What You Don't Know May Hurt You: A Revealed Preferences Approach*

Gonzalo R. Arrieta† Lukas Bolte‡

Preliminary Draft—Please do not circulate

Click here for the most recent version

October 22, 2023

Abstract

The dominant welfare approach is based on revealed preferences, which is restricted to settings where the individual knows their preferences have been fulfilled. We use a choosing-for-others framework to experimentally study welfare when what the individual believes to be true differs from what is actually true. We find substantial heterogeneity. About 40% of participants see welfare as independent of beliefs; 10% see welfare impact only via beliefs; and 35% exhibit mixed behavior. Our results suggest most people support the idea that welfare goes beyond awareness, which may inform media regulation, informational policies, and government communication.

1. Introduction

Welfare considerations are at the core of economic policy evaluations: a policy is good if it enhances welfare. The dominant approach to understanding what enhances welfare is to defer to choice and make welfare assessments based on revealed preferences. According to

*We are especially grateful to B. Douglas Bernheim for his guidance and encouragement and to Paul Milgrom and Alvin Roth for collaborating in the implementation of our experimental design. We also thank Muriel Niederle, Kirby Nielsen, and seminar participants at Stanford University for their helpful comments. Haley Hirokawa, Isha Patel, and Xiang Qing Wang provided excellent research assistance. This study is covered under Stanford University's IRB Protocol 44866 and Carnegie Mellon University IRB's #IRB-STUDY2015_00000482. The study was registered on the AEA RCT registry under ID AEARCTR-0011851 under the title "Red or Blue Pill? A Positive Welfare Analysis."

†Stanford University. E-mail: garrieta@stanford.edu.

‡Carnegie Mellon University. E-mail: lukas.bolte@outlook.com.

this criterion, an alternative x is deemed to be better than alternative y if the individual would choose x over y. Then, to decide between two policies, all that needs to be done is to find out which of the policies the individual (or individuals) would, in fact, choose x over y. Often, this assessment is done using past choices from related decision problems.

However, the reliance on choice data limits this welfare criterion based on revealed preference to questions where choice data is available. In particular, using choice data to assess individual welfare essentially restricts us to settings in which the individual *knows* that, say, alternative x is chosen. This is by virtue of the fact that an individual cannot consciously make a choice between alternative x and y without, in tandem, adjusting their beliefs about whether alternative x or y has been chosen.

But how then can we assess the welfare effect of alternative x over alternative y is *when the individual does not learn about the chosen alternative*. Such questions arise not only in the context of policy-making, where individuals may be unaware of the chosen policy but also in individual decision-making, where individuals may misunderstand the consequences of their actions.[1] For example, does the individual's welfare depend on what they believe the consequences of their actions are, on the actual consequences, or some combination?

This paper experimentally studies the welfare consequences of different alternatives when the individual remains unaware of which alternative was chosen; that is, we provide an answer to the previously mentioned questions. To overcome the aforementioned limitation of choice data, we resort to a choosing-for-others framework: an altruistic other party chooses between options affecting the individual but ones that hold what the individual believes fixed. Thus, we maintain the revealed preference paradigm as a basis for welfare assessment but extend the domain of applicable welfare questions through the choosing-for-others framework. In particular, in a situation where the individual does not learn about the chosen alternative, alternative x improves the individual's welfare relative to alternative y if an (altruistic) other party chooses alternative x over alternative y.

Our experimental design is as follows. We create a binary state of the world over which a particular participant (henceforth, the Receiver) has preferences: If asked, they would like the state of the world to be 1 (as opposed to 0). The state satisfies two crucial requirements:

[1] One illustrative example of a policy question with ambiguous welfare implications is given by the "The parable of the oblivious altruist" in Bernheim and Taubinsky (2018). *A small town in Arkansas experiences massive flooding, leaving many families homeless. To provide financial assistance for the impacted families, the government raises taxes, including a $100 levy on Norman. As a general matter, Norman thinks government spending is wasteful, but he is also an altruist and would gladly contribute $100 to the fund if he knew about it. However, he never learns about the flood or the relief effort. Does the government's policy make him better off or worse off?* Of course, it is impossible for Norman's choices to inform the planner in this setting: if he chooses the government policy, he would also learn that the government has such policy, contrary to the thought experiment's premise.

First, it is constructed so that it is impossible for the Receiver to know what the state is unless they are informed about it; second, we minimize that anyone else but the Receiver cares about the state directly. We then ask other participants (the "other parties") to trade off the amount of a surprise bonus given to the Receiver and whether the state is 1 or 0. (The bonus is a surprise to the Receiver to minimize concerns that the Receiver learns about the state from the bonus amount.) Since the state is constructed to be only (directly) meaningful to the Receiver, the participants make choices that, according to the revealed preference paradigm, maximize the Receiver's welfare. Critically, we elicit these welfare assessments for two cases: one, where the Receiver learns about the state—a relatively standard elicitation—and another, where they do not.

Our main findings are these. Consider participants who are willing to reduce the bonus of the Receiver so that the state is 1 for the case where the Receiver learns about the state, i.e., those participants who would choose qualitatively the same as the Receiver would in a standard choice problem. First, among those participants, around 40% make the same welfare assessment (they indicate indifference between the state 1 and state 0 for the same bonus attached to state 0) whether or not the Receiver learns about the state; around 10% act as if there is no welfare impact of the state on the Receiver when the Receiver does not learn about the state; and around 35% think that the welfare impact is less, but not zero when the Receiver does not learn about the state. (The remaining participants do not fall into these buckets; they seem to prefer state 0 when the Receiver does not learn the state, or they seem to value state 1 more when the Receiver does not learn the state.) Second, by exogenously varying what the Receiver believes when they do not learn about the state—they could think the state is most likely to be 1 or most likely to be 0—we also show that some participants act as if having accurate beliefs is important for welfare. These findings hold both in the full sample and in a subsample screened for quality.

We then correlate the incentivized welfare assessments with related unincentivized questions that similarly get at what type of welfare notion a participant subscribes to. In particular, we ask participants whether they would plug into the "Experience Machine" (Nozick, 1974) and two welfare assessments: the impact of a desired policy being implemented without the individual knowing about it (Bernheim and Taubinsky (2018); see Footnote 1), and the impact of receiving an original art piece over an indistinguishable (see details in section 3.1.4). We find a correlation between a composite measure of the unincentivized questions and the incentivized elicitation, giving us confidence in our welfare assessments. Moreover, this correlation suggests some context-independence of welfare notions and the possibility of a cheap and quick way to elicit empirical welfare types, which may be useful in assessing policy impacts for particular populations.

Our positive study of welfare contributes to a millennia-old question: "What is welfare?" There are two broad classes of theories conceptualizing welfare. The first is Welfare Hedonism, which proposes that "well-being consists solely in the presence of pleasure and the absence of pain" (e.g., Bentham (1789); Mill (1879)). A common variant of this notion of welfare is Mental Statism, which postulates that well-being is exclusively a reflection of mental states. The second is Preference (or Desire) Theory, which postulates that well-being consists of having one's preferences satisfied and is the dominant approach to welfare in economics. In its simplest form, this welfare notion asks whether the world is as the individual would like it to be and ignores whether the individual believes about the world; a more general form allows for the possibility that the individual's preferences encompass their own mental states. In this sense, it allows for mental states and those outside the individual's awareness (which we call external states) to affect welfare. However, as noted earlier, the question of which mental and external state combination enhances welfare cannot be deferred to the individual.[2]

We make progress in this respect, taking a Preference Theory approach and overcoming the practical complications by resorting to a choosing-for-others framework. We then ask how mental states and external states, and their combination, contribute to welfare.

There is literature in experimental philosophy and psychology that aims to elicit welfare notions positively. In particular, a great many studies have examined the extent to which participants would plug into a hypothetical machine that can simulate mental states—the "Experience Machine" (Nozick, 1974). Ignoring many subtleties, those prescribing to Mental Statism should, while those valuing "real" experiences should not, plug in.[3] We contribute to this experimental literature by constructing an incentivized measure of a participant's welfare notion and, furthermore, show that it correlates with the hypothetical Experience-Machine question.

Within economics, our paper relates and contributes to a growing literature on positive welfare economics, in particular to the branch which aims to determine how people evaluate the welfare of other individuals and groups, mainly from a paternalistic lens (e.g., Uhl (2011); Ambuehl et al. (2021, 2023); Bartling et al. (2023)). In particular, this paper improves our understanding of how individuals think about others' welfare, focusing on the

[2]A third class is "Objective theories" that we do not directly connect to, by which welfare is maximized when some objective criteria are met, irrespective of whether the individual prefers them (Aristotle, 2011; Sen, 1985).

[3]The experimental work has progressively refined experiments that try to account for potential biases when asking participants what they would do if facing the machine, with inconclusive evidence about the role of mental states (Baber, 2008; De Brigard, 2010; Smith, 2011; Weijers, 2013; Rowland, 2017; Hindriks and Douven, 2018).

welfare relevance of mental and external states, respectively.[4] An important finding in the paternalism literature is that individuals intervene in others' decision problems as if they seek to align others' choices with their own aspirations (Ambuehl et al., 2021); this finding supports an interpretation of our choosing-for-others results as reflective of how individuals think of their own welfare. To the extent that participants choose as if they were in the role of the Receiver, participants' views about others' welfare directly speak to their own preferences and, hence, their own welfare according to the revealed-preference paradigm.

Lastly, there is a large literature documenting that individuals may be motivated to hold particular beliefs (i.e., mental states) that may not be justified by actual evidence (i.e., the external state), at least by a Bayesian agent (Bénabou, 2015). For instance, individuals may be motivated to think highly of themselves as a source of ego utility (Köszegi, 2006), or about a state of the world to derive utility through anticipation (Bénabou and Tirole, 2002; Brunnermeier and Parker, 2005; Caplin and Leahy, 2001). These models are typically silent on whether such desire to manage one's mental state is welfare-enhancing or a mistake: is the agent's objective normative and corresponds to their welfare, or simply a positive theory of behavior? By showing that, for a large portion of participants, external states are not all that drive welfare, our results open the door for beliefs in general and biased ones in particular to have direct welfare implications.

The paper proceeds as follows. Section 2 presents a conceptual framework that makes the distinction between mental and external states, highlights the limitations of the existing revealed-choice paradigm as an underpinning of welfare economics, and shows which type of welfare assessments we study. We present our experimental design in Section 3. Section 4 gives the results, and Section 5 concludes.

2. Conceptual Framework

The revealed preference paradigm underlying welfare statements is only applicable when there are, in fact, decision problems that reveal preferences. However, some problems are impossible to state: we cannot elicit someone's choice between alternatives, *while holding their beliefs about which alternative is chosen fixed*.[5] In this section, we consider a simple framework of such choice problems, i.e., where alternatives consist of pairs of "mental

[4]We run a survey on the Social Science Predictions Platform (DellaVigna et al. 2019; Public Study ID sspp-2023-0032-v1 at www.socialscienceprediction.org) to capture our current understanding about the role of external and mental states for welfare. Predictions, while heterogeneous themselves, broadly capture the heterogeneity in welfare notions that we find in our data. Predictors underestimate, however, the degree to which participants behave as if external states drive welfare, conditional on beliefs.

[5]For example, in the exercise described in Footnote 1, we cannot have Norman choose between the government offering disaster relief or not, *while holding Norman's beliefs about whether there is a disaster relief or not fixed*.

states" (e.g., what the person believes) and "external states" (what is actually true). We first use this framework to highlight the limits of the revealed-preference paradigm as a foundation of welfare, to then use this framework to discuss preferences over this larger choice set (welfare types), whose prevalence we measure in our experiment in a choosing-for-others framework.

