is optimal, and 22% against 10% below it, where 1 is optimal and only 4% move down. They respond to the Misreport Up half of the draw and not to the Misreport Down half.

One qualification runs the other way: among those who solve the item, IC with Uncertainty Bias, where the token stays optimal, has a share at the token -11.5 points from the control's ($p = 0.049$, Holm $p = 0.099$), while IC with Upwards Bias has no detectable effect, so a draw in the menu may move a few reports off the token even where it should not.

Among the participants who fail the item, no treatment is distinguished from the control on $\tilde{d}$, and on the share at the token only IC with Upwards Bias is, by 18.8 points ($p = 0.040$, Holm $p = 0.242$), one of twelve comparisons: the noise they add is alike across treatments.

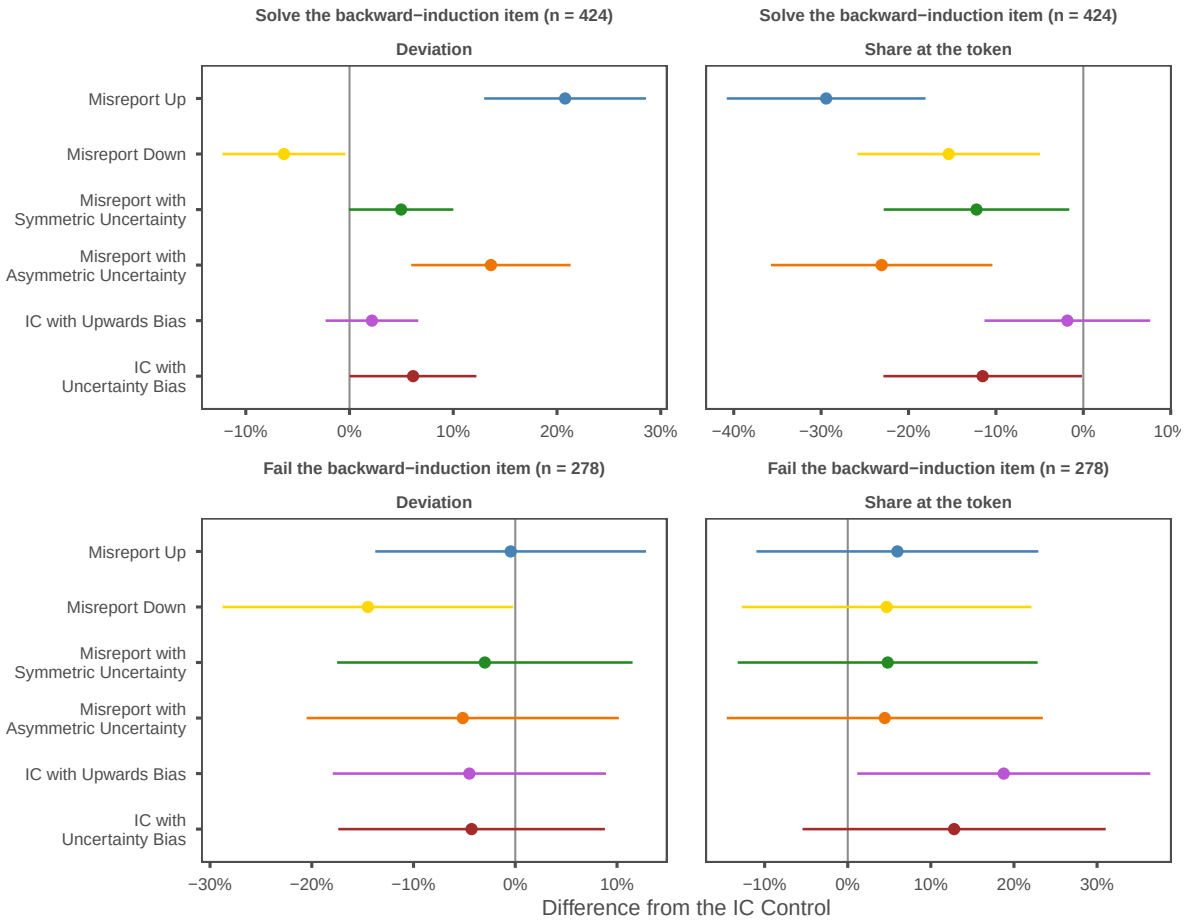
Lying aversion: a wheel selects a number from 1 to 9, each equally likely, which the participant is told to remember; on the next screen they report it by clicking one of ten boxes numbered 1 to 10 and are paid the number they report, so inflating the report pays strictly more. The indicator equals 1 if the report equals the draw; 71% of participants report the draw, and its correlation with reporting the token is $r = 0.176$ in the IC Control ($p = 0.084$) and $r = -0.049$ in Misreport Up ($p = 0.603$).

Table F2: Every treatment against the IC Control, within the participants who solve the backward-induction item and within those who fail it.

Treatment	N	Mean $\tilde{d}$	Difference	p	Share at token	Difference	p
Panel A: solve the backward-induction item							
IC Control	62	2%	—	—	88%	—	—
Misreport Up	71	23%	20.8	<0.001	58%	-29.4	<0.001
			(4.0)	[<0.001]		(5.8)	[<0.001]
Misreport Down	62	-5%	-6.3	0.038	72%	-15.4	0.005
			(3.0)	[0.154]		(5.3)	[0.018]
Misreport with Symmetric Uncertainty	63	7%	5.0	0.055	75%	-12.2	0.026
			(2.6)	[0.154]		(5.4)	[0.078]
Misreport with Asymmetric Uncertainty	51	15%	13.6	<0.001	64%	-23.1	<0.001
			(3.9)	[0.004]		(6.5)	[0.003]
IC with Upwards Bias	60	4%	2.2	0.344	86%	-1.8	0.707
			(2.3)	[0.344]		(4.8)	[0.707]
IC with Uncertainty Bias	55	8%	6.1	0.051	76%	-11.5	0.049
			(3.1)	[0.154]		(5.8)	[0.099]
Panel B: fail the backward-induction item							
IC Control	36	15%	—	—	47%	—	—
Misreport Up	45	15%	-0.5	0.947	53%	6.0	0.492
			(6.8)	[1.000]		(8.7)	[1.000]
Misreport Down	38	0%	-14.5	0.051	52%	4.7	0.600
			(7.3)	[0.304]		(8.9)	[1.000]
Misreport with Symmetric Uncertainty	40	12%	-3.0	0.688	52%	4.8	0.603
			(7.4)	[1.000]		(9.2)	[1.000]
Misreport with Asymmetric Uncertainty	35	10%	-5.2	0.512	52%	4.5	0.647
			(7.8)	[1.000]		(9.7)	[1.000]
IC with Upwards Bias	41	11%	-4.5	0.512	66%	18.8	0.040
			(6.9)	[1.000]		(9.0)	[0.242]
IC with Uncertainty Bias	43	11%	-4.3	0.523	60%	12.8	0.173
			(6.7)	[1.000]		(9.3)	[0.866]

Notes: As Table 1, fitted separately within the participants who solve the backward-induction item of the quiz (Panel A) and within those who fail it (Panel B): decisions on tokens 2 to 7, each Difference the treatment's coefficient in percentage points, with the IC Control omitted and token and trial fixed effects, CR2 standard errors clustered on the participant in parentheses, and p its p-value; in brackets, the p-value adjusted for the six treatments of the panel by Holm's step-down method.

Figure F1: Each treatment against the IC Control, within the participants who solve the backward-induction item and within those who fail it.



Notes: As Figure 8, from the two regressions of Table F2: the treatment coefficients on tokens 2 to 7, with the IC Control as the omitted category, token and trial fixed effects, and CR2 standard errors clustered on the participant, with 95% confidence intervals. Top row: the participants who reveal all four payoff boxes and choose Left in the backward-induction item; bottom row: the rest. Deviation is $\tilde{d}$ and share at the token an indicator for reporting it.

The weight placed on future consequences. The exit survey, not incentivized, carries the ultra-short Consideration of Future Consequences scale of Vilar et al. (2022): six statements, each on its own page in an order randomized across participants, answered from 1 ("Not at all like you") to 7 ("Very much like you"). Three statements form the CFC-Future subscale and three the CFC-Immediate subscale; each is averaged within participant, and the regression below uses CFC-Future, standardized.

The regression. For participant i in Misreport Up, with $\bar{d}_i$ their mean $\tilde{d}$ over the conflict tokens,

$$\bar{d}_i = \alpha + \sum_{g \neq \text{integrates}} \beta_g \, \mathbb{1}\{\text{group}_i = g\} + \gamma_1 \, \text{NF}_i + \gamma_2 \, \text{BI}_i + \gamma_3 \, \text{LA}_i + \gamma_4 \, \text{CFC}_i + u_i, \qquad \text{(F1)}$$

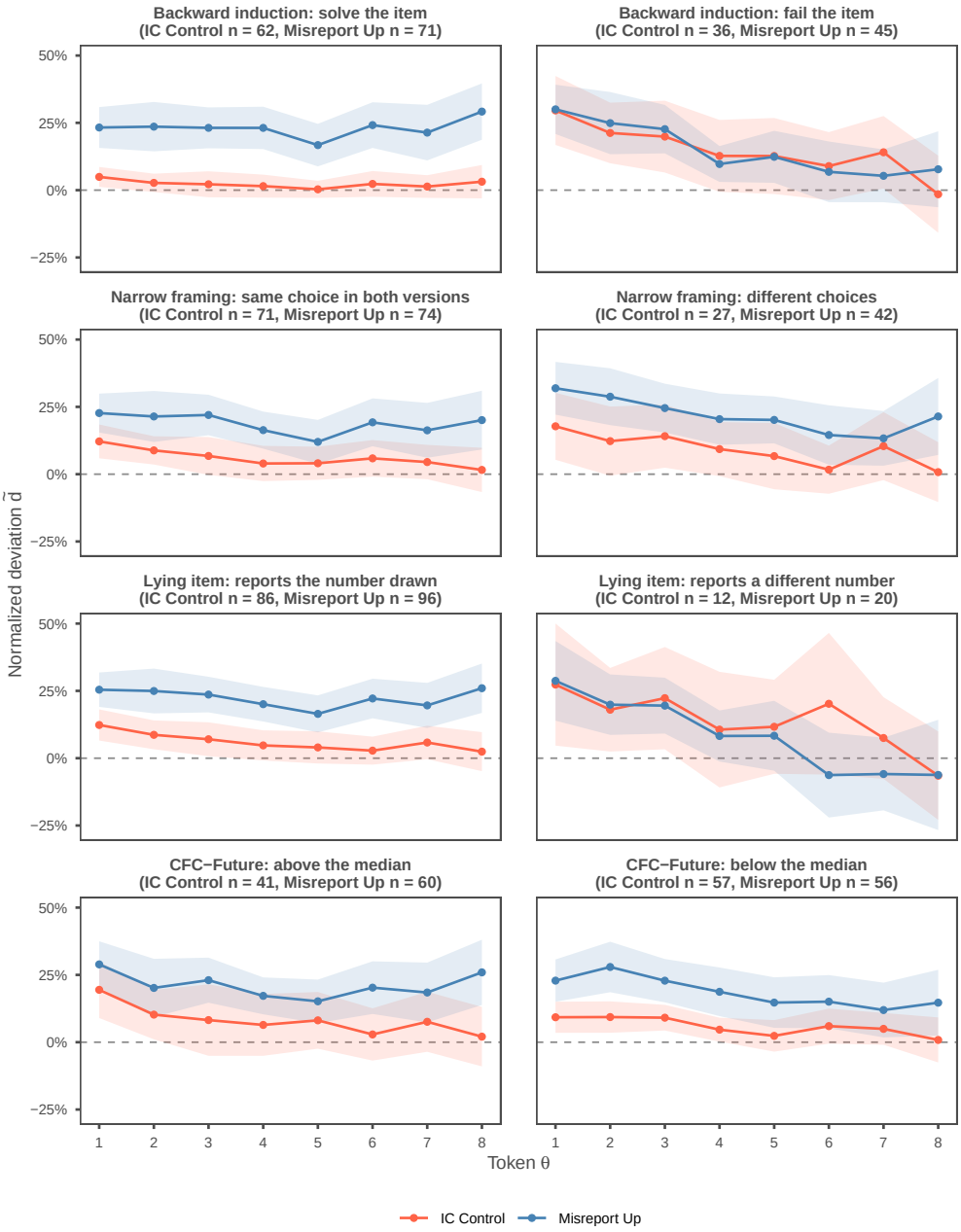
where the group indicators run over the five misconception groups other than the participants who integrate both stages, NF_i, BI_i, and LA_i are the three indicators above, and CFC_i is the standardized CFC-Future score. Table F3 fits the group indicators alone, the four traits alone, and both, one observation per participant with HC2 standard errors. Relative to the participants who integrate both stages, those who neglect the Personalized Menu and those who understand both and report accurately lie 44 and 38 percentage points closer to the token in column (3). Of the four traits, only reporting honestly on the lying item is distinguishable from zero, and it runs against its prediction: honest reporters lie 11 percentage points further from the token in the full specification, column (3) ($p = 0.029$).

Table F3: Participant-level $\tilde{d}$ in Misreport Up on the six misconception groups and the four general traits.

	(1)	(2)	(3)
Neglects the Personalized Menu	-44.7		-44.1
	(8.6)		(8.3)
Neglects the Measurement Stage	-25.9		-23.2
	(13.8)		(13.5)
Understands both, misreports partially	-3.3		-3.3
	(15.7)		(14.7)
Understands both, reports accurately	-38.0		-37.5
	(9.3)		(9.1)
Understands neither	-32.7		-31.0
	(9.5)		(9.3)
Narrow framing		1.9	1.9
		(5.1)	(4.2)
Backward induction fails		-5.5	-4.6
		(5.1)	(4.5)
Reported honestly		11.2	10.7
		(4.6)	(4.8)
CFC-Future (standardized)		3.2	3.2
		(2.1)	(1.8)
Joint test, five misconception-group dummies (p)	<0.001		<0.001
Joint test, four traits (p)		0.029	0.028
R^2	0.34	0.05	0.39
Participants	116	116	116

Notes: One observation per participant in Misreport Up: the participant's mean $\tilde{d}$ over the conflict tokens, in percentage points of the available distance. The omitted misconception group is the participants who integrate both stages, whose mean is 52.0. Narrow framing, failing backward induction and reporting honestly are indicators for the quiz items of Section 2.5; CFC-Future is the exit-survey subscale, standardized over the participants in the regression. HC2 standard errors in parentheses; the joint tests are Wald tests of the misconception-group dummies and of the general-trait terms.

Figure F2: Normalized deviation by token, within the two groups of each general-trait measure.

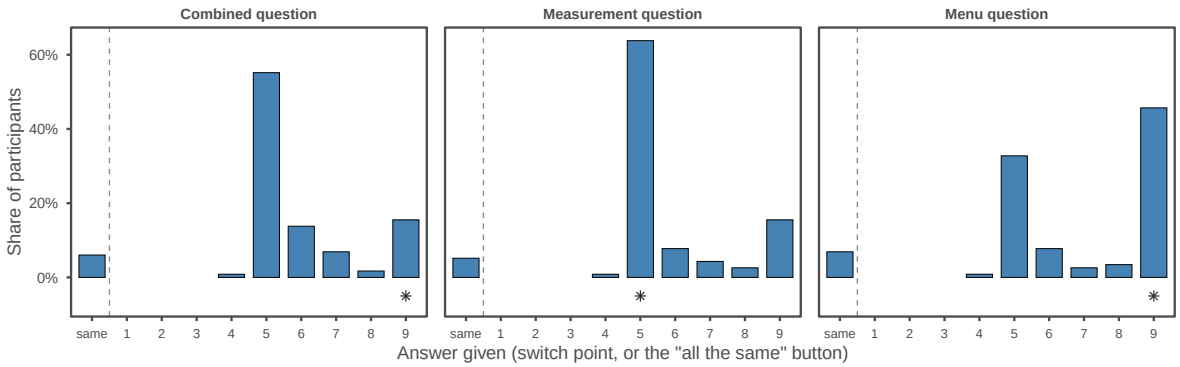


Notes: As panel (b) of Figure 5: the average normalized deviation $\tilde{d}$ at each of the eight conflict tokens, in each of the two main treatments, with every point a mean of participant means within the token cell and bands the 95% confidence intervals that go with it. Each row splits the participants by one of the four general-trait measures of Section 2.5, and the two panels of a row partition them, with the number of participants from each treatment in the panel's title. Top row: those who reveal all four payoff boxes and choose Left in the backward-induction item, and the rest. Second row: the same choice in the two versions of the coin item, or different choices. Third row: the lying item answered with the number drawn, or with another number. Bottom row: CFC-Future above or below its median over the two treatments. All panels share the vertical axis.

G The quiz answers in detail

Section 3.3 says how often participants are right; Figure G1 shows what they answer instead, and the wrong answers are not random. On the combined question the modal answer is 5, the token: the most common failure is to answer as if the Personalized Menu were absent. On the menu question the modal answer is 9, which is correct, so the same participants who cannot state the integrated optimum can state the Personalized Menu optimum on its own. This is also what motivates splitting the participants who understand both stages: wrong answers differ in direction, some pointing at the token and some away from it.

Figure G1: Misreport Up: the distribution of answers to the three questions.



Notes: Share of the 116 participants in Misreport Up giving each answer, by question. The star under the axis marks the correct answer. The button sits to the left of the switch points, behind a dashed rule, because it is a different kind of answer rather than a tenth position on the scale. Every position is shown even where nobody chose it.

Figure G2 starts from behavior instead: on the x axis is the distribution of the participant-level mean $\tilde{d}$ from Figure 6, in coarser bins, the width of each bar is the number of participants in it, and the stacked colors are the six groups. Section 3.3 reads its two ends, the zero bin, whose 30 participants are the 29 who report the token in every trial and one whose deviations offset, and the 16 who cover more than half the available distance; between them the groups are more evenly mixed.

The relation between the quiz and the trials does not depend on how the groups are drawn. Read simply as a number on the scale of the trials, with the token and the button at 0 and a report of 9 at 100%, each answer correlates with the participant's mean $\tilde{d}$ over the 116 participants in Misreport Up: $r = 0.59$ for the combined question, 0.42 for the menu question, and 0.38 for the measurement question, all with $p < 0.001$. Figure G3 shows the combined question participant by participant.

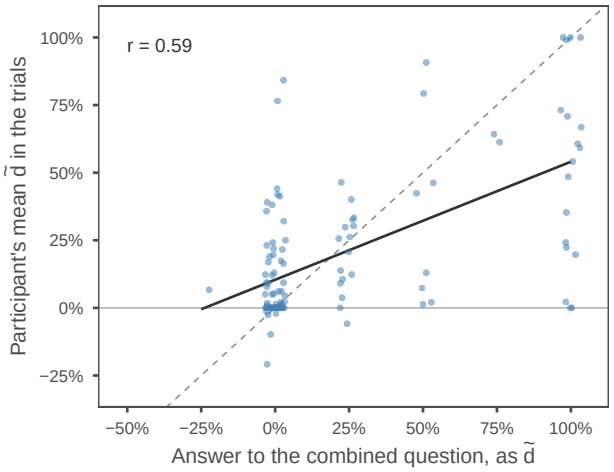
Panel (b) of Figure 7 pools all twenty trials, which leaves open whether the groups describe where participants ended up or where they began. Figure G4 plots each group trial by trial:

Figure G2: Misconception groups by participant deviation.



Notes: Misreport Up. Columns are bins of the participant mean $\tilde{d}$ from Figure 6, with width proportional to the number of participants, whose count is printed above. Within a column the fill gives the share in each misconception group. Bins are right-closed and zero is its own bin. Shares below 9% are unlabeled.

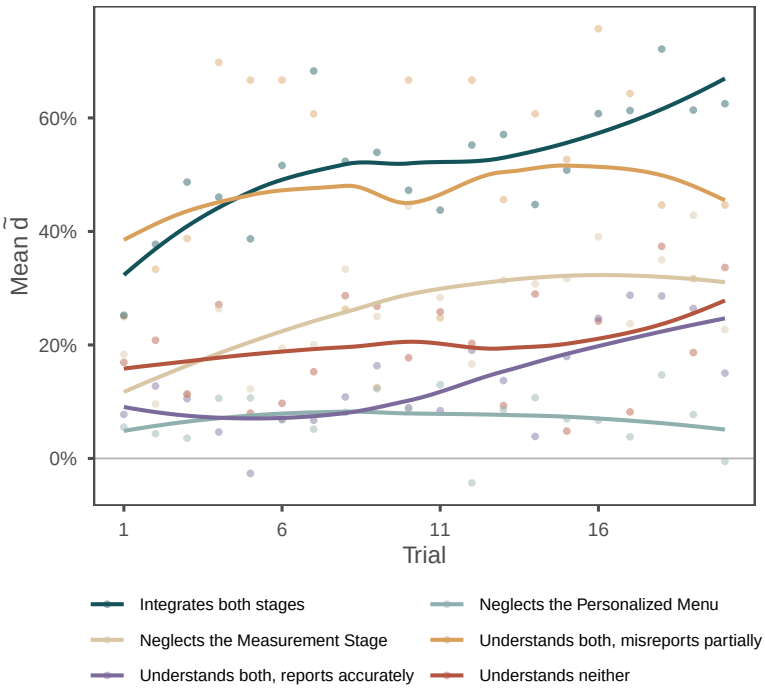
Figure G3: Misreport Up: the answer to the combined question and the participant's mean $\tilde{d}$.



Notes: Each point is a participant, with their answer to the combined question on the horizontal axis and their mean $\tilde{d}$ over their trials on the conflict tokens on the vertical axis. The answer is normalized as the trials are, at the quiz token of 5: an answer of 9 is 100%, and the token and the button reading "all switch points give the same earnings" are 0. Points are jittered sideways only. The dashed line is where the answer and the behavior coincide and the solid line is the least-squares fit.

the groups are already separated in the first trials, before much feedback has accumulated, and remain ordered throughout. Those who neglect the Personalized Menu barely move and stay close to the token; those who integrate both stages move the most. Whether the groups then move apart at different rates the data do not establish: fitting a linear trend within each group, with CR2 standard errors clustered on the participant, the joint test that the six slopes are equal does not reject ($F(5, 24.5) = 1.54$, $p = 0.216$), nor does a one-way F on one slope per participant ($F(5, 110) = 1.65$, $p = 0.152$); with 4 to 42 participants in a group, the slopes are not estimated precisely enough to separate.

Figure G4: Average $\tilde{d}$ by misconception group and trial.



Notes: Misreport Up, the six misconception groups of panel (b) of Figure 7 in the same colors. Each point is the mean $\tilde{d}$ of the group within the trial, with a loess through the points, since the per-trial cell means are noisy on groups this small.

H The role of feedback: costly misreports

A misreport costs in the Measurement Stage only when the amount on the drawn row lies between the token and the report, and the Interactive Feedback then shows it. Cason and Plott (2014) find that such a costly draw pulls the next offer onto the value, and read it as participants learning their way out of a misconception about the BDM. The pre-analysis plan includes their test: among pairs of consecutive trials that start from a misreport, compare

how the next report moves after a costly draw and after one that did not cost. Read at face value, in the unconditional rows of Table H1, a cost brings the next report back toward the token.

The comparison does not identify the effect of the cost, because the two groups of misreports differ before any feedback has acted. The drawn row is uniform over the ten rows and only the rows between the token and the report produce a cost, so a report d positions from the token costs with probability $|d|/10$: costly misreports are further from the token than non-costly ones, and distance predicts the next move on its own. A report at the far end of the scale can only stay or come back, while a report one position off has the whole scale to move away into and little room to move closer except by moving onto the token, and the two come from different participants: a one-position misreport is usually a slip by someone who reports the token in most trials, and a far one usually comes from someone who rarely does. The unconditional comparison therefore mixes the effect of the cost with the effect of the distance, in the direction that makes a cost look as if it pulled the report back.

Given the token and the report, the drawn row is the only thing that decides whether a misreport costs, and it is random. Within a cell defined by the token θ and the deviation d, every pair has the same chance of a cost and the same room to move, so costly and non-costly pairs in a cell are a randomized comparison. The corrected test compares them within cells and pools across cells by the Cochran-Mantel-Haenszel statistic. Table H1 sets the two tests side by side, in the layout of Cason and Plott's Table 3: the next report moves onto the token, moves toward it, stays as far, or moves away. Pooling the seven treatments, 22% of reports move away after a cost against 34% without one, and the unconditional test rejects ($p < 0.001$); within cells the difference is not distinguished from zero ($p = 0.566$). The share moving onto the token is 16% against 19% and does not differ under either test ($p = 0.248$ and $p = 0.434$). What survives the conditioning is a small partial return in the pooled sample: the share moving closer, onto or toward the token, has $p = 0.042$ within cells, and the share staying as far has $p = 0.039$. Treatment by treatment, the unconditional test rejects in several columns, but within cells no column separates in any of the seven treatments except the share staying as far in Misreport Down.

The same correction applied to Cason and Plott's own data gives a different answer: there the jump onto the value survives the conditioning and sharpens, from $p = 0.077$ to $p = 0.011$, while moving away less often does not survive it ($p = 0.408$). The confound produces the same raw pattern in both data sets; what remains once it is removed is theirs and not this experiment's, and the difference is where participants start. In Cason and Plott's data 16.7% of first offers are at the value and 31.1% of second ones, and the misses are consistent with the misconception that the BDM pays the offer rather than the posted price. Here 68.7% of

Table H1: Cason and Plott's Table 3 in this experiment.

BDM cost	N	Onto the token	Toward	Same	Away
All seven treatments (tokens 2 to 7)					
Costly	488	79 (16%)	177 (36%)	127 (26%)	105 (22%)
Not costly	1470	274 (19%)	410 (28%)	291 (20%)	495 (34%)
Fisher p		0.248	<0.001	0.004	<0.001
Within (θ, d) p		0.434	0.123	0.039	0.566
IC Control (tokens 1 to 8)					
Costly	100	17 (17%)	32 (32%)	31 (31%)	20 (20%)
Not costly	320	69 (22%)	88 (28%)	55 (17%)	108 (34%)
Fisher p		0.395	0.378	0.004	0.009
Within (θ, d) p		0.460	0.627	0.884	0.997
Misreport Up (tokens 1 to 8)					
Costly	198	36 (18%)	60 (30%)	55 (28%)	47 (24%)
Not costly	553	89 (16%)	146 (26%)	129 (23%)	189 (34%)
Fisher p		0.506	0.308	0.212	0.007
Within (θ, d) p		0.348	0.624	0.392	0.516
Misreport Down (tokens 2 to 7)					
Costly	66	14 (21%)	23 (35%)	11 (17%)	18 (27%)
Not costly	227	54 (24%)	59 (26%)	50 (22%)	64 (28%)
Fisher p		0.742	0.164	0.393	1.000
Within (θ, d) p		0.840	0.302	0.006	0.071
Misreport with Symmetric Uncertainty (tokens 2 to 8)					
Costly	68	17 (25%)	19 (28%)	17 (25%)	15 (22%)
Not costly	309	69 (22%)	78 (25%)	59 (19%)	103 (33%)
Fisher p		0.634	0.648	0.316	0.083
Within (θ, d) p		0.373	0.146	0.635	0.252
Misreport with Asymmetric Uncertainty (tokens 1 to 8)					
Costly	166	20 (12%)	66 (40%)	42 (25%)	38 (23%)
Not costly	343	57 (17%)	107 (31%)	57 (17%)	122 (36%)
Fisher p		0.190	0.059	0.023	0.004
Within (θ, d) p		0.649	0.825	0.841	0.434
IC with Upwards Bias (tokens 1 to 8)					
Costly	82	16 (20%)	38 (46%)	13 (16%)	15 (18%)
Not costly	238	57 (24%)	73 (31%)	30 (13%)	78 (33%)
Fisher p		0.449	0.015	0.457	0.016
Within (θ, d) p		0.181	0.350	0.206	0.188
IC with Uncertainty Bias (tokens 1 to 8)					
Costly	147	20 (14%)	54 (37%)	31 (21%)	42 (29%)
Not costly	332	54 (16%)	113 (34%)	44 (13%)	121 (36%)
Fisher p		0.496	0.604	0.040	0.096
Within (θ, d) p		0.953	0.051	0.702	0.054

Notes: The layout of Table 3 in Cason and Plott (2014), on the trial pairs defined in the text. Columns classify the trial-$(r+1)$ report by $|\tilde{d}_{r+1}|$ against $|\tilde{d}_r|$: zero (onto the token), smaller (toward), equal (same) or larger (away); shares are of the row. Fisher: their test, the exact test of costly against not costly on that column against the other three, unconditional. Within (θ, d): the same comparison stratified by the token and the signed deviation of the trial-r report, within which the draw assigns the cost at random.

participants in the three incentive compatible treatments report the token in the first trial, and over all nine tokens, 1,464 pairs, the jump onto the token after a cost is 20% against 23% without one ($p = 0.119$ within cells); the rows of Table H1 for the three treatments keep tokens 1 to 8 and give the same answer. A reading consistent with both data sets is that their participants arrive with a misconception about the BDM that a costly draw corrects, while most participants here report the token from the first trial, so that the misreports that remain are of a different kind. The comparison does not identify that difference: the within-cell tests here do not detect a return onto the token after a cost, but they cannot rule out a smaller one than theirs, and a test that rejects in one data set and not in the other does not establish that the effect differs.

I The screens

This section shows every page of the study as a participant saw it, in order, from the welcome page to the payment summary. The screenshots come from a test session of Misreport with Symmetric Uncertainty.[62] The interface numbers the stages and does not name them: Stage 1 is the Induced Value, Stage 2 the Measurement Stage, Stage 3 the Personalized Menu, and Stage 4 the Interactive Feedback. It calls the report the *Switch Point* and the quiz the *Follow-Up Questions*.[63]

The pages are the same in every treatment except where the personalization rule enters: the Stage 3 instruction page (Screen 17), the comprehension question on Stage 3 (Screen 18), practice question 2 (Screen 30), the Stage 3 choice itself (Screens 31 and 39), the feedback screen (Screen 40), and the rule reminders in the follow-up questions (Screens 46 to 54). Table I1 gives the sentence that the Stage 3 instruction page shows in each treatment; its example shows the two alternatives as buttons labeled as the rule defines them, a generic Token T, or \$10.00 in IC with Upwards Bias and IC with Uncertainty Bias, beside "\$5.00", "Switch Point", or "\$10.00 minus Switch Point", and in the three treatments with a draw it shows both possible pairs, each with its probability.

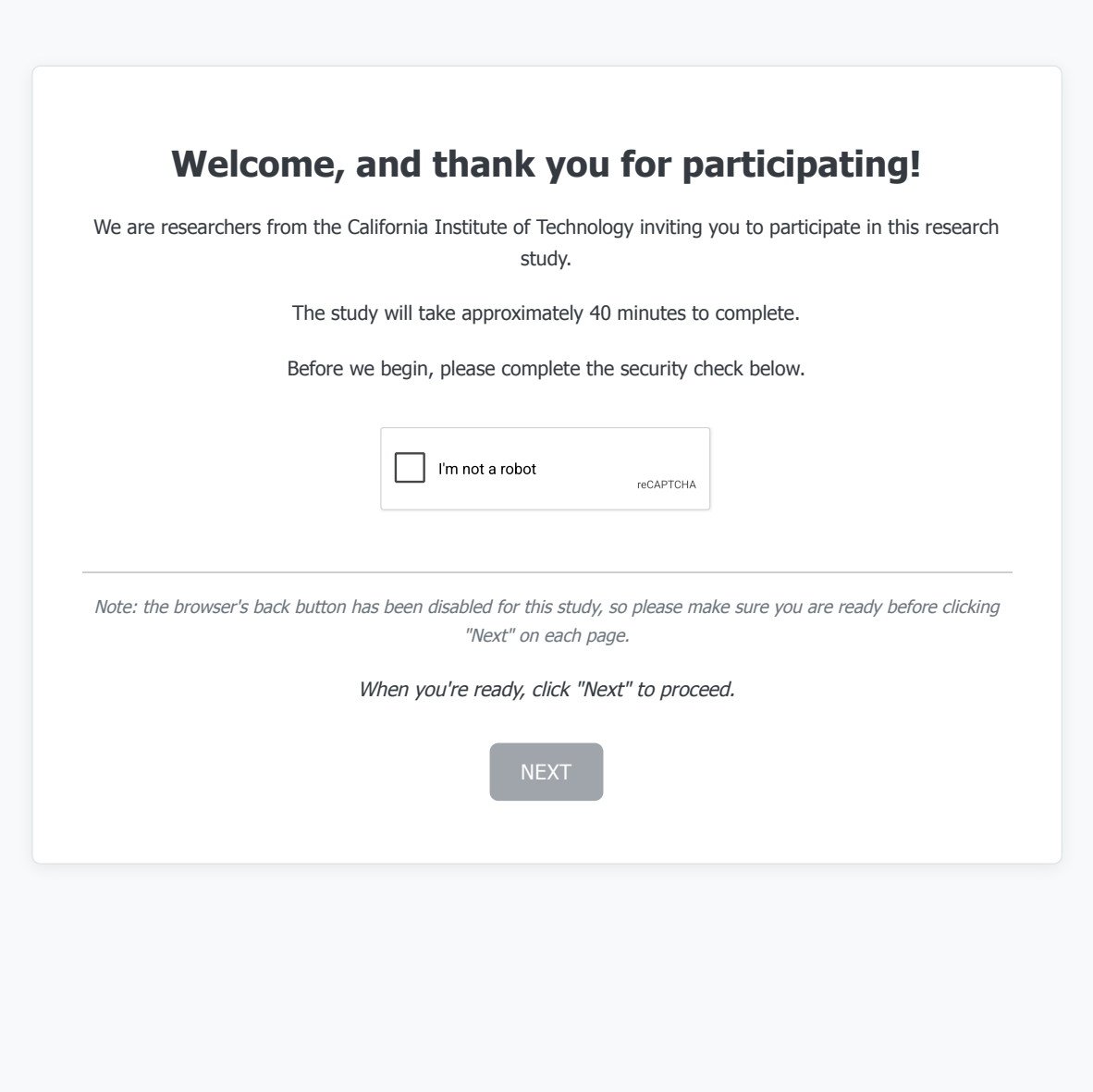
[62] The session had a single paid trial, so the trial pages read "Trial 1 of 1" where participants read "Trial k of 20"; everything else is as in the experiment.

[63] Each instruction page reveals its blocks one at a time with a "Show More" button, every page ends with a "Next" button, and the browser's back button is disabled.

