Welfare and the Act of Choosing

B. Douglas Bernheim, Kristy Kim, and Dmitry Taubinsky*

December 2023

Abstract

The standard revealed preference approach to welfare economics encounters fundamental difficulties when the act of choosing directly affects welfare through emotions such as guilt, pride, and anxiety. We address this problem by developing an approach that redefines consumption bundles in terms of the sensations they produce, and measures welfare by blending choice-based methods with self-reported well-being techniques. In applications to classic social preferences paradigms, our approach shows that standard revealed-preference methods, including choices over menus, mismeasure welfare due to several types of choice set dependence of preferences, while self-reported happiness and satisfaction are not sufficient statistics for welfare.

*Bernheim: Stanford and NBER (bernheim@stanford.edu). Kim: UC Berkeley (kristykim@berkeley.edu). Taubinsky: UC Berkeley and NBER (dmitry.taubinsky@berkeley.edu). We thank seminar participants at Berkeley, Harvard, and Chicago for helpful comments. This project was supported by the Alfred P. Sloan Foundation. The experiment was determined exempt by the UC Berkeley IRB (CPHS protocol number 2021-07-14494).

1 Introduction

A growing literature in Behavioral Welfare Economics seeks to devise methods for assessing economic well-being in settings where decision making does not conform to the standard model of rational choice (Bernheim and Taubinsky, 2018). For example, prior studies address the possibilities that people may make mistakes, and that even their properly informed judgments may not admit coherent preference representations. Far less attention has been given to the possibility that people may care about the *experience of choosing*—for example, due to feelings of guilt, pride, the joy of exercising autonomy, or anxiety over responsibility. When the act of choosing directly affects welfare, choices made by an individual fundamentally differ from choices made by a social planner. Due to the resulting *Non-Comparability Problem*, welfare may not be recoverable from standard choice data.

An example from Koszegi and Rabin (2008) illustrates the Non-Comparability Problem. Suppose we task Norma with dividing a sum of money between herself and a friend. Norma is averse to bearing responsibility for leaving her friend with nothing when other options are available. Consequently, no matter how the task is presented, she divides the money equally. However, she is inherently selfish and fervently wishes someone would take the decision out of her hands, as long as they give her the entire prize. Thus, none of Norma's choices can directly reveal her true preference. In particular, if we ask her to choose between the original decision problem and a setting in which a third party reliably decides to give her everything, she will still feel responsible for the outcome, and consequently choose to divide the money herself, splitting it equally, despite her true preference for the alternative. A social planner guided by Norma's choices would thus incorrectly conclude that equally sharing is the best choice to make on Norma's behalf.

In this paper, we provide a general account of the Non-Comparability Problem, and offer a conceptual solution that involves redefining consumption bundles in terms of the mental states they produce. Implementation involves a blend of choice-based and self-reported well-being (SWB) methods. We use SWB methods (e.g., Benjamin et al., 2012, 2014) to generate proxies for mental state bundles, and choice-based methods to reveal preferences over the mental state bundles. We provide proof-of-concept for our solution through an experiment involving social preferences. Our results uncover new insights about social preferences, while illustrating important limitations of both standard choice-based and SWB methods. These findings suggest that our approach is potentially applicable to other domains that we summarize toward the end of the paper.

Prior work has relied on two strategies for avoiding the Non-Comparability Problem. One is simply to assume that the experience of choosing is not welfare-relevant when people make meta choices from higher-order menus. In the social preferences domain, such an assumption rules out our motivating example. Even so, this analytic strategy is popular in the branch of the social preferences literature that investigates the avoidance of opportunities to give (Dana et al., 2006a; Broberg et al., 2007; DellaVigna et al., 2012; Lazear et al., 2012). Specifically, the pertinent studies assume, either implicitly or explicitly, that if a person is indifferent between having an opportunity to share and

paying $X to avoid it, then creating the sharing opportunity reduces the sharer's money-metric well-being by $X. Similar assumptions are also implicit in the literature on the value of authority, autonomy, and control (e.g., Fehr et al., 2013; Owens et al., 2014; Bartling et al., 2014), which elicits intrinsic preferences for decision power by offering inconsequential opportunities to expand menus. We provide a formal theoretical result showing that such assumptions are both untestable and indispensable for the measurement of welfare when one relies on standard choice data.

Another way to avoid the Non-Comparability Problem is to jettison the choice-based welfare paradigm in favor of a competing tradition that employs self-reported measures of well-being. Such measures can encompass the experience of choosing as well as feelings about the outcome. However, they introduce other conceptual difficulties, such as the *Aggregation Problem*. Subjective experience is disaggregated over time, states of nature, and categories (for example, hunger, anger, and elation). To assess welfare based on measured subjective sensations, one must therefore introduce a principle of aggregation that is not itself a sensation. Aggregating based on each individual's linguistic construction of a word or phrase such as "happiness" or "life satisfaction" is normatively arbitrary. Furthermore, people's choices imply that they sometimes reject this standard of evaluation (Benjamin et al., 2014).