The revealed preference paradigm and its limitations as a welfare criterion. Let $x \in X$ be a set of goods and $\mu \in \Delta(X)$ be a distribution over this set of goods. We refer to x as the external state and μ as the mental state (of the individual). We want to know how the bundle (x, μ) affects the individual's welfare, i.e., we want to learn about some function $\mathscr{W} : X \times \Delta(X) \to \mathbb{R}$. (Of course, mental states may go beyond beliefs; for instance, the mental state may be a function of the change in beliefs. Our experiment accounts for such a possibility.)

The revealed preference paradigm underlying much of welfare economics assumes that individuals know what is best for them, and so $\mathscr{W}$ is estimated by giving the individual choice problems. A typical problem involves choosing from a set $A' \equiv \{(x, \delta_x) \in A \times \Delta(A)\}$, for some $A \subseteq X$, where δ_x places all weight on x. Here, the individual chooses their preferred good, and whatever they choose, they must believe.[6] These choice problem can then be used to estimate $\mathscr{W}$; however, only for a restricted domain. To illustrate, suppose $X = \{0, 1\}$ and consider choices over $\{(1, \mu), (0, \mu)\}$, where μ denotes the probability of the external state being 1. Giving this choice problem to the individual is infeasible: When the individual chooses $(1, \mu)$, they know that the external state is 1, and so their mental state cannot be μ (unless $\mu = 1$, but then choosing $(0, \mu)$ is not possible).

We overcome the problem of missing data using a choosing-for-others framework. Suppose an altruistic and otherwise disinterested third party makes the choice on behalf of the individual. In particular, we assume that they, too, maximize $\mathscr{W}$ (our experimental design aims to minimize all other concerns). Then we can, in fact, study such choice problems: The third party can make choices over $\{(1, \mu), (0, \mu)\}$, with $\mu \neq 1$—i.e., those choices that the individual themselves cannot make.

Preferences over mental and external states. We focus on the case $\mathscr{W}(1,1) > \mathscr{W}(0,0)$, i.e., the individual would choose $(1, 1)$ over $(0, 0)$, which, given the assumed revealed-preference paradigm, is equivalent to $(1, 1)$ increasing the individual's welfare relative to $(0, 0)$. We are

[6] More generally, the individual may choose a lottery over some subset of $A \times \Delta(A)$. This general form allows for the possibility that the individual's mental state is non-degenerate, i.e., the individual does not learn the external state for sure. However, if the individual updates their belief correctly, then all available lotteries need to be such that the expected mental state must coincide with the distribution over external states.

interested in decomposing this welfare gain since two arguments are changing: the individual's external state and mental state. (In our experiment, the choices of a small group of participants imply $\mathscr{W}(1,1) = \mathscr{W}(0,0)$; for those participants, there is no welfare gain to decompose, and so we exclude them for some of our analyses.)

Welfare Hedonism and Mental Statism imply that all of the welfare gains come from the change in mental state; conversely, it may be that only external states matter. Lastly, it may be a combination of mental and external states.

> *Mental state is all that matters.* $\mathscr{W}(x,\mu)$ is independent of x. This functional form corresponds to Mental Statism. The intuition is simple: *what you don't know can't hurt you*. We refer to this welfare notion as pure mental statism.
>
> *External state is all that matters.* $\mathscr{W}(x,\mu)$ is independent of μ, i.e., only x. In some sense, this is the opposite of Mental Statism. Here, the individual's welfare is affected by what is actually true and not what the individual believes to be true. In this case, *what you don't know can hurt you*. Analogous to the previous case, we refer to this welfare notion as pure external statism.
>
> *Mental and external states matter.* Of course, there are many ways in which mental and external states jointly affect welfare. We focus on two cases. First, it may be that mental and external states *matter independently*, i.e., $\mathscr{W}(x,\mu) = ES(x) + MS(\mu)$ (where both ES and MS are strictly increasing). Second, external and mental states may be complements, i.e., $\mathscr{W}$ has increasing differences.[7] One interpretation here is that "incorrect beliefs," e.g., believing the external state is 1 whereas it is actually 0, are bad for welfare. Intuitively, according to such $\mathscr{W}$, it is inherently bad *to live a lie*. In this case, choosing 0 (the "bad" external state) can actually improve welfare if μ is low enough.[8]

Our experiment, described in the next section, allows us to estimate $\mathscr{W}$. In particular, we go beyond the usual elicitation of $\mathscr{W}(1,1) - \mathscr{W}(0,0)$ and additionally consider $\mathscr{W}(1,\mu) - \mathscr{W}(0,\mu)$, for varying μ.

3. Experimental design

We conduct an incentivized online experiment to measure how external and mental states determine welfare. A controlled experiment allows us to create an external state that only

[7] For $\mu' > \mu$, $\mathscr{W}(1,\mu') + \mathscr{W}(0,\mu) \geq \mathscr{W}(1,\mu) + \mathscr{W}(0,\mu')$.

[8] The welfare impact of a mismatch between external and mental state could be asymmetric: E.g., overly optimistic beliefs could be more or less harmful than pessimistic beliefs relative to accurate ones.

directly affects a particular participant (the Receiver) and that can be changed without them knowing. Moreover, it enables us to vary the mental state of the Receiver, illuminating interaction effects between external and mental states and lets us ask (unincentivized) welfare-related survey questions to validate our incentivized measures. It seems difficult to achieve all this with naturally occurring data.

3.1. Environment and treatments

The experiment consists of all but one participant making surrogate choices for one particular participant—the Receiver. We create a binary state over which the Receiver plausibly has some preference, whereas all other participants are unlikely to value the state other than through their altruism. Participants then make choices involving the state. Crucially, they make their choices for two cases: when the Receiver learns the state (the *Learns* case; which is relatively standard), and when the Receiver does not learn the state, i.e., their mental state is fixed and only the external state changes (the *NotLearns* case).

In our baseline experiment, the Receiver's initial mental state is ambiguous (i.e., the Receiver has some prior belief on the likelihood that their preferences are satisfied, which our participants do not know). We conduct two additional treatments to study how, for a fixed mental state, the implied welfare impact of changing the external state depends on the mental state, varying the Receiver's mental state.

Lastly, we ask participants some unincentivized questions related to welfare notions, allowing us to assess consistency in our participants' answers and the external validity of the incentivized measures.

3.1.1. The state

The state must be such that the Receiver has strict preferences over it.[9] Additionally, to be able to interpret participants' choices as maximizing the Receiver's welfare, the state must also satisfy the following two requirements. Firstly, only the Receiver has direct preferences over the state, and so participants, who we presume to be altruistic, act to maximize the Receiver's welfare. Otherwise, there would be a confounding motive in interpreting the participants' choices. In particular, a participant might seem like they value changing the external state, holding the mental state of the Receiver fixed, while they actually just care about the state for selfish reasons. Formally, we assume that there is a term $\alpha\mathscr{W}$ in the

[9]We convey to the participants that the Receiver would choose a particular state in a standard choice problem. In doing so, we were careful not to suggest whether the Receiver preferences were driven by external or mental states, which, as discussed in Section 2 is difficult to interpret anyways (See instructions in the appendix section B for the exact implementation).

participants' objective where $\alpha > 0$ is the weight on the Receiver's welfare $\mathscr{W}$, and the external and the (Receiver's) mental state do not enter elsewhere. (Note that this includes preferences such as "warm-glow," as long as these preferences operate via the Receiver's welfare.

Secondly, the external state cannot reveal the Receiver's mental state; otherwise, we would again be restricted to standard decision problems (see Section 2). In practice, we need to convince participants that the Receiver will never know the external state unless we inform them.

The state we constructed reasonably satisfies these requirements. We purchased four books by two Nobel Laureates in economics: two copies of "Discovering Prices" by Paul Milgrom and two copies of "Who Gets What and Why" by Alvin Roth. Each book has a handwritten note dedicated to the Receiver.[10] For one copy of each of the types of books, the note was handwritten by the author, and we refer to these two copies as the "books with the original notes." For the other two copies, the handwritten notes were copied from the original note, and we refer to these two copies as the "books with the fake notes." The books with the fake notes are indistinguishable from the books with the original notes, which we ensured by having Paul Milgrom and Alvin Roth themselves "copy" their respective notes. The Receiver will receive two books for sure, which may be either the books with the original or the fake notes.

The state is defined by whether the notes are original (the preferred state) or not.[11] We strongly suggest that the Receiver prefers the original notes (see the experimental instructions in Appendix B). Moreover, it seems implausible that any individual other than the Receiver has strict preferences over the state other than those operating via the Receiver's welfare. Participants could have been concerned with the cost of producing the book with the notes; we addressed this by producing both sets of books. Participants could also have been concerned with the books not gifted to the Receiver; we minimized this concern by returning those books to the authors, who can always create new notes.

Our second requirement is satisfied by virtue of the original and the fake notes being indistinguishable.

[10]The participant behind the role of the Receiver is called Alex, and the notes say, "To Alex: I hope you enjoy reading about auctions! Paul Milgrom" and "For Alex: I hope you enjoy reading about market design. Alvin E. Roth," respectively.

[11]The Receiver also receives a surprise monetary bonus that is determined by the participants' choices, which is also part of the Receiver's external state and mental state but ignored in this section for clarity.

3.1.2. Preference elicitation

Our primary outcome measure is the amount of money (as a surprise bonus to the Receiver) added to the books with the fake notes for the participant to be indifferent between giving the Receiver the books with the original and fake notes. We interpret this measure as the Receiver's welfare gain in Receiver-dollars from getting the books with the original notes instead of the fake notes.

We use the Receiver's monetary payoff instead of the participant's monetary payoff for three reasons. First, varying levels of altruism on the participants' side add noise to the data. Second, altruism levels in an online experiment may be low, i.e., many participants may be approximately indifferent about the state (see Section 3.2 for details about the participant population). Thirdly, participants may resort to a welfare notion, particularly mental statism, as an excuse not to reduce their monetary payoff, biasing our measure (Exley, 2016). A concern with using the Receiver's monetary payoff as a scale is that the Receiver might make an inference about the external state from the payment they receive, and so our control over the Receiver's mental state would be lost. To minimize this concern, we frame the monetary bonus the Receiver receives as a surprise bonus that they always Receive but with an unknown amount.

We determine the amount of money leaving the participant indifferent using a modified version of a multiple price list. In particular, our procedure consists of three steps:

1. We ask participants whether they would rather give the Receiver the books with original notes, the ones with fake notes, or whether they are indifferent.

2. We ask participants who do not select indifference to choose between the notes they chose in the first step and those they did not choose with an added $1 monetary payment to the Receiver.

3. Participants who chose the books with the original notes in both previous steps are presented with a multiple-price list where the monetary payment to Receiver when given the books with the fake notes spans $2, $3, $5, $7, $10, $15, $25, $45, $70, $100, $140, $200.

We chose such a procedure, in particular, prefacing the multiple price list with two relatively simple questions since those two questions already allow us to identify participants who are indifferent, those who prefer to give the original notes and those who, for some reason, prefer to give the fake notes. With this typification, we can already answer which share of participants sees welfare-relevant value in changing the external state while holding the mental state fixed.

We conduct this elicitation for two cases for each participant, presented in random order: In *Learns*, the Receiver will be informed of the external state; in *NotLearns*, the Receiver will not be informed of the external state. The three-step elicitation procedures happen simultaneously; that is, participants go through each step for both cases first before proceeding to the next step.