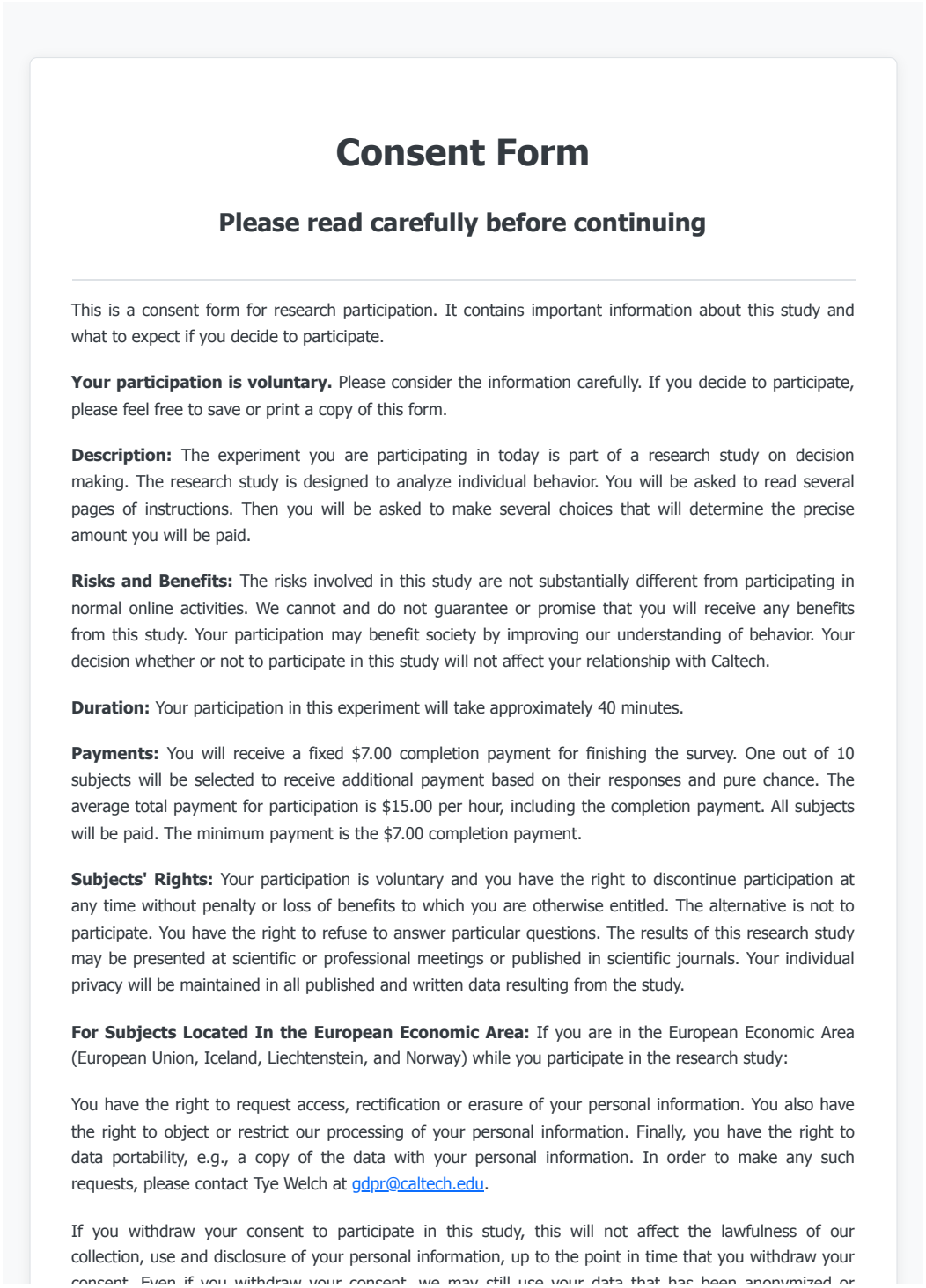
Table I1: The Stage 3 instruction, by treatment.

Treatment	"In this stage, you will choose between..."
IC Control	the Trial's Token and $5.00.
Misreport Up	the Trial's Token and your Switch Point from Stage 2.
Misreport Down	the Trial's Token and an amount calculated as $10.00 minus your Switch Point.
Misreport with Symmetric Uncertainty	the Trial's Token and an amount determined by a random draw. With 50% chance, the amount is your Switch Point. With 50% chance, the amount is $10.00 minus your Switch Point.
Misreport with Asymmetric Uncertainty	the Trial's Token and an amount determined by a random draw. With 90% chance, the amount is your Switch Point. With 10% chance, the amount is $10.00 minus your Switch Point.
IC with Upwards Bias	$10.00 and your Switch Point from Stage 2.
IC with Uncertainty Bias	$10.00 and an amount determined by a random draw. With 50% chance, the amount is your Switch Point. With 50% chance, the amount is $10.00 minus your Switch Point.

I.1 Instructions and comprehension questions



Screen 1: Welcome page.



Consent Form

Please read carefully before continuing

This is a consent form for research participation. It contains important information about this study and what to expect if you decide to participate.

Your participation is voluntary. Please consider the information carefully. If you decide to participate, please feel free to save or print a copy of this form.

Description: The experiment you are participating in today is part of a research study on decision making. The research study is designed to analyze individual behavior. You will be asked to read several pages of instructions. Then you will be asked to make several choices that will determine the precise amount you will be paid.

Risks and Benefits: The risks involved in this study are not substantially different from participating in normal online activities. We cannot and do not guarantee or promise that you will receive any benefits from this study. Your participation may benefit society by improving our understanding of behavior. Your decision whether or not to participate in this study will not affect your relationship with Caltech.

Duration: Your participation in this experiment will take approximately 40 minutes.

Payments: You will receive a fixed $7.00 completion payment for finishing the survey. One out of 10 subjects will be selected to receive additional payment based on their responses and pure chance. The average total payment for participation is $15.00 per hour, including the completion payment. All subjects will be paid. The minimum payment is the $7.00 completion payment.

Subjects' Rights: Your participation is voluntary and you have the right to discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled. The alternative is not to participate. You have the right to refuse to answer particular questions. The results of this research study may be presented at scientific or professional meetings or published in scientific journals. Your individual privacy will be maintained in all published and written data resulting from the study.

For Subjects Located In the European Economic Area: If you are in the European Economic Area (European Union, Iceland, Liechtenstein, and Norway) while you participate in the research study:

You have the right to request access, rectification or erasure of your personal information. You also have the right to object or restrict our processing of your personal information. Finally, you have the right to data portability, e.g., a copy of the data with your personal information. In order to make any such requests, please contact Tye Welch at gdpr@caltech.edu.

If you withdraw your consent to participate in this study, this will not affect the lawfulness of our collection, use and disclosure of your personal information, up to the point in time that you withdraw your consent. Even if you withdraw your consent, we may still use your data that has been anonymized or

Screen 2: Consent form. Upper half; the page continues on the next screen.

collection, use and disclosure of your personal information, up to the point in time that you withdraw your consent. Even if you withdraw your consent, we may still use your data that has been anonymized or pseudonymized so that the data does not identify you, as permitted by applicable law for the purposes of: (a) the public interest, (b) scientific research, and (c) archiving in the public interest. Further, we will maintain your data in fully identifiable form if required by law.

You consent to the collection, use and disclosure of your personal information, which may include health and other sensitive personal information, for the purpose of carrying out the research study and confirming the accuracy of the study, with the lawful basis to comply with legal and regulatory requirements. You may withdraw your consent at any time, and we will stop processing (e.g., analyzing) your personal information, except as described above.

Please view Caltech's General Data Protection Regulation Notice at the following website: https://www.caltech.edu/general-data-protection-regulation-notice

Contacts and Questions: For questions, concerns, or complaints about the study you may contact the Protocol Director, Camila Farres (e-mail: cfarresr@caltech.edu, Phone: (626) 563-7478) in the Division of Humanities and Social Sciences.

Independent Contact: If you are not satisfied with how this study is being conducted, or if you have any concerns, complaints, or general questions about the research or your rights as a participant, please contact the Caltech Institutional Review Board (IRB) to speak to someone independent of the research team at (626) 395-4699 or via email at irb@caltech.edu.

Note: All payments in this study are listed in US Dollars ($). If you are using a different local currency, it will be converted by Prolific from US Dollars to your currency at the current exchange rate.

Do you consent to participate?

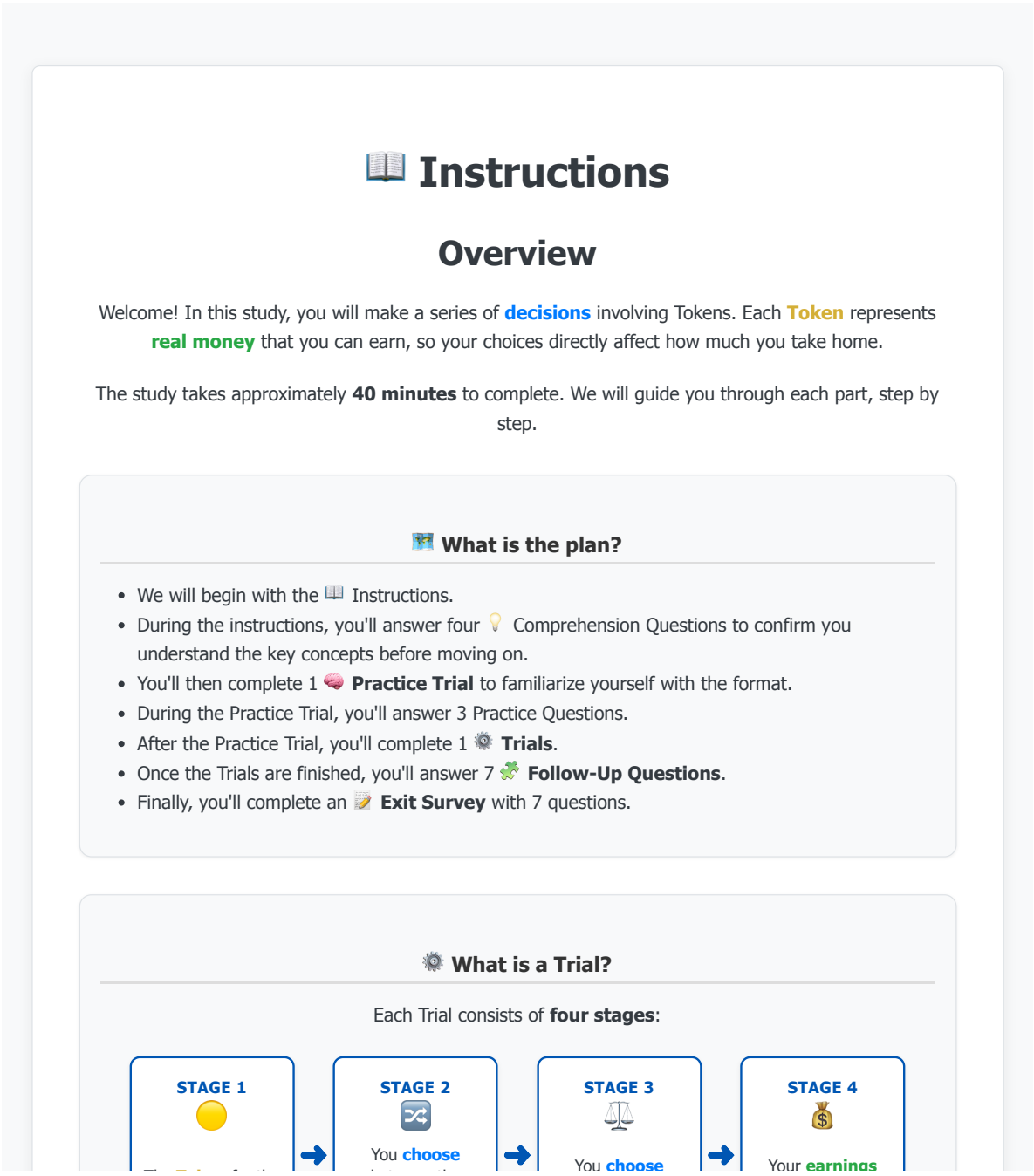
Yes | No

Please enter your Prolific ID:

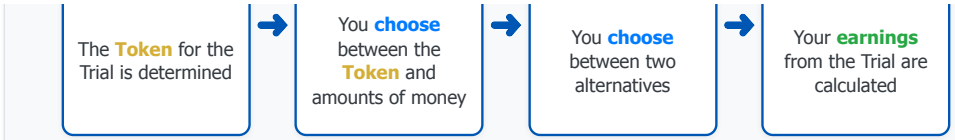
If you consent, click "Next" to begin the study.

NEXT

Screen 3: Consent form. Lower half.



Screen 4: Instructions: overview. Upper half; the page continues on the next screen.



How can you make money?

Everyone receives a **completion payment** of **$7.00** for finishing the study. On top of that, you may earn additional money in two ways:

- **Practice Question Payment:** during the Practice Trial, for each of the 3 Practice Questions you answer correctly on your first attempt, you earn an extra **$0.50** (up to **$1.50** total). This payment is independent of the Bonus Payment below — everyone who answers correctly earns it.
- **Bonus Payment:** at the end of the study, 1 out of every 10 participants is randomly selected for a **Bonus Payment**. If you are selected, your bonus depends on your answers in one randomly chosen Trial or Follow-Up Question. The Practice Trial is not part of this selection.

Throughout the study, a **green box** at the top of each page will remind you how the decisions you make on that page affect your **payment**. Here is an example of what it will look like:

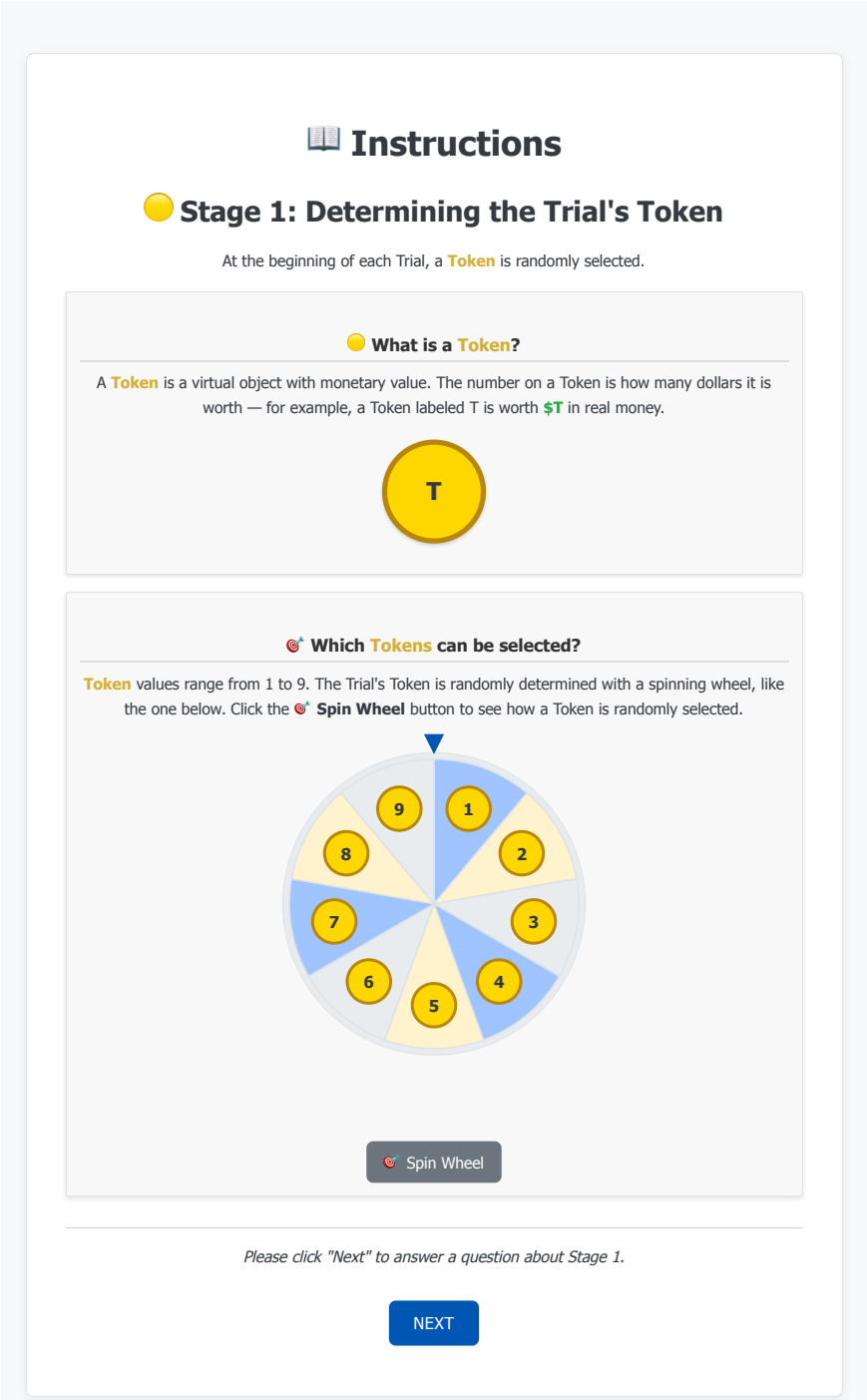
Example: Your decisions on this page affect your **payment** in the following way...

The following pages will explain each stage in more detail.

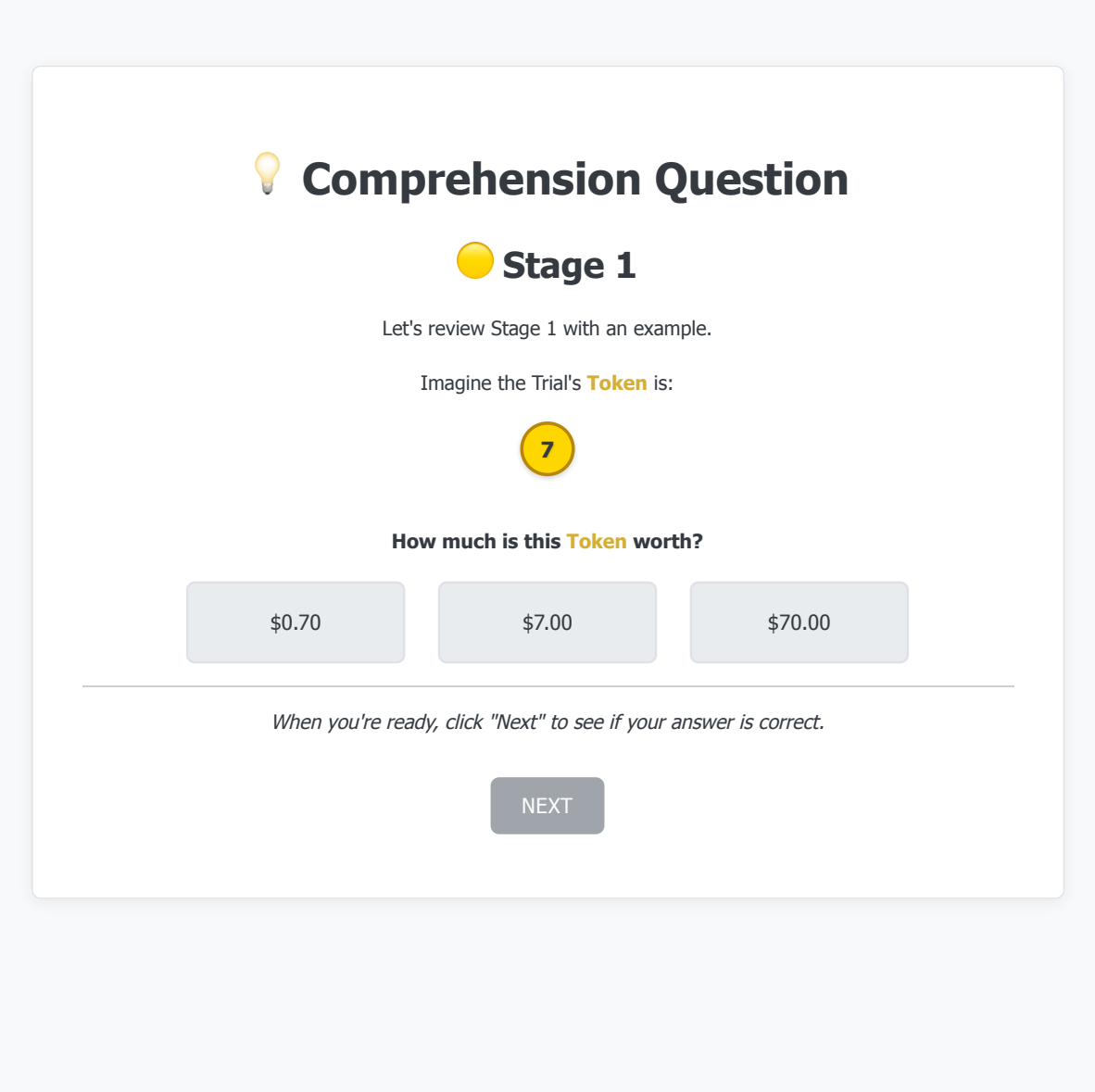
Please click "Next" to continue.

NEXT

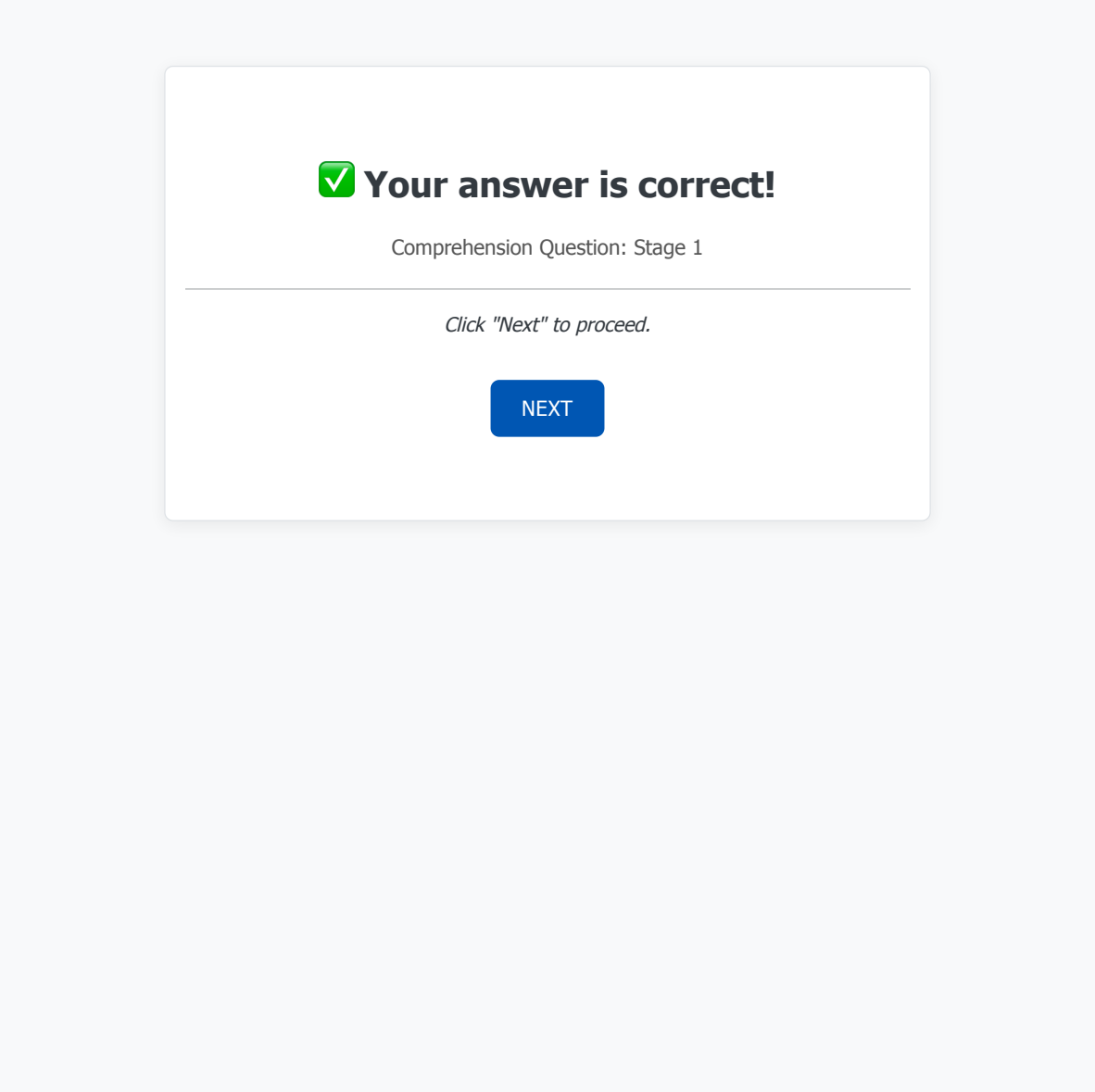
Screen 5: Instructions: overview. Lower half.



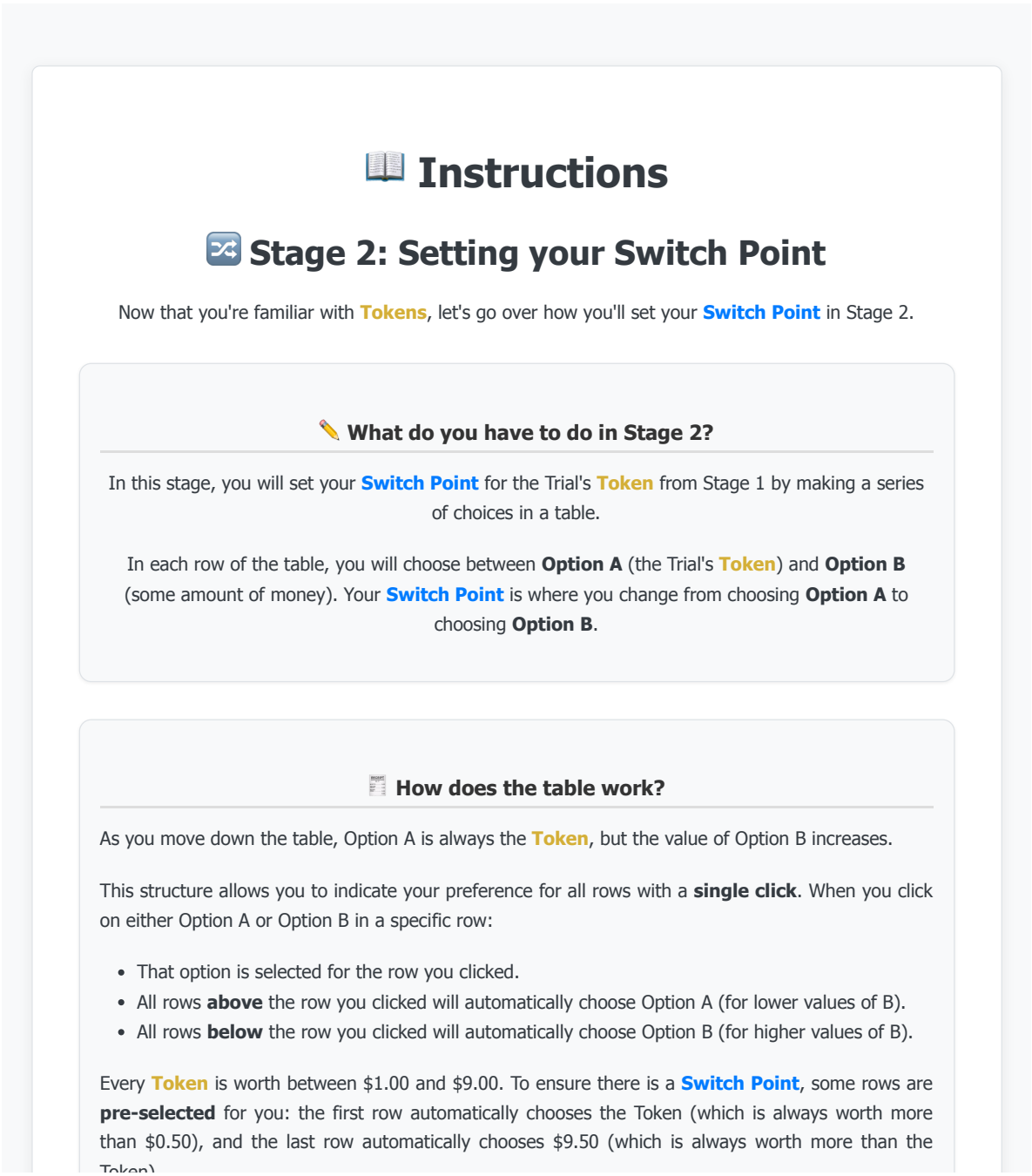
Screen 6: Instructions: Stage 1, the Induced Value.



Screen 7: Comprehension question on Stage 1.



Screen 8: Feedback after a correct answer to the comprehension question on Stage 1. The same page follows each correct answer.



Screen 9: Instructions: Stage 2, the Measurement Stage, first part. Upper half; the page continues on the next screen.

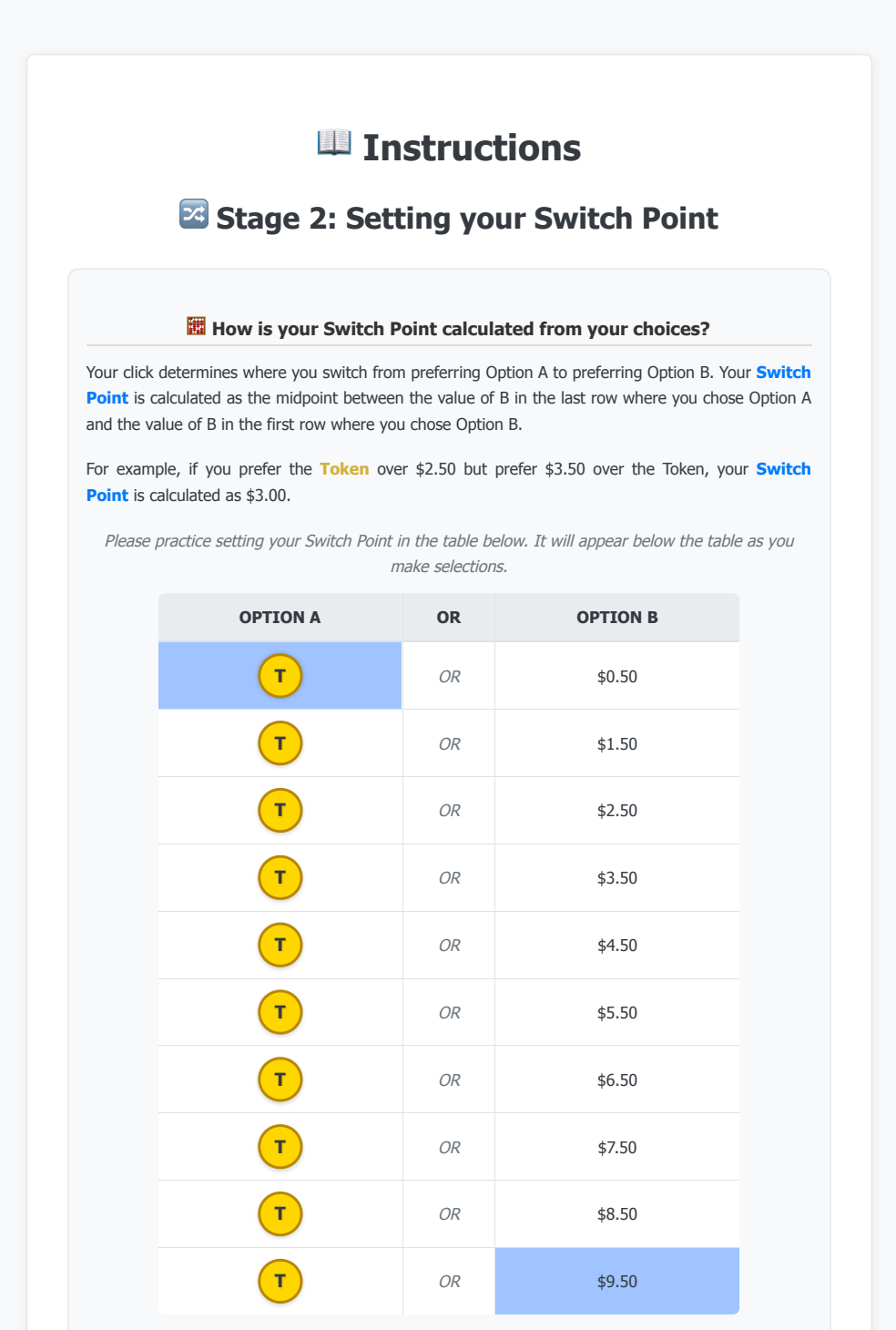
than $0.50), and the last row automatically chooses $9.50 (which is always worth more than the Token).

In the table below, the Trial's **Token** is shown as a generic Token labeled T.

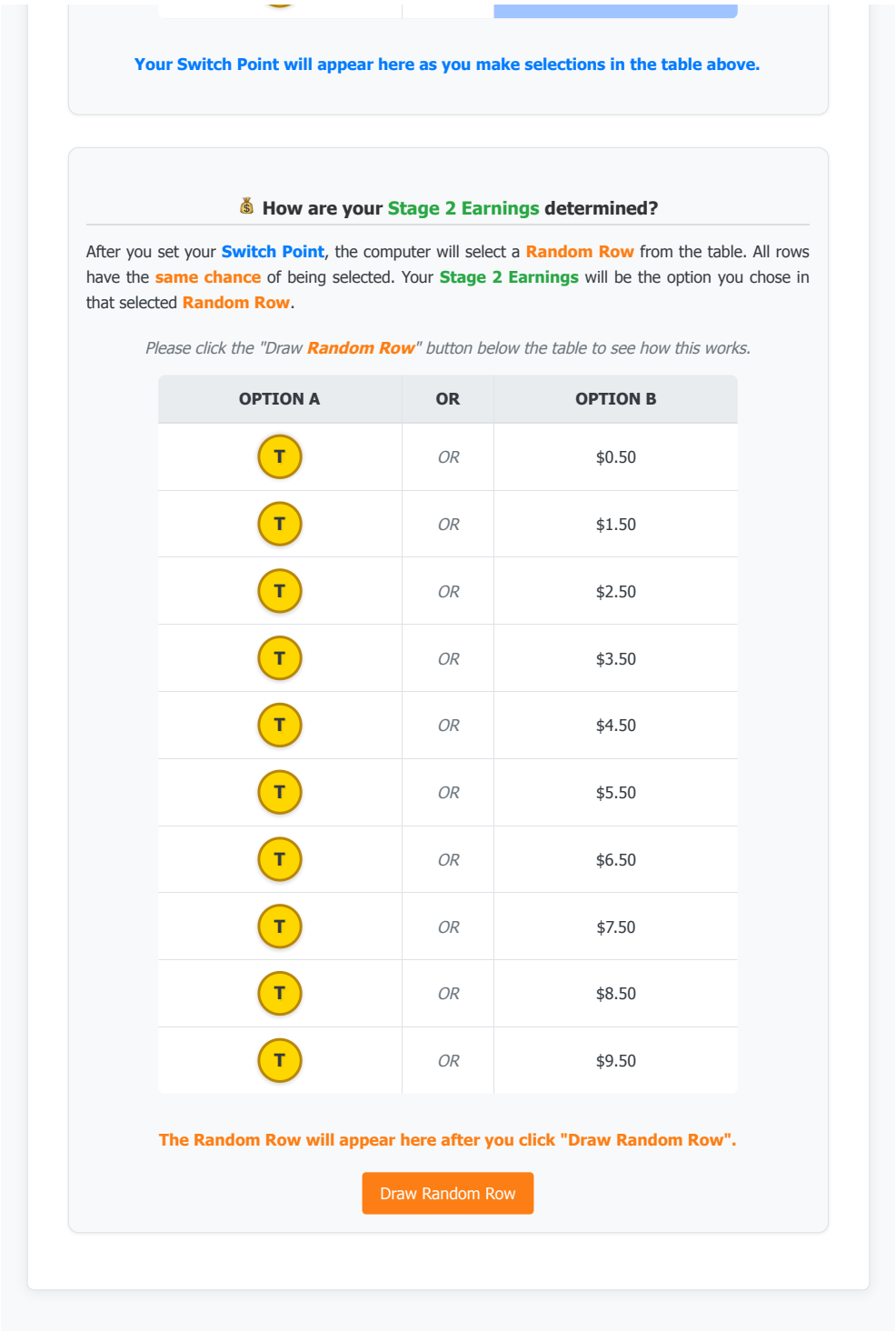
Please practice making selections to get used to how the table works.