Our proposed approach is a hybrid that draws on both the choice-based tradition and the SWB tradition. Briefly, for each option in a given decision problem, we elicit proxies for the disaggregated bundle of anticipated mental states it induces. Using standard econometric methods, we then use choice data to estimate preferences over mental state bundles. This step builds on Benjamin et al. (2014), who constructed an index of well-being by relating choices to a collection of subjective evaluations. The recovered preferences allow us to make welfare comparisons between choice situations, such as those associated with different policy options, based on the mental-state bundles they induce. Using those preferences, we can construct estimates of the money-metric consumer surplus derived from any given decision problem by deploying standard concepts such as equivalent variation. Because mental states reflect not only the selected outcome but also the experience of choosing, the method overcomes the Non-Comparability Problem that otherwise afflicts choice-based approaches. And because the method uses choice-based techniques to aggregate over mental states, it overcomes the Aggregation Problem that afflicts the SWB approach.

We study three classes of two-person allocation problems in an online experiment with 2,740 participants: standard dictator games (DG module), settings where the computer chooses the allocation (CC module), and opt-out games where the decision maker has the option to avoid a dictator game (OO module). One version of the CC module simply presents the computer-generated allocation without any mention of alternatives, while another version explains that the computer randomly selected an option from a DG choice set.

We begin by estimating a model of preferences over (anticipated) mental-state bundles using data from the DGs. We verify that the relationship between choice and anticipated mental state bundles

is stable with respect to different types of decision problems, including opt-out games; that our elicitations span the relevant set of mental states; and that various other potential confounds we highlight do not come into play. We find that the Aggregation Problem is important in practice by showing that the recovered preferences rank mental state bundles differently than self-reported happiness and satisfaction, and that various other mental state proxies help predict choice even when controlling for those measures. We rule out the possibility that this conclusion is an artifact of measurement error. Consequently, happiness and satisfaction are not sufficient statistics for welfare.

Next, we use the recovered preferences to evaluate welfare from the same allocation when the participant chooses it and when the computer chooses it for them, based on the mental-state bundles associated with each. To make the comparisons economically meaningful, we convert differences in welfare into dollars. We find that participants derive greater well-being from any given option when it is chosen by the computer (in the primary module) than when they choose it themselves. For the less-equitable options, this finding is consistent with the hypothesis that people experience disutility from sensations such as guilt when treating their partner unfairly. Surprisingly, we observe similar patterns for the more equitable options. Consistent with intuition, people experience greater pride when participants choose the more equitable allocation themselves than when the computer assigns it. However, anticipated happiness, financial satisfaction, and overall satisfaction are all higher when the computer exogenously assigns the more equitable allocation than when the participants choose it, and these effects drive the overall welfare effect. Interestingly, however, in the "choice set" version of the CC module, participants derive the same utility from the less-equitable allocation regardless of who selects it. This finding suggests that the mere awareness of a more financially advantageous option reduces participants' well-being.

Importantly, the discrepancy between utility for the computer-chosen and participant-chosen allocations is greater for less equitable options than for more equitable options. It follows that standard choice-based welfare analysis *overstates* the net benefit the individual derives from the more equitable option relative to the less equitable option. Indeed, this overstatement can lead one to conclude that the individual prefers being assigned the more equitable option rather than the less equitable option, when in fact the opposite is true. In this way, we demonstrate that the Non-Comparability Problem is important in standard choice problems.

The final portion of the paper examines opt-out games. We find that people once again experience outcomes differently depending on the process that generates them. On the one hand, participants derive less utility from the opt out allocation when they choose it themselves than when the computer assigns it. This result is consistent with participants experiencing negative sensations such as guilt when they opt-out, as hypothesized in our motivating example. On the other hand, we also find that having an outside option reduces the utility associated with the other available options. Except in knife-edge cases where these two effects exactly offset each other, the practice of using avoidance designs to "price out" the value of encountering a decision problem (as in Broberg et al., 2007, Lazear

et al., 2012, DellaVigna et al., 2012) measures welfare incorrectly. In our setting, we find that in the majority of cases, the pricing-out approach understates the social value of creating sharing opportunities. This conclusion is also a practical manifestation of the Non-Comparability Problem.