The monetary value elicited in *NotLearns* allows us to test whether participants see welfare value in changing the external state while holding the mental state fixed. The analogous outcome in *Learns* is also important, serving two purposes: First, it allows us to identify participants who do not value satisfying the Receiver's preferences even when the Receiver learns about it, e.g., because of a lack of altruism, confusion, or because these participants do not assign value to an original over an indistinguishable fake. We will exclude such participants since there is no welfare to decompose, and so we cannot study welfare notions. Second, it serves as a benchmark to compare the outcome in *NotLearns* to. For instance, loosely speaking, we can ask how much of the welfare gain from changing both the external and mental state comes from changing the external state only.

3.1.3. Varying the Receiver's initial mental state

In our baseline experiment, the Receiver's initial mental state is ambiguous. More specifically, when they are not informed about which books they received, they have some belief about which books—the ones with the original or fake notes—they received, and this belief may be optimistic, or it may be pessimistic. Varying this initial mental state allows us to test whether participants' welfare assessments of changing the external state (in *NotLearns*) depend on the level of the fixed mental state. For instance, whether participants assign a welfare gain when external and mental states are close to each other.

Therefore, we randomly assign participants in addition to the baseline treatment to the *HighMS* or the *LowMS* treatment. In the *HighMS* treatment, participants are informed that "[the Receiver] knows that if we don't tell [them] which books [they] got, there is at least a 75% chance that they are the ones with the original notes." In the *LowMS* treatment, we tell participants that "[the Receiver] knows that if we don't tell [them] which books [they] got, there is at least a 75% chance that they are the ones with the fake notes."[12]

[12]We implemented these and the baseline treatments by randomly selecting one before approaching Alex, the Receiver. We truthfully informed Alex that with, e.g., 75% chance, he will have gotten the books with the original notes if we do not reveal the books, and with complementary probability, which books he will have gotten will have been determined in some other way.

3.1.4. Unincentivized welfare-related questions

At the end of the experiment, we asked participants three unincentivized questions. The first question is geared to help us relate participants' incentivized choices affecting others with what they would choose for themselves. Two are asking similar questions as our incentivized welfare assessments in more naturalistic settings but unincentivized, thus speaking to the external validity of our study. Jointly, the questions allow us to study the consistency of the types of welfare assessments our participants make across contexts.

Our first hypothetical question presents participants with the Experience Machine, a canonical thought experiment long discussed in the psychology and philosophy literature on welfare notions and the role of mental states (Nozick (1974); see the instructions in Appendix Section B for the exact wording of this and the other two questions). The second question directly relates to policy preferences by asking participants whether the policy described in "The parable of the oblivious altruist" by Bernheim and Taubinsky (2018) given in Footnote 1 leaves the protagonist, whom we call John instead of Norman, better or worse off. The third question takes advantage of a real scenario where 1000 individuals received a drawing, knowing only one got the original copy and the rest had indistinguishable fakes. We ask participants whether the person who has the original drawing is better off getting the original one instead of a fake, even if they—or anyone else—will never know which they have.[13] Participants answer these questions by selecting one of two answers to each question. Note that the latter two questions are similar to our experimental setup: they consider the welfare impact of changes in the external state while the mental is held fixed. The Experience Machine studies the welfare impact of changing the mental state while the external state is held fixed. However, it can still be related to our incentivized experimental measure, e.g., those subscribing to mental statism should plug into the machine.

3.2. Implementation and recruitment details

We recruited our main sample of 1478 participants on Prolific, an online platform frequently used for research studies, and randomized them into one of our three treatments. We recruited all participants on August 23 of 2023. In order to qualify for our study, participants were required to be located in the USA and have a minimum of 100 prior submissions on Prolific, with a perfect approval rate. The experiment was implemented using the oTree platform (Chen et al., 2016). Each participant received a $3 completion payment, and the median completion time was around 16.2 minutes. We pre-registered this study in the AEA RCT Registry (AEARCTR-0011851).

[13] See www.moforgeries.org for a detailed description by the organizers of this project.

After participants read through the instructions, they were required to correctly answer understanding questions before proceeding to the main parts of our study. Rather than excluding participants, they are given as many times as needed to correctly answer the understanding questions. For full experimental instructions of all study versions that we run, see Section B in the Appendix.

4. Results

This section presents our experimental results. We first show the distribution of welfare assessments, two for each participant, uncovering heterogeneity and strong support for welfare effects beyond what is captured by mental states (Section 4.1). We then discuss the result from two additional treatments that vary the Receiver's fixed mental state; here, we find evidence of a preference to match mental and external state, which we relate to loss aversion (Section 4.2). Lastly, we correlate the incentivized welfare assessments with related unincentivzed questions, uncovering some degree of consistency and external validity.

4.1. Empirical types

We first report the distribution of the kinds of welfare assessments participants make. We find substantial heterogeneity with these three largest groups: For the modal participant, the (Receiver's) welfare gain from getting the books with the original as opposed to the fake notes is independent of whether the Receiver learns about the books (pure "external statists"); a somewhat smaller group reports a partial reduction in welfare gains when the Receiver does not learn about the books; a smaller but significant group reports no welfare gain unless the Receiver learns about the books (pure "mental statists").

Table 1 shows our main classification. Motivated by the conceptual framework in Section 2, our focus is on those participants who assign a strictly positive welfare gain to the Receiver getting the books with the original notes for the case *Learns* where the Receiver is informed about which books they get; we then group participants by how the inferred welfare gain differs in *Learns* vs *NotLearns* (where the Receiver is not informed). Throughout this section, we include all participants; in Section A of the appendix, we report the same analysis for a quality-restrict sample, where the results are qualitatively similar.[14]

Consider the *Baseline* treatment, i.e., Table 1's top row. Briefly, a negligible share (0.60%)

[14]We use participants' open-ended responses to a question about why they responded the same or differently for both cases (i.e., *Learns* and *NotLearns*) to flag those who exhibit an evident misunderstanding of instructions (in particular, a misunderstanding of the difference between *Learns* and *NotLearns*), those who seem to be using artificial intelligence, and those whose quantitative answers are inconsistent with their written reasoning.

	$WTP<0$			$WTP=0$			$WTP>0$				
	$ES<0$	$ES=0$	$ES>0$	$ES<0$	$ES=0$	$ES>0$	$ES<0$	$ES=0$	$ES<WTP$	$ES=WTP$	$ES>WTP$
Baseline N=497	0.20%	0.20%	0.20%	0.41%	14.40%	3.04%	3.65%	10.14%	29.01%	33.67%	5.07%
Low N=497	0.80%	0.00%	1.01%	0.80%	17.91%	2.21%	6.44%	8.45%	24.75%	31.99%	5.63%
High N=488	0.20%	0.00%	0.41%	0.20%	12.70%	3.48%	4.51%	7.58%	27.25%	35.86%	7.79%

Table 1: Classification of welfare assessments

Note: WTP is the willingness-to-pay to give Receiver the books with original notes in *Learns* when he learns about it. ES is the willingness-to-pay to give Receiver the books with original notes in *NotLearns* when he does not learn about it.

of our participants assign a welfare loss to the books with the original notes in case *Learns*, i.e., they strictly prefer the Receiver to receive the books with fake notes over the books with real notes when they learn about it. Relatedly, also in case *Learns*, 17.85% of participants see no welfare effect; they are indifferent between the Receiver getting the books with the original and the fake notes.[15] And most of these participants also see no welfare gain in the *NotLearns* case.

Next, consider our group of interest displayed in the last set of columns. Most participants (81.54%) assign a strictly positive welfare gain to the Receiver getting the books with the original notes in the *Learns* case. We find that 10.14% of our participants do not see a welfare gain in the *NotLearns* case (while seeing one in the *Learns* case), and 33.67% see the same (positive) welfare gains in both cases. Although other interpretations may be possible, the participants may be interpreted as *only* seeing welfare gains in variations of mental and external states, respectively; they are pure mental statists and pure external statists.

34.08% of participants assign strictly positive welfare gains in both cases but of different amounts, and most of these participants assign a lower welfare gain in the *NotLearns* case. For this group, both mental and external states matter for welfare. They may matter independently or because these participants see a welfare gain when mental and external states match (see Section 2). In the next section, we report the results of two additional treatments varying the Receiver's mental state, gauging the extent to which this latter possibility matters.

We note that the heterogeneity in the kinds of welfare assessment implies that distinguishing between external and mental states for welfare analysis is important. Moreover, most participants see welfare effects beyond those stemming from awareness.

[15]We include participants who choose one of the non-indifference options in Step 1 of our multiple-price list procedure, and in Step 2, the other option, which now has a 1$ added to the Receiver's payoff. See a low welfare impact of the Receiver getting one or the other set of books, including no impact.

4.2. Testing independence of mental and external state for welfare

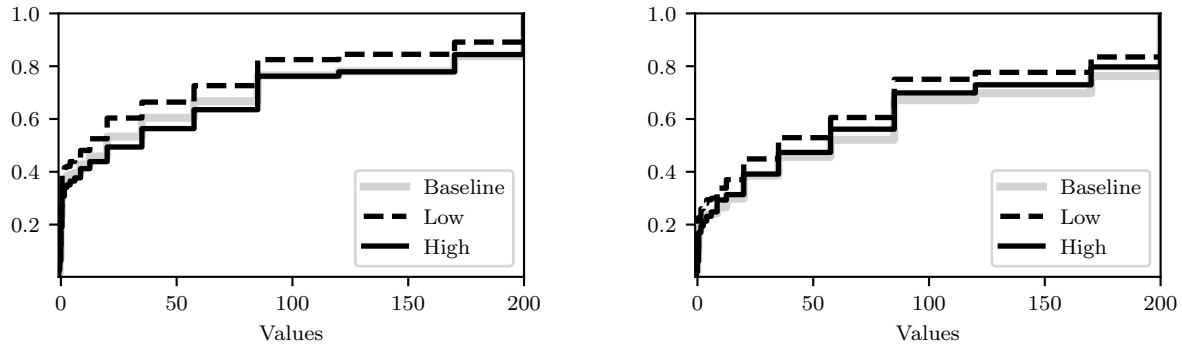
In this section, we report on two additional treatments varying the Receiver's mental state, i.e., the Receiver's belief that they will get the books with the original notes. In the *Learns* case, this mental state variation can be interpreted as a reference point manipulation; in the *NotLearns* case, it allows us to test for preferences to bring mental and external states close together. For both cases, we find that, indeed, participants assign a higher welfare gain (from the Receiver getting the books with the original notes) when their initial mental state (which is also their final mental state in *NotLearns*) is high, i.e., they expect to receive the books with original notes.

We begin by replicating our classification exercise from Section 4.1 for the two treatments (*LowMS* and *HighMS*) in the second and third rows of Table 1. Recall that in *low* (*HighMS*), participants know the Receiver is told there is at least a 75% chance (at most a 25% chance) they receive the books with fake notes. While the share of participants classified into each group is similar across treatments, suggesting our classification is approximately invariant to variations in the Receiver's initial mental state, we note two differences: First, the choices of more participants imply a strictly positive welfare gain from the Receiver getting the books with the original over fake notes in *MSHigh* compared to *MSLow*, suggesting, e.g., some degree of loss aversion. Second, participants seem to assign a larger welfare gain for the *NotLearns* case in *HighMS* in a first-order stochastic dominance sense, suggesting some preference to match the mental state with the external state. This pattern holds regardless of whether one conditions on a strictly positive assigned welfare gain in the *Learns* case.