OPTION A	OR	OPTION B
T	*OR*	$0.50
T	*OR*	$1.50
T	*OR*	$2.50
T	*OR*	$3.50
T	*OR*	$4.50
T	*OR*	$5.50
T	*OR*	$6.50
T	*OR*	$7.50
T	*OR*	$8.50
T	*OR*	$9.50

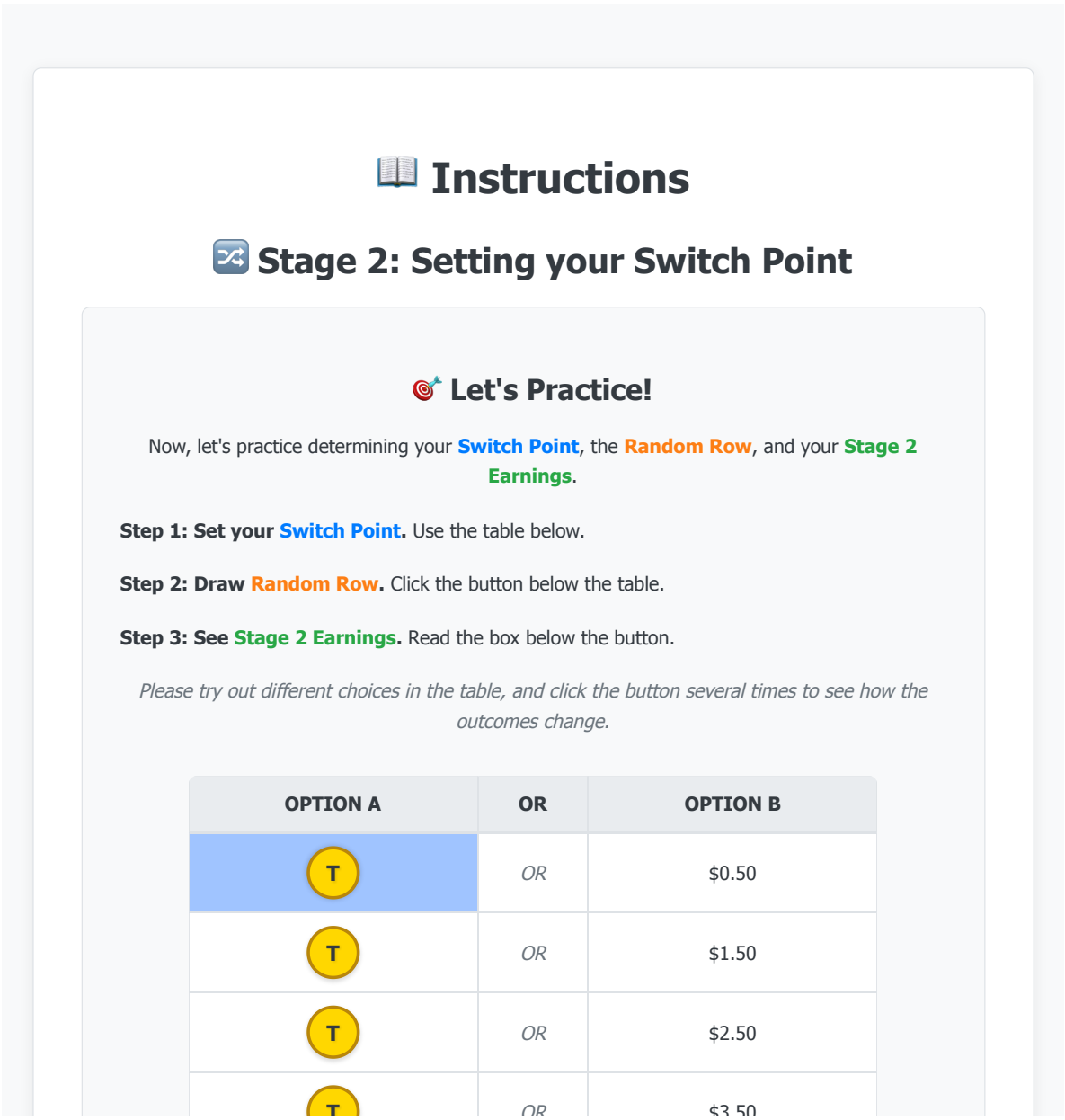
Screen 10: Instructions: Stage 2, the Measurement Stage, first part. Lower half.



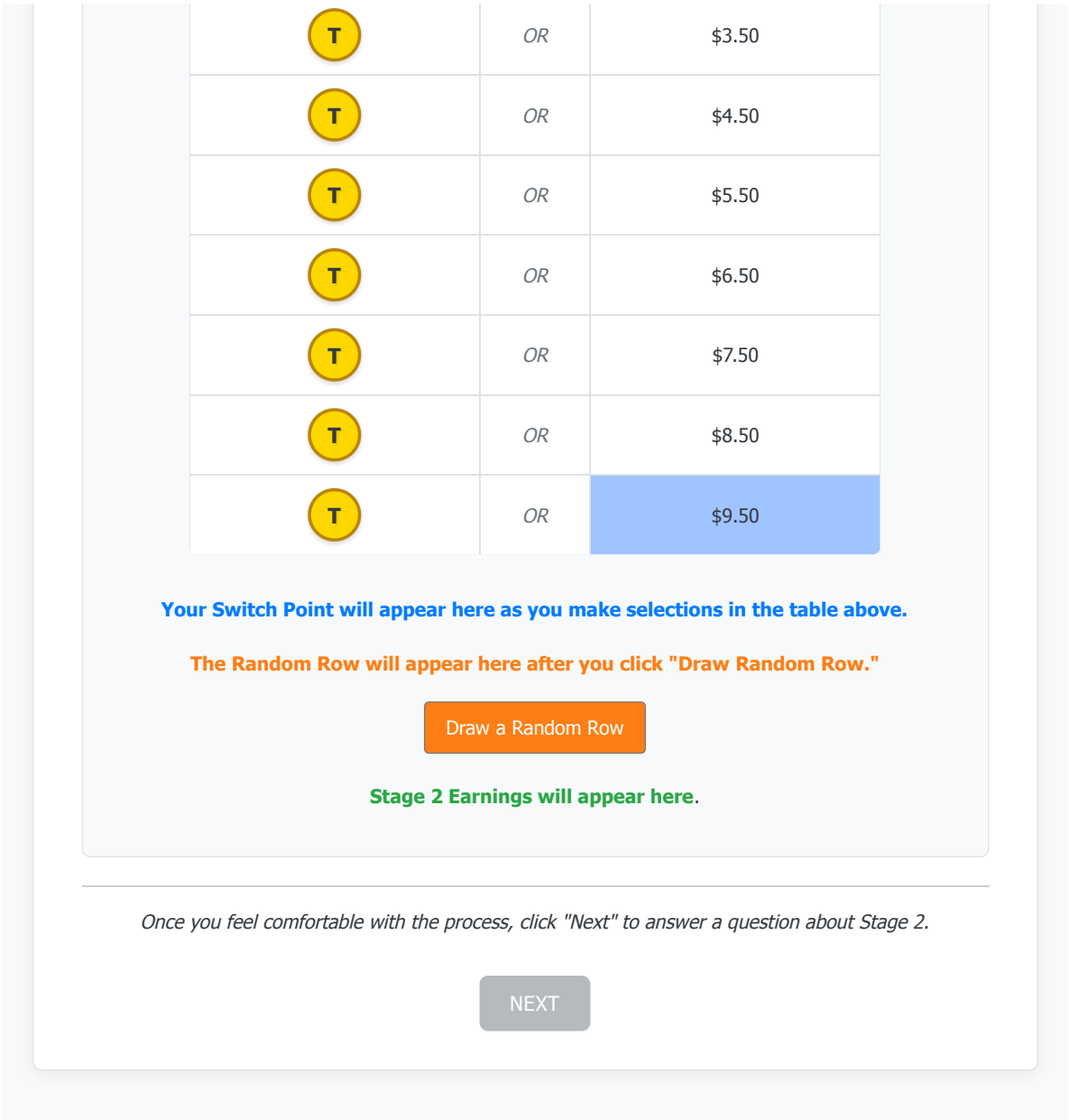
Screen 11: Instructions: Stage 2, the Measurement Stage, second part. Upper half; the page continues on the next screen.



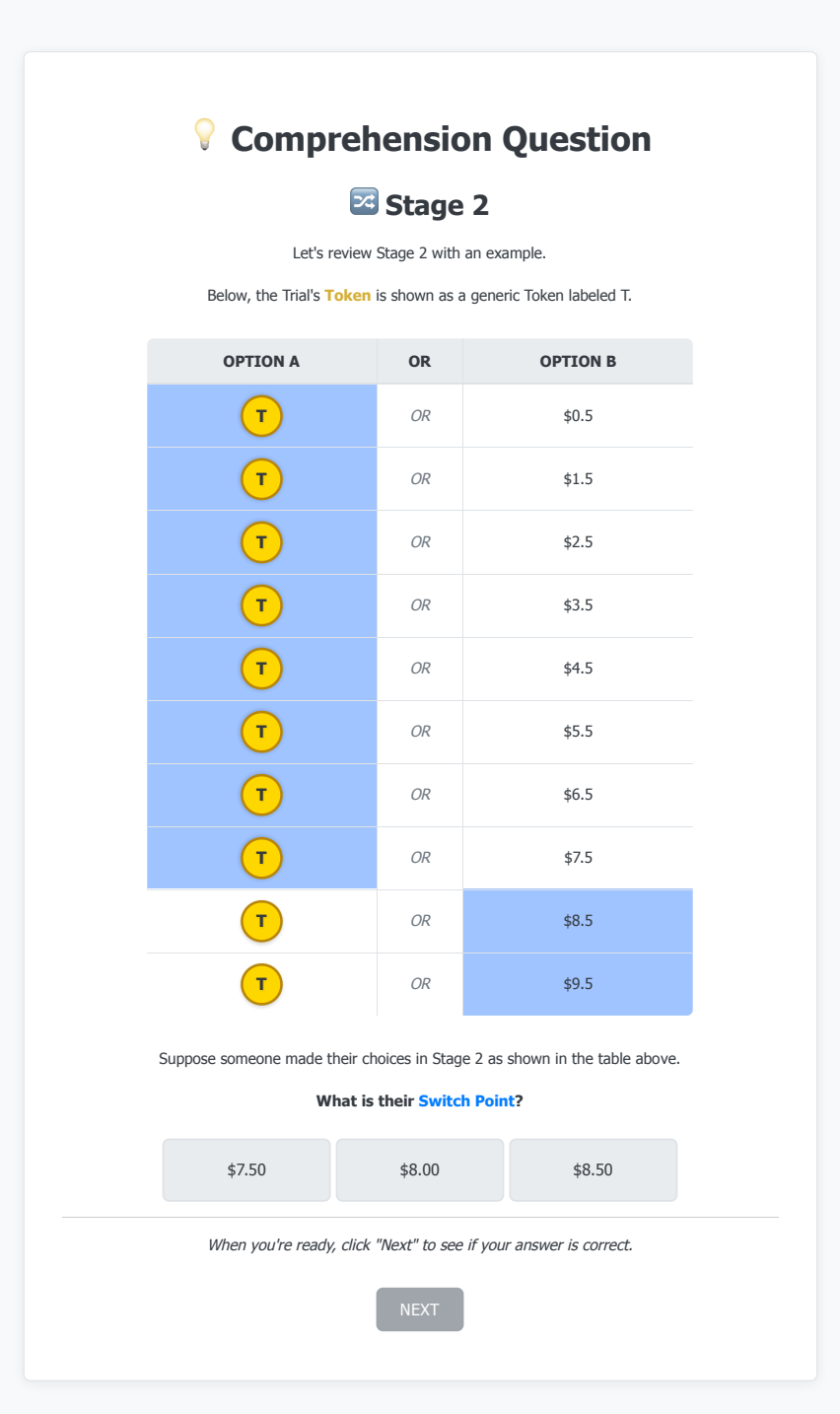
Screen 12: Instructions: Stage 2, the Measurement Stage, second part. Lower half.



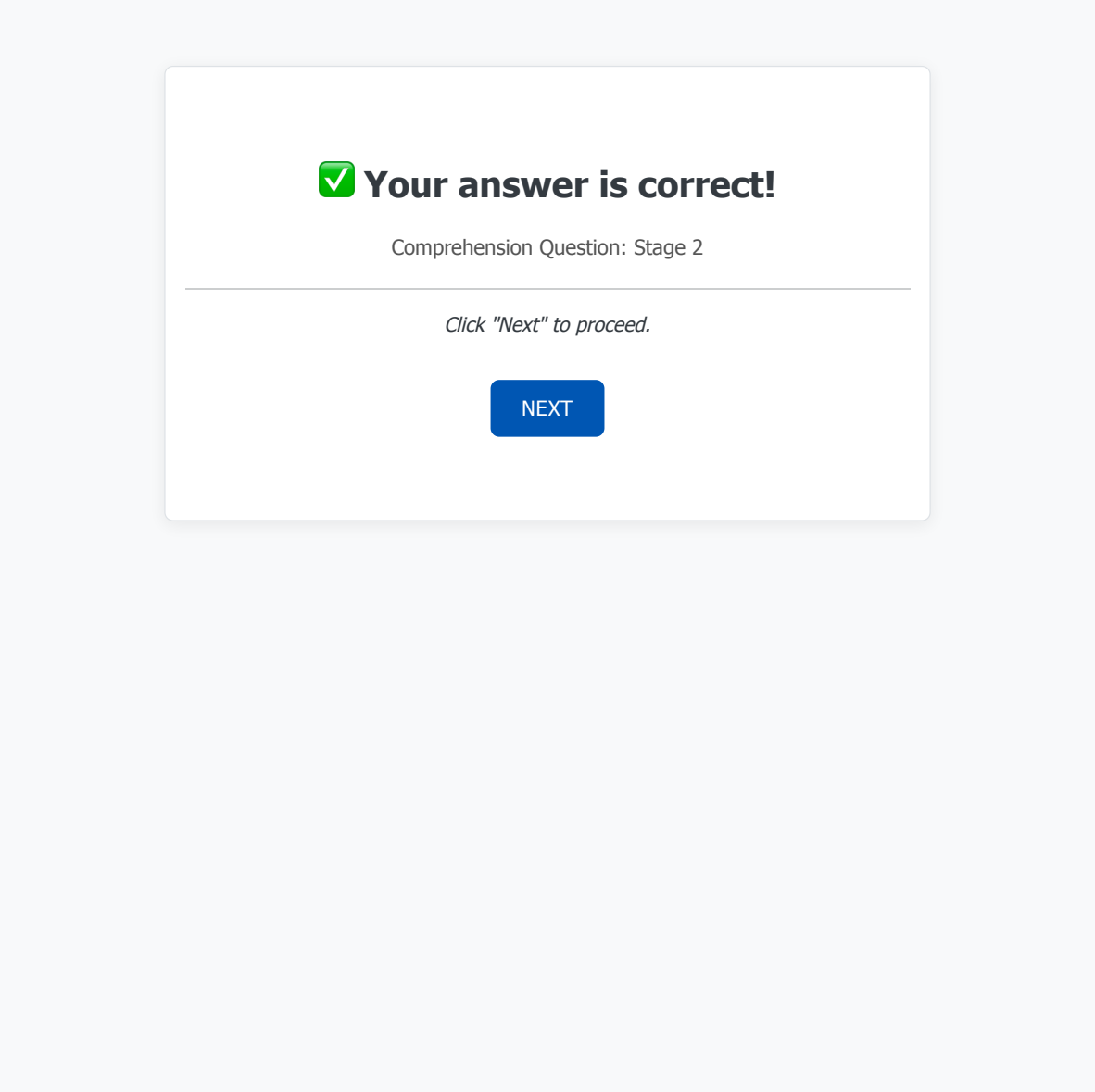
Screen 13: Instructions: Stage 2, the Measurement Stage, third part. Upper half; the page continues on the next screen.



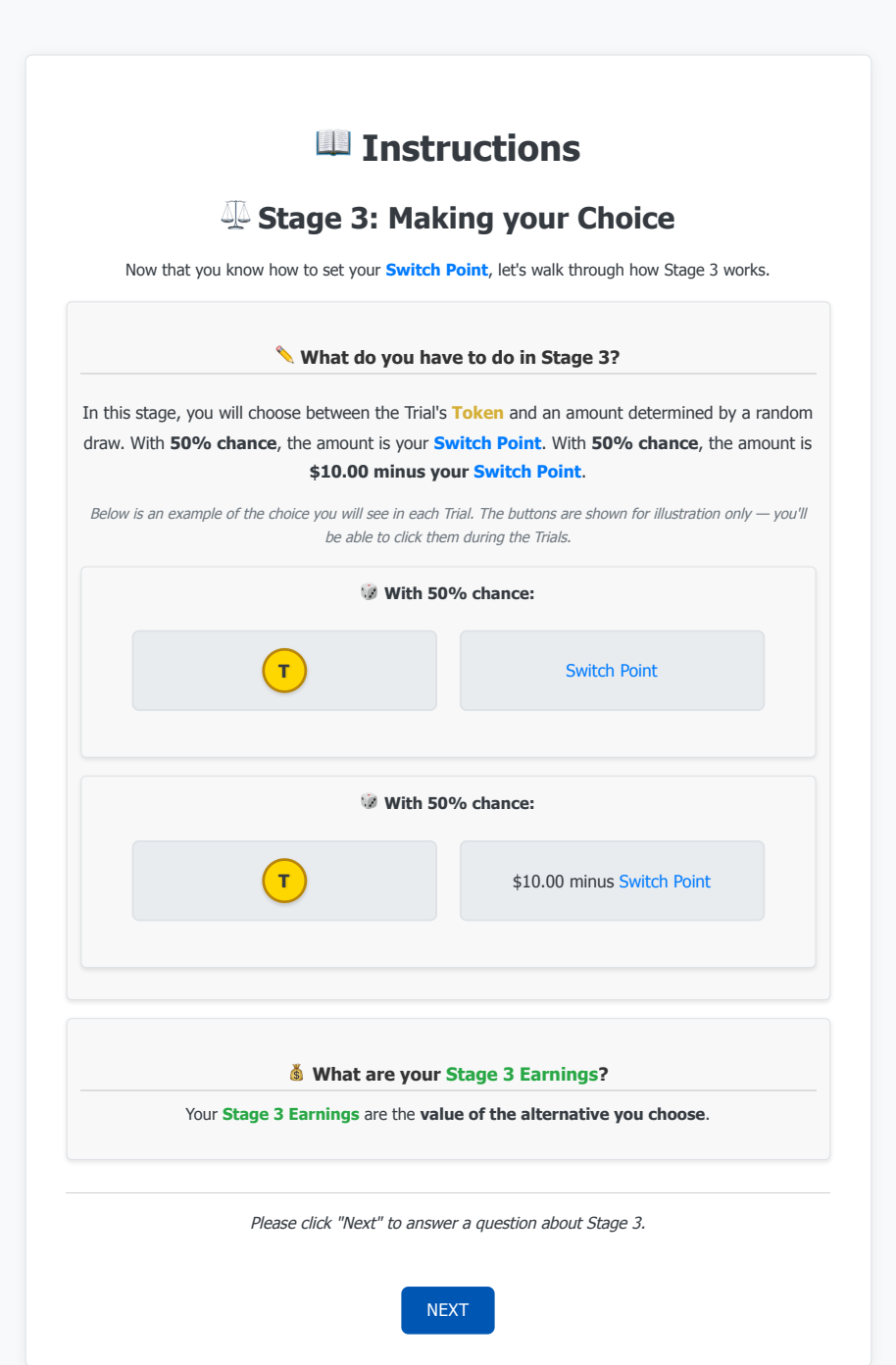
Screen 14: Instructions: Stage 2, the Measurement Stage, third part. Lower half.



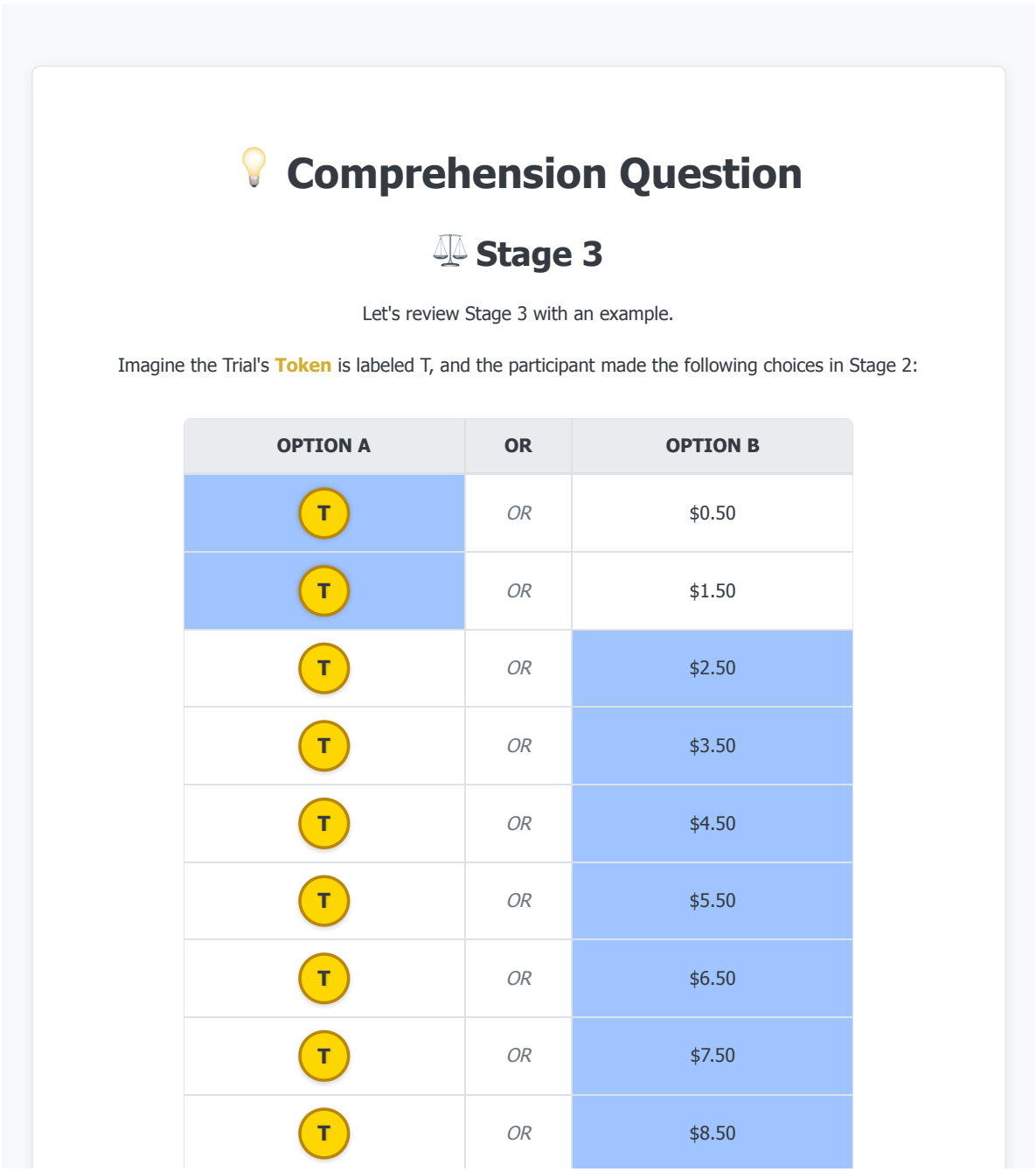
Screen 15: Comprehension question on Stage 2.



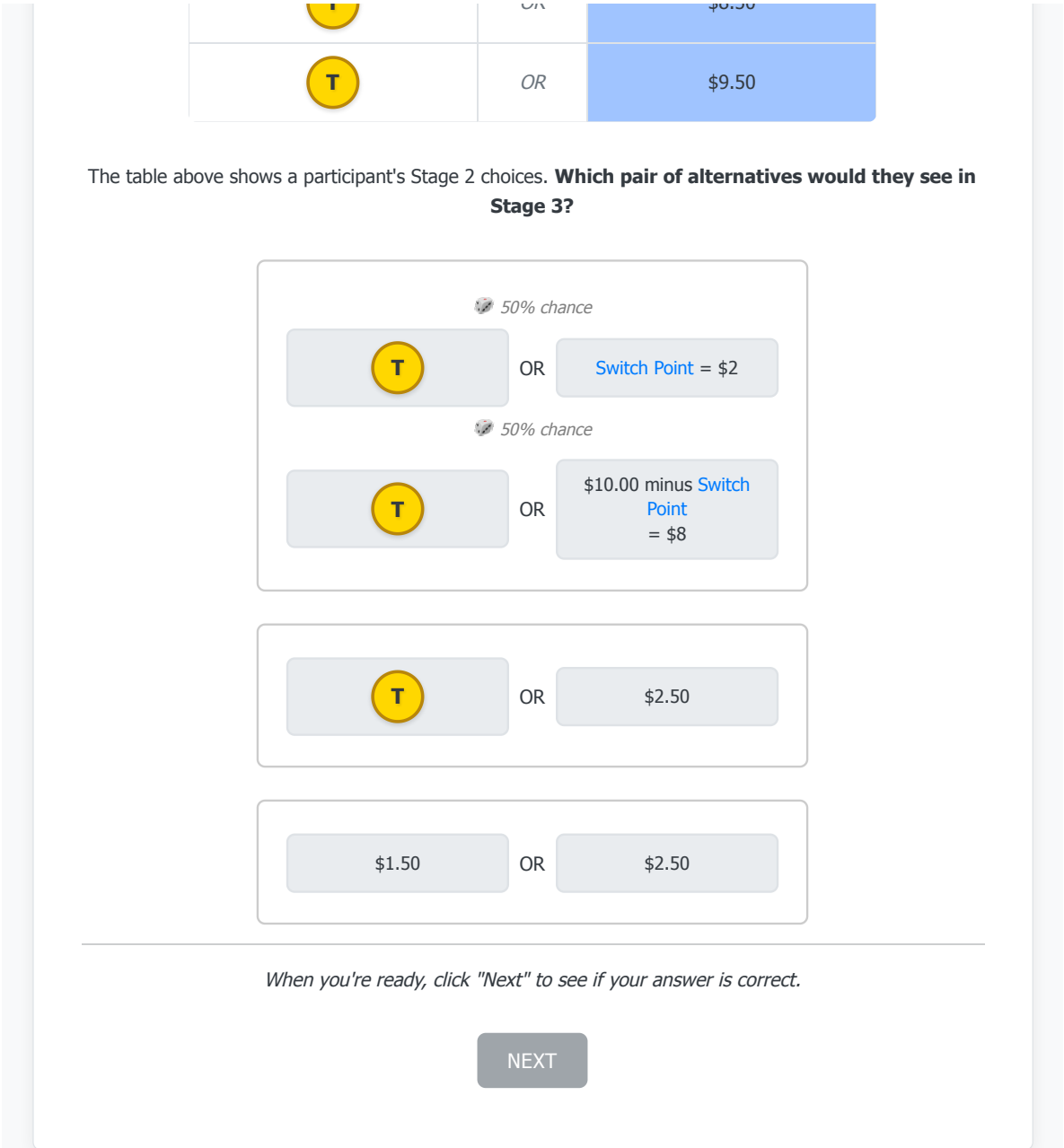
Screen 16: Feedback after a correct answer to the comprehension question on Stage 2.



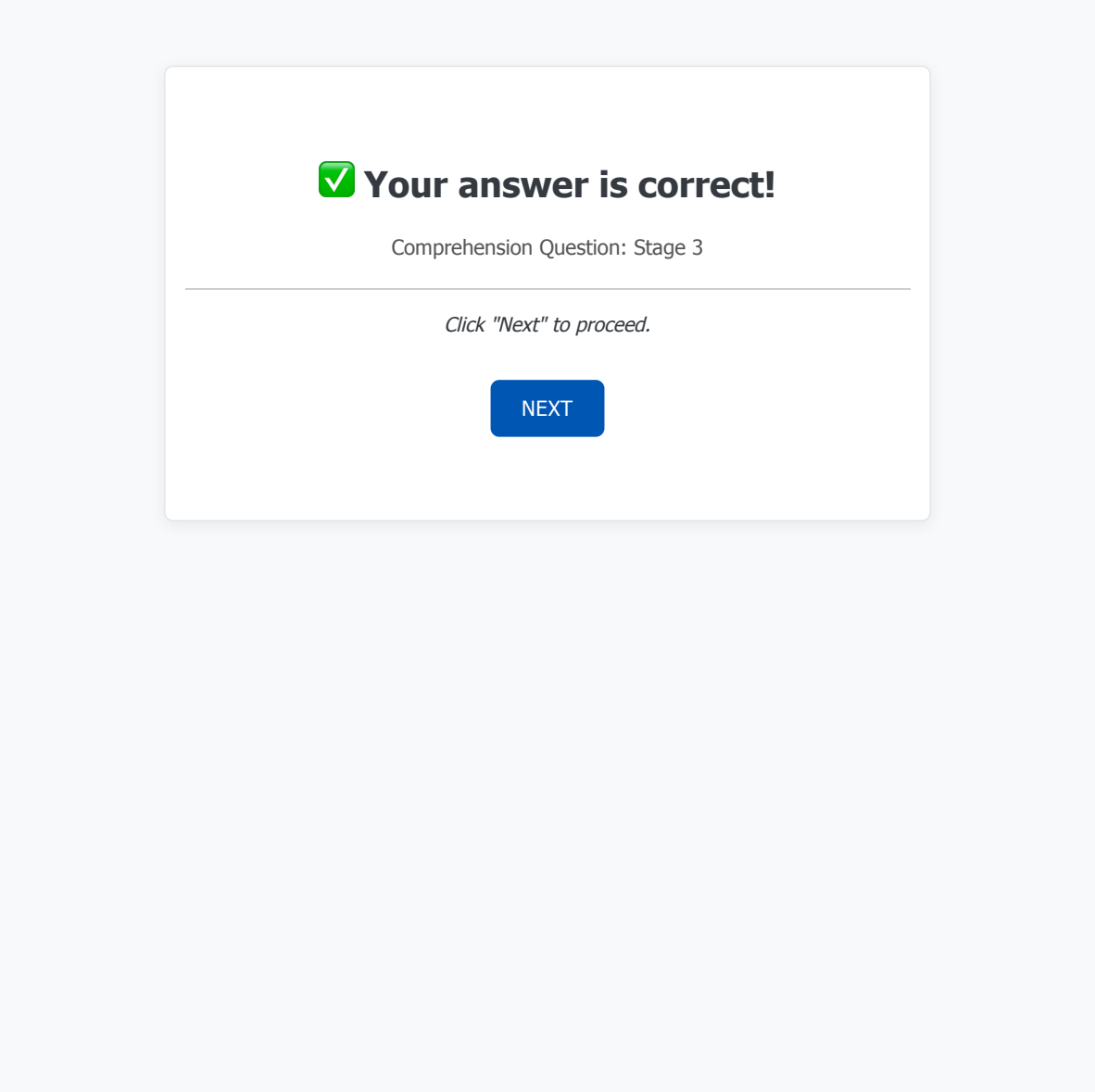
Screen 17: Instructions: Stage 3, the Personalized Menu, as shown in Misreport with Symmetric Uncertainty. Table I1 gives the sentence and the example of each treatment.



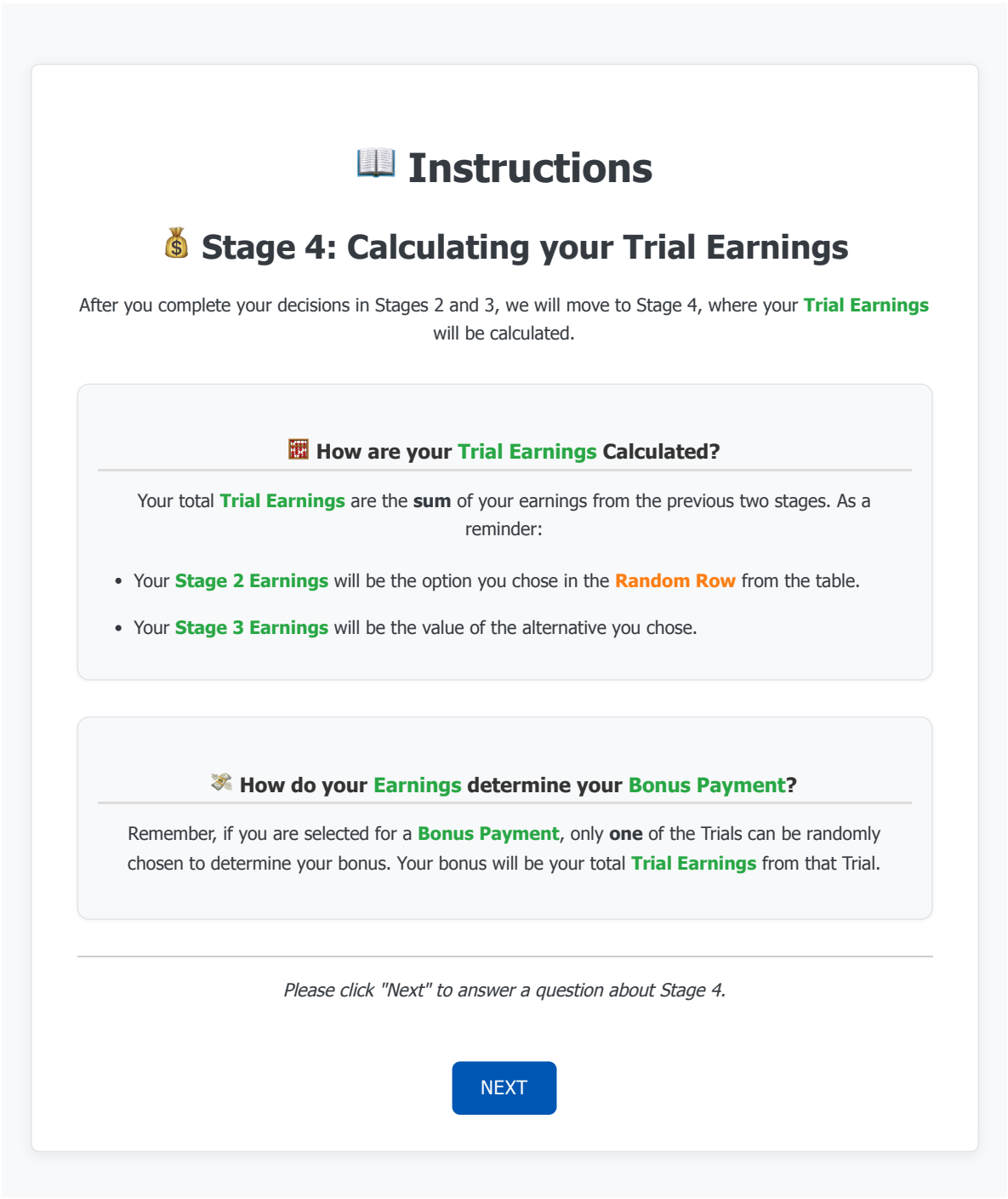
Screen 18: Comprehension question on Stage 3. Upper half; the page continues on the next screen.



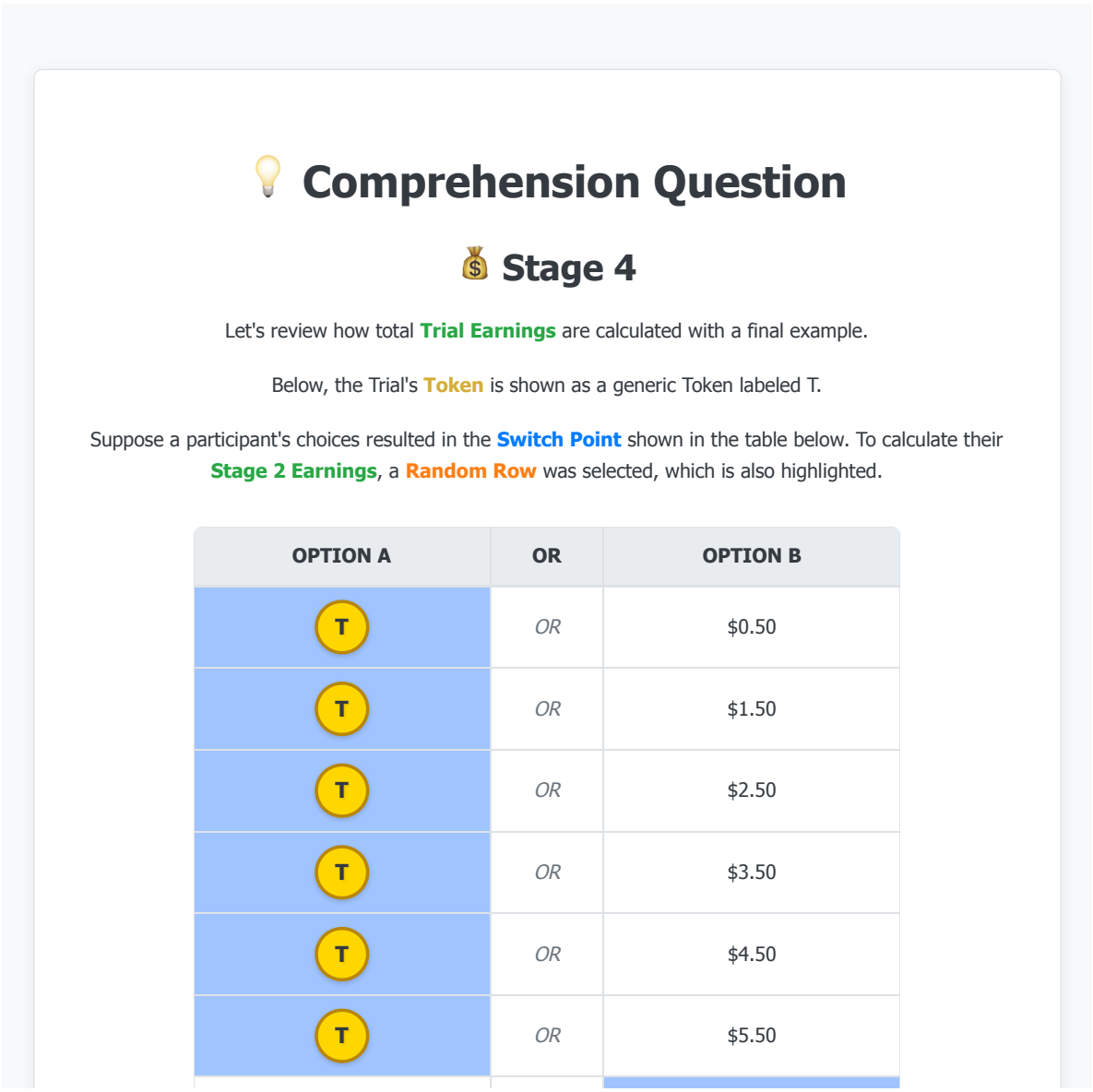
Screen 19: Comprehension question on Stage 3. Lower half.