Our analysis draws on, extends, and connects several important literatures. We contribute to the literature on choice-based behavioral welfare economics by providing more general formalizations of the Non-Comparability Problem (Koszegi and Rabin, 2008; Bernheim, 2016; Bernheim and Taubinsky, 2018), by proposing a conceptual solution that avoids potentially problematic and difficult-to-test identifying assumptions, and by providing proof-of-concept for the solution using data from a laboratory experiment. Because our methods involve a hybrid of choice-based methods and SWB, we also contribute to the literature that studies the relationship between choice and SWB, including the extent to which factors beyond happiness and satisfaction predict choice (Benjamin et al., 2012, 2014), as well as to the broader literature on the use of SWB for welfare measurement and policy evaluation (e.g., Gruber and Mullainathan, 2005; DiTella et al., 2001; Ludwig et al., 2012; Deaton, 2018). Finally, our analysis yields new insights concerning particular applications studied in the literature on "reluctant" giving, which investigates the reasons people may avoid or excuse themselves from situations where they may feel obligated to act prosocially (Dana et al., 2006a,b; Broberg et al., 2007; Lazear et al., 2012; DellaVigna et al., 2012; Exley, 2015).

The paper is organized as follows. Section 2 provides conceptual foundations. We describe the Non-Comparability Problem and Aggregation Problem in greater detail, and then propose our solution. Section 3 covers experimental design, and Section 4 explains our empirical methods. Sections 5 examines the relationships between choices and proxies for mental states. Section 6 contains our welfare analyses, including our evaluations of standard methods that attempt to infer well-being from choices or SWB. Section 7 concludes by suggesting a variety of other domains and difficult welfare questions to which our approach is potentially applicable.

2 Conceptual Framework

In the section, we elaborate on the nature of the Non-Comparability Problem for choice-based methods (Section 2.1), and of the Aggregation Problem for SWB methods (Section 2.2). Then we describe our proposed solution to both problems (Section 2.3).

2.1 The Non-Comparability Problem

The Non-Comparability Problem arises when people care about the experience of choosing. In principle, their preferences could encompass any feature of a decision problem, including the constraint set, the decision tree, the path taken through the tree, and other details such as how information is presented. We begin by exhibiting the problem in settings where people care about the constraint set but no other aspect of the decision problem, and then explain why alternative preference formulations do not offer escape routes.

2.1.1 The Non-Comparability Problem with Constraint-Set Dependence

For the first portion of our theoretical analysis, we follow Koszegi and Rabin (2008) by assuming that preferences are defined over bundles of the form (X, x), where X is the constraint set and x is the selected option. The phrase *constraint set* references the set of all possible *outcomes*,[1] rather than the set of all possible paths through a decision tree. For example, consider an individual who must choose one of two possible allocations of money between themselves and another person: an equitable allocation e, and an inequitable allocation s that provides the other person with a much smaller share. When the individual chooses between e and s directly, $X = \{e, s\}$. Alternatively, they might first choose among the menus $\{s\}$, $\{e\}$, and $\{e, s\}$, and then choose from the selected menu. That structure yields a richer set of possible paths through the decision tree, but the constraint set is still $X = \{e, s\}$.

When the individual chooses $x^*(X)$ from the constraint set X, we can conclude that $(X, x^*(X)) \succeq (X, x)$ for all $x \in X$. However, for two distinct constraint sets X and X', their choices do not reveal whether they prefer $(X, x^*(X))$ or $(X', x^*(X'))$. Consequently, we cannot determine whether a policy that changes the constraint set helps or hurts them. This observation is the essence of the Non-Comparability Problem.

Adapting an example from Koszegi and Rabin (2008), consider the special case in which $X = \{e, s\}$, where e and s are once again the equitable and selfish outcomes, respectively, and assume the individual's preferences have the following form:

$$U(X, x) = u(x) - \left[\max_{y \in X} v(y) - v(x)\right] \tag{1}$$

We interpret $v(x)$ as an index of the "virtuousness" of x, so that the individual enjoys intrinsic utility, $u(x)$, minus a penalty for selecting anything other than the most virtuous outcome.[2] Now suppose they intrinsically want the selfish allocation $(u(s) > u(e))$, but that equity is more virtuous $(v(e) > v(s))$. When $v(e) - v(s) > u(s) - u(e)$, they choose e from the constraint set $\{e, s\}$ *no matter how the decision problem is structured.* And yet,

$$U(\{s\}, s) = u(s) > u(e) = U(\{e\}, e) = U(\{e, s\}, e). \tag{2}$$

If we think think of $X = \{x\}$ as a government policy that mandates x (an *x-mandate*), expression (2) implies that an individual prefers an s-mandate not only to a e-mandate, but also to a policy that gives her a choice between s and e.

Critically, this individual's choices cannot reveal their underlying preference for an s-mandate. One might hope to detect this preference by offering the following two-stage choice: they first select a

[1] For settings with risk or uncertainty, one can think of X as a set of lotteries. The "outcome" then consists of the selected lottery, rather than a realization from that lottery.

[2] This model is a variation on the Gul and Pesendorfer (2001) formulation of temptation preferences.

menu, the alternatives being $\{e, s\}, \{e\}$, and $\{s\}$, and then select an option from the chosen menu. However, the constraint set for this two-stage decision problem is still $X = \{e, s\}$. Consequently, since the individual's preferences are constraint-set-dependent, they will either select $\{s, e\}$ followed by e, or select $\{e\}$ (followed mechanically by e). In either case, their choice does not reveal their underlying preference for an s-mandate.