Next, we study these patterns in greater detail, looking at the changes in the distribution of welfare gain more granularly. Figure 1 displays the cumulative distribution function of the elicited welfare gains by treatment (*Baseline*, *LowMS*, *HighMS*) and by case (*NotLearns* in the left panel and *Learns* in the right panel). For both panels, i.e., for whether or not the Receiver learns which set of books they receive, we observe a similar pattern on the intensive margin of welfare assessments as we did on the extensive margin (the classification in Table 1): the distribution of welfare gains is increasing (in the first-order stochastic dominance sense) in the Receiver's mental state, i.e., comparing *HighMS* and *LowMS*. (We refrain from analyzing differences to the *Baseline* treatment since the Receiver's mental state as perceived by the participants is ambiguous.)

For the case *NotLearns* (left panel), this pattern implies a preference to match the external state to the mental state. In other words, participants assign a welfare gain to the Receiver having accurate beliefs.[16] An implication of this finding is illustrated by "The

[16]The welfare costs of a mismatch between mental and external state, together with a general preference for the Receiver to get the books with the original notes, could imply a "net welfare gain" of 0 if the participant



(a) *NotLearns* (b) *Learns*

Figure 1: The two panels display the cumulative distributions of the Receiver's welfare gain by treatment for the *NotLearns* and *Learns* cases.

parable of the oblivious altruist" given the Footnote 1. Suppose that in one scenario, Norman expects the government to provide relief effort; in another, he does not. The elicited welfare gain from the Receiver having accurate beliefs implies that the welfare impact of the government providing the relief effort (even if Norman never learns about it or what made it necessary) is larger in the first scenario.

For the case *Learns* (right panel), the interpretation of the pattern differs. The participants' decisions do not affect the (mis)match between the final mental and external states since the Receiver will always be informed about the external state, and so there is always a perfect match. Instead, the observed pattern can be rationalized with participants assigning some (welfare-relevant) loss aversion to the Receiver (Tversky and Kahneman, 1991): Losses loom larger than gains, in a welfare sense, and when the initial mental state is high, say, for illustrative purposes, certain that the books have the original notes, then the welfare gain for the Receiver from getting the books with original notes is from avoiding a loss and hence large.

4.3. External relevance and choosing-for-self

In this section, we correlate the incentivized welfare assessments from Section 4.1 with related unincentivized questions (see Section 3.1.4) that similarly get at what type of welfare notion a participant subscribes to. We find a correlation between the incentivized and unincentivized measures, suggesting some context-independence of the welfare types, as well as a cheap way to elicit them. In particular, we observe a correlation between a hypothetical

assigns the Receiver a relatively low initial mental state. Thus, some participants who in the *NotLearns* case do not see a welfare gain might, in fact, not be pure mental statist but rather have more nuanced preferences that, in particular, depend on the external state. However, this case is non-generic.

choosing-for-self question, suggesting that our choosing-for-others framework speaks to the participants' own welfare.

The first hypothetical question presents participants with the Experience Machine (Nozick, 1974); the second asks participants whether the policy described in 'The parable of the oblivious altruist" (Bernheim and Taubinsky (2018); see Footnote 1) leaves the protagonist better or worse off; the final question presents participants with a real scenario where 1000 individuals received a drawing knowing only one had the original copy and the rest have indistinguishable fakes and asks them whether this person is better off by getting the original one instead of a fake. For each question, participants provide a binary answer, where one answer is more aligned with pure mental statism and pure external statism (see Section 2), and are given the opportunity to explain their answer in the free-form text field.[17]

Consider two groups of participants: those who consistently choose the pure-mental-statism option ($N = 276$) and those who consistently choose the pure-external-statism option ($N = 114$).

Figure 2 shows the cumulative distribution function of the ratio of welfare assessment for the *NotKnows* case to the one for the *Knows* case for each of these two groups. We restrict the sample to those participants whose welfare assessment for the *Knows* case is strictly positive. Note that the "external statists," according to the unincentivized questions, exhibit a larger ratio across the whole distribution compared to the "mental statists," suggesting they indeed value changes in the external state more.

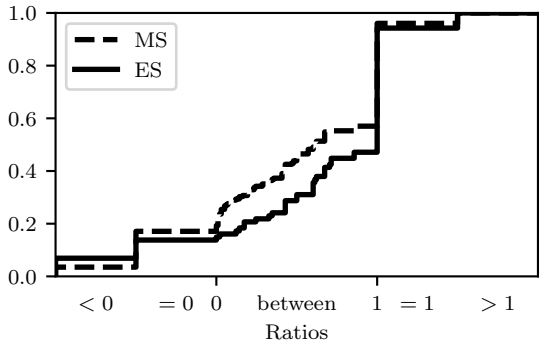


Figure 2: Caption

Note: CDF of the ratio of ES-WTP to Trad WTP, conditional on $WTP > 0$, for participants in baseline who give responses to unincentivized questions that are consistent with welfare notions that assign more value to the MS or to the ES. 28.2% of participants in baseline who have $WTP > 0$ give such responses.

[17]The distributions of the answers for each question are as follows. Experience Machine (question: "Would you go into the machine?"): "Yes" 51%; "No" 49%. Oblivious altruist (question: "Does the government raising taxes to provide financial relief make John better or worse off?"): "Better off" 63%; "Worse off" 37%. Fake-vs-original painting (question: "Is this person better off by getting the original one instead of a copy?"): "Yes" 57%; "No" 43%.

This evidence suggests that the elicited welfare assessments are somewhat stable, albeit far from perfect. Figure 3 in Appendix A reproduces Figure 2 but for each of the three questions separately. There, we observe a pairwise correlation of the incentivized welfare assessment with the answers to the Experience Machine question and the question about the fake drawings, respectively, but not with the answers to the oblivious altruist question.

5. Conclusion

This paper uses a choosing-for-others framework to experimentally study the roles of mental and external states for welfare. We find that the modal participant acts as if welfare goes beyond awareness, rejecting pure mental statism as a welfare notion. We also find substantial heterogeneity, with some participants acting as pure mental statists. Moreover, in the aggregate, participants assign welfare gains to having accurate beliefs (i.e., matching mental and external states).

Understanding the role of mental states and beliefs for welfare is critical to understanding the value of non-instrumental information and, hence, may prove important to media regulation, informational policies, and government communication. For example, our findings allow for a re-interpretation of the welfare impact of surveillance programs: While the discussion of surveillance versus privacy usually focuses on its instrumental implications, our results suggest that individuals unaware of being surveilled may suffer welfare costs. More generally, the same point applies to a broader set of government policies that rely on an oblivious population, like those practiced by intelligence agencies and the military.

One particular policy implication has to do with taxation. Chetty et al. (2009) documents that obfuscated taxes can lead to higher spending, i.e., they are less distortionary, suggesting that the government should use obfuscation as a tool. However, our findings suggest that to many, external states (e.g., what the taxes actually are) matter per se, independent of what the perceived taxes are (beyond budgetary concerns)—there is a welfare cost from the mismatch of mental and external states.

Our work broadly speaks to the value of information in the functioning of society. Consider a society of altruists, those who value their neighbor's (direct) payoff as much as their own. In this society, members have the occasional opportunity to do a costly favor for another member. If a member is a pure mental statist, then they only take the costly action if the beneficiary learns about it. In contrast, a pure external statist would take the costly action regardless.

Lastly, the heterogeneity we observe in our results provokes questions about the stability of welfare notions and their determinants. Future work could explore such questions,

for example, considering how culture and norms determine support for different welfare notions.

References

Ambuehl, S., Bernheim, B. D., and Ockenfels, A. (2021). What motivates paternalism? an experimental study. *American economic review*, 111(3):787–830.

Ambuehl, S., Blesse, S., Doerrenberg, P., Feldhaus, C., and Ockenfels, A. (2023). Politicians' social welfare criteria: An experiment with german legislators. Technical report, CESifo Working Paper.

Aristotle (2011). *Aristotle's Nicomachean ethics*. Translation by Bartlett, Robert C and Collins, Susan D and others. University of Chicago Press.

Baber, H. (2008). The experience machine deconstructed. *Philosophy in the Contemporary World*, 15(1):132–137.

Bartling, B., Cappelen, A. W., Hermes, H., and Tungodden, B. (2023). Free to fail? paternalistic preferences in the united states. Technical report, NHH Dept. of Economics Discussion Paper.

Bentham, J. (1789). Pml. *An Introduction to the Principles of Morals and Legislation*.

Bernheim, B. D. and Taubinsky, D. (2018). Behavioral public economics. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 1, pages 381–516. Elsevier.

Brunnermeier, M. K. and Parker, J. A. (2005). Optimal expectations. *American Economic Review*, 95(4):1092–1118.

Bénabou, R. (2015). The Economics of Motivated Beliefs. *Revue d'économie politique*, 125(5):665–685.

Bénabou, R. and Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics*, 117(3):871–915.

Caplin, A. and Leahy, J. (2001). Psychological expected utility theory and anticipatory feelings. *The Quarterly Journal of Economics*, 116(1):55–79.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Chetty, R., Looney, A., and Kroft, K. (2009). Salience and taxation: Theory and evidence. *American Economic Review*, 99(4):1145–77.

De Brigard, F. (2010). If you like it, does it matter if it's real? *Philosophical Psychology*, 23(1):43–57.

DellaVigna, S., Pope, D., and Vivalt, E. (2019). Predict science to improve science. *Science*, 366(6464):428–429.

Exley, C. L. (2016). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies*, 83(2):587–628.

Hindriks, F. and Douven, I. (2018). Nozick's experience machine: An empirical study. *Philosophical Psychology*, 31(2):278–298.

Köszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, 4(4):673–707.

Mill, J. (1879). *Utilitarianism*. Longmans, Green and Company.

Nozick, R. (1974). *Anarchy, state, and utopia*. John Wiley & Sons.

Rowland, R. (2017). Our intuitions about the experience machine. *J. Ethics & Soc. Phil.*, 12:110.

Sen, A. (1985). *Commodities and Capabilities*. North-Holland, Amsterdam. New Delhi: Oxford University Press, 1987; Italian translation: Giuffre Editore, 1988; Japanese translation: Iwanami, 1988.

Smith, B. (2011). Can we test the experience machine? *Ethical Perspectives*, 18(1):29–51.

Tversky, A. and Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The quarterly journal of economics*, 106(4):1039–1061.

Uhl, M. (2011). Do self-committers mind other-imposed commitment? an experiment on weak paternalism. *Rationality, markets, and morals*.

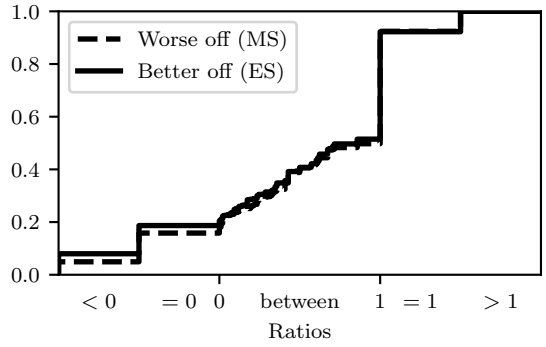
Weijers, D. (2013). Intuitive biases in judgements about thought experiments: The experience machine revisited. *Philosophical Writings*, 41(1).