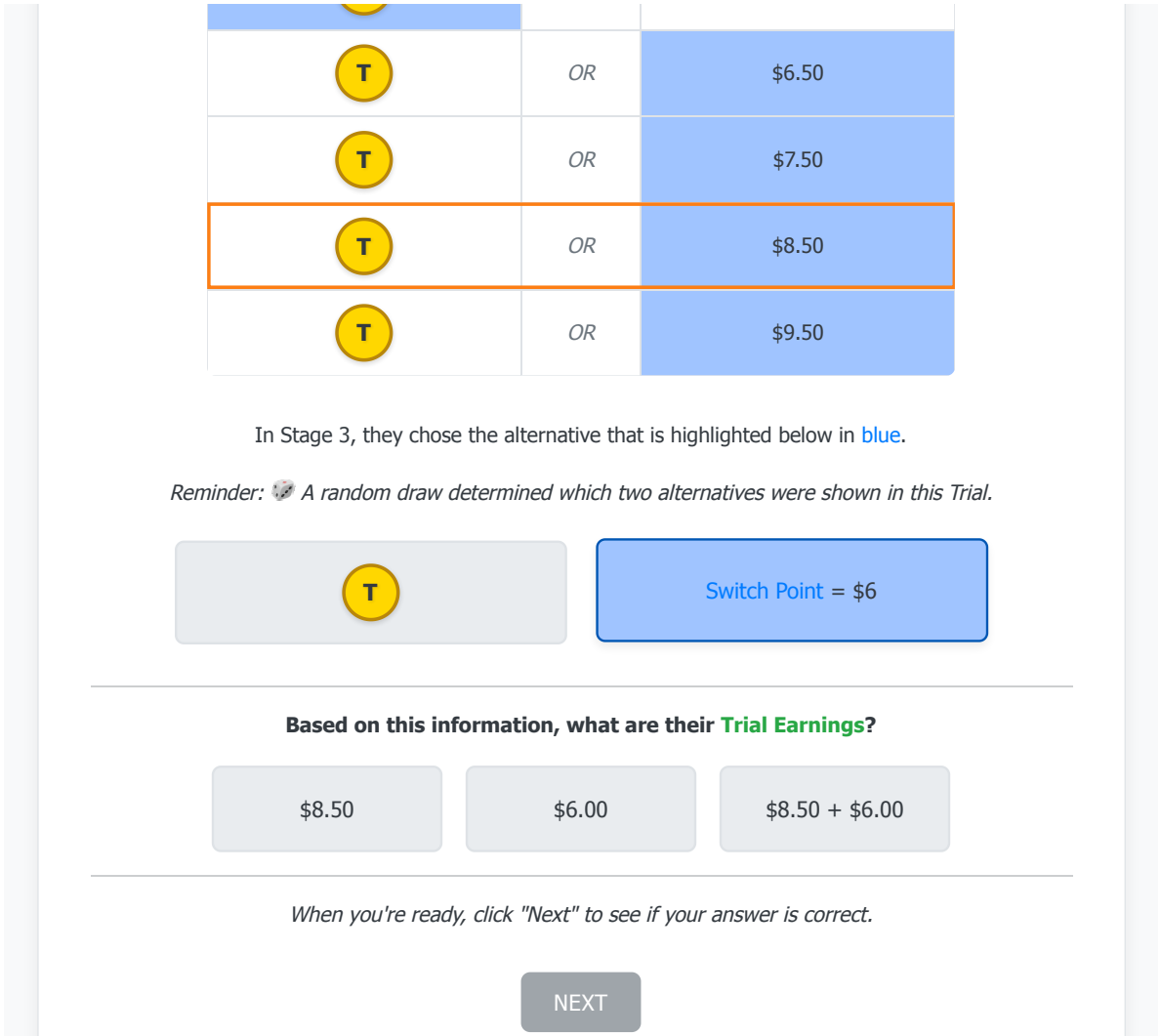
Screen 20: Feedback after a correct answer to the comprehension question on Stage 3.



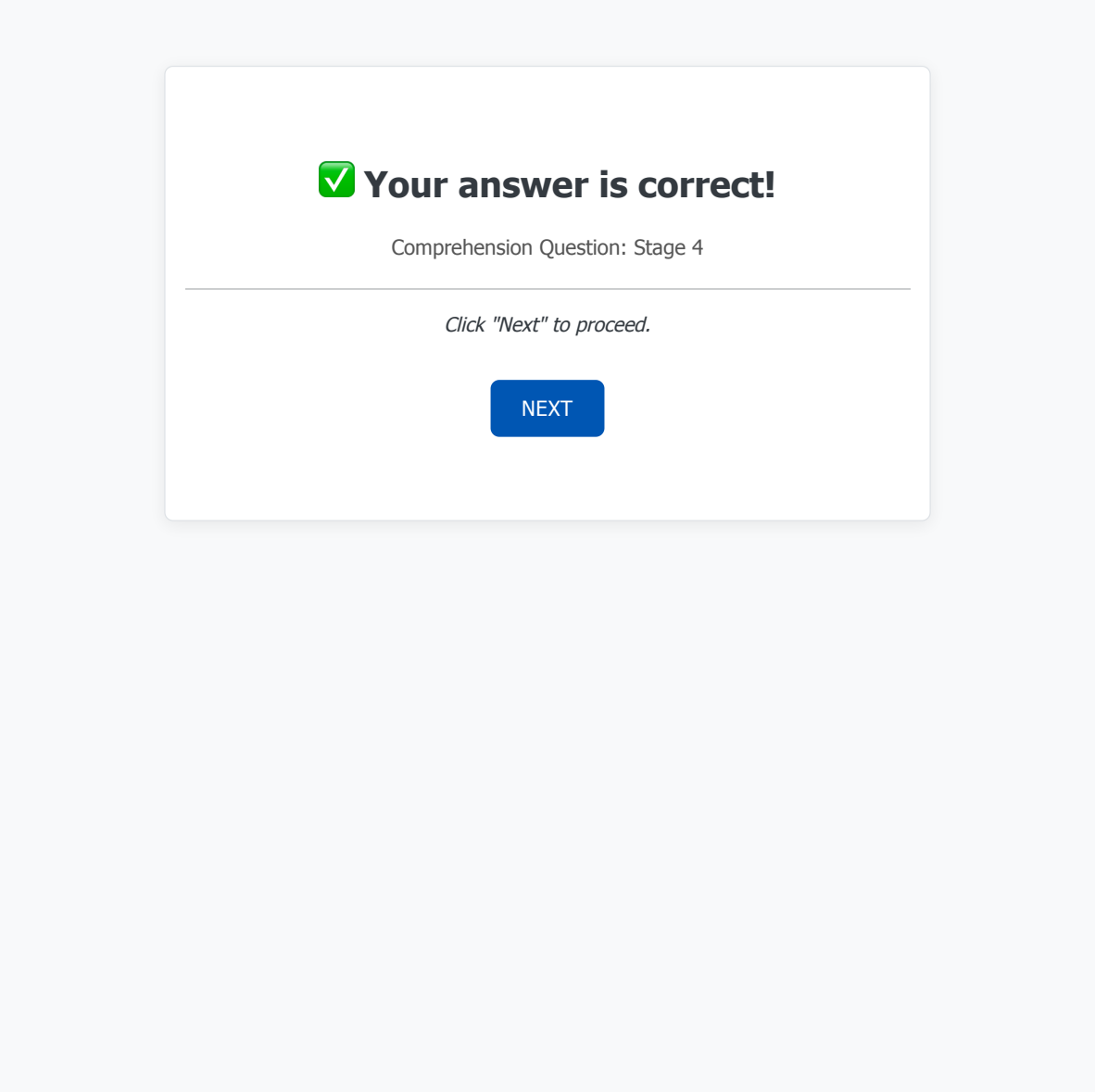
Screen 21: Instructions: Stage 4, the Interactive Feedback.



Screen 22: Comprehension question on Stage 4. Upper half; the page continues on the next screen.

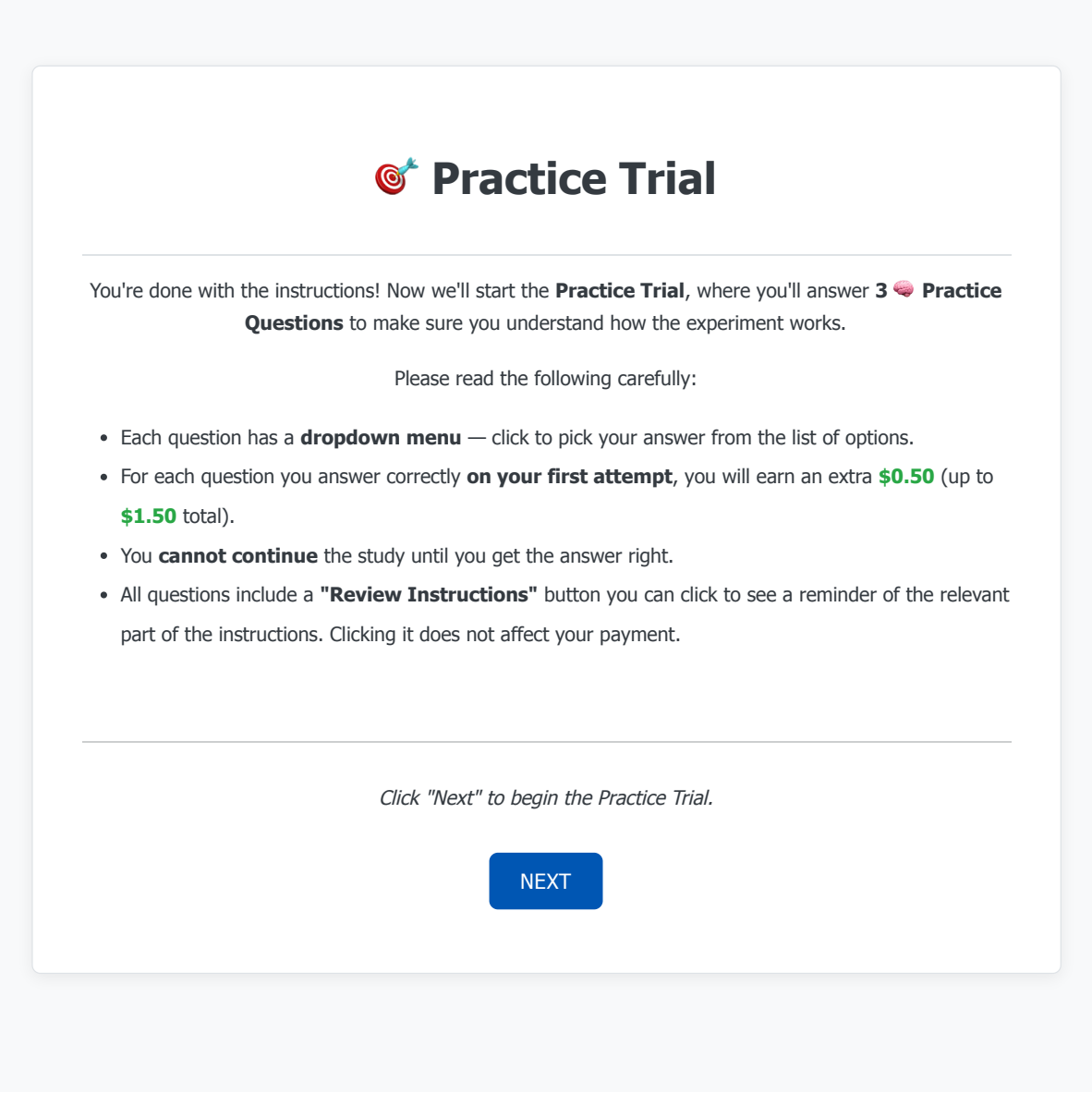


Screen 23: Comprehension question on Stage 4. Lower half.

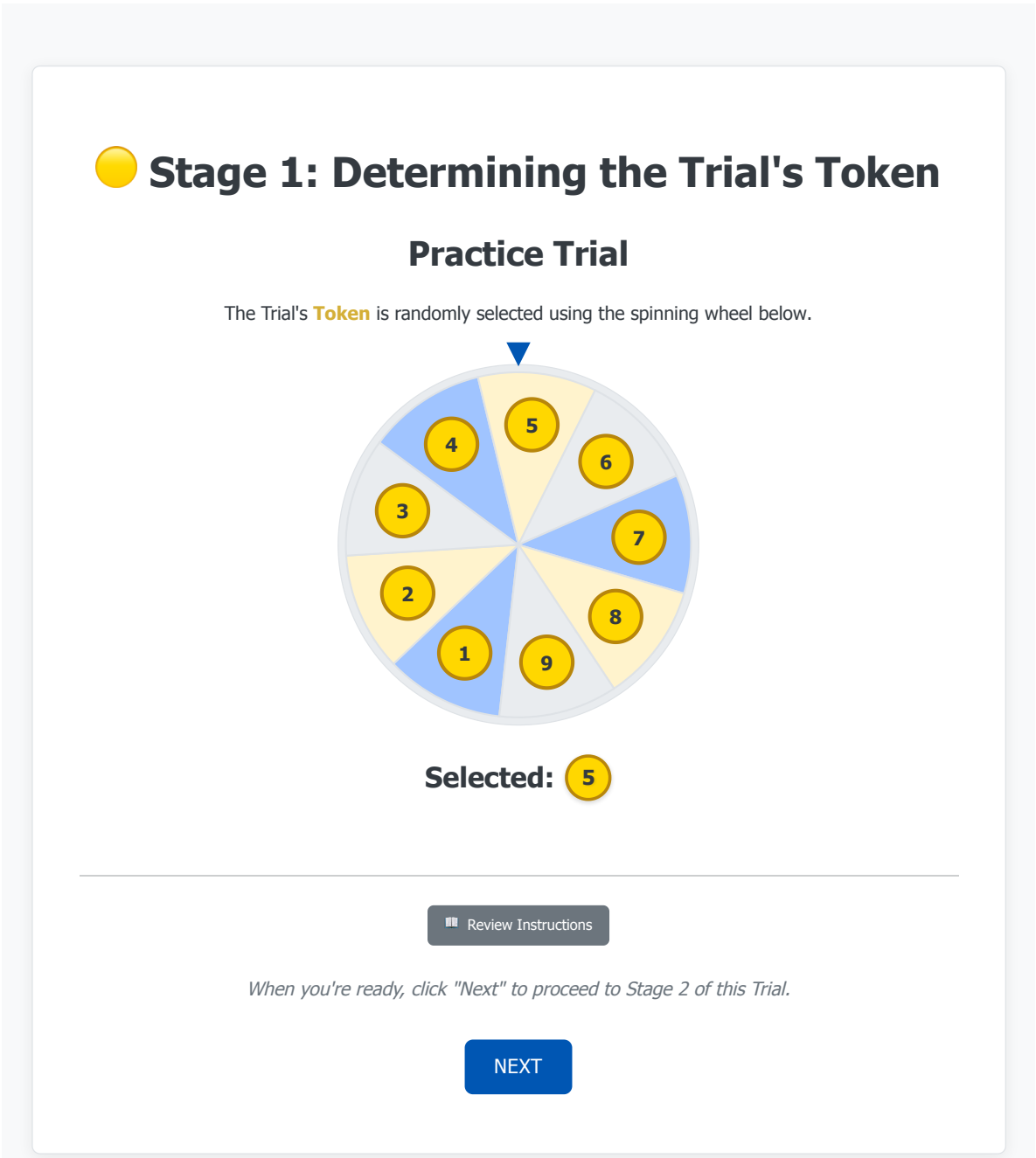


Screen 24: Feedback after a correct answer to the comprehension question on Stage 4.

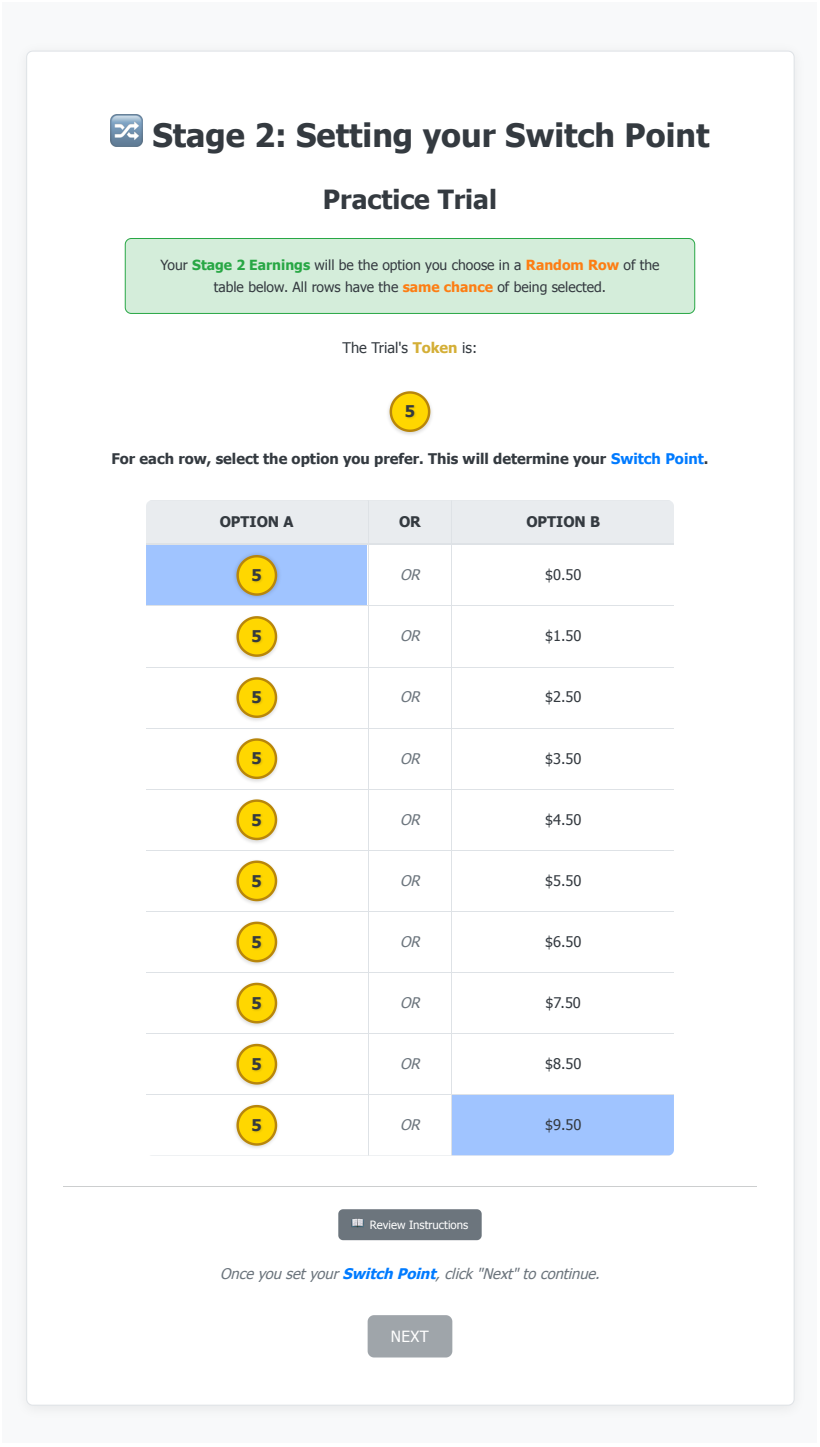
I.2 The practice trial



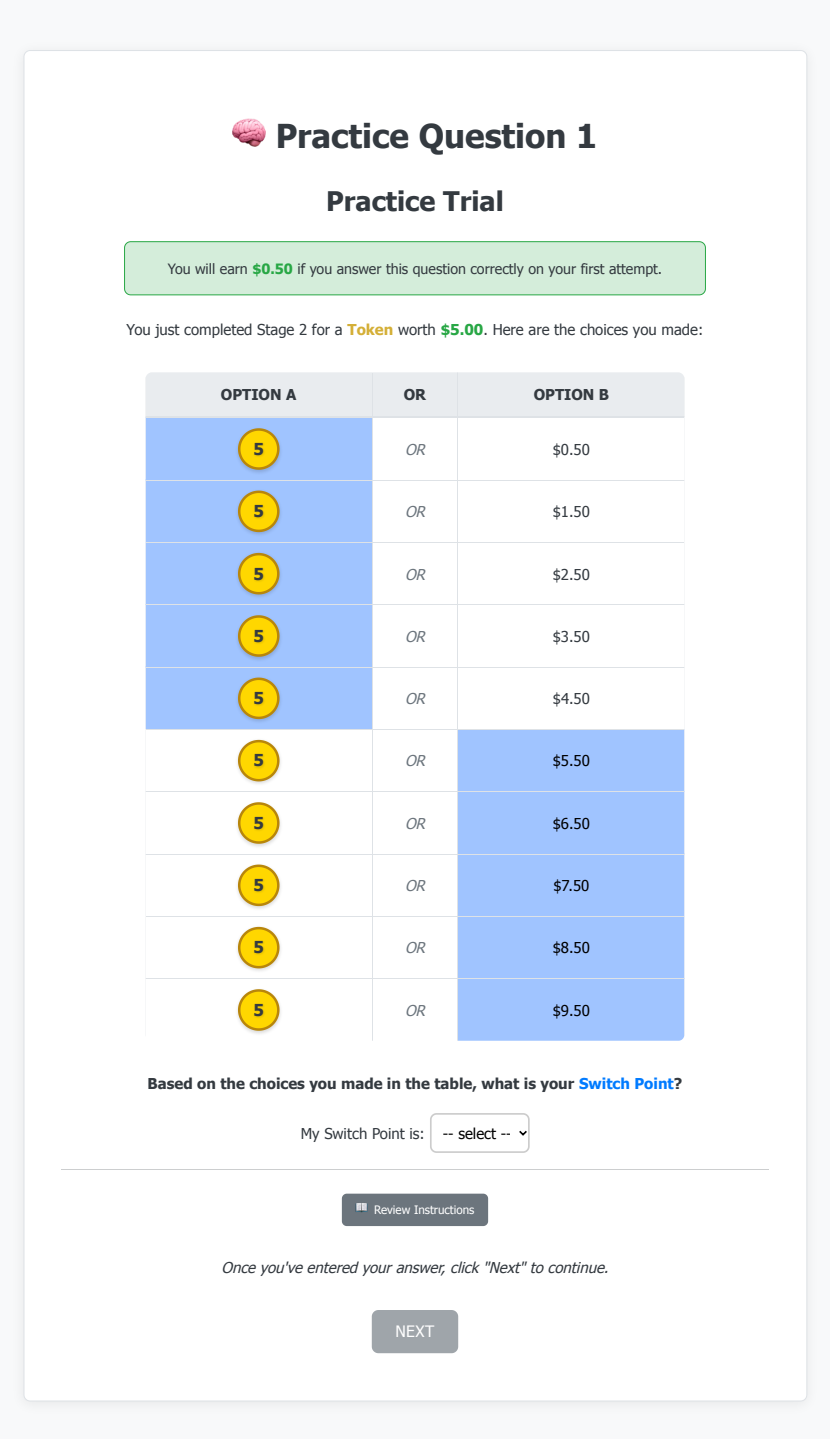
Screen 25: Start of the practice trial.



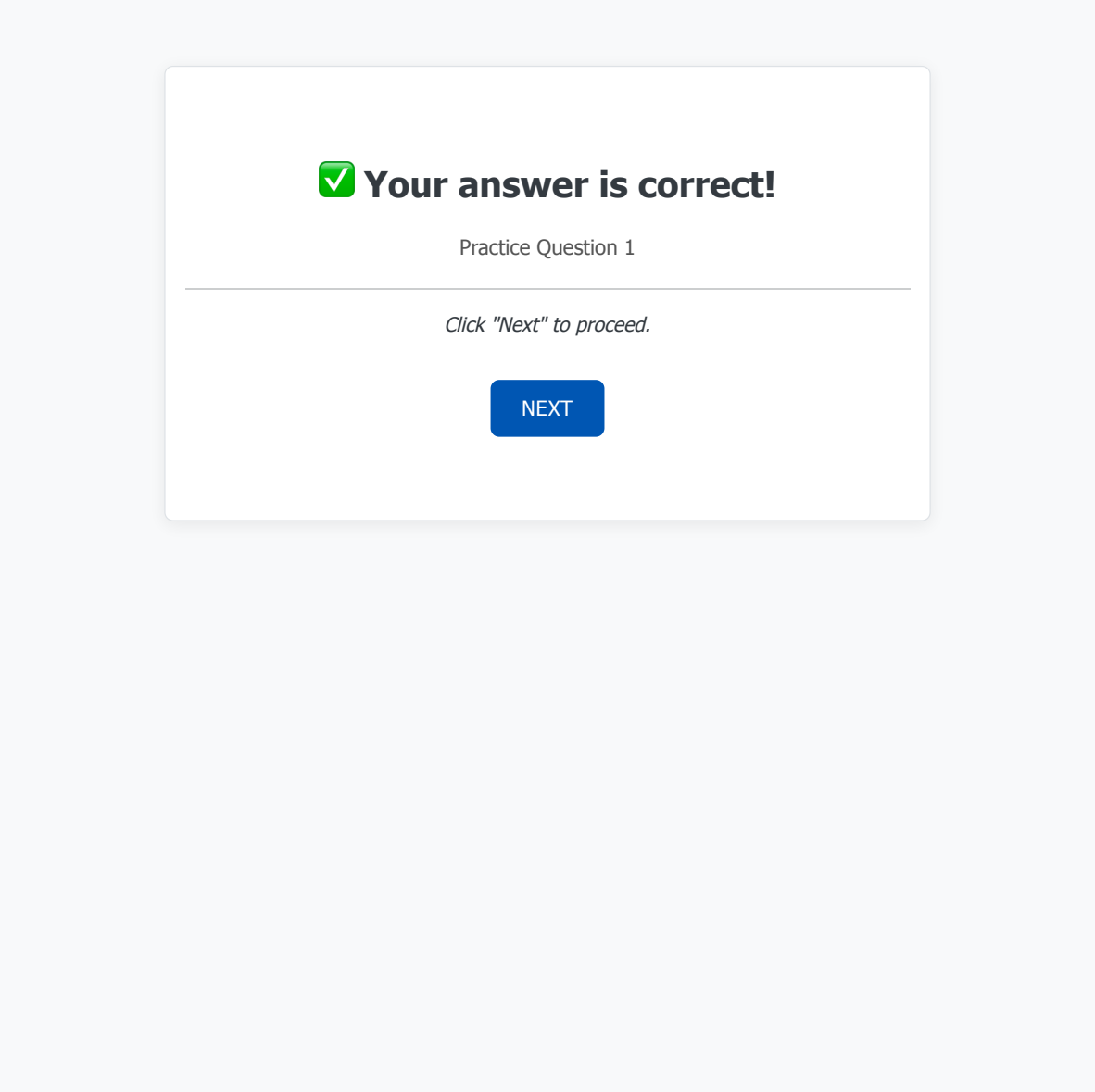
Screen 26: Practice trial, Stage 1: the token.



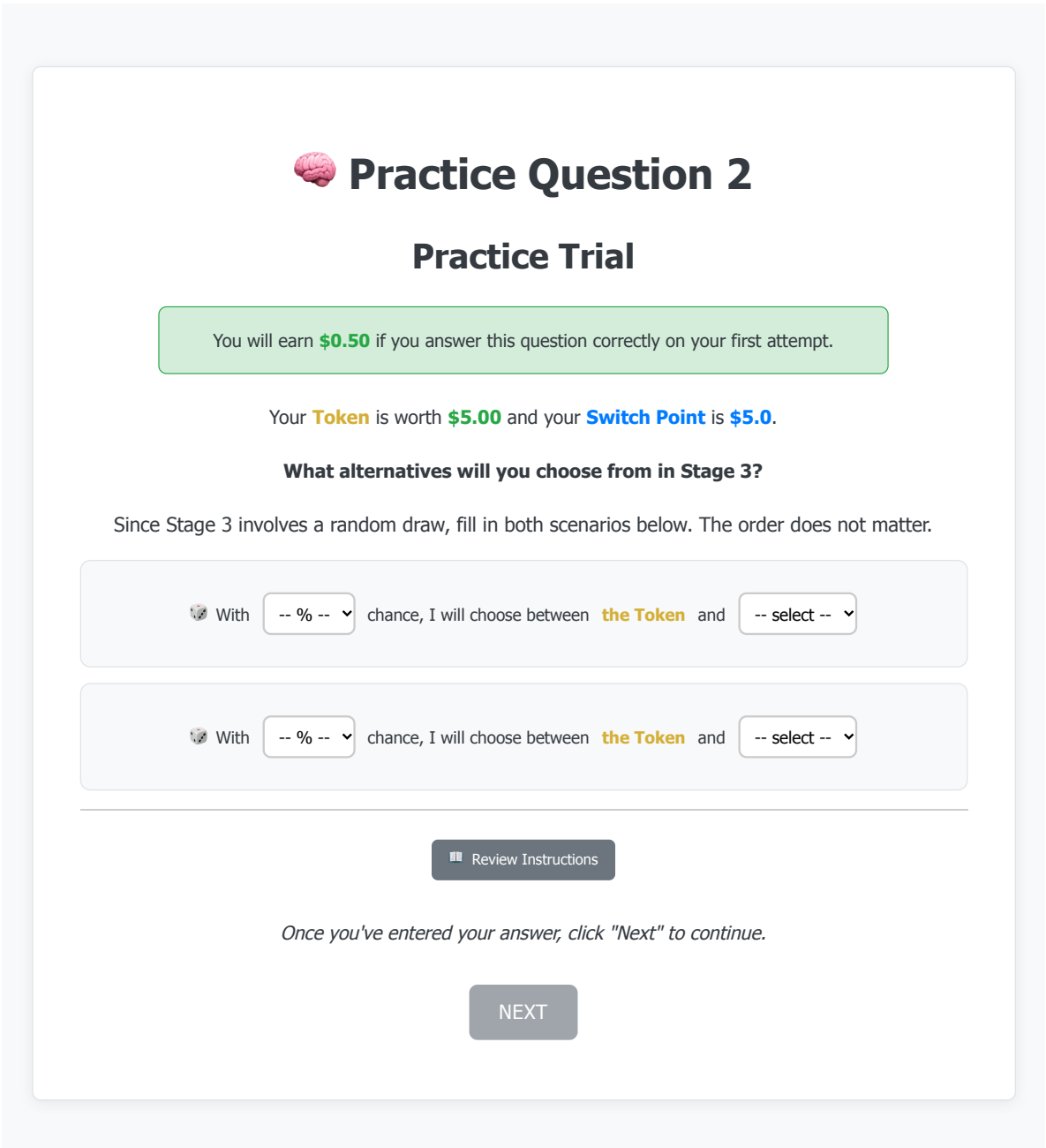
Screen 27: Practice trial, Stage 2: the price list.



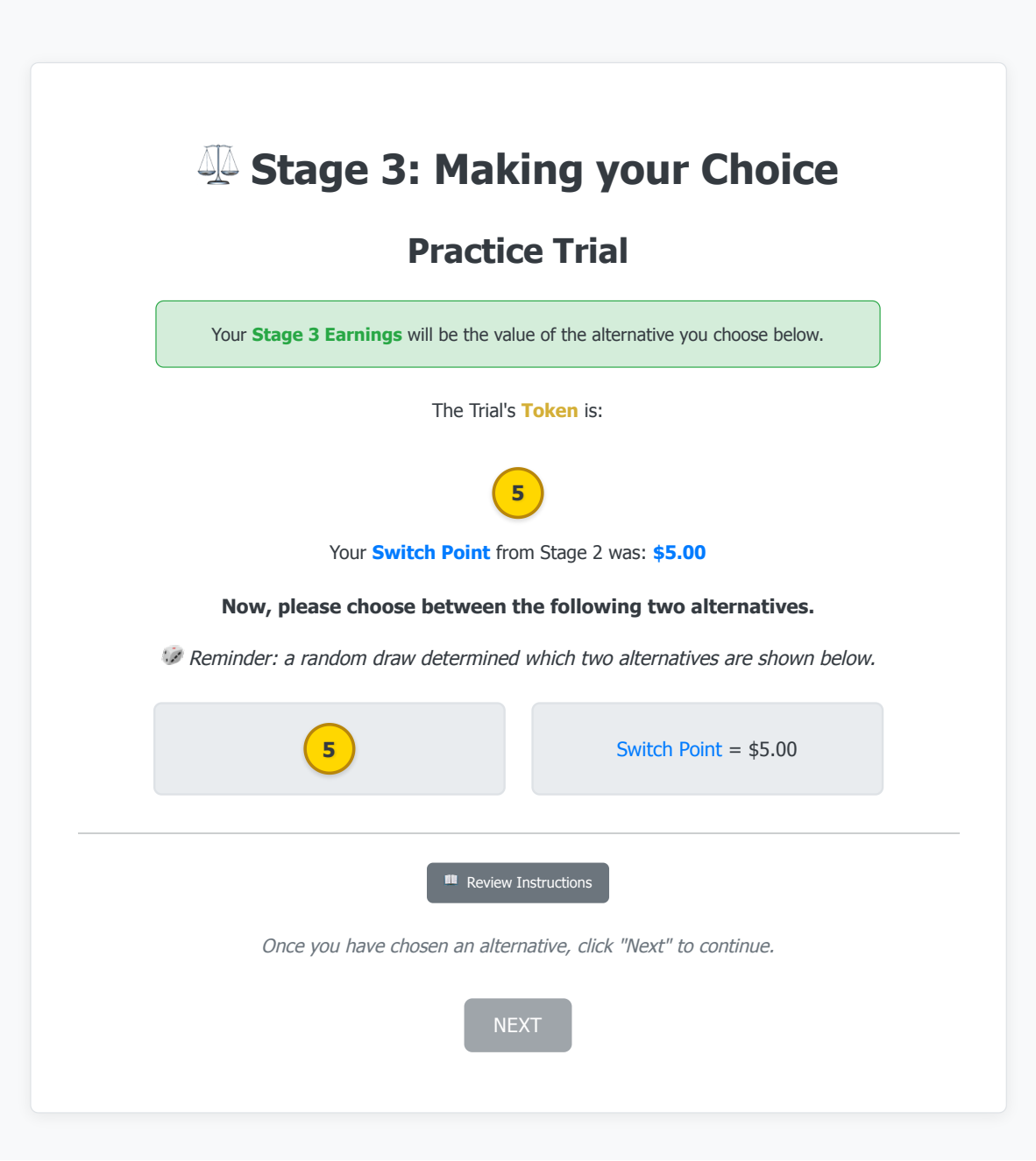
Screen 28: Practice question 1: the switch point.