More generally, when the individual encounters a constraint set X, the selected outcome always maximizes $u + v$. The functions u and v are not separately identified from choice data. Offering a choice between constraint sets X' and X'' does not help resolve this identification problem, because by definition the constraint set for this new problem is then $X = X' \cup X''$, which adds additional data about $u + v$, but still does not help distinguish u from v.

2.1.2 The Non-Comparability Problem with Menu Dependence

Is there another way to formulate preferences over the experience of decision making that avoids the Non-Comparability Problem? One alternative is to assume that people care about the *menus* they encounter in the course of navigating through a multi-stage decision problem, rather than the constraint sets. This assumption is implicit in the Krusell et al. (2010) welfare analysis of the Gul and Pesendorfer (2001) temptation model, and it appears to lie behind the use of avoidance designs and other metachoices in empirical work. In this section, we provide a general formulation for menu-dependent preferences, and prove that it also encounters the Non-Comparability Problem, in that the solution it offers involves the imposition of arbitrary and untestable assumptions.

We consider settings in which an individual selects an item x from a constraint set X through a sequence of choices—for example, by first choosing a cuisine, then a restaurant, then an entree from its menu. The final stage of such processes necessarily presents a conventional menu $\mathcal{M}^1$ consisting of a collection of items, which we call a *level-1 menu*. In the penultimate stage, the individual selects $\mathcal{M}^1$ from a collection of level-1 menus $\mathcal{M}^2$, which we call a *level-2 menu*. Recursively, in a K-stage decision problem, the first stage involves the choice of $\mathcal{M}^{K-1}$ from a collection of level-$(K-1)$ menus $\mathcal{M}^K$, which we call a *level-K menu*.[3]

To illustrate, suppose the individual makes selections from the constraint set $\{e, s\}$ in two steps: first they decide whether each of the options should be available; then they choose from the available options. In that case, the level-2 menu $\mathcal{M}^2$ consists of the level-1 menus $\mathcal{M}^1_1 = \{s\}$, $\mathcal{M}^1_2 = \{e\}$, and $\mathcal{M}^1_3 = \{e, s\}$.

Within this framework, how might we resolve the Non-Comparability Problem while allowing for the possibility that the act of choosing generates welfare-relevant sensations? The simplest solution is to assume that the individual's preferences depend only on the items they receive and the level-1 menus they face. Under this assumption, choosing $\mathcal{M}^1_2 = \{e\}$ or $\mathcal{M}^1_3 = \{e, s\}$ over $\mathcal{M}^1_1 = \{s\}$

[3]Conventional decision trees for deterministic choice problems induce such menu hierarchies. With randomness, a level-k menu would consist of lotteries over level-$(k-1)$ menus.

implies that, in a setting where the level-1 menu is exogenously assigned, they would indeed prefer to have e on that menu. Furthermore, individuals who choose $\mathcal{M}_2^1 = \{e\}$ over $\mathcal{M}_1^1 = \{s\}$ prefer an e-mandate over an s-mandate, and hence are better off with an e-mandate.

This simple solution attempts to resolve the problem by fiat, through the assumption that choices over menus do not generate welfare-relevant sensations, even though choices over the elements of those menus may. But if choosing an entree from a restaurant's menu is an emotionally consequential process, then it is difficult to justify how, as a matter of principle, one can rule out the possibility that the same is true for the choice of a restaurant, or even the choice of a cuisine. There is little if any psychological foundation, either theory or evidence, for the general claim that only the final stage of a multi-stage decision process is emotionally consequential. And yet, without that assumption, the Non-Comparability Problem resurfaces. In the example given above, an individual's choice of $\{e\}$ from $\mathcal{M}^2 = \{\{s\}, \{e\}, \{e, s\}\}$ simply reveals that they prefer the bundle $\{e, \{e\}, \mathcal{M}^2\}$ over $\{s, \{s\}, \mathcal{M}^2\}$, $\{e, \{e, s\}, \mathcal{M}^2\}$, and $\{s, \{e\}, \mathcal{M}^2\}$. It does not follow that they prefer an e-mandate, which corresponds to the bundle $\{e, \{e\}, \{\{e\}\}\}$, over an s-mandate, which corresponds to the bundle $\{s, \{s\}, \{\{s\}\}\}$.