A. Additional figures and tables

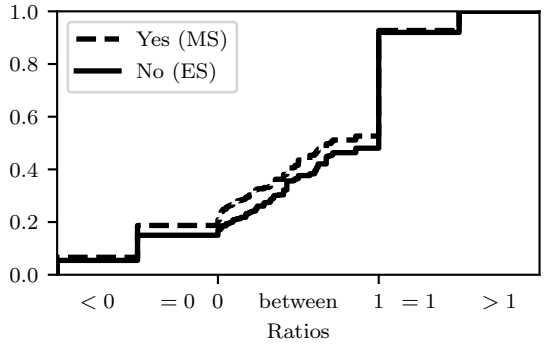
	$WTP<0$			$WTP=0$			$WTP>0$				
	$ES<0$	$ES=0$	$ES>0$	$ES<0$	$ES=0$	$ES>0$	$ES<0$	$ES=0$	$ES<WTP$	$ES=WTP$	$ES>WTP$
High-quality data											
Baseline N=406	0.00%	0.00%	0.00%	0.25%	10.10%	0.99%	2.46%	10.84%	33.50%	36.95%	4.93%
Low N=393	0.25%	0.00%	0.25%	0.25%	12.72%	1.53%	6.62%	9.16%	28.24%	35.62%	5.34%
High N=399	0.00%	0.00%	0.00%	0.00%	10.53%	1.50%	3.26%	8.77%	30.08%	40.10%	5.76%

Table 2: Share of responses by type for the quality-restricted sample

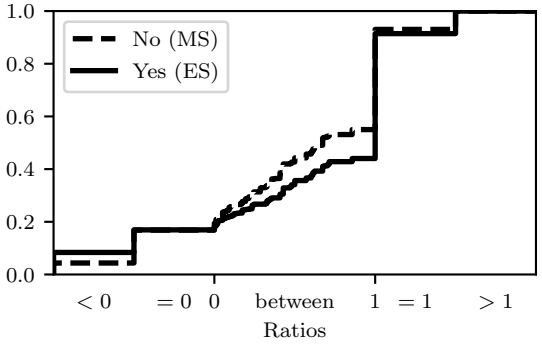
Note: WTP is the willingness-to-pay to give Receiver the books with original notes in *Learns* when he learns about it. ES is the willingness-to-pay to give Receiver the books with original notes in *NotLearns* when he does not learn about it.



(a) Figure 3 only for Arkansas only



(b) Figure 3 only for Experience Machine



(c) Caption

Figure 3: Figure 3 only for Warhol

	$WTP<0$			$WTP=0$			$WTP>0$				
	$ES<0$	$ES=0$	$ES>0$	$ES<0$	$ES=0$	$ES>0$	$ES<0$	$ES=0$	$ES<WTP$	$ES=WTP$	$ES>WTP$
warhol_Yes	0.16%	0.0%	0.62%	0.93%	17.13%	2.96%	6.54%	6.7%	21.18%	37.07%	6.7%
warhol_No	0.6%	0.12%	0.48%	0.12%	13.4%	2.87%	3.59%	10.29%	31.46%	31.34%	5.74%
arkansas_Better off	0.18%	0.0%	0.73%	0.73%	14.99%	3.11%	6.4%	8.59%	26.33%	32.72%	6.22%
arkansas_Worse off	0.54%	0.11%	0.43%	0.32%	15.04%	2.79%	3.97%	8.81%	27.39%	34.48%	6.12%
experience_Yes	0.26%	0.0%	0.66%	0.53%	16.05%	3.03%	5.26%	9.61%	26.97%	31.84%	5.79%
experience_No	0.56%	0.14%	0.42%	0.42%	13.93%	2.79%	4.46%	7.8%	27.02%	35.93%	6.55%
'all_questions_ES'	0.88%	0.0%	1.75%	1.75%	14.91%	4.39%	5.26%	5.26%	25.44%	35.96%	4.39%
'all_questions_MS'	0.72%	0.0%	0.36%	0.0%	13.04%	3.26%	2.9%	11.23%	32.97%	32.25%	3.26%

Table 3: Data are percentages (fractions of the total. Letter values must sum up to 100%

Note: This table considers the whole sample. WTP is the willingness-to-pay to give Receiver the books with original notes in *Learns* when he learns about it. ES is the willingness-to-pay to give Receiver the books with original notes in *NotLearns* when he does not learn about it.

	$WTP<0$			$WTP=0$			$WTP>0$				
	$ES<0$	$ES=0$	$ES>0$	$ES<0$	$ES=0$	$ES>0$	$ES<0$	$ES=0$	$ES<WTP$	$ES=WTP$	$ES>WTP$
'warhol_Yes'	0.0%	0.0%	0.0%	0.94%	15.49%	1.88%	3.76%	7.98%	26.29%	36.62%	7.04%
'warhol_No'	0.36%	0.36%	0.36%	0.0%	13.57%	3.93%	3.57%	11.79%	31.07%	31.43%	3.57%
'arkansas_Better off'	0.0%	0.0%	0.0%	0.54%	13.98%	2.15%	4.3%	9.14%	31.18%	33.87%	4.84%
'arkansas_Worse off'	0.33%	0.33%	0.33%	0.33%	14.66%	3.58%	3.26%	10.75%	27.69%	33.55%	5.21%
'experience_Yes'	0.42%	0.0%	0.42%	0.42%	13.33%	3.33%	4.17%	12.08%	27.5%	34.58%	3.75%
'experience_No'	0.0%	0.4%	0.0%	0.4%	15.42%	2.77%	3.16%	8.3%	30.43%	32.81%	6.32%
'all_questions_ES'	0.0%	0.0%	0.0%	2.27%	15.91%	0.0%	2.27%	2.27%	31.82%	40.91%	4.55%
'all_questions_MS'	1.08%	0.0%	1.08%	0.0%	10.75%	4.3%	3.23%	11.83%	27.96%	38.71%	1.08%

Table 4: Data are percentages (fractions of the total. Letter values must sum up to 100%

Note: This table considers the sample in the baseline treatment only. WTP is the willingness-to-pay to give Receiver the books with original notes in *Learns* when he learns about it. ES is the willingness-to-pay to give Receiver the books with original notes in *NotLearns* when he does not learn about it.

	$WTP<0$			$WTP=0$			$WTP>0$				
	$ES<0$	$ES=0$	$ES>0$	$ES<0$	$ES=0$	$ES>0$	$ES<0$	$ES=0$	$ES<WTP$	$ES=WTP$	$ES>WTP$
'warhol_Yes'	0.19%	0.0%	0.0%	0.39%	12.09%	2.14%	5.65%	7.6%	23.59%	42.5%	5.85%
'warhol_No'	0.0%	0.0%	0.15%	0.0%	10.36%	0.73%	2.92%	11.09%	35.91%	33.87%	4.96%
'arkansas_Better off'	0.23%	0.0%	0.23%	0.0%	10.59%	1.58%	5.63%	9.23%	30.18%	36.26%	6.08%
'arkansas_Worse off'	0.0%	0.0%	0.0%	0.27%	11.41%	1.19%	3.18%	9.81%	30.9%	38.33%	4.91%
'experience_Yes'	0.0%	0.0%	0.17%	0.33%	11.3%	1.5%	4.32%	10.63%	30.56%	36.05%	5.15%
'experience_No'	0.17%	0.0%	0.0%	0.0%	10.91%	1.17%	3.86%	8.56%	30.7%	39.09%	5.54%
'all_questions_ES'	1.05%	0.0%	0.0%	0.0%	13.68%	2.11%	5.26%	5.26%	26.32%	42.11%	4.21%
'all_questions_MS'	0.0%	0.0%	0.0%	0.0%	10.09%	0.88%	2.19%	12.28%	35.96%	35.53%	3.07%

Table 5: Data are percentages (fractions of the total. Letter values must sum up to 100%

Note: This table considers the high-quality sample for all treatments. WTP is the willingness-to-pay to give Receiver the books with original notes in *Learns* when he learns about it. ES is the willingness-to-pay to give Receiver the books with original notes in *NotLearns* when he does not learn about it.

	$WTP<0$			$WTP=0$			$WTP>0$				
	$ES<0$	$ES=0$	$ES>0$	$ES<0$	$ES=0$	$ES>0$	$ES<0$	$ES=0$	$ES<WTP$	$ES=WTP$	$ES>WTP$
'warhol_Yes'	0.0%	0.0%	0.0%	0.58%	9.25%	0.58%	2.31%	8.67%	30.06%	41.62%	6.94%
'warhol_No'	0.0%	0.0%	0.0%	0.0%	10.73%	1.29%	2.58%	12.45%	36.05%	33.48%	3.43%
'arkansas_Better off'	0.0%	0.0%	0.0%	0.0%	9.49%	0.63%	3.8%	8.86%	36.08%	36.08%	5.06%
'arkansas_Worse off'	0.0%	0.0%	0.0%	0.4%	10.48%	1.21%	1.61%	12.1%	31.85%	37.5%	4.84%
'experience_Yes'	0.0%	0.0%	0.0%	0.51%	8.72%	1.03%	3.08%	12.82%	31.28%	38.46%	4.1%
'experience_No'	0.0%	0.0%	0.0%	0.0%	11.37%	0.95%	1.9%	9.0%	35.55%	35.55%	5.69%
'all_questions_ES'	0.0%	0.0%	0.0%	0.0%	12.82%	0.0%	2.56%	2.56%	33.33%	46.15%	2.56%
'all_questions_MS'	0.0%	0.0%	0.0%	0.0%	9.88%	1.23%	2.47%	13.58%	29.63%	41.98%	1.23%

Table 6: Data are percentages (fractions of the total. Letter values must sum up to 100%

Note: This table considers the high-quality sample in the baseline treatment only. WTP is the willingness-to-pay to give Receiver the books with original notes in *Learns* when he learns about it. ES is the willingness-to-pay to give Receiver the books with original notes in *NotLearns* when he does not learn about it.

We are giving a present to someone!

The questions we will ask you to answer involve another person. His name is Alex.

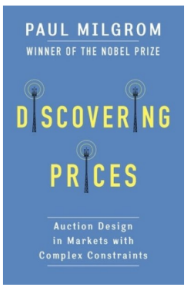
Alex loves economics, and we are going to give him a present!

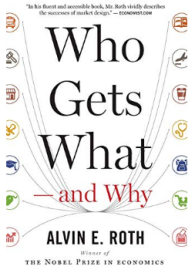
He is going to get two books by two Nobel laureates in economics—Professors Paul Milgrom and Alvin Roth (you can click each of their names to open their Wikipedia page). Professors Milgrom and Roth are professors at our university and have agreed to help with the study.

Alex has already read some of their work and told us he has great admiration for them.

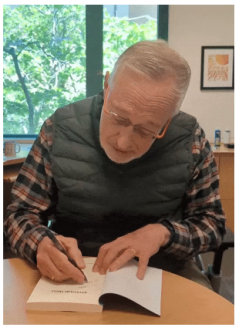
The two books come with handwritten notes. But here is the twist! **We have two copies of each of these books.** One with **original notes** from the famous authors themselves (Profs. Milgrom and Roth), and one with **fake notes** written by someone excellent at copying their handwriting.

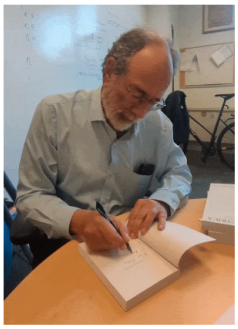
Here are the two books





Here are videos of the professors writing the notes





The fake versions of the handwritten notes are indistinguishable from the original ones. Professors Milgrom and Roth themselves could not tell which is which!

Alex will receive two books, either the two with the original handwritten notes or the two with the fake ones. We will return the two books that we do not give to Alex back to Professors Milgrom and Roth.

When you are ready, click "Next."

Next

Your task

In addition to the books, Alex may also receive a **surprise bonus**.

Your task in this study is to make choices that we will use to determine:

- which two books (original or fake) Alex receives,
- his surprise bonus.

In particular, we will randomly choose one participant like yourself and actually implement what they chose for Alex.

There is one more important detail: **We might or might not tell Alex whether the two books he gets are the ones with the original or fake notes.**

Case 1: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes

Case 2: we WILL tell Alex whether the books he got are the ones with the original or fake notes

We ask you to make your choices for each case. We will then choose a case randomly and implement that case.