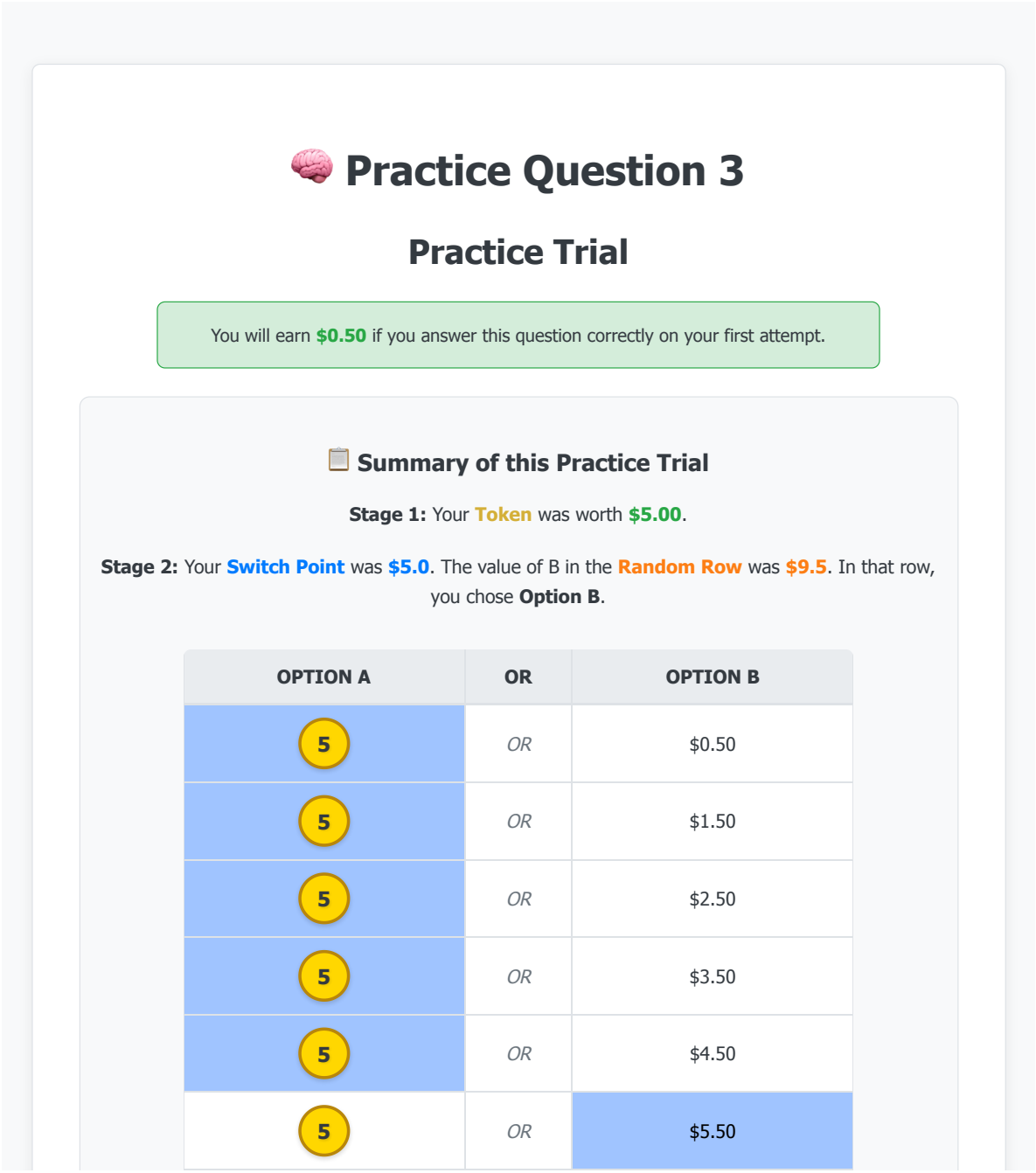
Screen 29: Feedback after a correct answer to practice question 1. The same page follows each correct practice answer.



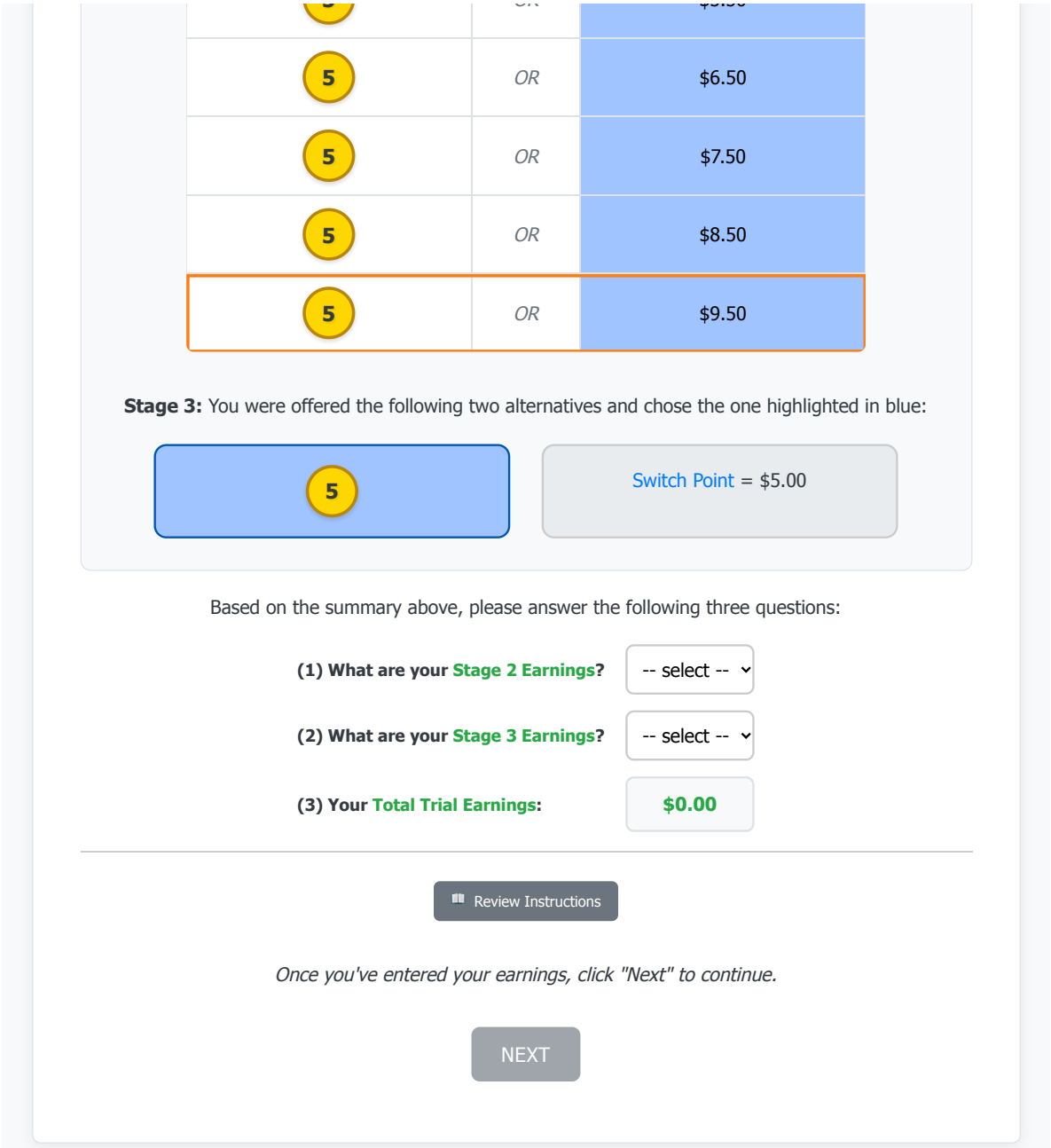
Screen 30: Practice question 2: the Stage 3 alternatives.



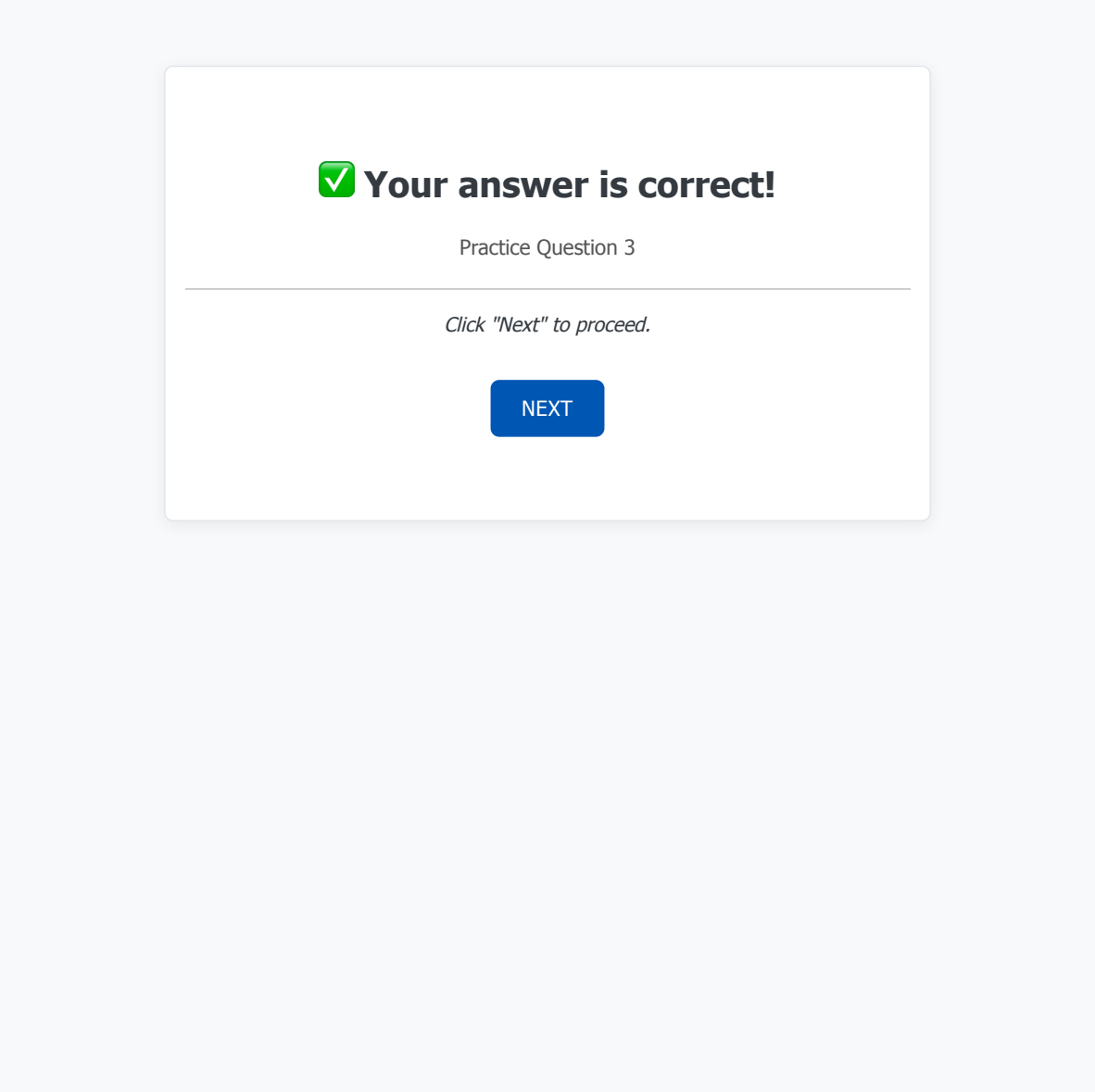
Screen 31: Practice trial, Stage 3: the choice.



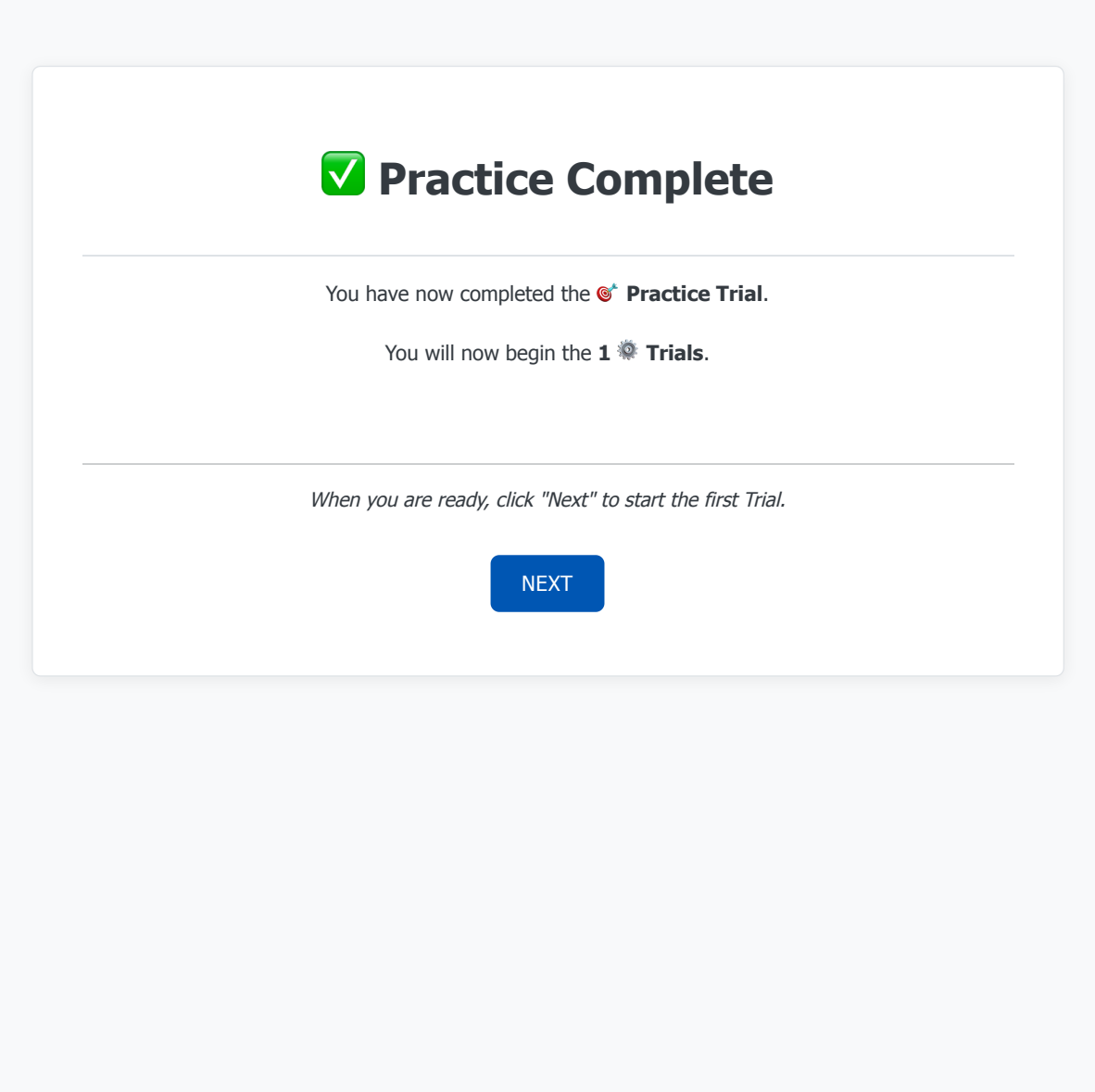
Screen 32: Practice question 3: the trial earnings, on the feedback screen. Upper half; the page continues on the next screen.



Screen 33: Practice question 3: the trial earnings, on the feedback screen. Lower half.

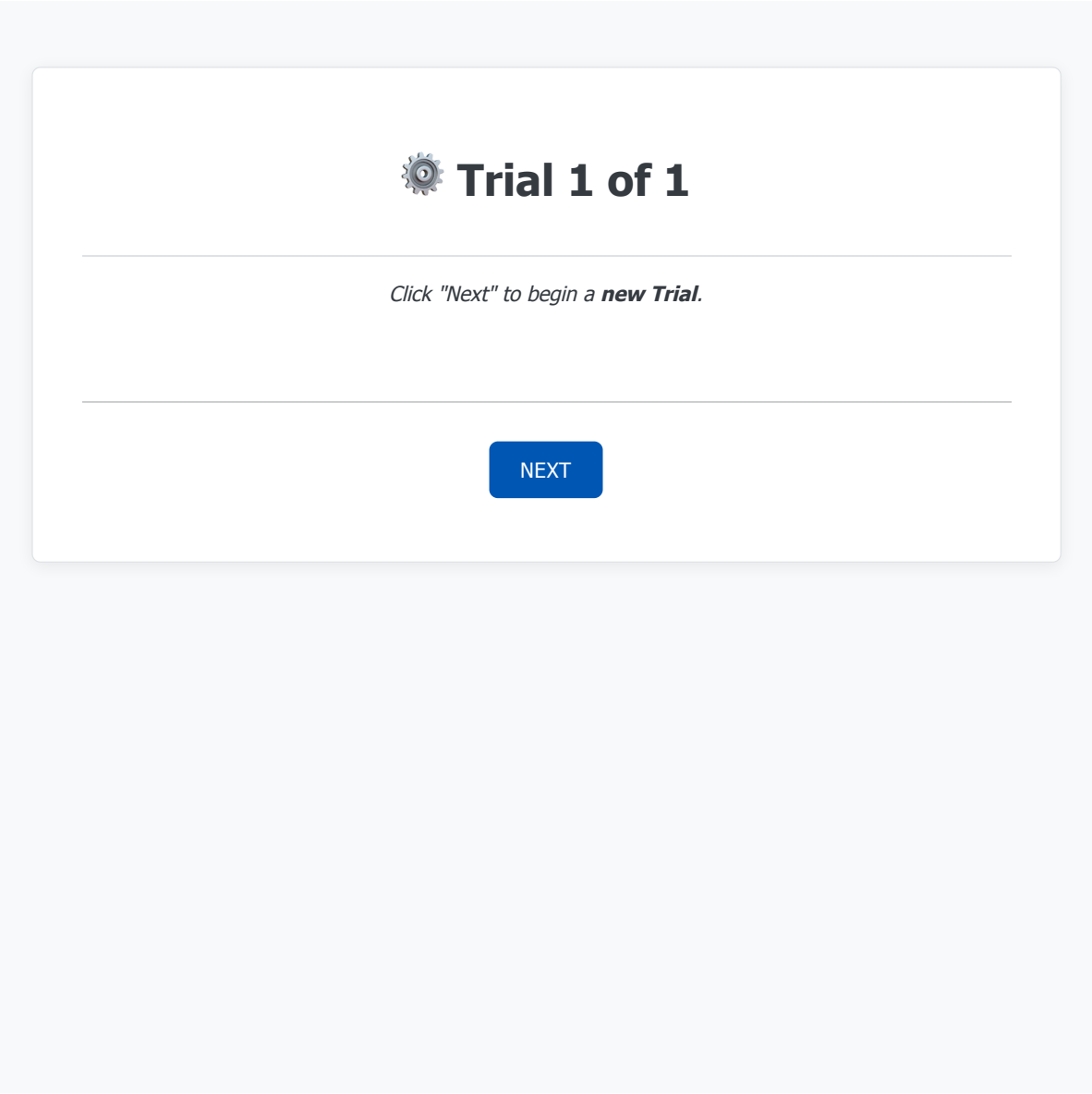


Screen 34: Feedback after a correct answer to practice question 3.

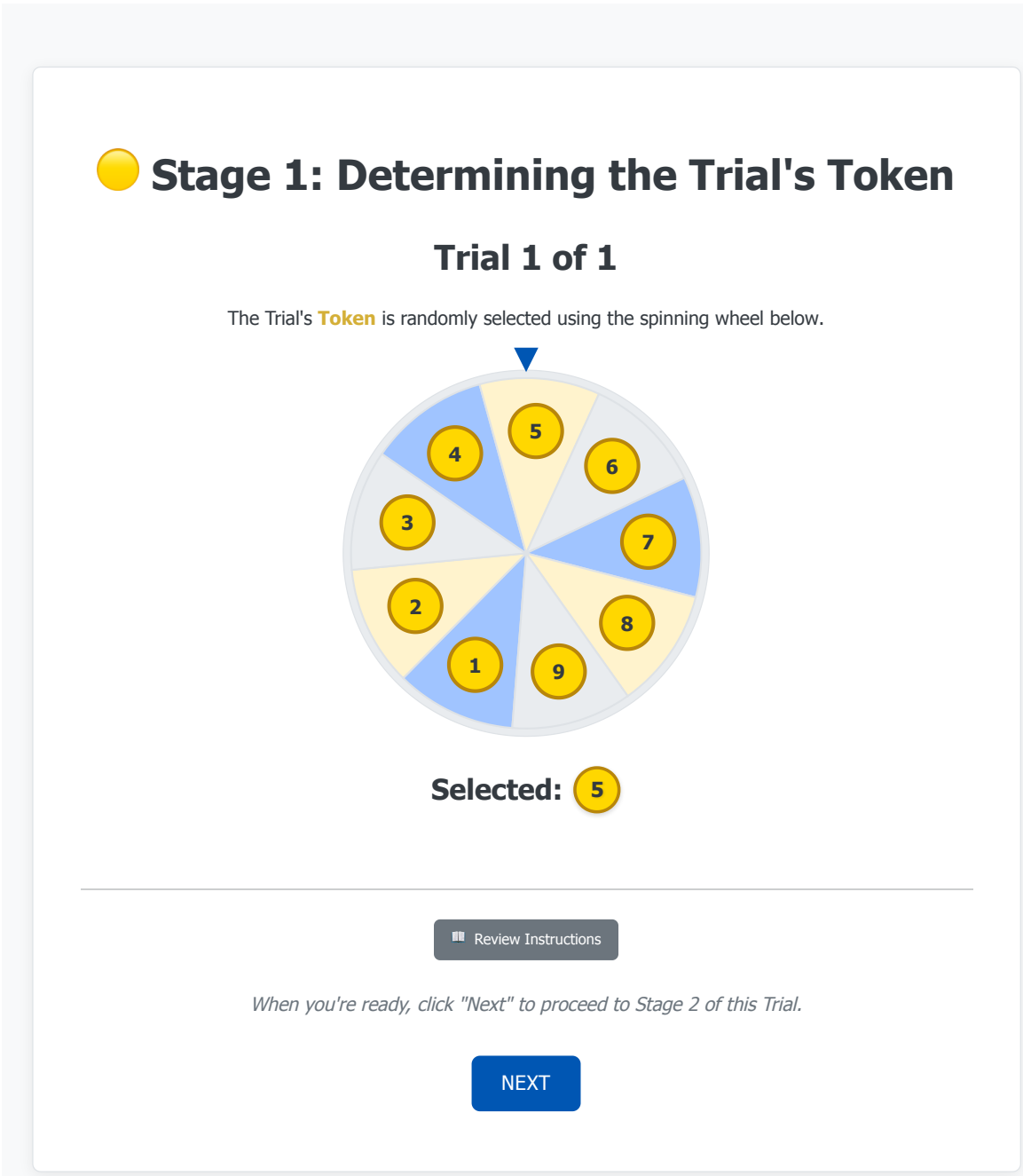


Screen 35: Start of the paid trials.

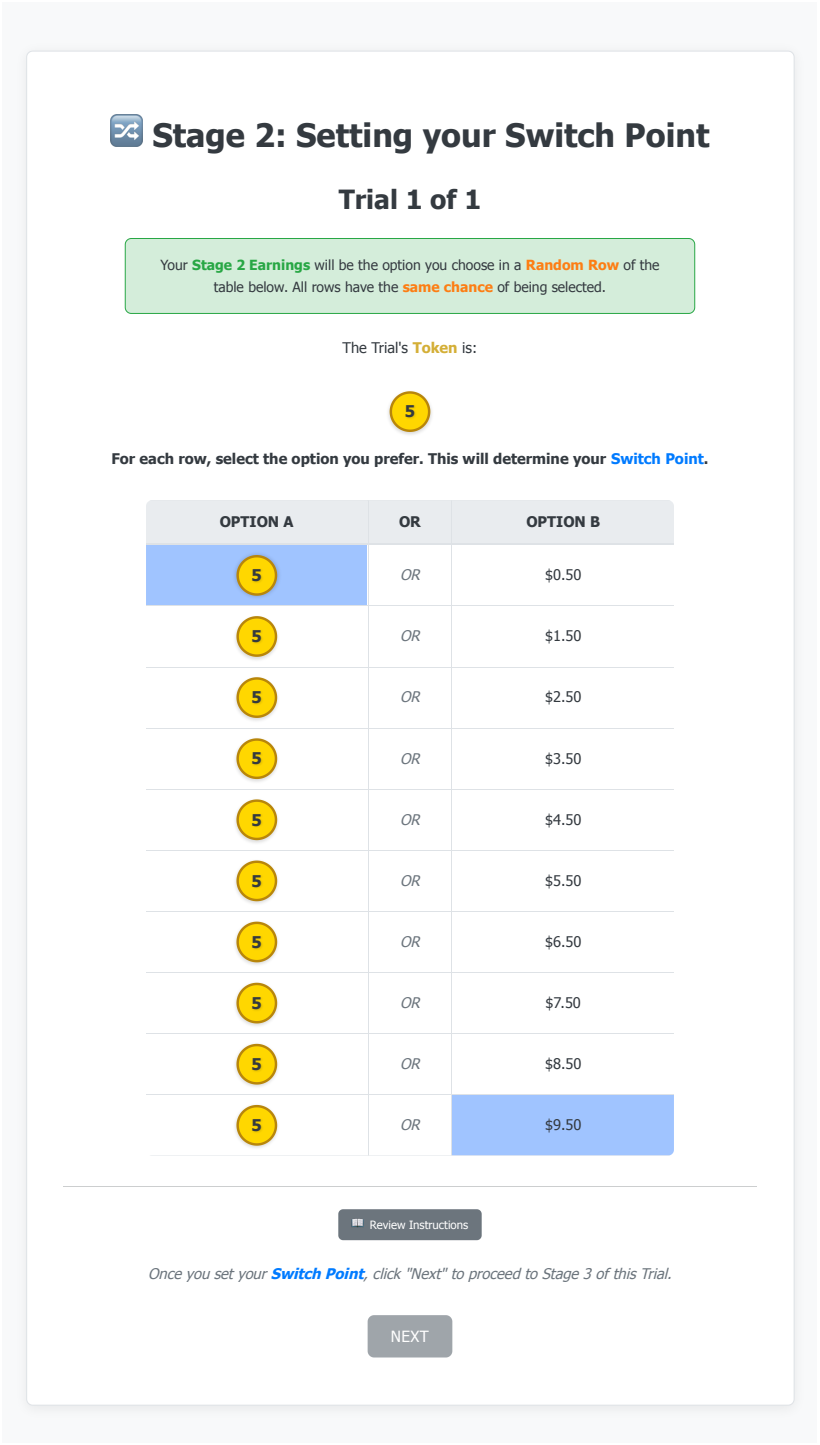
I.3 A trial



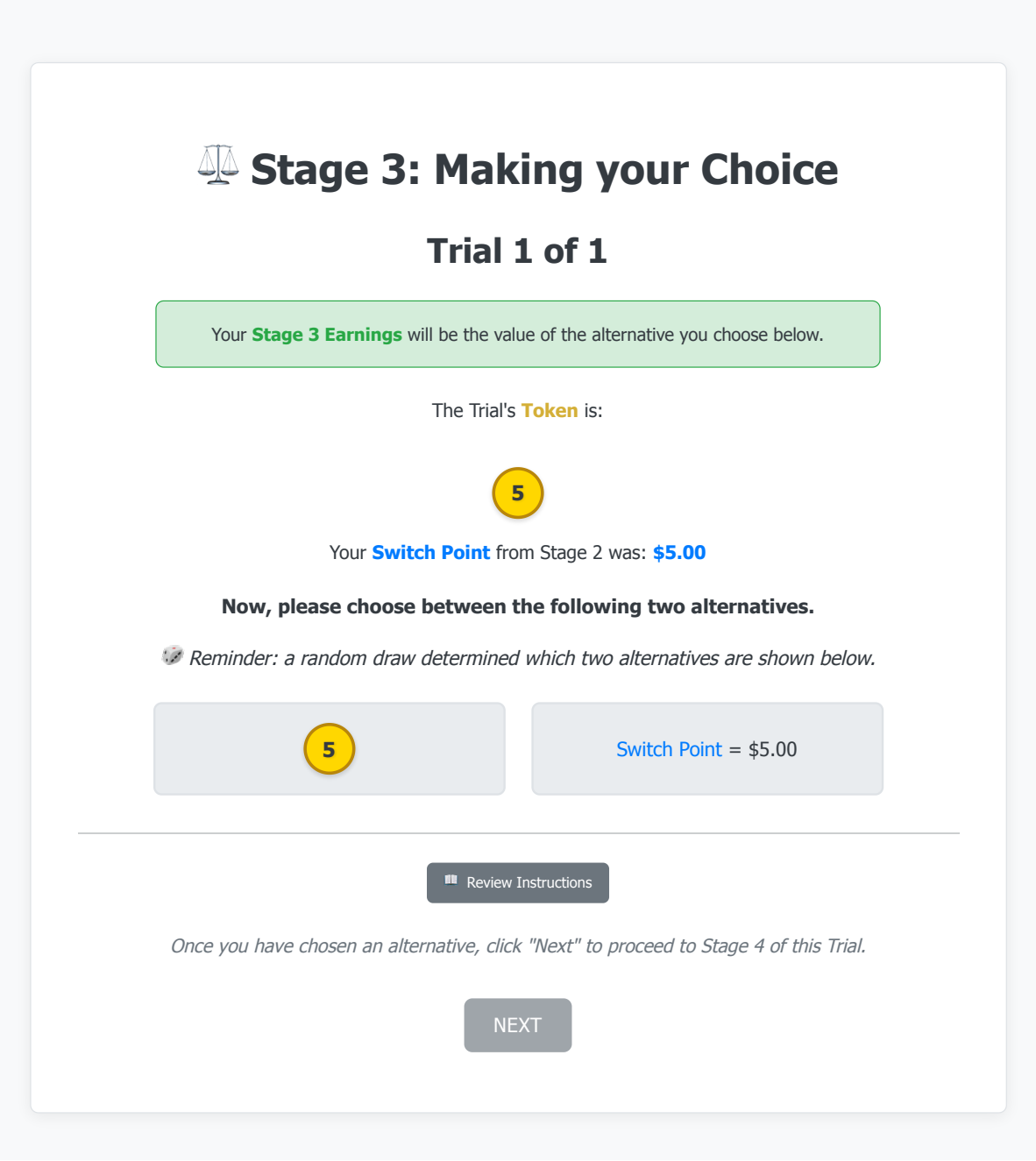
Screen 36: Start of a trial.



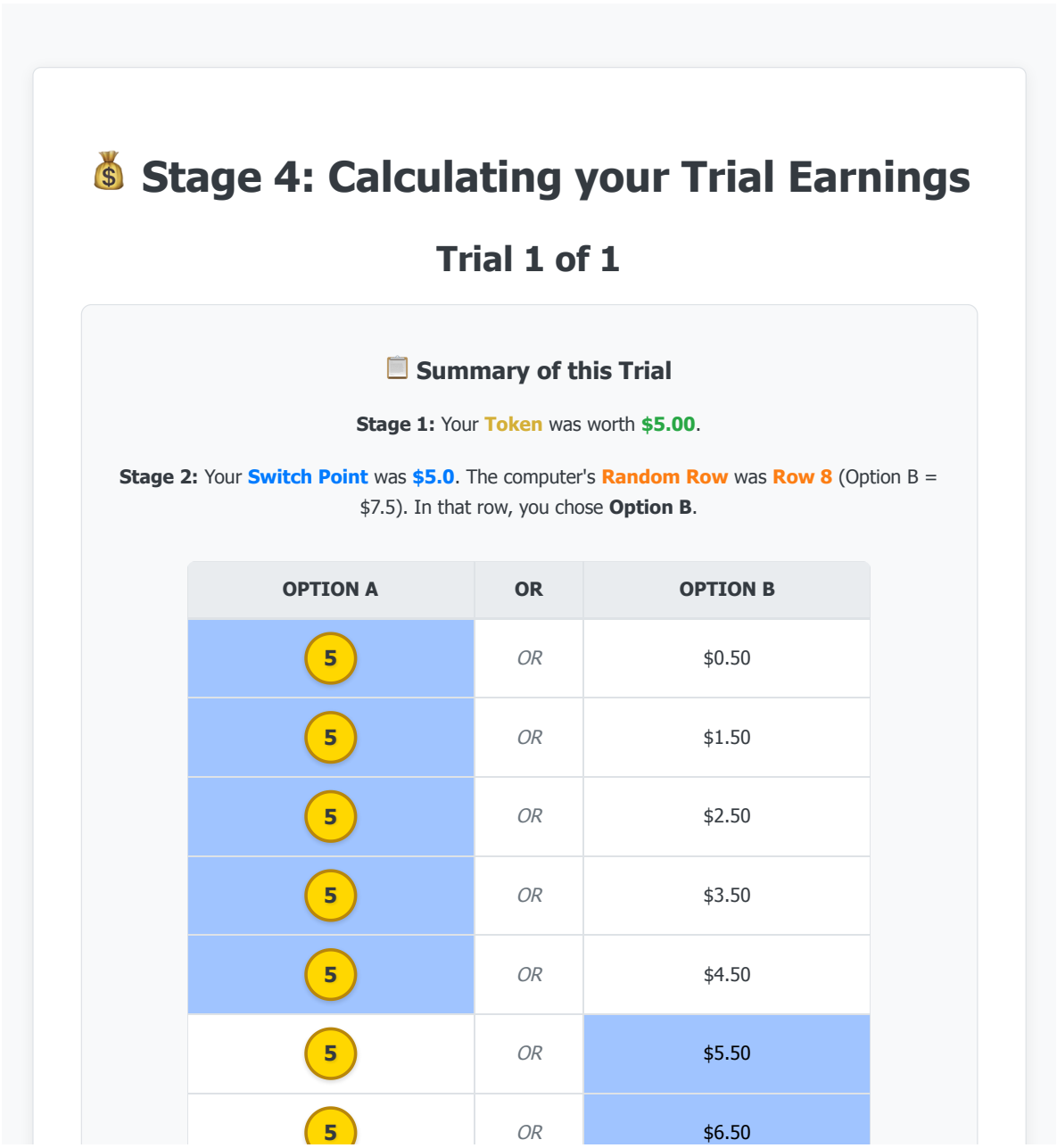
Screen 37: Stage 1: the token.



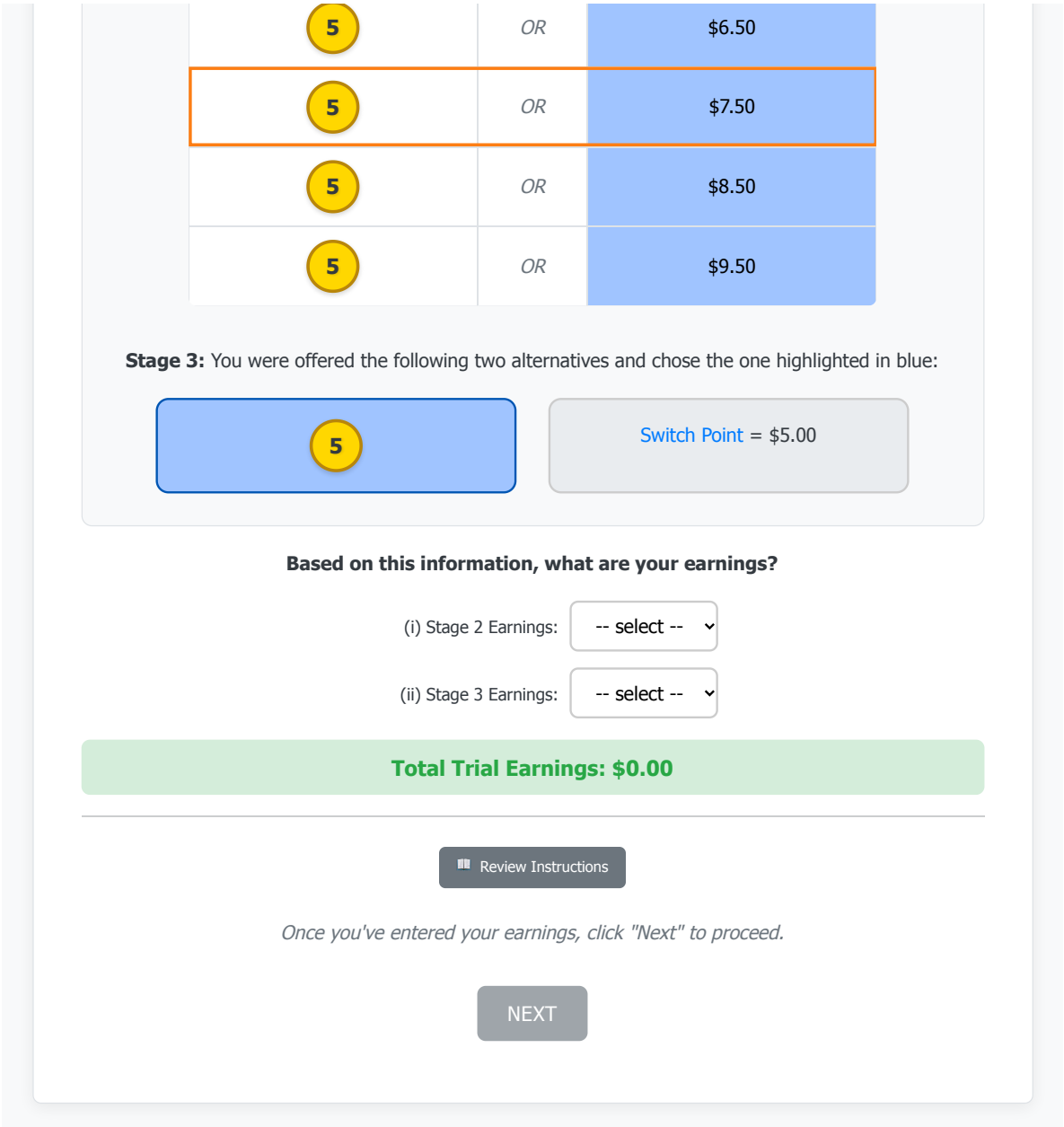
Screen 38: Stage 2: the price list.