One might nevertheless hope to devise a test for the hypothesis of interest, that utility depends only on the level-1 menu and the selected item. As an example, we might try to infer preferences over $(x, \mathcal{M}^1, ..., \mathcal{M}^{K-1})$ vectors by examining K-stage decisions, which start with choices from level-K menus. If those revealed preferences imply indifference toward $(\mathcal{M}^2, ..., \mathcal{M}^{K-1})$ (for fixed $(x, \mathcal{M}^1)$), one might be tempted to conclude that level-2 through level-$(K-1)$ menus are not directly welfare-relevant. While it is impossible to perform that test for $K = \infty$, doing so for some suitably large value of K might appear to give the analyst reason for comfort,

Unfortunately, this simple test is flawed. Intuitively, the problem is that K-stage decisions depend on preferences over $(x, \mathcal{M}^1, ..., \mathcal{M}^K)$ vectors, not over $(x, \mathcal{M}^1, ..., \mathcal{M}^{K-1})$ vectors. Consequently, one cannot validly recover preferences over $(x, \mathcal{M}^1, ..., \mathcal{M}^{K-1})$ by varying $\mathcal{M}^K$ across K-stage decision problems unless one maintains the assumption that $\mathcal{M}^K$ is not directly welfare-relevant. If that assumption is wrong, then the proposed test is invalid, as are any conclusions concerning the direct welfare-relevance of $(\mathcal{M}^2, ..., \mathcal{M}^{K-1})$.

We formalize this under-identification principle, along with the shortcomings of the proposed test, for a simple but reasonably general model of menu-dependence (1). Suppose we have data on choices for levels 1 through K (i.e., choices that begin with some $\mathcal{M}^k$ for $k = 1, ..., K$). We assume the preferences governing these choices correspond to a utility function within the following class:

$$V(x, \mathcal{M}^1, ..., \mathcal{M}^K) = u(x) + \pi^1(x, \mathcal{M}^1) + \sum_{k=2}^{K} \pi^k(\mathcal{M}^{k-1}, \mathcal{M}^k) \tag{3}$$

where

$$\pi^k(x,\{x\}) = 0 \text{ for all } x, \text{ and } \pi^k(\mathcal{M}^{k-1}, \left\{\mathcal{M}^{k-1}\right\}) = 0 \text{ for } k = 2, ..., K \text{ and all } \mathcal{M}^{k-1}. \quad (4)$$

In other words, the individual receives intrinsic utility from the item x, and also derives utility or disutility $\pi^k(\mathcal{M}^{k-1}, \mathcal{M}^k)$ when selecting from a level k menu. Significantly, where there is no level-k choice (i.e., the level-k menu is degenerate in the sense that either $\mathcal{M}^1 = \{x\}$ or $\mathcal{M}^k = \left\{\mathcal{M}^{k-1}\right\}$ for some $k \in \{2, ..., K\}$), the individual experiences no level-k utility. It follows that the individual's preferences over x-mandates depend only on u. Thus, to conduct welfare analysis for mandates, we need to identify the function u. It is therefore important to establish whether the hypothesized data allow us to do so.

Suppose the function V rationalizes the data. Consider any alternative to u, call it $\tilde{u}$. Define

$$\tilde{\pi}^1(x, \mathcal{M}^1) = \pi^1(x, \mathcal{M}^1) + [u(x) - \tilde{u}(x)] + f^1(\mathcal{M}^1)$$

for any function f^1 such that $f^1(\{x\}) = \tilde{u}(x) - u(x)$ for all $x \in X$. Recursively, for $k = 2, ..., K$, likewise define

$$\tilde{\pi}^k(\mathcal{M}^{k-1}, \mathcal{M}^k) = \pi^k(\mathcal{M}^{k-1}, \mathcal{M}^k) - f^{k-1}(\mathcal{M}^{k-1}) + f^k(\mathcal{M}^k)$$

for any function f^k such that $f^k(\{\mathcal{M}^{k-1}\}) = f^{k-1}(\mathcal{M}^{k-1})$ for all $\mathcal{M}^{k-1}$.

With this construction, we have

$$\begin{aligned} \tilde{V}(x, \mathcal{M}^1, ..., \mathcal{M}^K) &= \tilde{u}(x) + \tilde{\pi}^1(x, \mathcal{M}^1) + \textstyle\sum_{k=2}^K \tilde{\pi}^k(\mathcal{M}^{k-1}, \mathcal{M}^k) \\ &= V(x, \mathcal{M}^1, ..., \mathcal{M}^K) + f^K(\mathcal{M}^K) \end{aligned} \quad (5)$$

Furthermore, the $\tilde{\pi}^k$ functions have the same property as the π^k function: when there is no level-k choice, the individual experiences no level-k utility (condition (4)). Consequently, $\tilde{V}$ belongs to the same class of utility representations as V.

The critical point is that, if V rationalizes choices for levels 1 through K, so does $\tilde{V}$. In other words, for choices below level $K+1$, V and $\tilde{V}$ are observationally equivalent. The reason is that such choices do not involve the selection of $\mathcal{M}^K$. Even in a level-K choice, $\mathcal{M}^K$ is fixed. Consequently, adding $f^K(\mathcal{M}^K)$ to V changes none of its implications for behavior. Recalling that the construction produces an observationally equivalent utility function for *any* $\tilde{u}$, we then see that u is fundamentally unidentified. The following proposition summarizes this conclusion.