Alex knows that if we don't tell him which books he got, he may have got the ones with the original or the fake notes.

When ready, click "Next."

Next

Baseline treatment

Your task

In addition to the books, Alex may also receive a **surprise bonus**.

Your task in this study is to make choices that we will use to determine:

- which two books (original or fake) Alex receives,
- his surprise bonus.

In particular, we will randomly choose one participant like yourself and actually implement what they chose for Alex.

There is one more important detail: **We might or might not tell Alex whether the two books he gets are the ones with the original or fake notes.**

Case 1: we WILL tell Alex whether the books he got are the ones with the original or fake notes

With a 75% chance, Alex will get the books with the fake notes, and with a 25% chance, **you will determine which books he gets.**

Case 2: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes

With a 75% chance, Alex will get the books with the fake notes, and with a 25% chance, **you will determine which books he gets.**

We ask you to make your choices for each case. We will then choose a case randomly and implement that case.

Alex knows that if we don't tell him which books he got, there is at least a 75% chance that they are the ones with the fake notes.

When ready, click "Next."

Next

LowMS treatment

Your task

In addition to the books, Alex may also receive a **surprise bonus**.

Your task in this study is to make choices that we will use to determine:

- which two books (original or fake) Alex receives,
- his surprise bonus.

In particular, we will randomly choose one participant like yourself and actually implement what they chose for Alex.

There is one more important detail: **We might or might not tell Alex whether the two books he gets are the ones with the original or fake notes.**

Case 1: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes

With a 75% chance, Alex will get the books with the original notes, and with a 25% chance, **you will determine which books he gets.**

Case 2: we WILL tell Alex whether the books he got are the ones with the original or fake notes

With a 75% chance, Alex will get the books with the original notes, and with a 25% chance, **you will determine which books he gets.**

We ask you to make your choices for each case. We will then choose a case randomly and implement that case.

Alex knows that if we don't tell him which books he got, there is at least a 75% chance that they are the ones with the original notes.

When ready, click "Next."

Next

HighMS treatment

Well done!

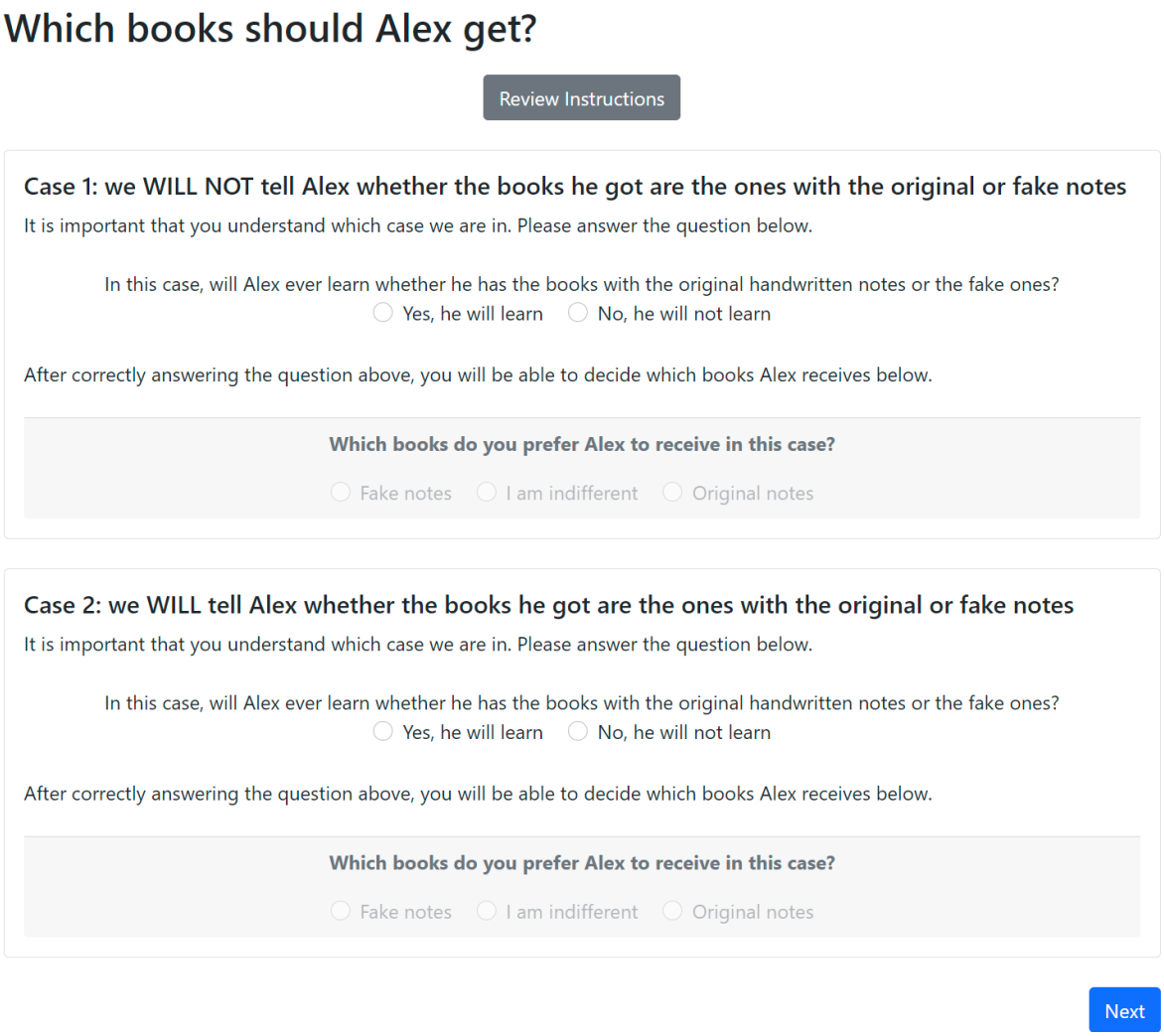
On the next pages, we will ask you 15 questions to determine which books Alex gets.

For example, one of them will be: *which books do you prefer Alex to receive? The ones with the original or fake notes?* There are other questions where we add a bonus for Alex to one of the options. You can click on this button to see all questions: Questions

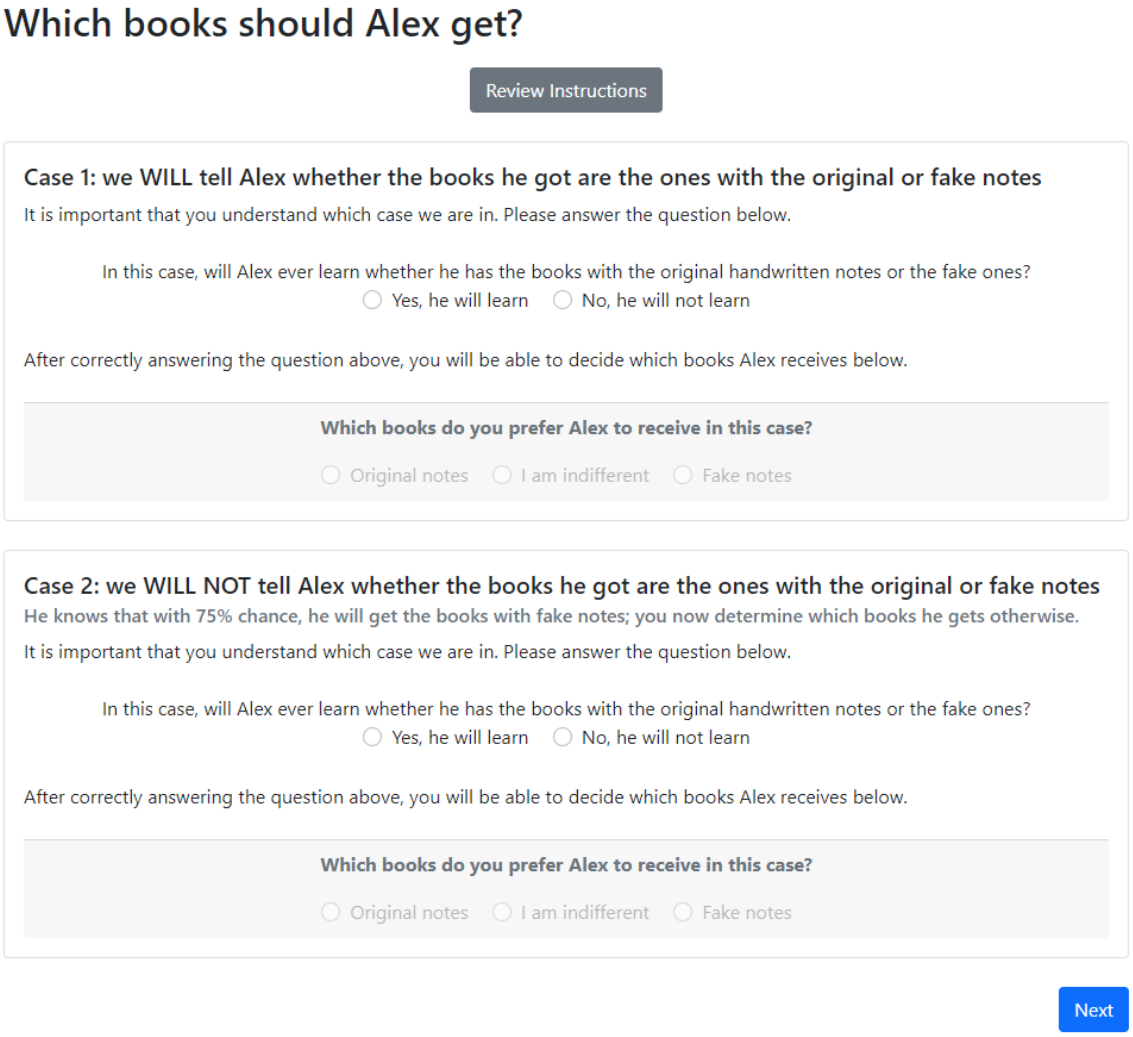
We will randomly pick one of the questions and implement whatever option you choose. This means that any question can be the one that determines what Alex gets, so please answer them carefully.

When you are ready, click "Next."