Screen 39: Stage 3: the choice.

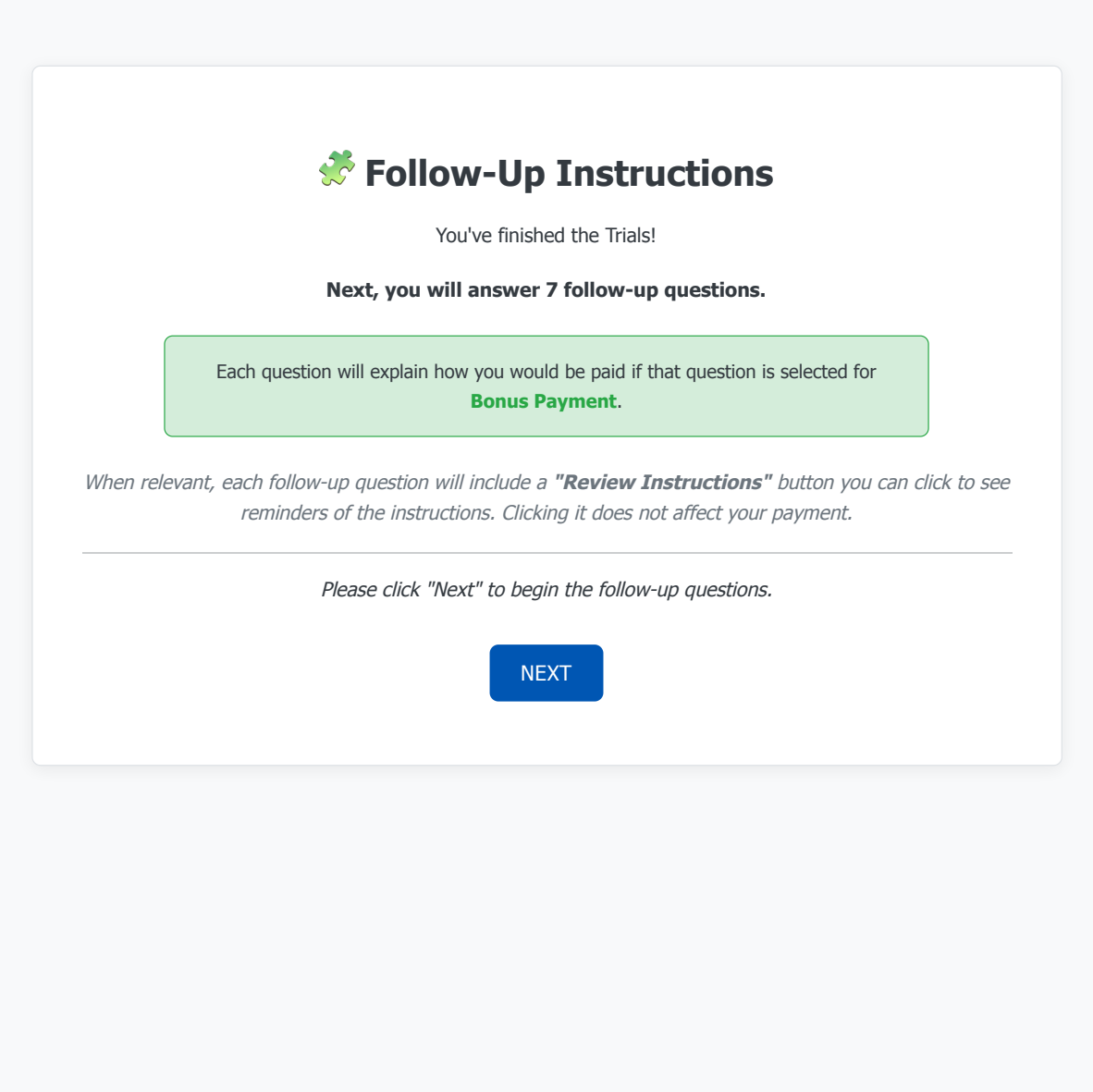


Screen 40: Stage 4: the Interactive Feedback. Upper half; the page continues on the next screen.

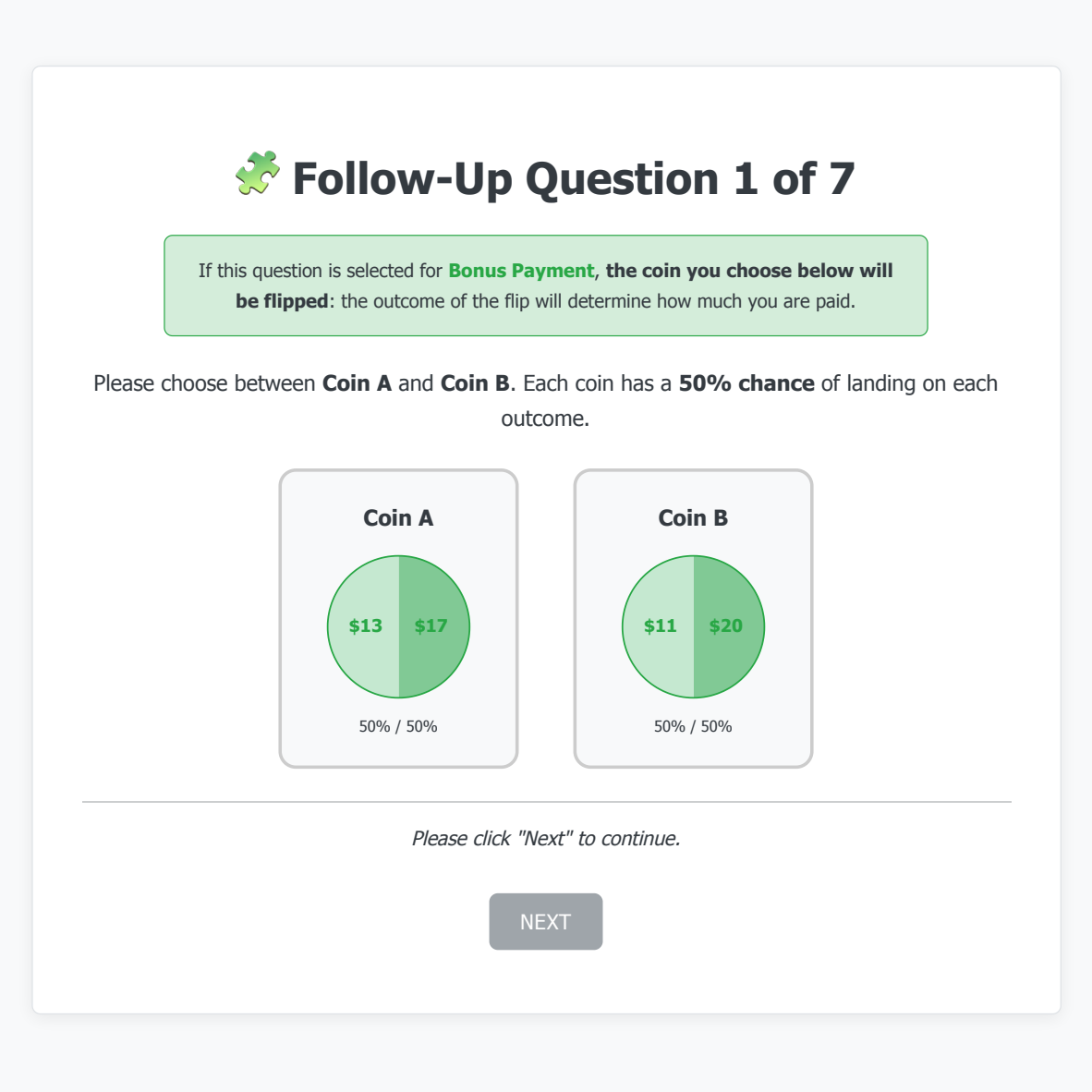


Screen 41: Stage 4: the Interactive Feedback. Lower half.

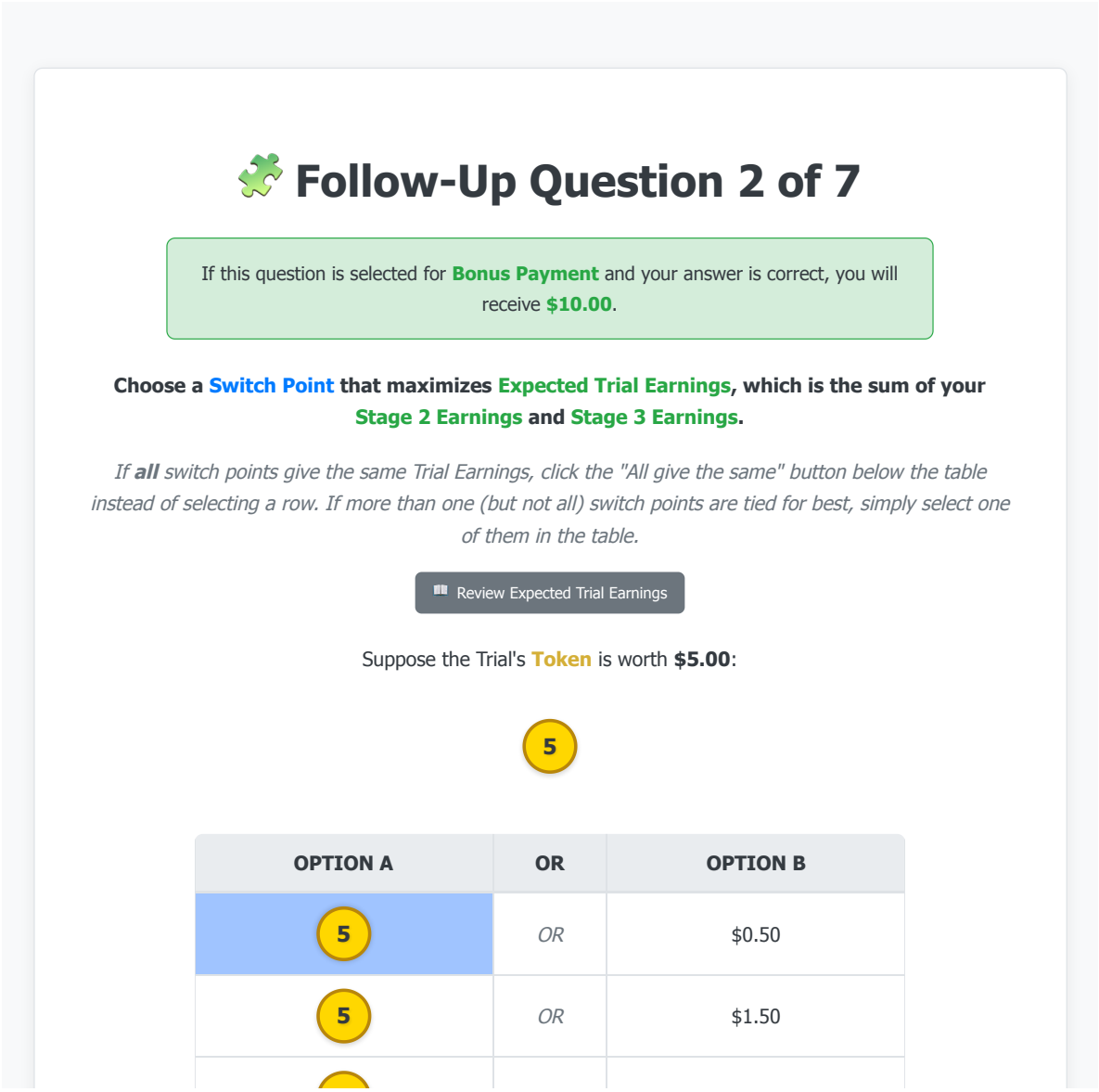
I.4 The follow-up questions



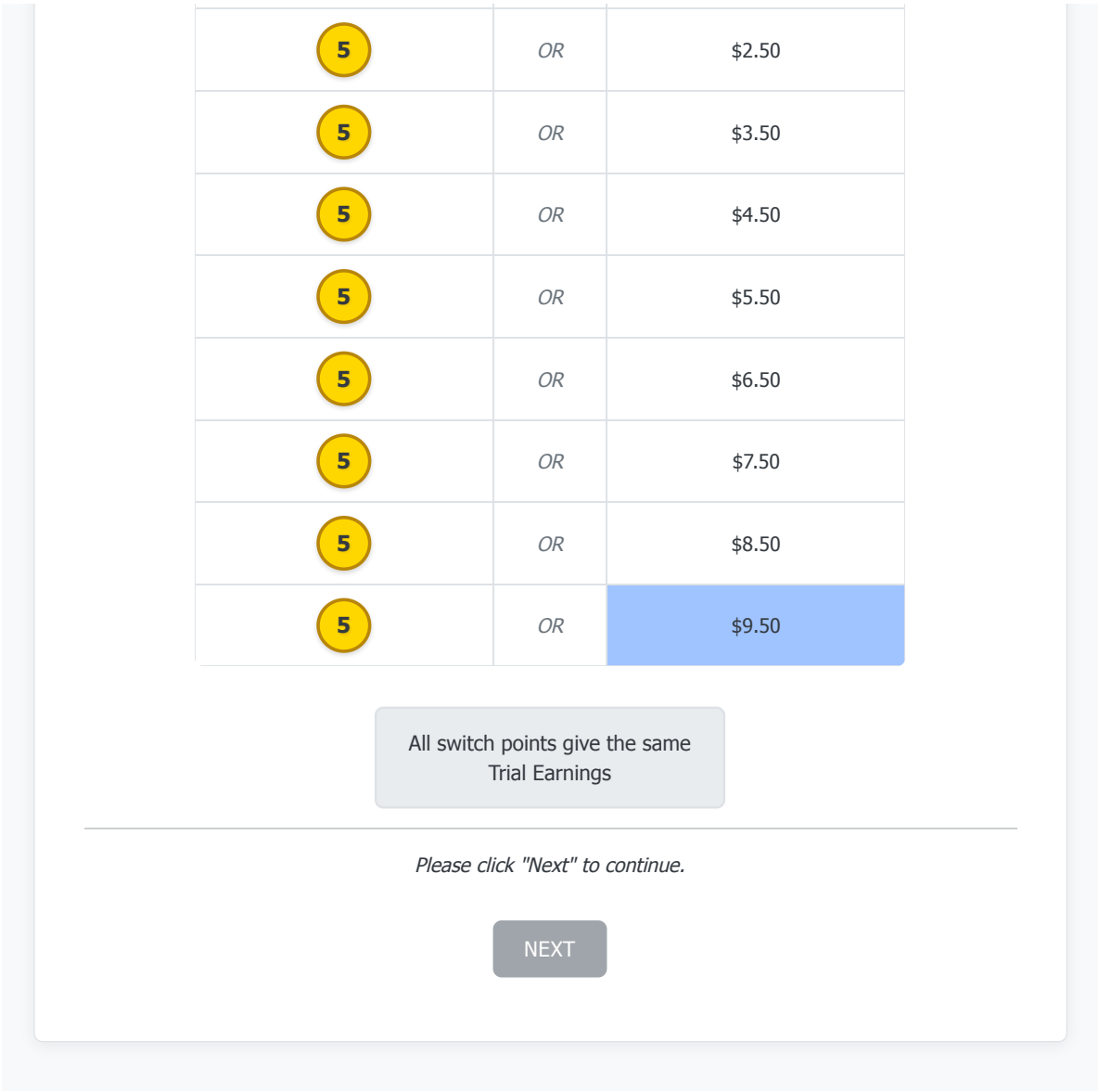
Screen 42: Introduction to the follow-up questions.



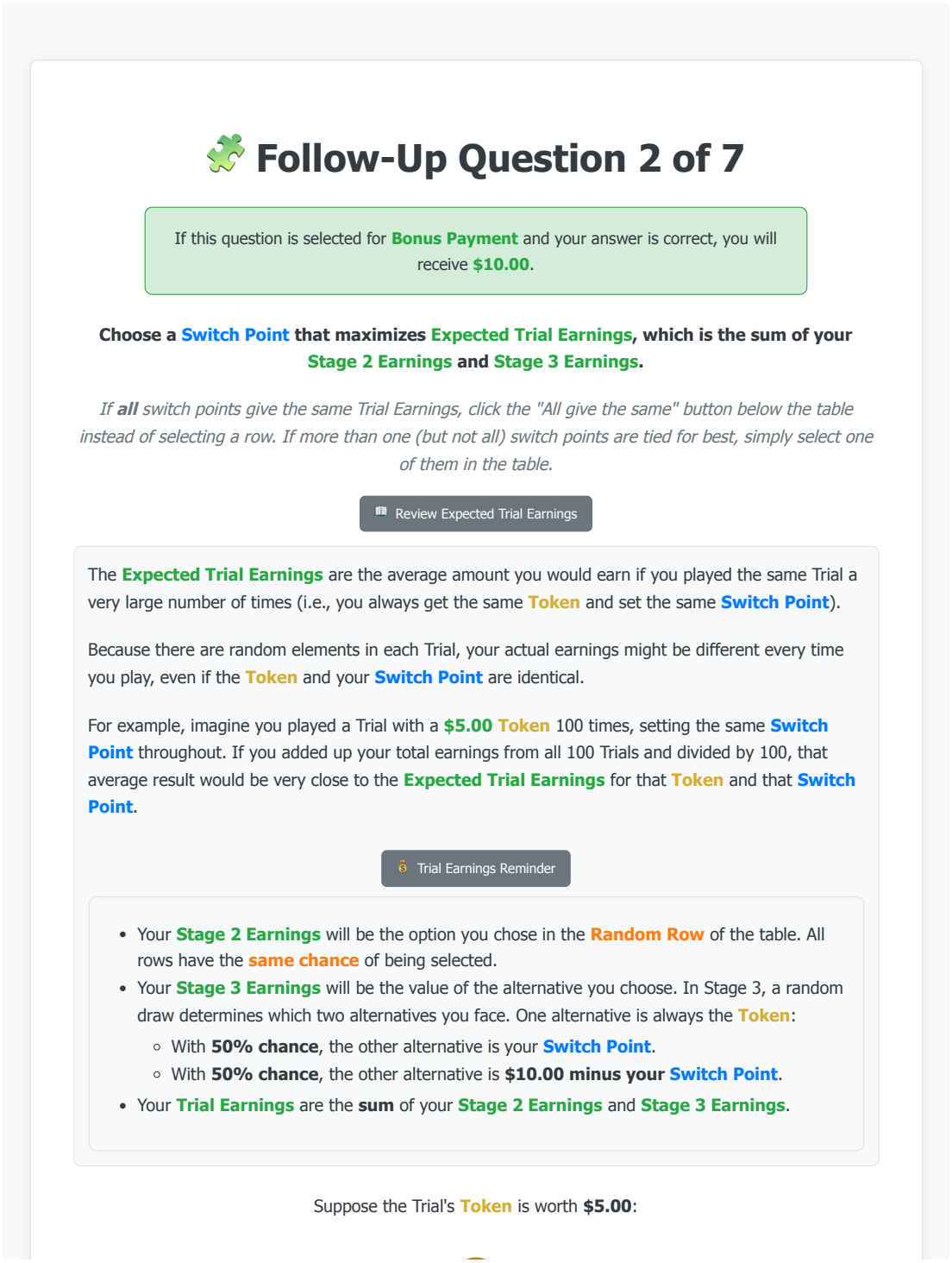
Screen 43: Follow-up question 1: the narrow-framing choice, integrated version. The order of the two versions is randomized across participants; here the integrated version came first.



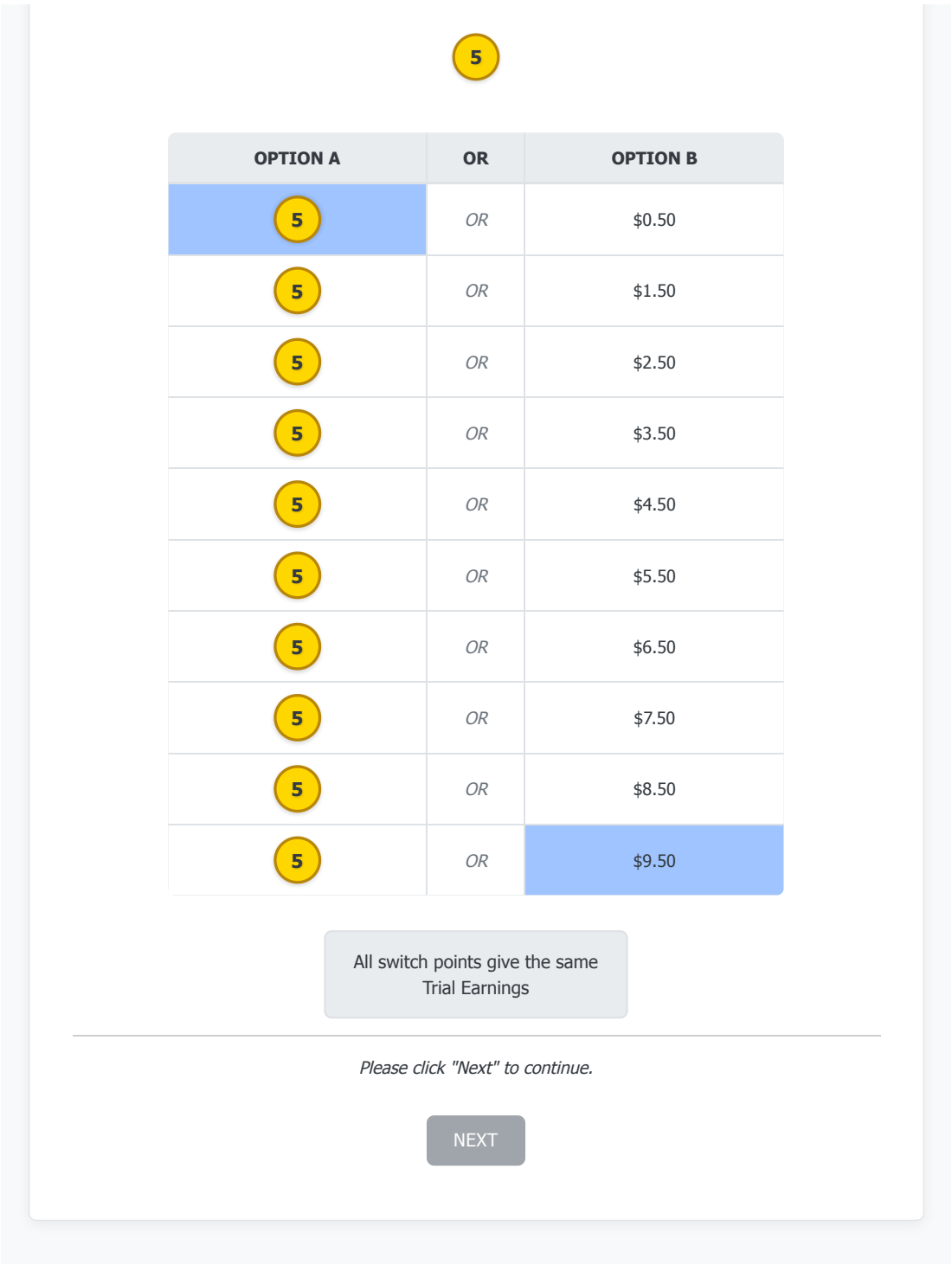
Screen 44: Follow-up question 2: the combined question. Upper half; the page continues on the next screen.



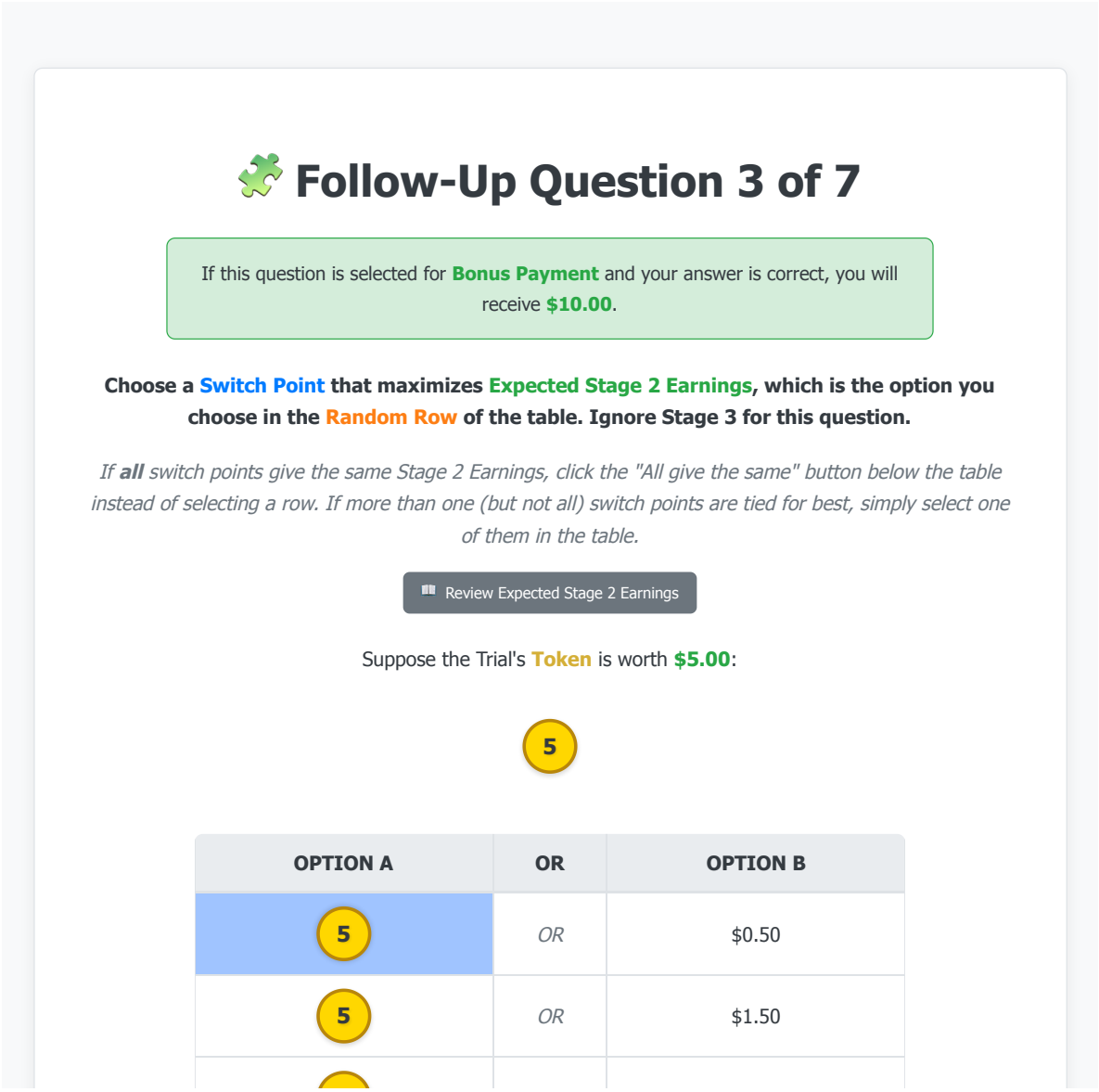
Screen 45: Follow-up question 2: the combined question. Lower half.



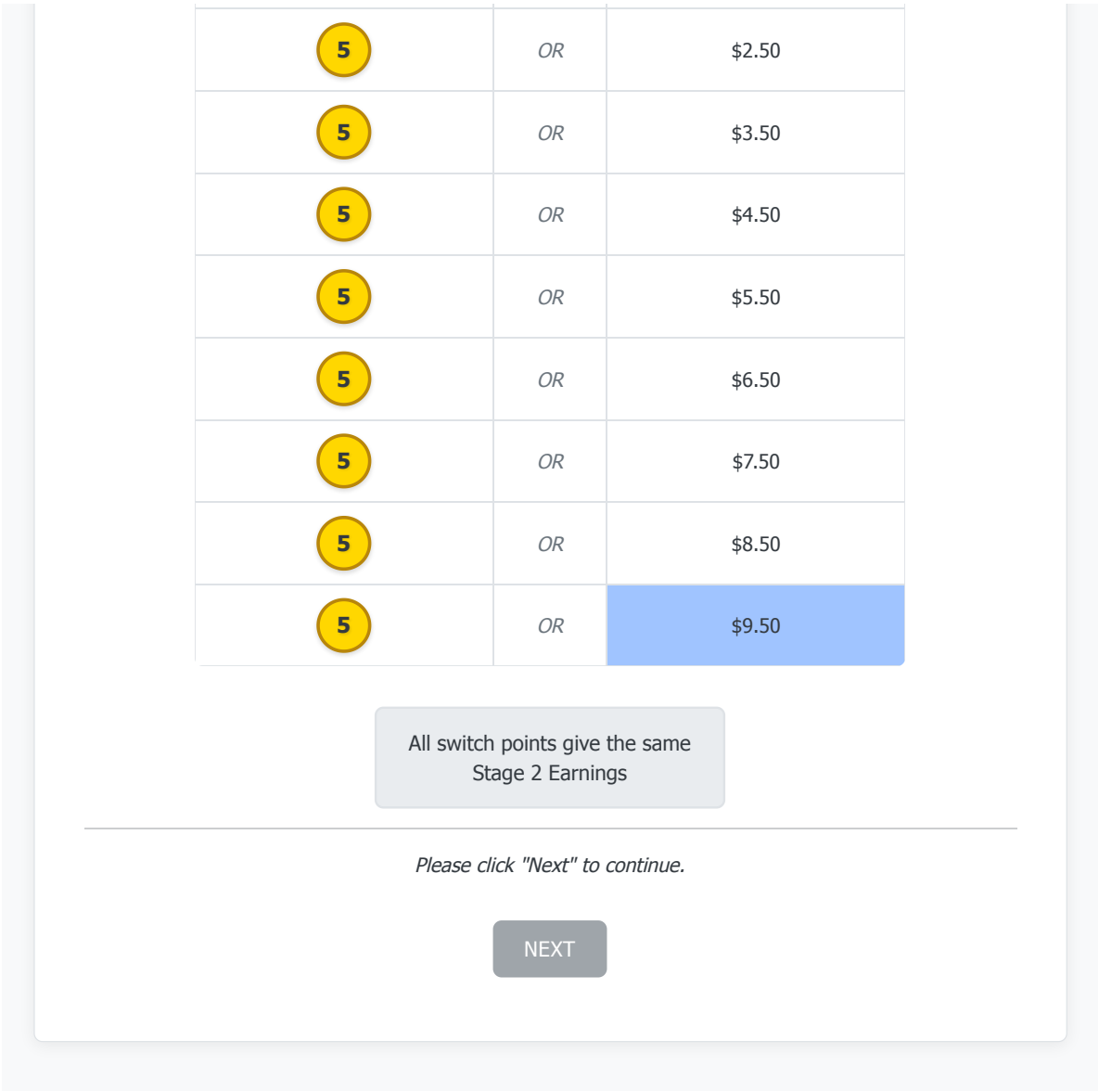
Screen 46: Follow-up question 2 with the review box expanded. Upper half; the page continues on the next screen.



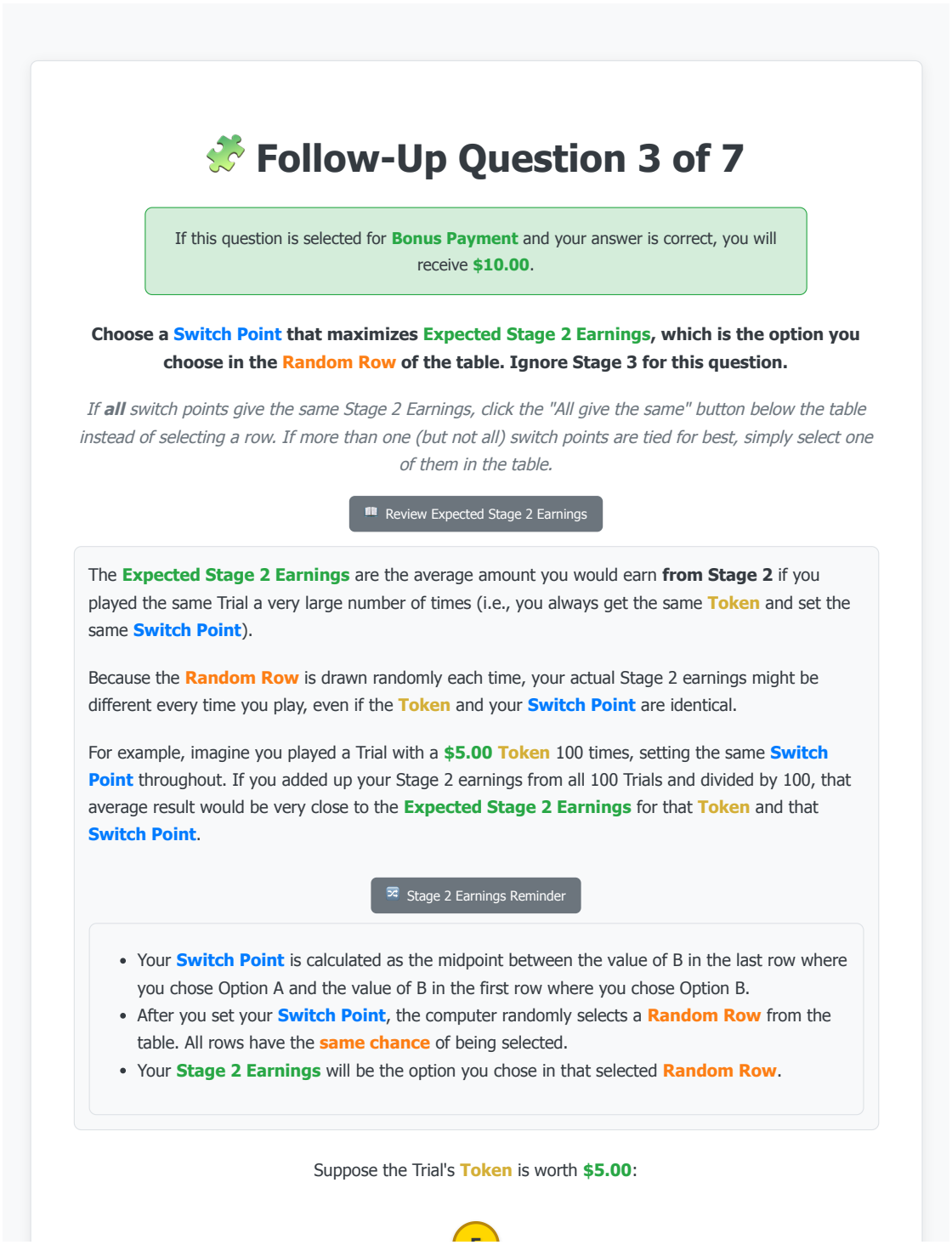
Screen 47: Follow-up question 2 with the review box expanded. Lower half.