Proposition 1. *For some fixed K, consider any utility function V with component functions $\left(u, \pi^1, ..., \pi^K\right)$ from the class described by equations (3) and (4). For all functions $\tilde{u} : X \to \mathbb{R}$,*

there exists a utility function $\hat{V}$ from the same class with component functions $\left(\tilde{u}, \tilde{\pi}^1, ..., \tilde{\pi}^K\right)$ such that, for all $k \leq K$, V and $\hat{V}$ are observationally equivalent with respect to all stage-k choices.

Identification of u is of course possible if one imposes additional restrictions. One such restriction is level-K menu independence: $\pi^K(\mathcal{M}^{K-1}, \mathcal{M}^K) = 0$ for all $(\mathcal{M}^{K-1}, \mathcal{M}^K)$ where $\mathcal{M}^{K-1} \in \mathcal{M}^K$. In the preceding construction, if V is level-K menu-indendent, then $\tilde{\pi}^K(\mathcal{M}^{K-1}, \mathcal{M}^K) = -f^{K-1}(\mathcal{M}^{K-1}) + f^K(\mathcal{M}^K)$, which means $\tilde{V}$ is not. Consequently, *assuming* level-K menu independence can resolve identification. However, there is no way to *test* level-K menu independence with level-K data.

As an alternative, when the choice data pertain to levels 1 through K, one might hope to test level-k menu independence for some $k < K$. Failing to reject that hypothesis, one might then identify u from choice data pertaining to levels 1 through k. Unfortunately, the preceding construction shows that strategy cannot work. For if V satisfies level-k menu dependence, then $\tilde{\pi}^k(\mathcal{M}^{k-1}, \mathcal{M}^k) = -f^{k-1}(\mathcal{M}^{k-1}) + f^k(\mathcal{M}^k)$, which means $\tilde{V}$ does not, and we have already established that V and $\tilde{V}$ are observationally equivalent for the hypothesized data.

Most policies are not x-mandates. Instead, they induce the constraints on choice (i.e., the menu structure) each individual confronts. But it should be clear from the foregoing that the same identification problem applies.

To illustrate, suppose policy options induce decisions with non-degenerate menus for levels 1 through s, and degenerate menus for higher levels. Suppose the utility function V rationalizes choices for levels 1 through $K \geq s$. For any function f^s, define

$$\tilde{\pi}^s(\mathcal{M}^{s-1}, \mathcal{M}^s) = \pi^s(\mathcal{M}^{s-1}, \mathcal{M}^s) - f^{s-1}(\mathcal{M}^{s-1}) + f^s(\mathcal{M}^s),$$

where $f^{s-1}(\mathcal{M}^{s-1}) := f^s(\{\mathcal{M}^{s-1}\})$. Using the same construction as before, we can construct $\tilde{\pi}^k(\mathcal{M}^{k-1}, \mathcal{M}^k)$ for $k > s$. Reversing the recursive construction, we obtain $\tilde{\pi}^k(\mathcal{M}^{k-1}, \mathcal{M}^k)$ for $k < s$, and ultimately $\tilde{u}$.[4]Putting these components together, we obtain a utility function $\tilde{V}$ within the pertinent class that is related to V, as before, by equation (5). Hence, the hypothesized data cannot distinguish between them. Furthermore, when evaluating the benefits of switching from a policy that induces $\mathcal{M}^s$ to one that induces $\mathcal{N}^s$ (in each case with degenerate higher-order menus), one will conclude that the gain is greater (or less) by $f^s(\mathcal{N}^s) - f^s(\mathcal{M}^s)$ when using $\tilde{V}$ than when using V. Because we are free to select any function f^s, one can therefore say nothing about the relative merits of the two policies.

[4]Starting at $k = s - 1$ and moving downward to $k = 2$, we define $\tilde{\pi}^k(\mathcal{M}^{k-1}, \mathcal{M}^k) = \pi^k(\mathcal{M}^{k-1}, \mathcal{M}^k) - f^{k-1}(\mathcal{M}^{k-1}) + f^k(\mathcal{M}^k)$, where $f^{k-1}(\mathcal{M}^{k-1}) := f^k(\{\mathcal{M}^{k-1}\})$. Then we define $\tilde{\pi}^1(x, \mathcal{M}^1) = \pi^1(x, \mathcal{M}^1) - f^o(x) + f^1(\mathcal{M}^1)$, where $f^0(x) := f^1(\{x\})$, and $\tilde{u}(x) = u(x) + f^0(x)$.

2.1.3 Other manifestations of the Non-Comparability Problem

The Non-Comparability Problem is not limited to settings with constraint-set dependence and menu dependence. It also applies when other aspects of the decision process, such as features of choice architecture that affect the salience of information, cause the decision maker to experience welfare-relevant sensations.