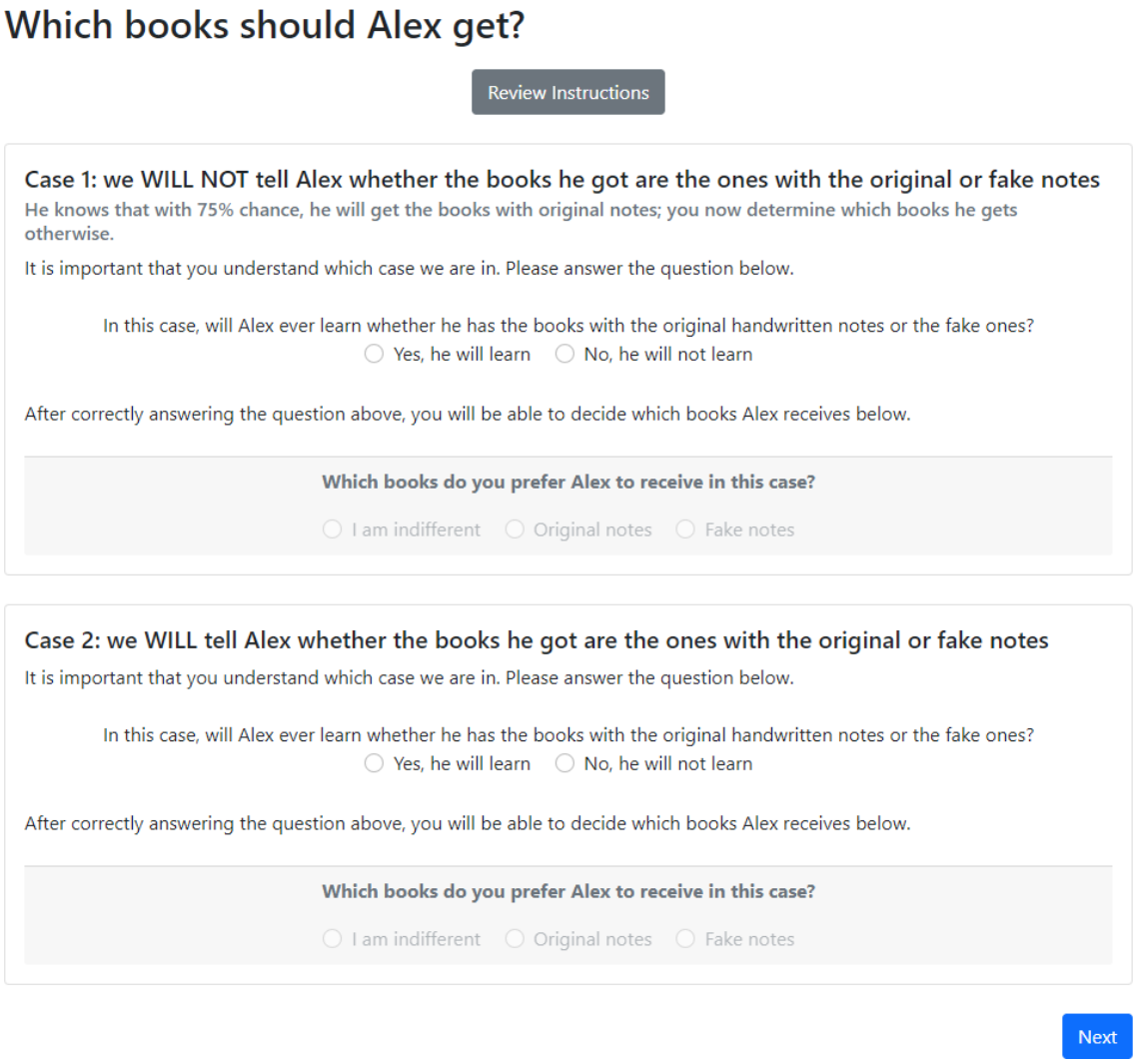
Next



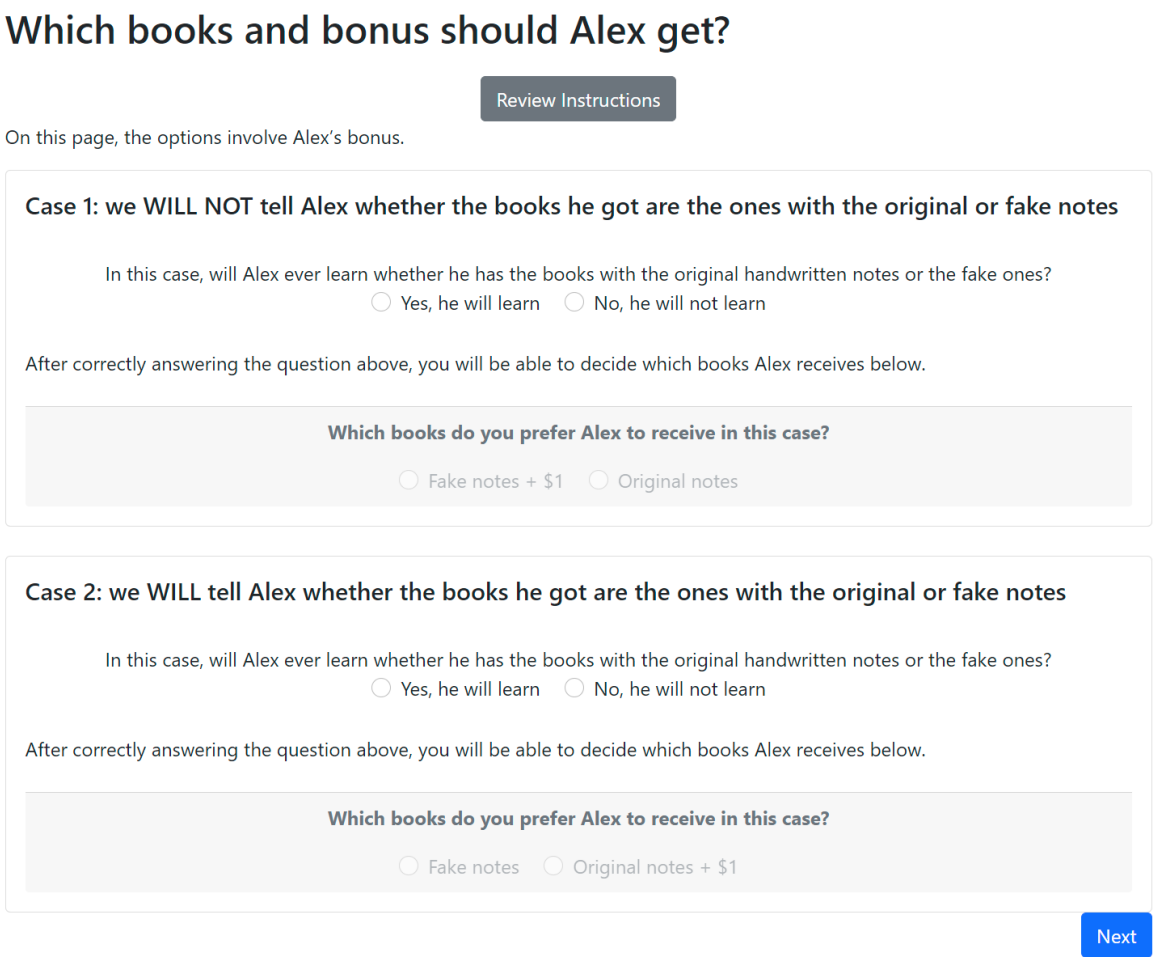
Baseline treatment



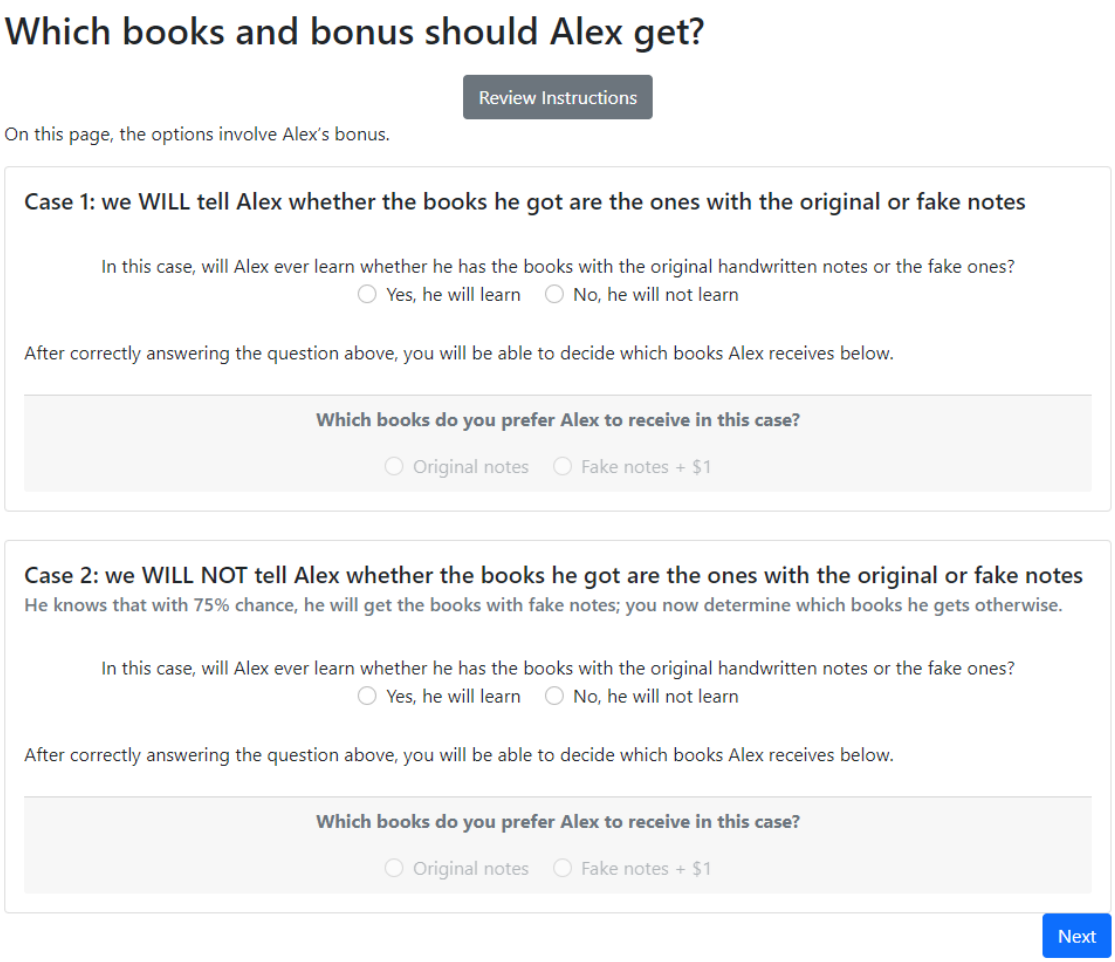
LowMS treatment



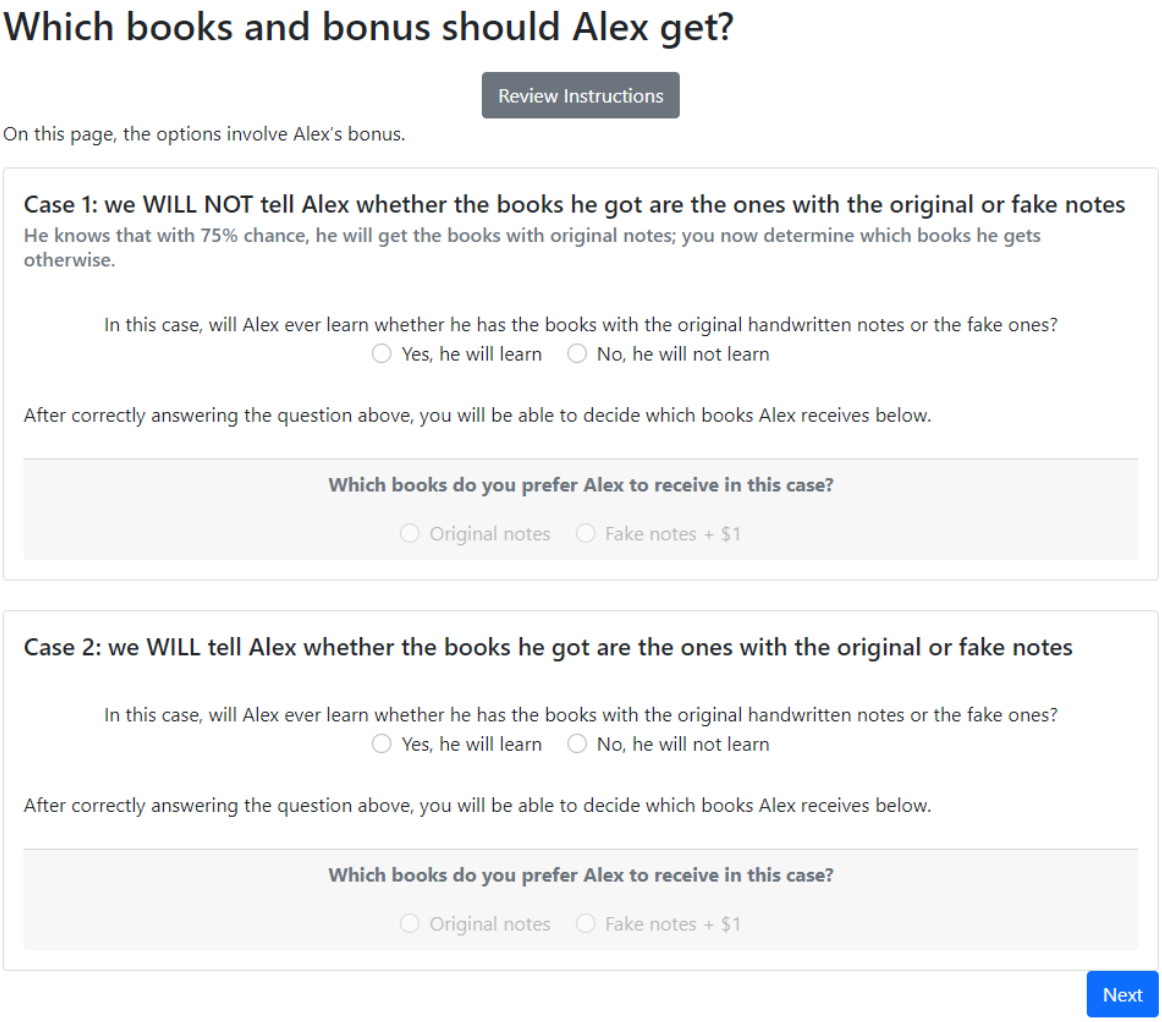
HighMS treatment



Baseline treatment



LowMS treatment



HighMS treatment

Which books and bonus should Alex get?

You make several choices again involving Alex's bonus, filling in a table like the one below. For each row, we ask you to choose between the original notes and fake notes with a bonus for Alex.

Fake notes and...		Original notes and...
...$2	OR	...$0
...$3	OR	...$0
...$5	OR	...$0
...$7	OR	...$0
...$10	OR	...$0
...$15	OR	...$0
...$25	OR	...$0
...$45	OR	...$0
...$70	OR	...$0
...$100	OR	...$0
...$140	OR	...$0
...$200	OR	...$0

We assume that once you choose the fake notes for one row, you will choose the fake notes for all rows below because the rows below simply make the fake notes better by increasing the bonus. You will only need to choose the row in which you switch from preferring the original notes to preferring the fake notes. You do that by clicking on the row.

Once you are confident that you understand how the table works, continue to give your answers.

Next

Which books and bonus should Alex get?

Review Instructions

Case 1: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

○ Yes, he will learn ○ No, he will not learn

After correctly answering the question above, please tell us which option you prefer by clicking the table below.

Fake notes and...		Original notes and...
...$2	OR	...$0
...$3	OR	...$0
...$5	OR	...$0
...$7	OR	...$0
...$10	OR	...$0
...$15	OR	...$0
...$25	OR	...$0
...$45	OR	...$0
...$70	OR	...$0
...$100	OR	...$0
...$140	OR	...$0
...$200	OR	...$0

Case 2: we WILL tell Alex whether the books he got are the ones with the original or fake notes

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

○ Yes, he will learn ○ No, he will not learn

After correctly answering the question above, please tell us which option you prefer by clicking the table below.

Fake notes and...		Original notes and...
...$2	OR	...$0
...$3	OR	...$0
...$5	OR	...$0
...$7	OR	...$0
...$10	OR	...$0
...$15	OR	...$0
...$25	OR	...$0
...$45	OR	...$0
...$70	OR	...$0
...$100	OR	...$0
...$140	OR	...$0
...$200	OR	...$0

Which books and bonus should Alex get?

Review Instructions

Case 1: we WILL tell Alex whether the books he got are the ones with the original or fake notes

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

○ Yes, he will learn ○ No, he will not learn

After correctly answering the question above, please tell us which option you prefer by clicking the table below.

Original notes and...		Fake notes and...
...$0	OR	...$2
...$0	OR	...$3
...$0	OR	...$5
...$0	OR	...$7
...$0	OR	...$10
...$0	OR	...$15
...$0	OR	...$25
...$0	OR	...$45
...$0	OR	...$70
...$0	OR	...$100
...$0	OR	...$140
...$0	OR	...$200

Case 2: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes

He knows that with 75% chance, he will get the books with fake notes; you now determine which books he gets otherwise.

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

○ Yes, he will learn ○ No, he will not learn

After correctly answering the question above, please tell us which option you prefer by clicking the table below.

Original notes and...		Fake notes and...
...$0	OR	...$2
...$0	OR	...$3
...$0	OR	...$5
...$0	OR	...$7
...$0	OR	...$10
...$0	OR	...$15
...$0	OR	...$25
...$0	OR	...$45
...$0	OR	...$70
...$0	OR	...$100
...$0	OR	...$140
...$0	OR	...$200

Next

Which books and bonus should Alex get?

Review Instructions

Case 1: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes

He knows that with 75% chance, he will get the books with original notes; you now determine which books he gets otherwise.

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

○ Yes, he will learn ○ No, he will not learn

After correctly answering the question above, please tell us which option you prefer by clicking the table below.

Original notes and...		Fake notes and...
...$0	OR	...$2
...$0	OR	...$3
...$0	OR	...$5
...$0	OR	...$7
...$0	OR	...$10
...$0	OR	...$15
...$0	OR	...$25
...$0	OR	...$45
...$0	OR	...$70
...$0	OR	...$100
...$0	OR	...$140
...$0	OR	...$200

Case 2: we WILL tell Alex whether the books he got are the ones with the original or fake notes

In this case, will Alex ever learn whether he has the books with the original handwritten notes or the fake ones?

○ Yes, he will learn ○ No, he will not learn

After correctly answering the question above, please tell us which option you prefer by clicking the table below.

Original notes and...		Fake notes and...
...$0	OR	...$2
...$0	OR	...$3
...$0	OR	...$5
...$0	OR	...$7
...$0	OR	...$10
...$0	OR	...$15
...$0	OR	...$25
...$0	OR	...$45
...$0	OR	...$70
...$0	OR	...$100
...$0	OR	...$140
...$0	OR	...$200

Please review your responses for the two cases

Case 1: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes

1. I prefer Alex to receive the ones with the original notes and $1 over the ones with the fake notes
2. I prefer Alex to receive the ones with the original notes over the ones with the fake notes
3. I prefer Alex to receive the ones with the original notes over the ones with the fake notes and $1
4. I prefer Alex to receive the ones with the original notes over the ones with the fake notes and $2
5. I prefer Alex to receive the ones with the original notes over the ones with the fake notes and $3
6. I prefer Alex to receive the ones with the original notes over the ones with the fake notes and $5
7. I prefer Alex to receive the ones with the original notes over the ones with the fake notes and $7
8. I prefer Alex to receive the ones with the original notes over the ones with the fake notes and $10