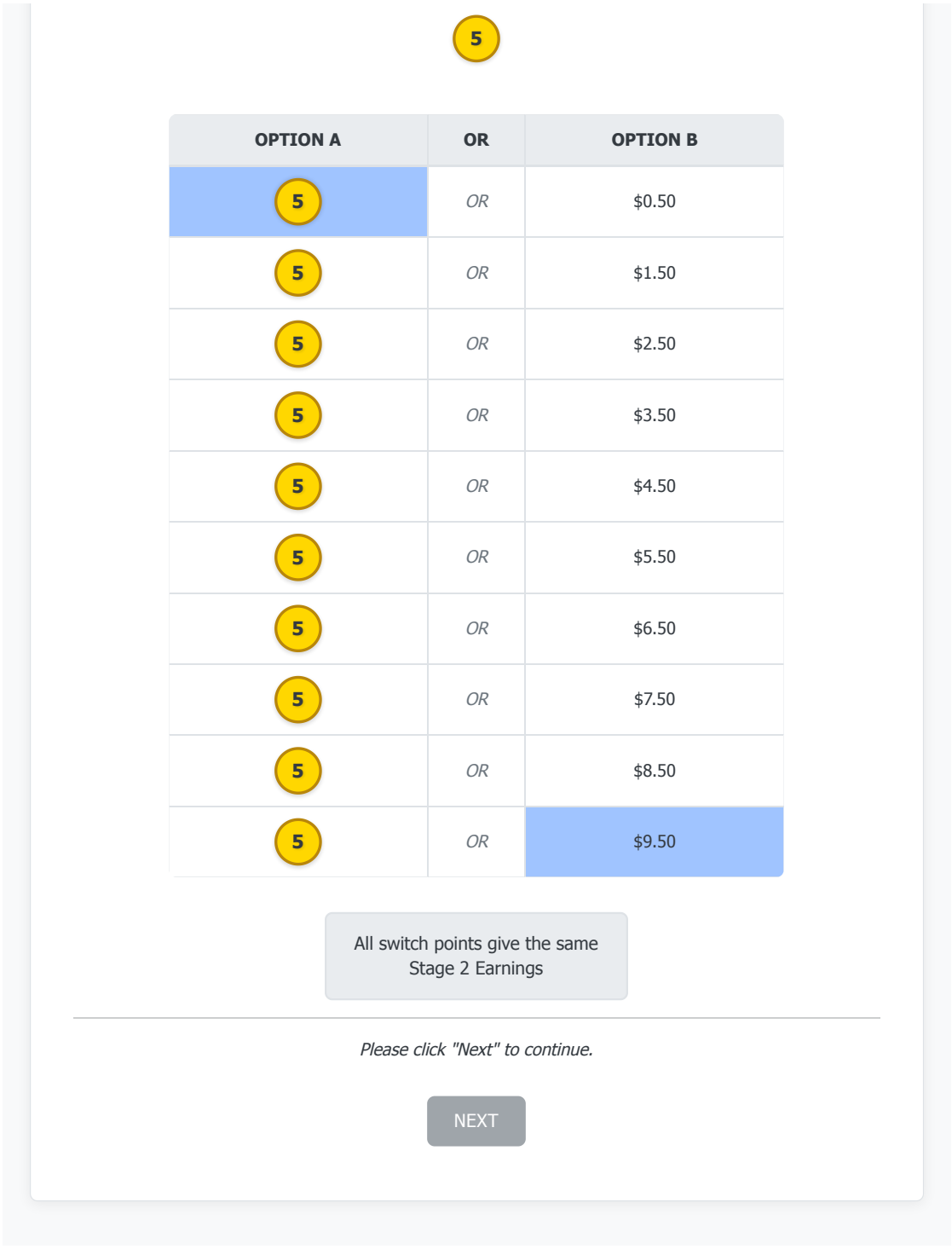
Screen 48: Follow-up question 3: the measurement question. Upper half; the page continues on the next screen.



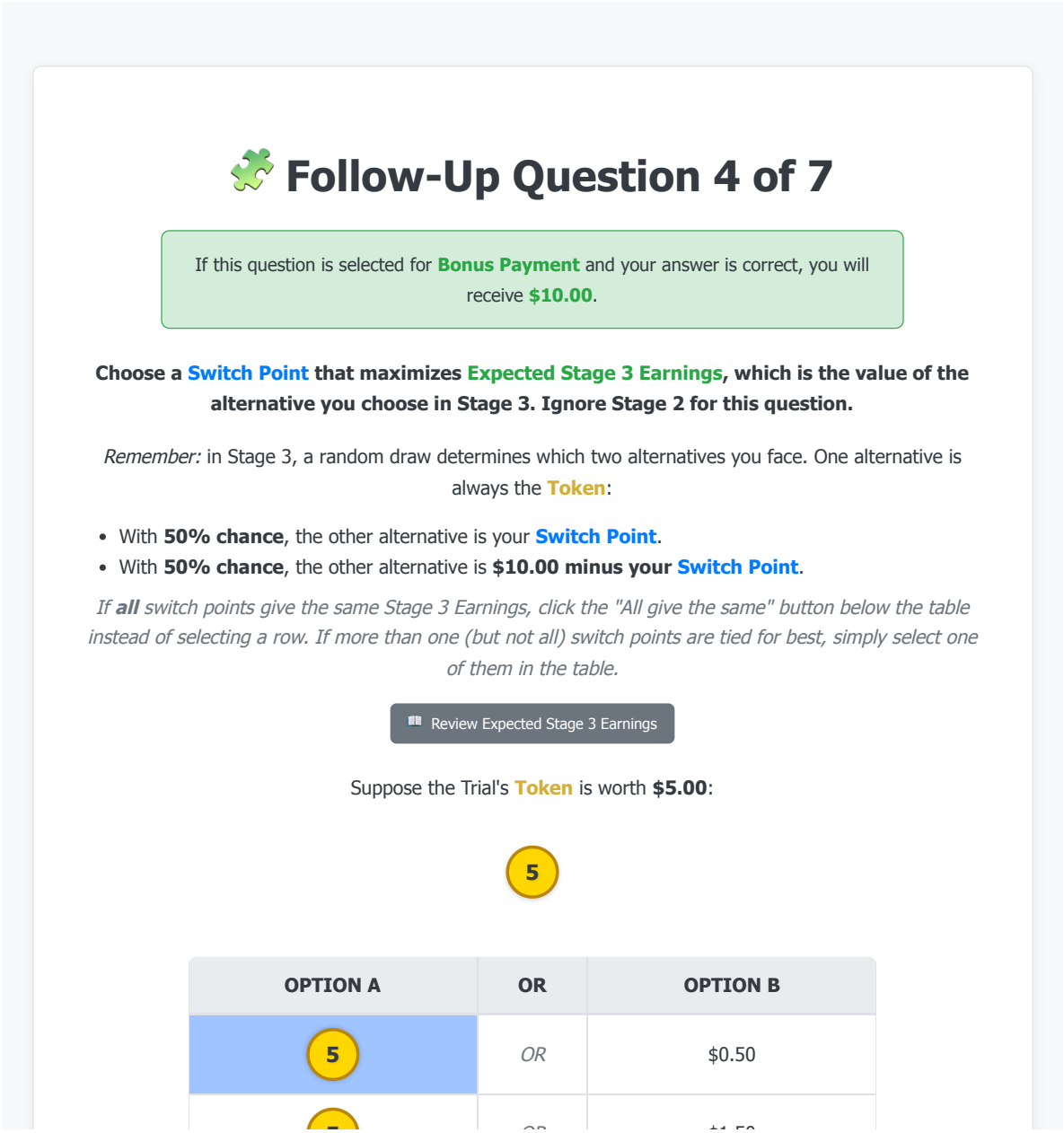
Screen 49: Follow-up question 3: the measurement question. Lower half.



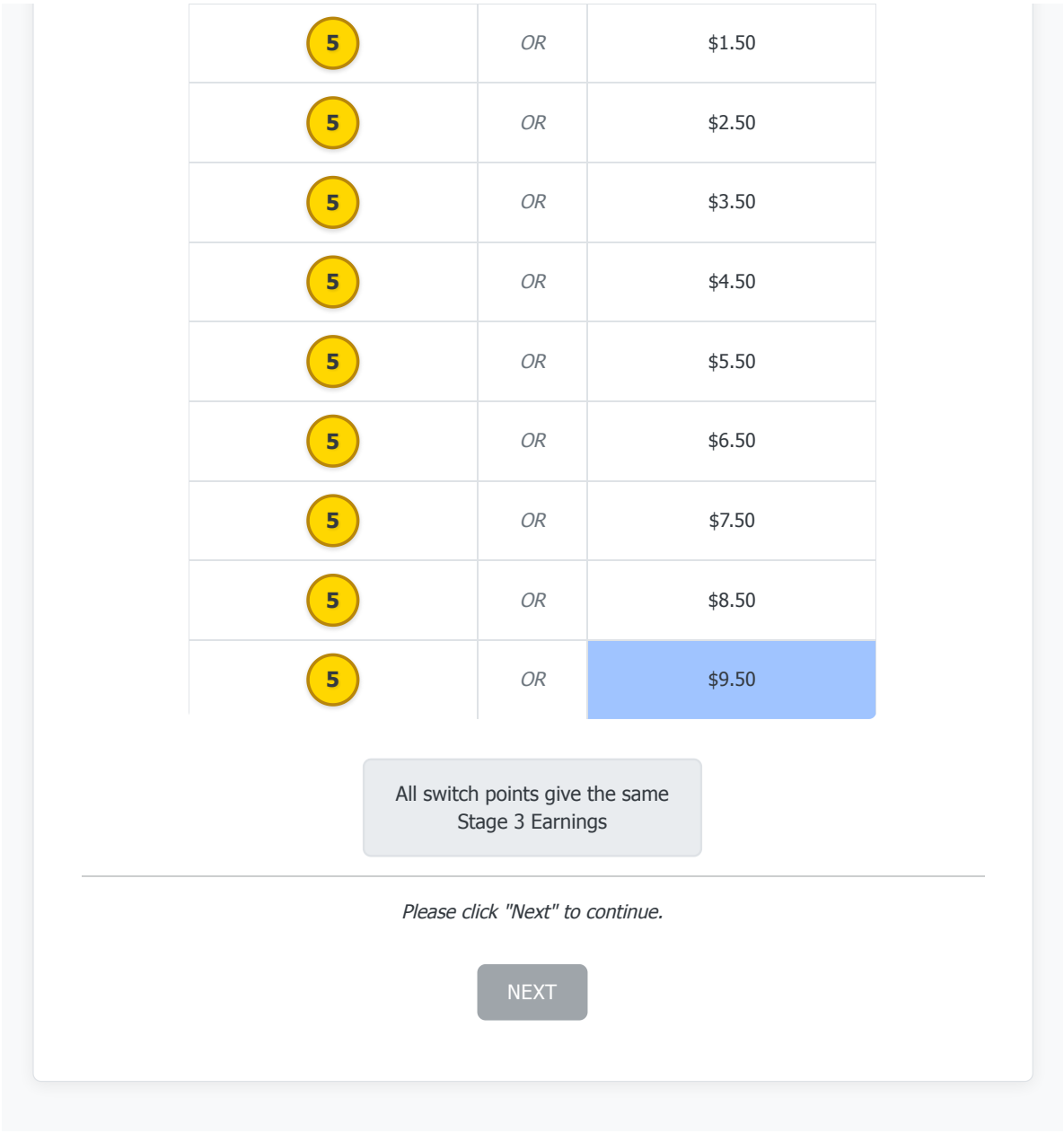
Screen 50: Follow-up question 3 with the review box expanded. Upper half; the page continues on the next screen.



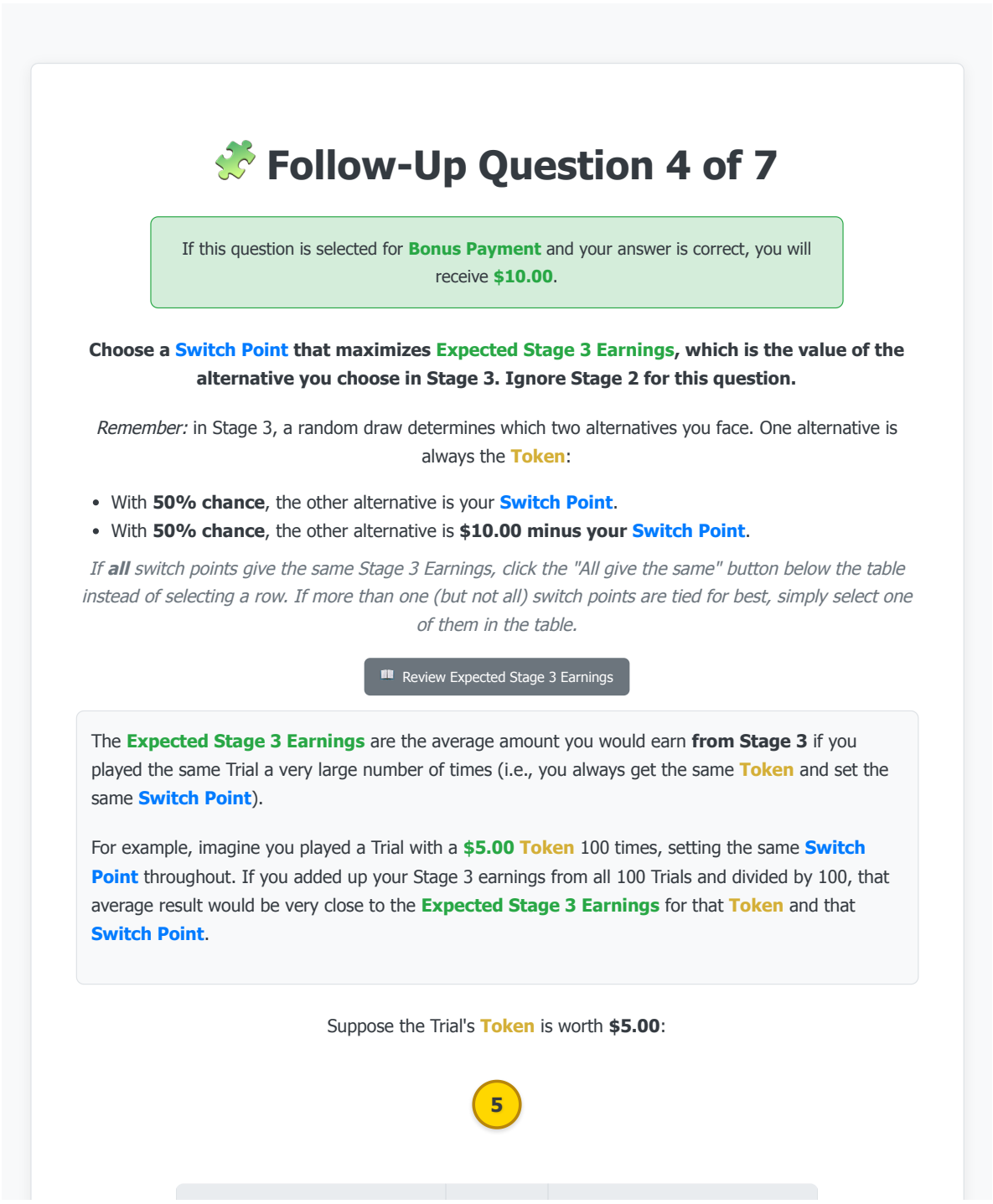
Screen 51: Follow-up question 3 with the review box expanded. Lower half.



Screen 52: Follow-up question 4: the menu question. Upper half; the page continues on the next screen.



Screen 53: Follow-up question 4: the menu question. Lower half.



Screen 54: Follow-up question 4 with the review box expanded. Upper half; the page continues on the next screen.

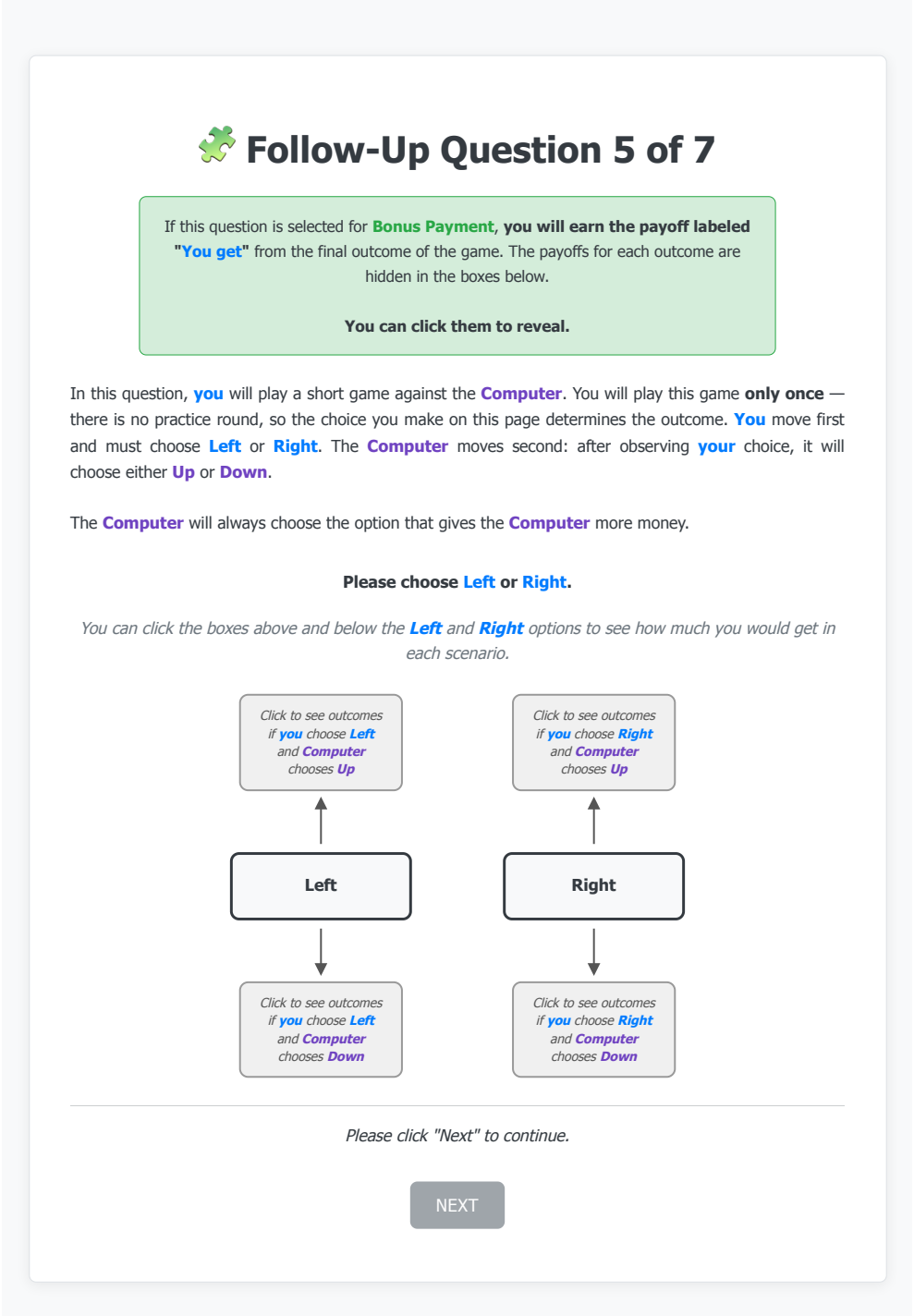
OPTION A	OR	OPTION B
5	*OR*	$0.50
5	*OR*	$1.50
5	*OR*	$2.50
5	*OR*	$3.50
5	*OR*	$4.50
5	*OR*	$5.50
5	*OR*	$6.50
5	*OR*	$7.50
5	*OR*	$8.50
5	*OR*	$9.50

All switch points give the same Stage 3 Earnings

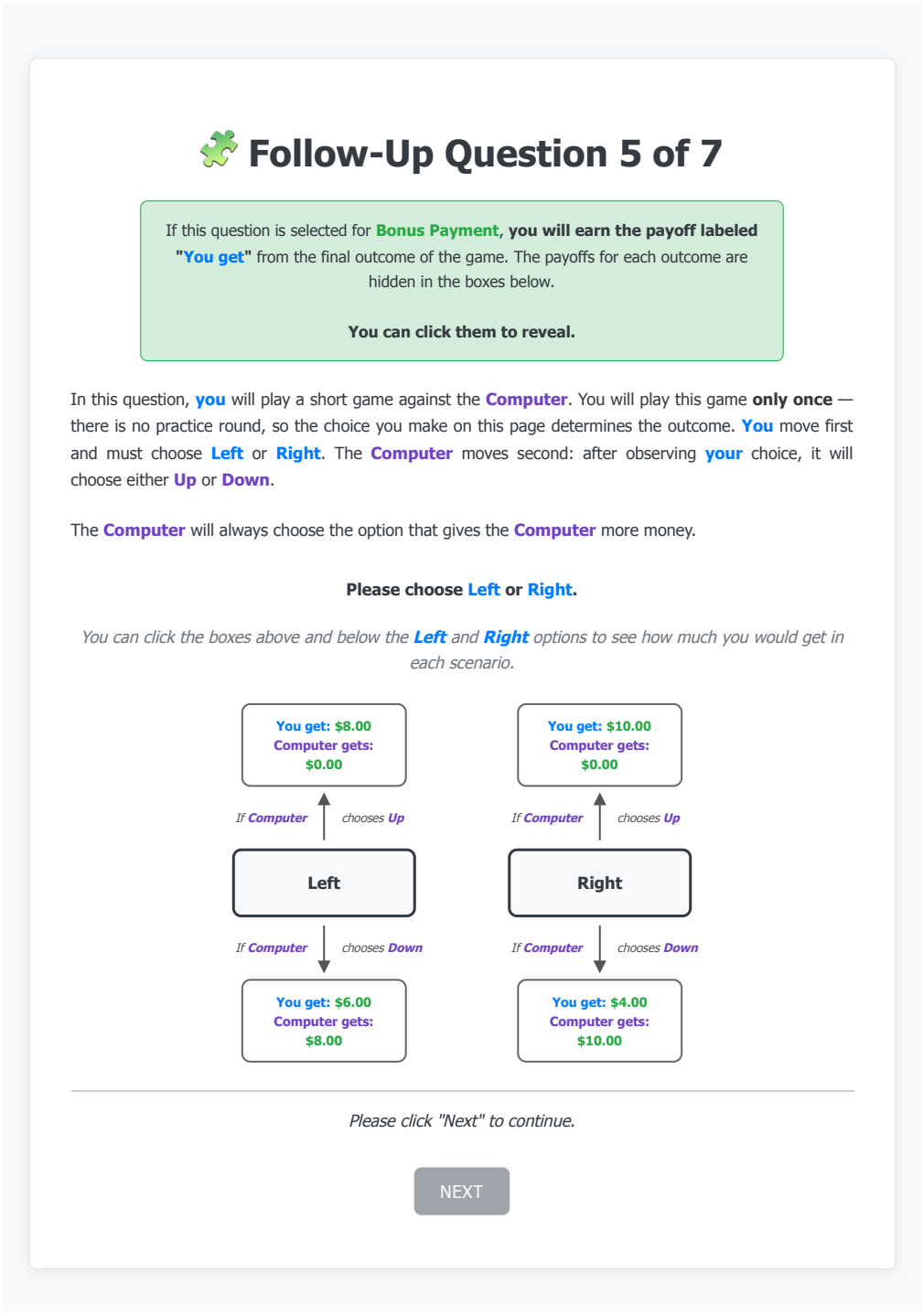
Please click "Next" to continue.

NEXT

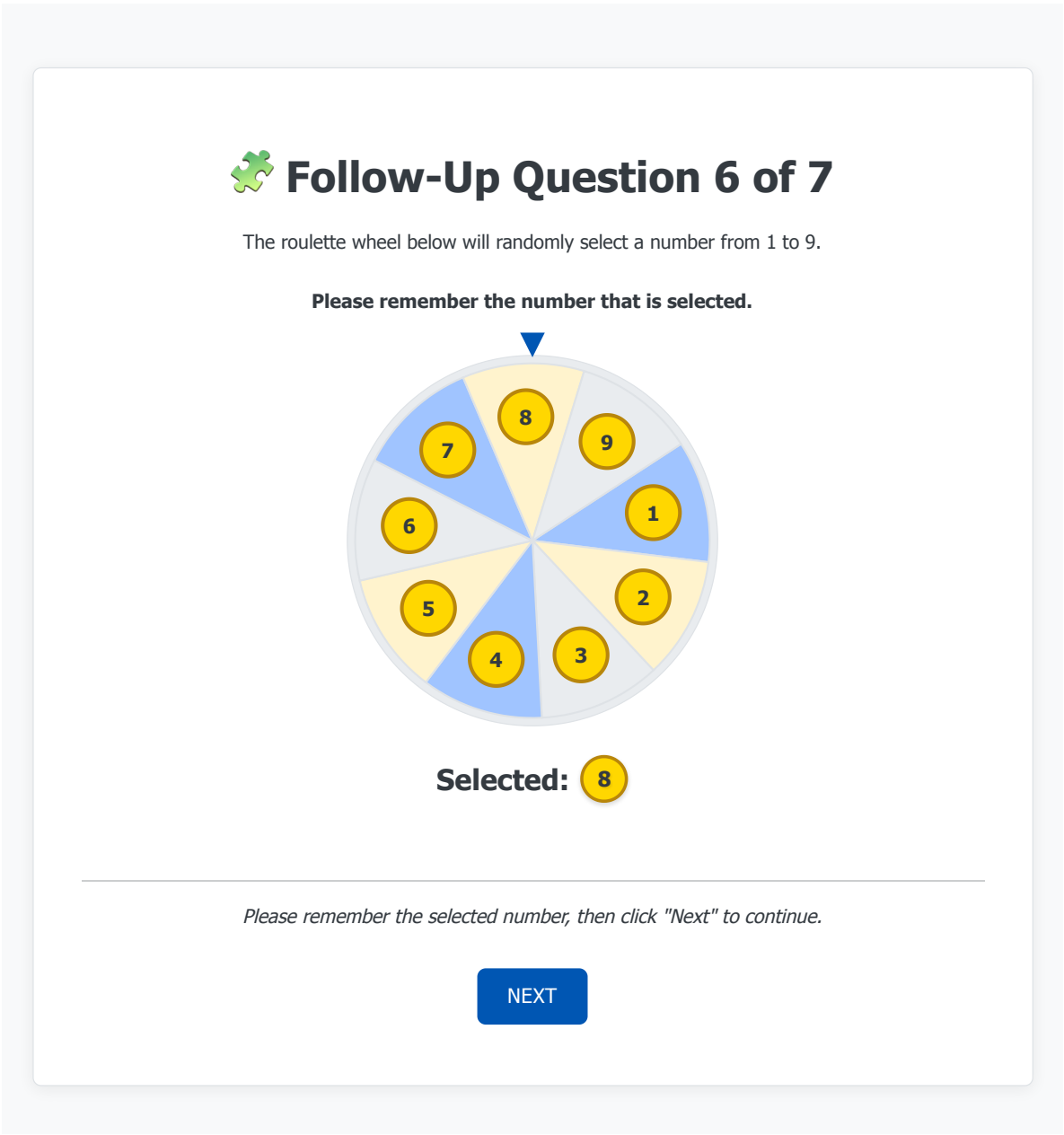
Screen 55: Follow-up question 4 with the review box expanded. Lower half.



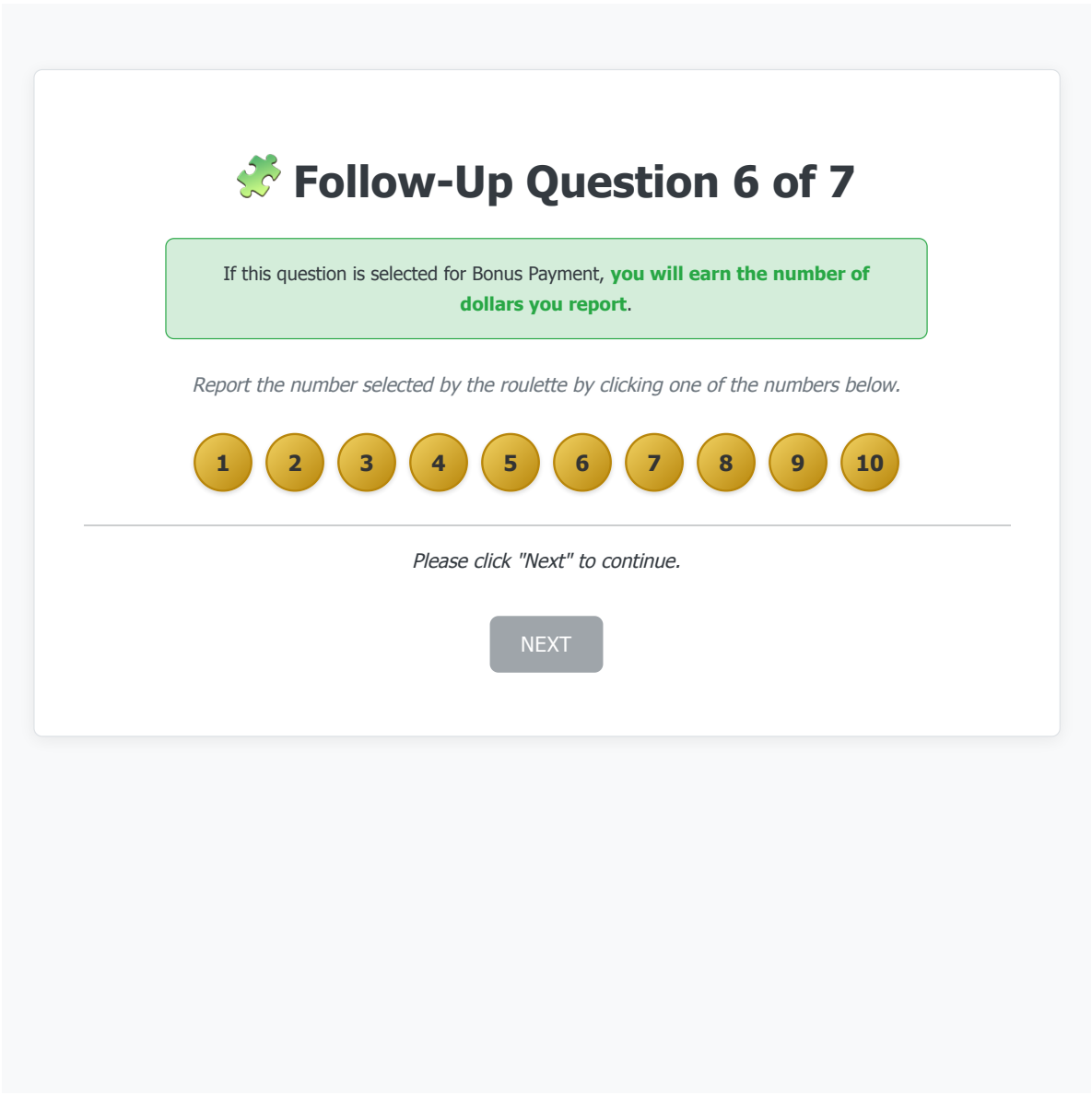
Screen 56: Follow-up question 5: the backward-induction game, payoffs hidden.



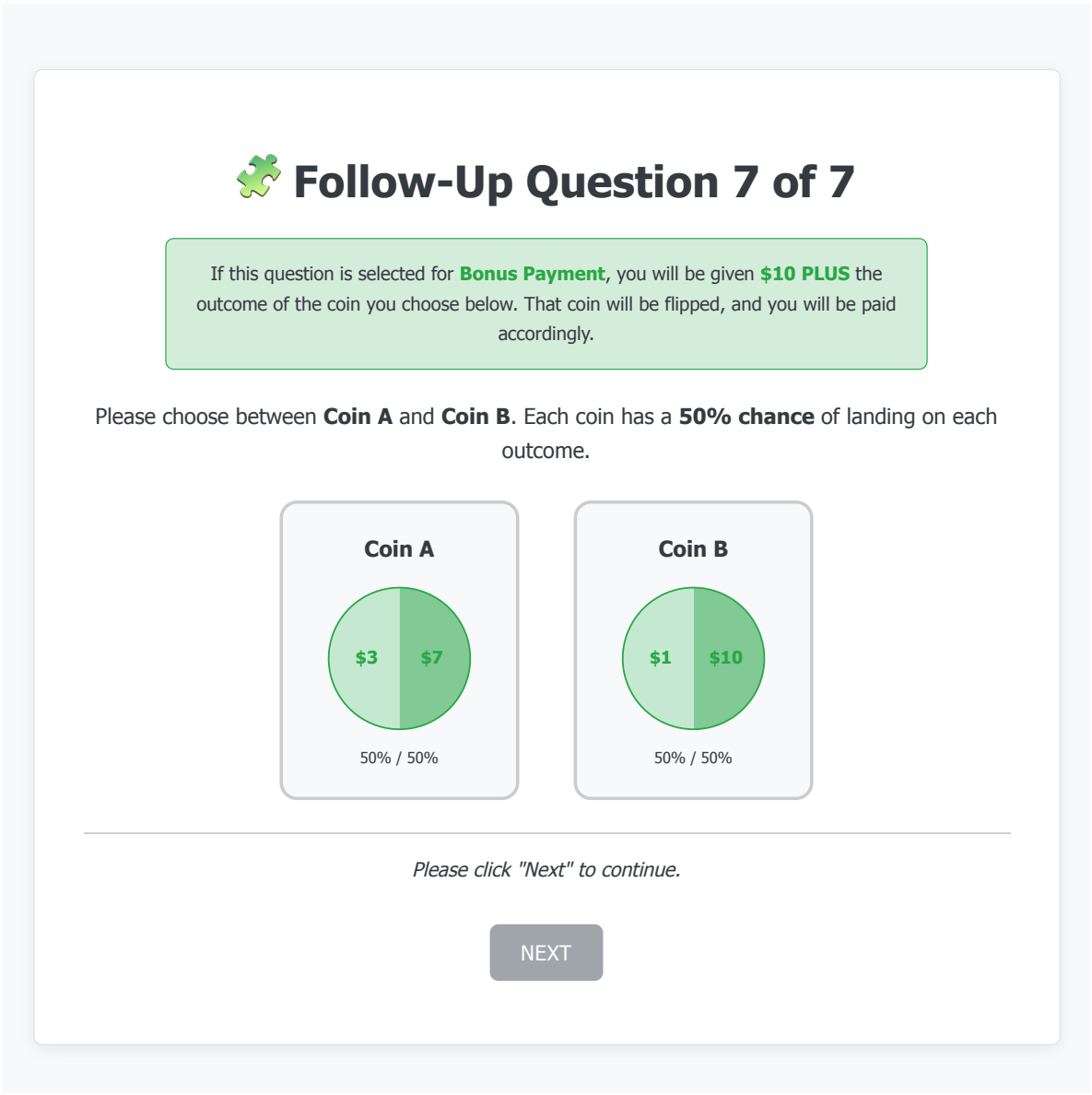
Screen 57: Follow-up question 5 with the four payoff boxes revealed.



Screen 58: Follow-up question 6: the lying-aversion item, the wheel.

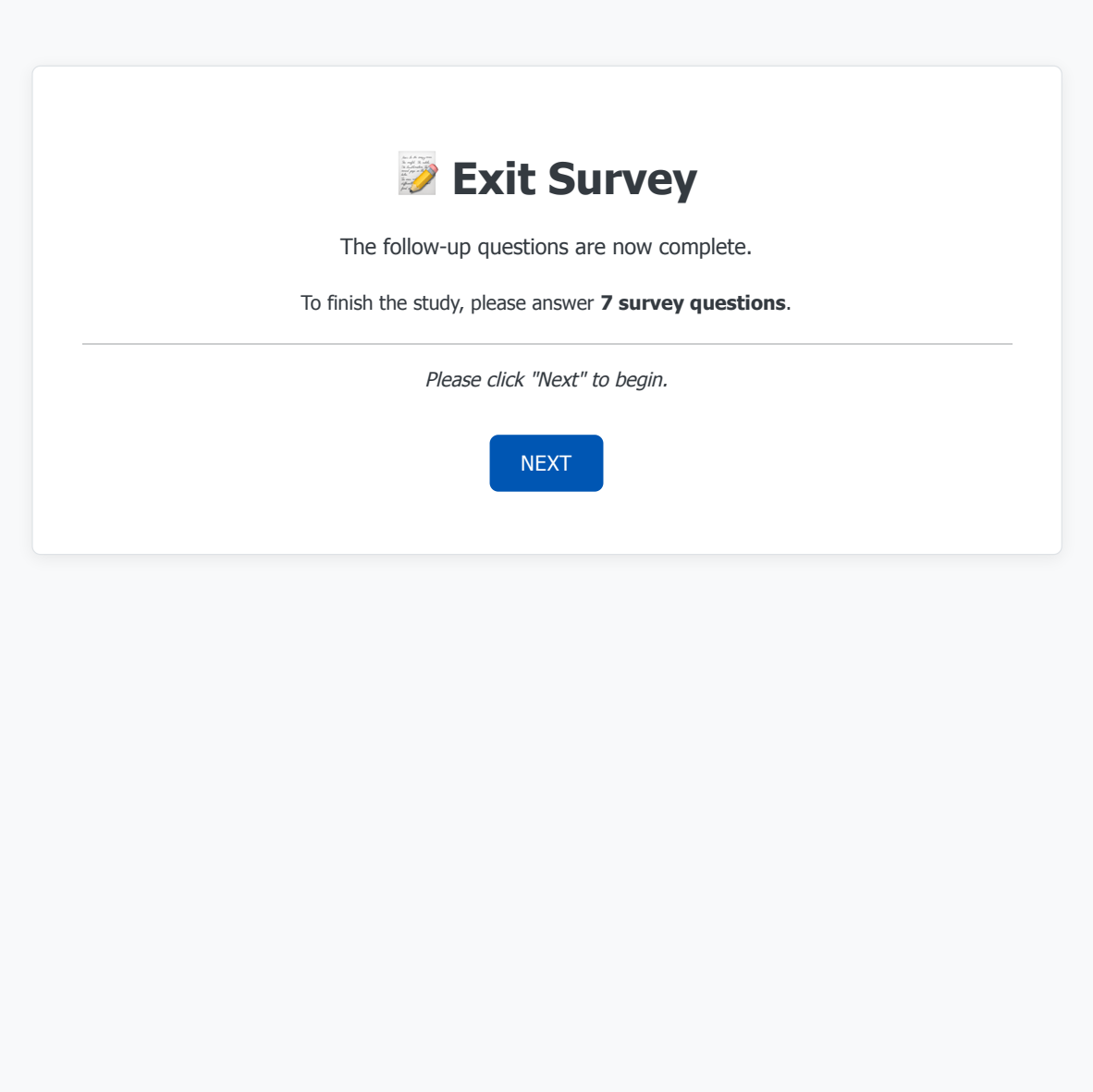


Screen 59: Follow-up question 6: the report.