The following concrete example illustrates this point. When making informed decisions about saving for retirement, a worker necessarily attends to the potential consequences of poor planning. Suppose this attention causes anxiety, which the worker alleviates by electing high contributions to retirement accounts. A planner charged with designing a social security system might look to the worker's decisions for guidance. Moreover, if the worker remains attentive to the same information (and consequently experiences the same anxiety) when the government provides for retirement, that guidance is potentially appropriate. However, it is entirely possible that the worker would prefer to live in a society with a social security system that provides for a more modest retirement while avoiding anxiety by ignoring the issue entirely. No informed choice can reveal that preference, nor (by definition) any other preference that is conditional on inattentiveness. Consequently, the Non-Comparability Problem once again arises.

2.2 Self-Reported Well-Being and the Aggregation Problem

One way to escape the Non-Comparability Problem is to jettison the choice-based welfare paradigm in favor of a competing tradition that employs self-reported measures of well-being. Such measures potentially offer a solution because they can encompass the experience of choosing as well as feelings about the outcome. However, they introduce other conceptual difficulties, including the *Aggregation Problem*, which we describe below; see also Bernheim (2016) and Bernheim and Taubinsky (2018).

To understand the nature of this problem as well its severity, it is useful to focus on the best-case scenario for SWB methods: (i) each of us has an internal cardinal "register" that measures our overall hedonic well-being, aggregating sensations associated with the various hedonic dimensions of our present circumstances (e.g., hunger, fatigue, anxiety, etc.), memories of the past, and expectations of the future; (ii) we are capable of reading this register when asked; (iii) we understand certain natural-language questions (e.g., concerning "life satisfaction") as requests to do so; and (iv) we always report these readings honestly. A conceptual problem nevertheless arises because the register may well yield different readings at different points in time and in different states of nature. For example, even if someone experiences emotions associated with memories and expectations, they almost certainly enjoy greater overall hedonic well-being when passing through more pleasurable phases of life. Likewise, to the extent the register reflects what has happened rather than what could have happened, it presumably yields higher readings when advantageous states of nature materialize. Consequently, even if we assume the existence of an internal, readable, cardinal register measuring overall hedonic well-being, we cannot avoid the need for additional aggregation principles, nor the implication that the aggregate well-being index we seek is not something people experience

hedonically.

The need for such principles becomes even more acute as one steps away from the best-case scenario. For example, if (as seems likely) people only have internal registers for disaggregated feelings such as hunger, fatigue, and anxiety, then even a measure of momentary (rather than lifetime) well-being would involve aggregation over multiple hedonic sensations, rather than a single "reading" of an aggregative register.

Accordingly, if we interpret SWB questions as eliciting "register readings," then it is in inappropriate to treat the responses as pertaining to the level of aggregation that normative economic analysis requires. To justify using those responses as welfare measures, one must instead interpret the questions as asking respondents to consider readings of multiple hedonic registers—i.e., various types of sensations for current, past and potential future experiences, as well as for counterfactual events that might have materialized—and to construct an overall assessment of well-being based on their own judgment concerning appropriate aggregation principles. Unfortunately, this interpretation of SWB is problematic.

To understand the conceptual difficulty, consider a simple example: the reading on Norman's hunger register is 3, and the reading on his fatigue register is 6 (where higher numbers indicate better hedonic states). For simplicity, assume these are the only sensations that contribute to well-being, and that no internal hedonic register aggregates them. If we ask Norman to report the equally weighted average of these readings, his answer will be 4.5. If we ask him to place three times as much weight on hunger, he will report 4, and if we ask him to place three times as much weight on fatigue, he will report 5. In other words, his answers to these questions reflect the weights we ask him to use when aggregating, rather than a defensible normative principle.

Now imagine that Norman reports 4 when we ask him about his overall happiness, accounting for both hunger and fatigue. From that response, we learn that, based on his linguistic associations (for example, between "happiness" and satiation), he interprets the question as directing him to place three times as much weight on hunger as on fatigue. Similarly, if he reports 5 when we ask him about his overall satisfaction, we learn that, based on other linguistic associations (for example, between "satisfaction" and energy), he interprets that question as directing him to place three times as much weight on fatigue as on hunger. Thus, our use of natural language does not change the fact that our questions implicitly instruct Norman to use one set of weights rather than another. Rather, our instructions concerning the weights simply become less clear. It follows that the principle of aggregation for SWB is linguistic (and therefore arbitrary), rather than normative.

To defend SWB against this critique, one would have to argue that a particular combination of words and phrases unambiguously evokes an appropriate normative ideal. But if (as we have posited) there is no internal register measuring a hedonic sensation at the same level of aggregation as the SWB response, what is the normative ideal to which the SWB method conceptually aspires? What criterion would one use to evaluate, objectively, whether a particular set of words and phrases implicitly

instructs respondents to apply normatively appropriate weights to distinct hedonic experiences?