9. I prefer Alex to receive the ones with the original notes over the ones with the fake notes and $15
10. I prefer Alex to receive the ones with the original notes over the ones with the fake notes and $25
11. I prefer Alex to receive the ones with the original notes over the ones with the fake notes and $45
12. I prefer Alex to receive the ones with the original notes over the ones with the fake notes and $70
13. I prefer Alex to receive the ones with the original notes over the ones with the fake notes and $100
14. I prefer Alex to receive the ones with the fake notes and $140 over the ones with the original notes
15. I prefer Alex to receive the ones with the fake notes and $200 over the ones with the original notes

Case 2: we WILL tell Alex whether the books he got are the ones with the original or fake notes

1. I prefer Alex to receive the ones with the original notes and $1 over the ones with the fake notes
2. I prefer Alex to receive the ones with the original notes over the ones with the fake notes
3. I prefer Alex to receive the ones with the original notes over the ones with the fake notes and $1
4. I prefer Alex to receive the ones with the original notes over the ones with the fake notes and $2
5. I prefer Alex to receive the ones with the original notes over the ones with the fake notes and $3
6. I prefer Alex to receive the ones with the original notes over the ones with the fake notes and $5
7. I prefer Alex to receive the ones with the fake notes and $7 over the ones with the original notes
8. I prefer Alex to receive the ones with the fake notes and $10 over the ones with the original notes
9. I prefer Alex to receive the ones with the fake notes and $15 over the ones with the original notes
10. I prefer Alex to receive the ones with the fake notes and $25 over the ones with the original notes
11. I prefer Alex to receive the ones with the fake notes and $45 over the ones with the original notes
12. I prefer Alex to receive the ones with the fake notes and $70 over the ones with the original notes
13. I prefer Alex to receive the ones with the fake notes and $100 over the ones with the original notes
14. I prefer Alex to receive the ones with the fake notes and $140 over the ones with the original notes
15. I prefer Alex to receive the ones with the fake notes and $200 over the ones with the original notes

Do the above answers reflect what you intended to answer, or do you want to give your answers again?

- ○ Yes, the answers above reflect what I intended to answer
- ○ No, I want to give my answers again

Please click "Next" to proceed to the next page.

Next

Thank you for your responses

Reminder:

- Case 1: we WILL NOT tell Alex whether the books he got are the ones with the original or fake notes
- Case 2: we WILL tell Alex whether the books he got are the ones with the original or fake notes

You gave different responses in Case 1 and Case 2. Why? Please tell us in approximately 1-3 sentences.
(There is nothing wrong with your answers! We are just interested in your reasoning)

In the remaining pages, we will present you with scenarios and ask you questions about them. Your responses are very important for our study; please think about the scenarios and answer carefully.

Next

The Experience Machine

If given the option, would you choose to plug into an experience machine that could provide you with an entirely immersive, simulated reality where you can experience any desirable scenario, despite not being real? Keep in mind that while plugged in, you would never be aware that you are in the experience machine and would believe that the simulated reality is real.

Suppose there was an experience machine that would give you *any* experience you desired (eating good food, having a successful career, making meaningful connections, etc.). While in the machine, you would not know that you are in it; you would think that what you are experiencing is actually happening.

Would you go into the machine?

- ○ Yes
- ○ No

Why? Answer in approximately 1-2 sentences.

Next

Flood in Arkansas

A small town in Arkansas experiences massive flooding, leaving many families homeless. To provide financial relief to the impacted families, the government temporarily increases taxes, including a $100 levy on John. John lives far away and *will never learn about the flooding or the relief effort*. However, he cares about helping others and would gladly contribute $100 to the relief effort if he knew about the flood.

Please tell us what you think using the information provided above. This is not a trick question; we want to understand what you think about the impact that the policy has on John.

Does the government raising taxes to provide financial relief make John better or worse off?

- ○ Better off
- ○ Worse off

Why? Answer in approximately 1-2 sentences.

Next

Andy Warhol

Hundreds of Andy Warhol fakes, and one original drawing worth $20k, sold for $250 each. An art collective purchased an original Warhol drawing and copied it 999 times. The copies are carefully created so that not even their creators can tell them apart from the original drawing. They then mixed the original together with the copies and sold the 1000 drawings.

Please watch the video (1min 1sec) the art collective made (audio is not needed). You can read more about this story here.

Someone got the original Andy Warhol drawing. Since the original drawing and the copies are indistinguishable, please assume that neither the person who got it nor anyone else will ever know which is the original drawing or who has it.

Is this person better off by getting the original one instead of a copy?

- ○ Yes
- ○ No

Why? Answer in approximately 1-2 sentences.

Next