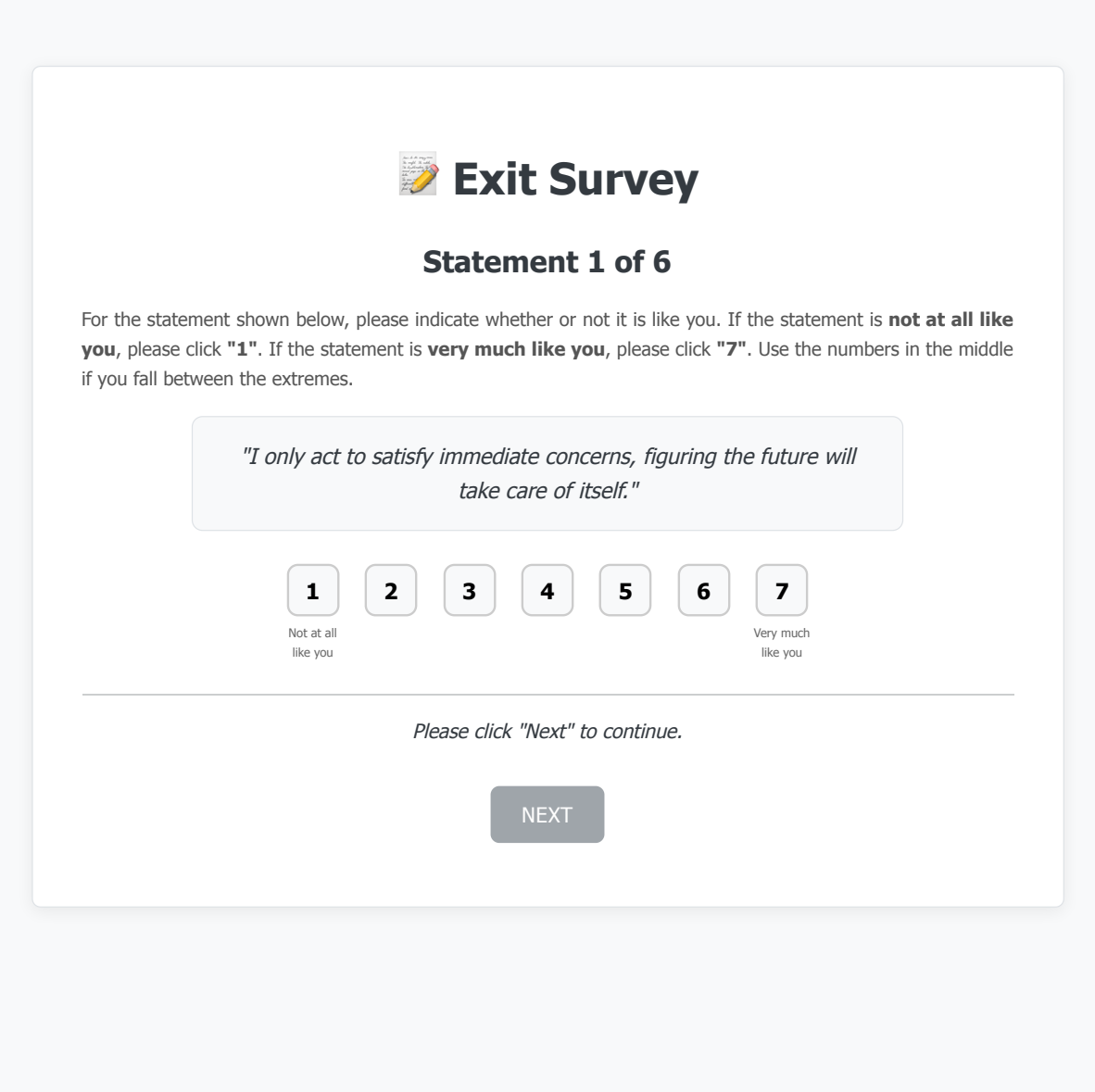


Screen 60: Follow-up question 7: the narrow-framing choice, background version.

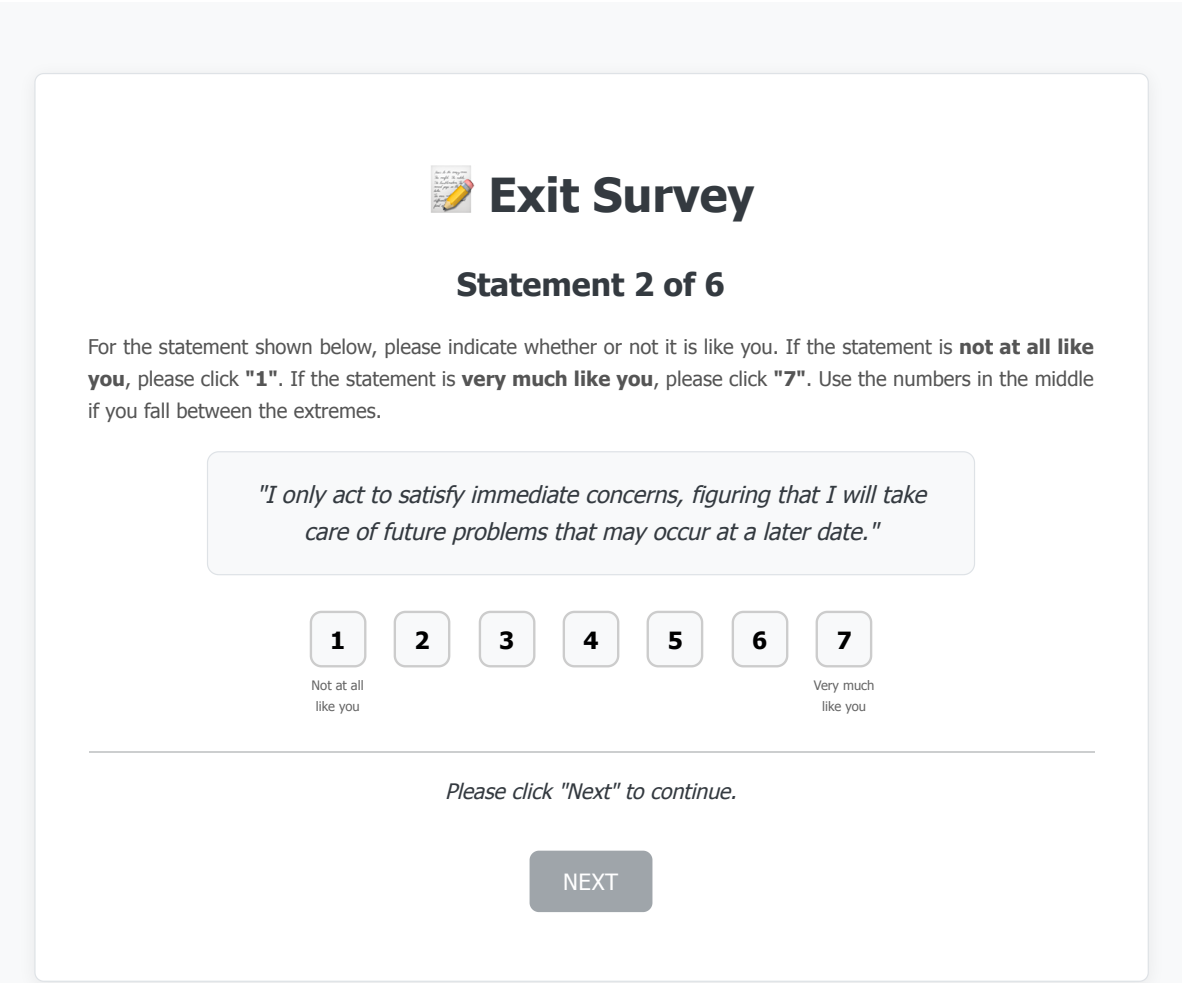
I.5 The exit survey and the payment



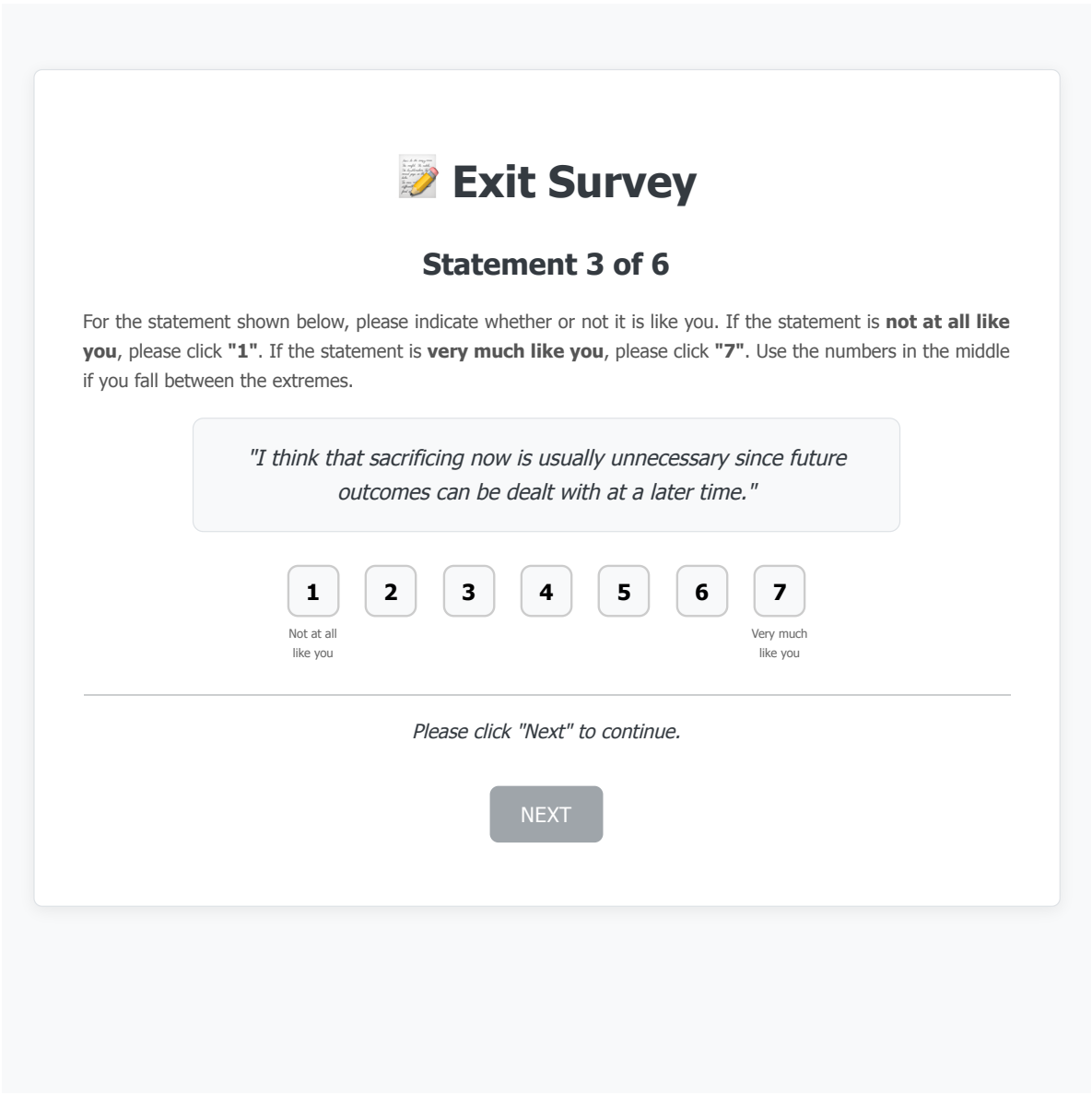
Screen 61: Introduction to the exit survey.



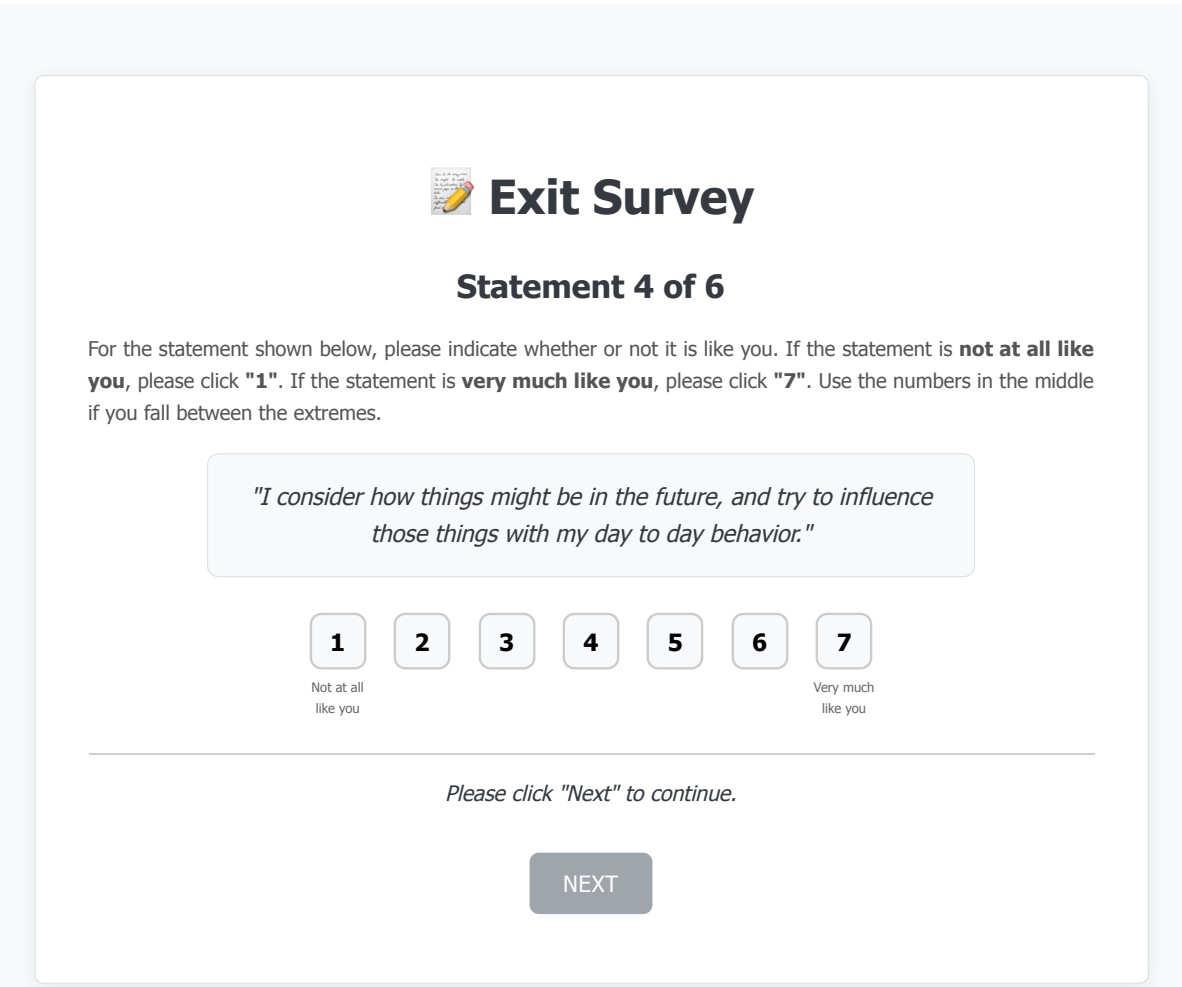
Screen 62: Exit survey, statement 1 of 6. The six statements of the Consideration of Future Consequences scale appear one per page, in an order randomized across participants.



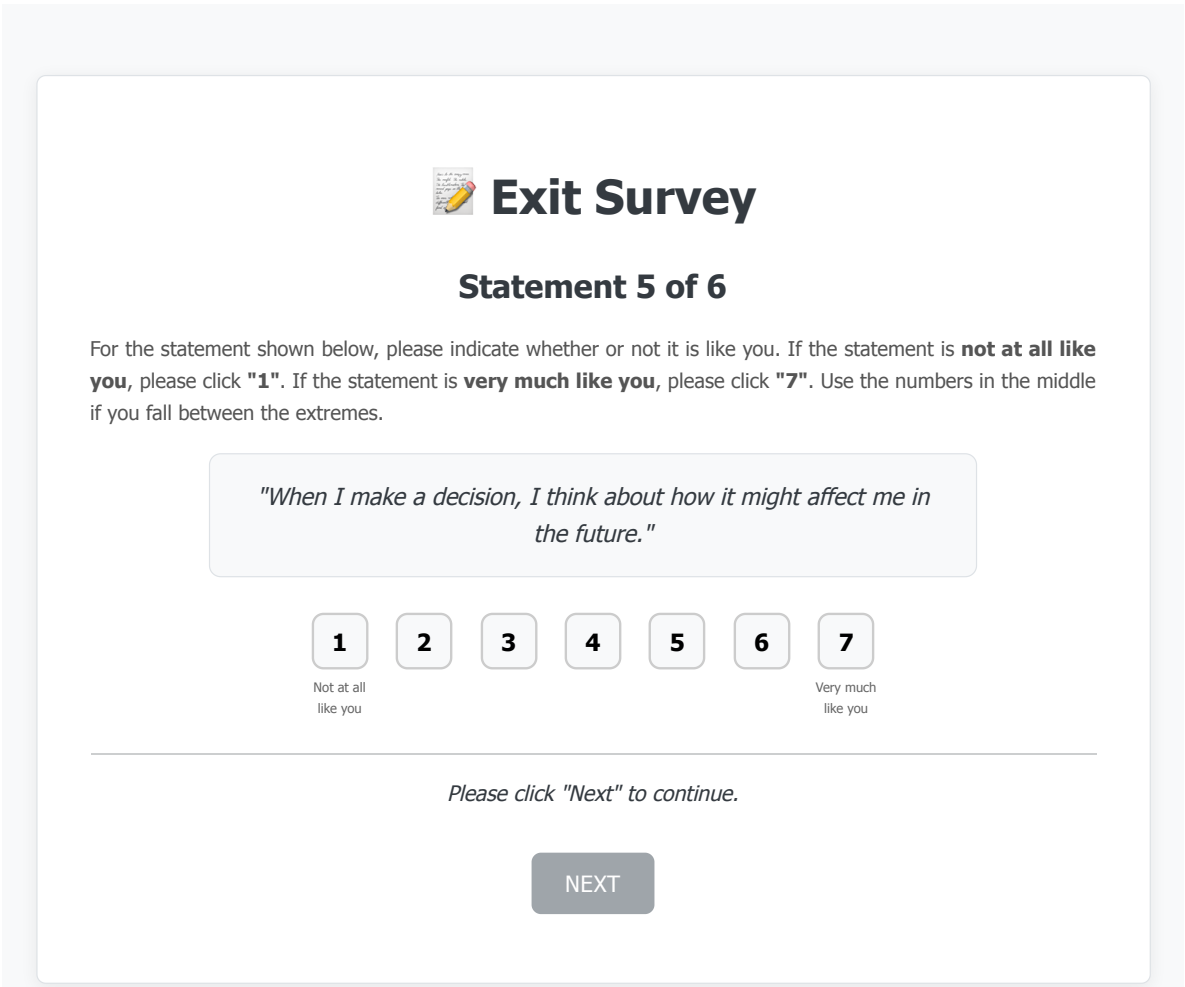
Screen 63: Exit survey, statement 2 of 6.



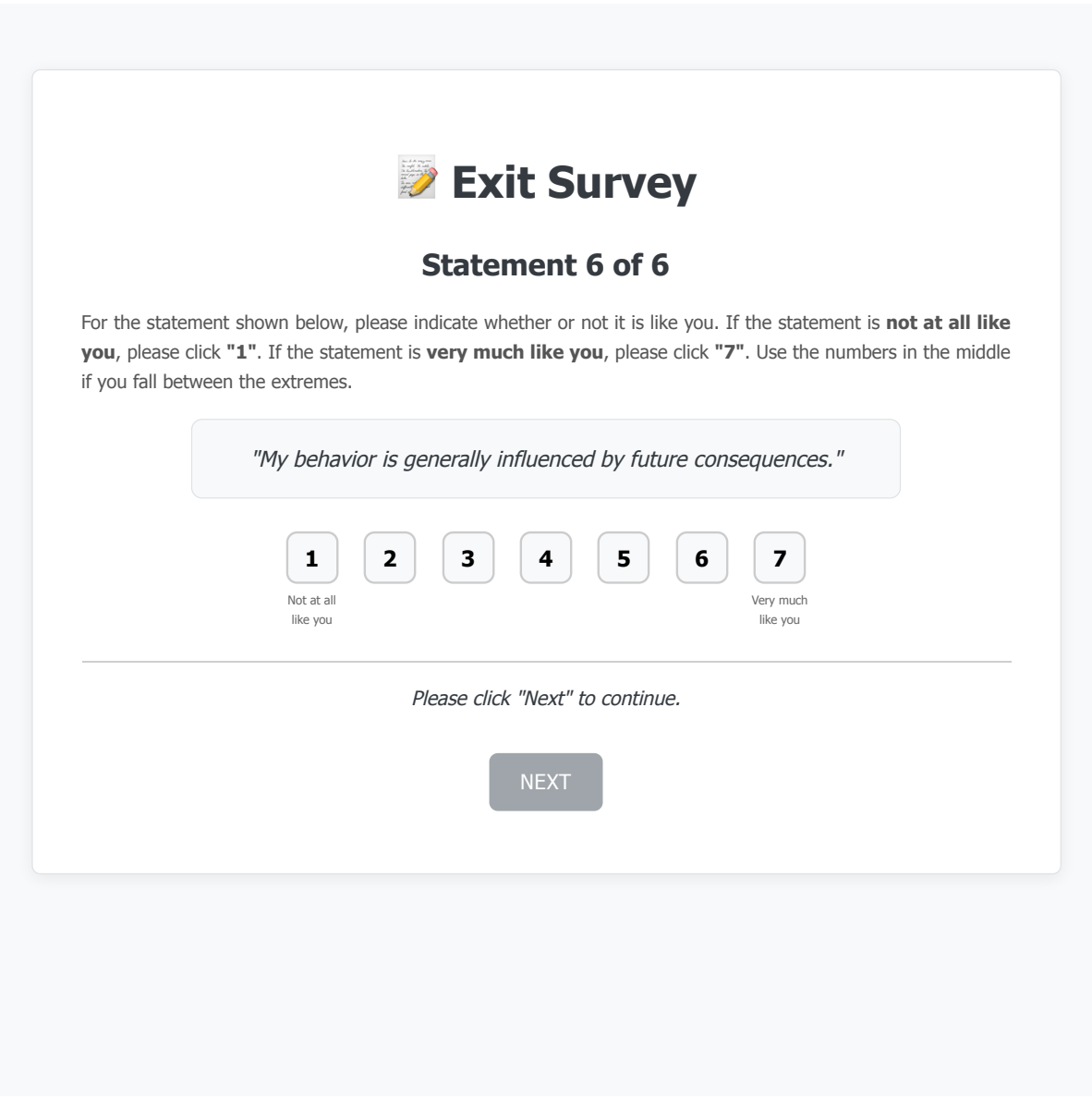
Screen 64: Exit survey, statement 3 of 6.



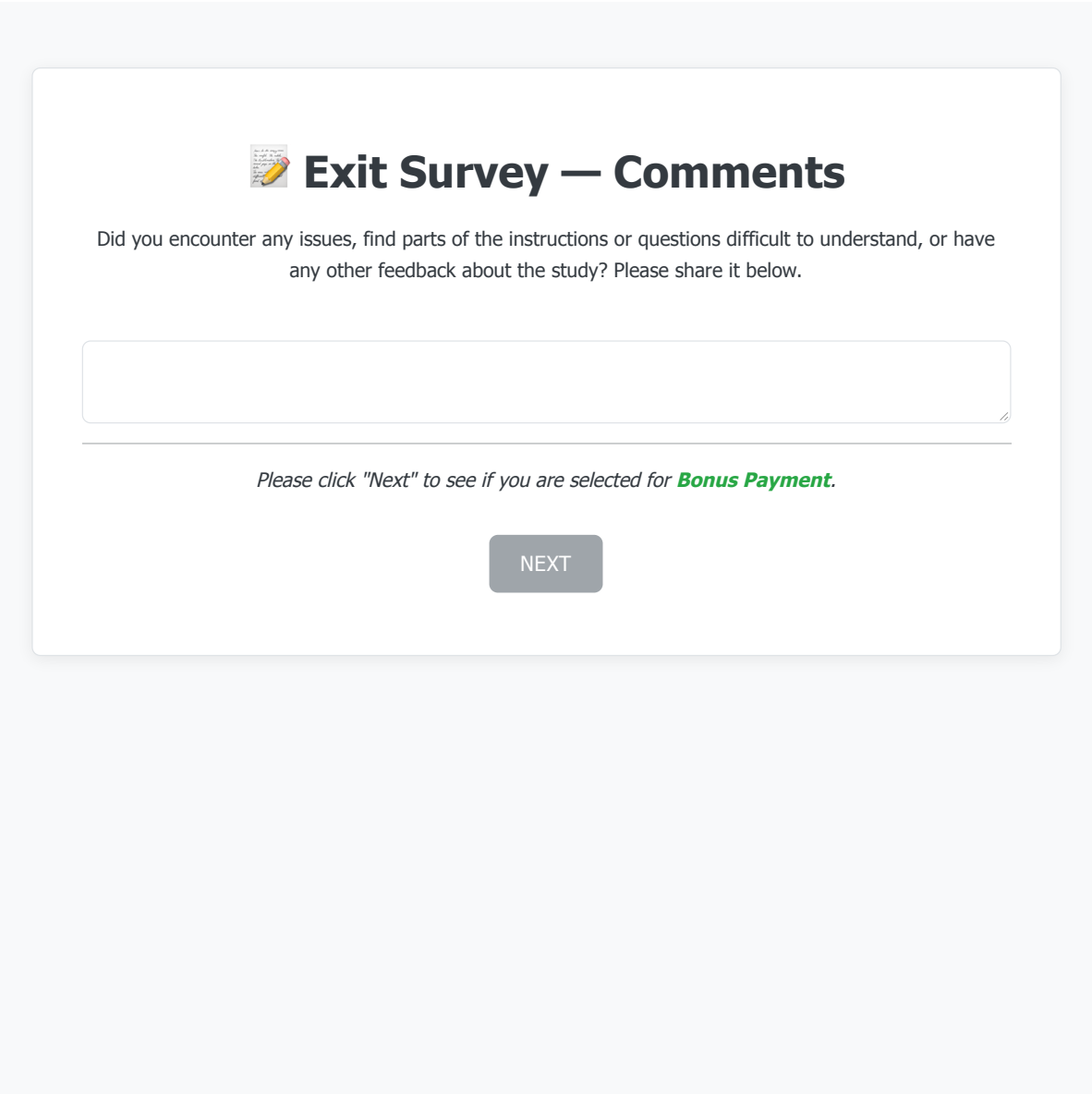
Screen 65: Exit survey, statement 4 of 6.



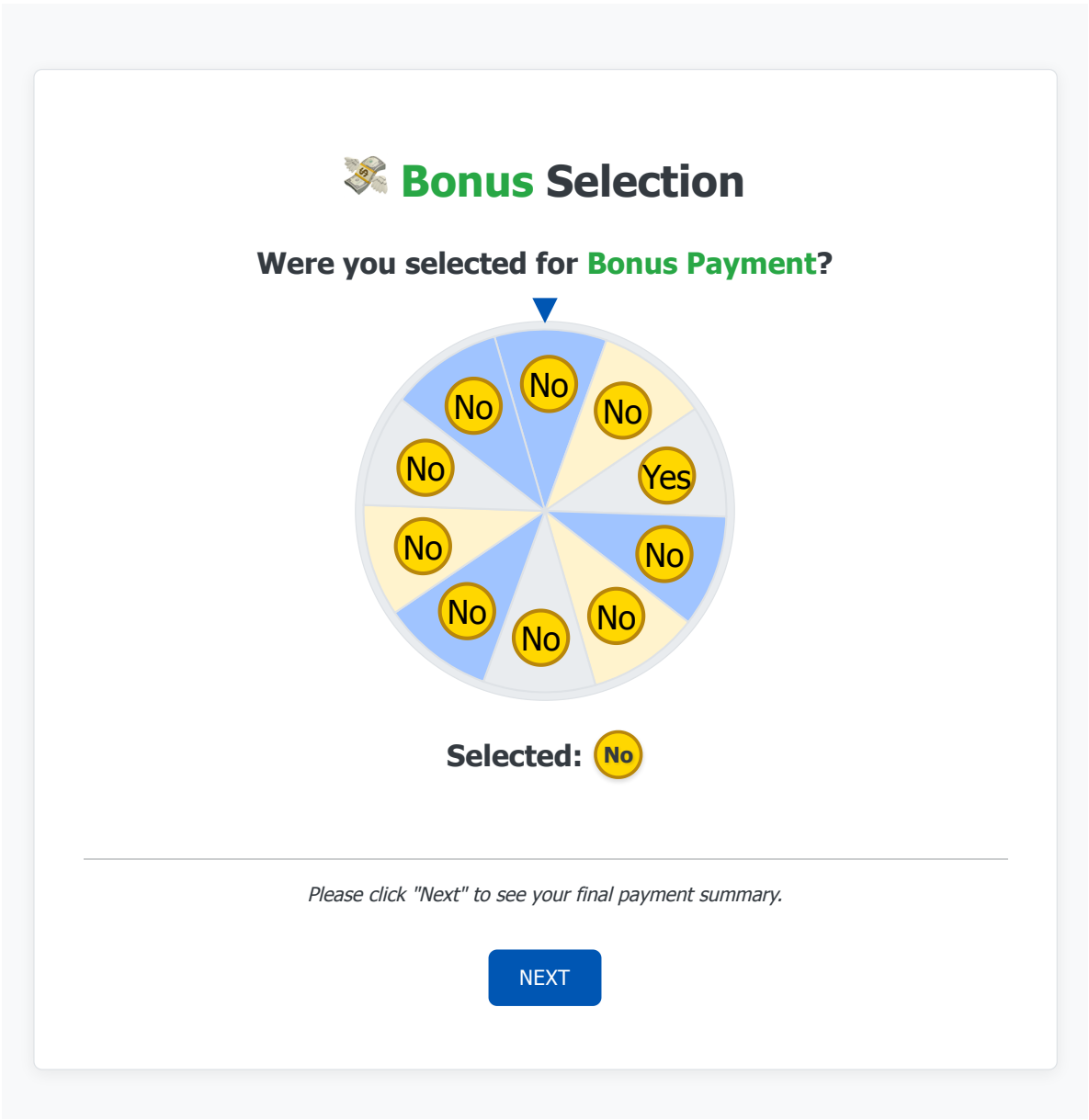
Screen 66: Exit survey, statement 5 of 6.



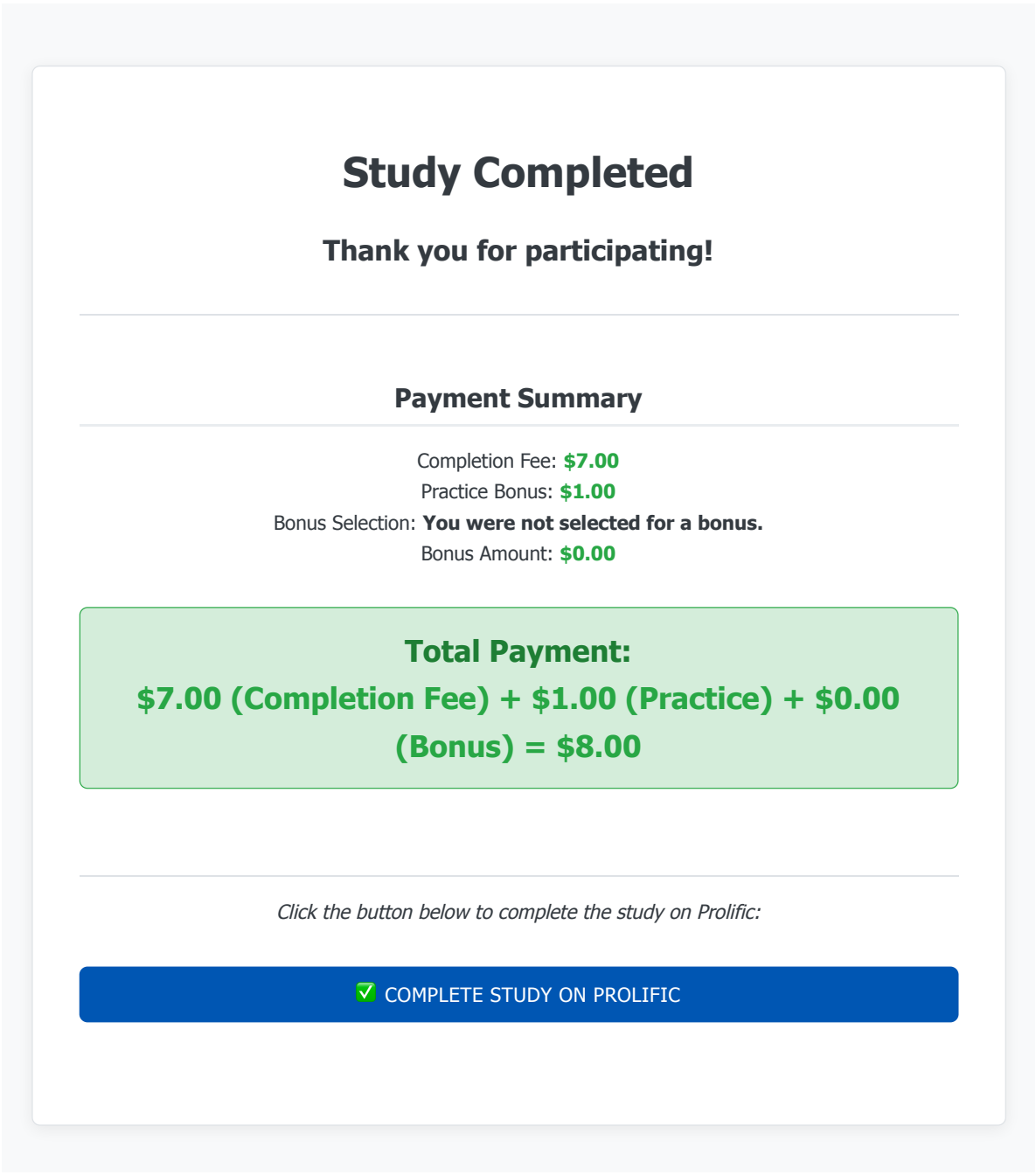
Screen 67: Exit survey, statement 6 of 6.



Screen 68: Exit survey: the comments box.



Screen 69: The bonus selection.



Screen 70: The payment summary.