While choice-based welfare methods also rely on people's judgments concerning aggregation over hedonic sensations, they nevertheless avoid the arbitrariness of SWB. When conducting normative analysis, deference to any aggregate judgment is warranted only if the purposes of the analysis and the judgment are conformable. Economists typically see normative analysis as a tool for guiding policy makers when they select among alternatives, under the assumption that the objective is to promote the well-being of those affected by the selection. Significantly, when people make choices for themselves, they aggregate over the many dimensions of their experience for precisely the same purpose—that is, to make a selection that promotes their well-being. Thus, the purposes of constructing judgments for normative analysis on the one hand, and for making choices on the other, are conformable. When advising policy makers on the selection of an alternative that affects some individual, we defer to the individual's choices because they reveal the alternatives that, in the individual's judgment, would provide them with the greatest overall benefit if selected.[5]

In contrast, when we ask a question about overall well-being, truthful respondents construct their judgments for the purpose of providing answers, not for making selections. Moreover, those judgments conform to their operational interpretations of words and phrases, rather than to a planner's objectives.[6] In short, with respect to aggregation, choice invokes a defensible normative ideal, but SWB does not.

2.3 The Proposed Solution

Our proposed solution to the conceptual problems described in the preceding subsections involves a hybrid approach. We use choice-based methods to overcome the Aggregation Problem that afflicts the SWB approach, while exploiting SWB methods to overcome the Non-Comparability Problem that afflicts choice-based approaches.

We proceed from the assumption that people care only about their own subjective experiences, or *mental states*.[7] To be clear, this perspective does not presuppose selfishness; it subsumes the possibility that mental states depend on outcomes for others.

Formally, let z denote a vector (or bundle) of mental states, and let Z be the set of feasible mental-

[5]Choice provides a normatively compelling aggregation principle within the philosophical tradition of Desire-Fulfillment Theory, which holds that well-being consists in having one's desires satisfied (e.g., Parfit, 1984; Heathwood, 2016). In effect, choice is the operational expression of desire.

[6]In principle, one could attempt to implement Desire-Fulfillment Theory by eliciting desires through SWB-style questions, e.g., "To what extent does this outcome satisfy your desires?" But the word "desire" is also open to a variety of colloquial interpretation, for example concerning the relative weighting of "needs" and "wants." Merely using a term of art cannot resolve the Aggregation Problem.

[7]Thus, we follow the philosophical tradition of mental statism (Mill, 2012). This conception of welfare excludes the possibility that an individual's well-being depends on considerations about which they are and always will be entirely unaware. While that restriction is plausibly contentious (see, e.g., Nozick, 1974, who criticizes utilitarianism by describing a thought experiment involving an "Experience Machine"), no definition of well-being avoids controversy entirely.

state bundles. We assume that an individual's preferences correspond to a well-behaved binary relation, $\succsim$, over the elements of Z. To allow for constraint-set dependence (Section 2.1.1), menu dependence (Section 2.2.2), and other considerations that can give rise to the Non-Comparability Problem (Section 2.2.3), we assume that we can write the mental states induced by choices in a K-stage decision problem as $z = \zeta(x, \mathcal{M}^1, ..., \mathcal{M}^K, d)$, where d subsumes details of choice architecture, such as features that guide the decision maker's attention. For a given choice problem D (consisting of a choice set X, a decision tree, and details d), we can think of an individual as selecting an anticipated mental state from the set $\{\zeta(x, \mathcal{M}^1, ..., \mathcal{M}^K, d)\}_{(x,M^1,...,\mathcal{M}^K)\in\mathbb{M}(D)}$, where $\mathbb{M}(D)$ is the set of sequences $(x, \mathcal{M}^1, ..., \mathcal{M}^K)$ that can arise in D.

Now imagine the planner's task is to place an individual in one of two choice situations, A or B. Because of the Non-Comparability Problem, a meta-choice between these two decision problems does not tell us which would make the individual better off if they were simply assigned to it. However, if we knew the preference relation $\succsim$ over mental state bundles, and we if we could observe the mental states both choice situations induce (z_A and z_B), we could resolve this ambiguity. In particular, we could determine which choice situation makes the individual better off by asking whether $z_A \succeq z_B$ or $z_B \succeq z_A$.

The key challenge, then, is to recover preferences over mental state bundles. We propose accomplishing this task in two steps, building on both choice-based and SWB methods. The first step is to elicit the dimensions of anticipated sensations for various tuples $(x, \mathcal{M}^1, ..., \mathcal{M}^K, d)$. Each such elicitation provides a proxy for the mental state bundle $\zeta(x, \mathcal{M}^1, ..., \mathcal{M}^K, d)$. The second step is to observe choices. Choosing $(x_1, \mathcal{M}_1^1, ..., \mathcal{M}_1^{K-1}, \mathcal{M}^K, d)$ over $(x_2, \mathcal{M}_2^1, ..., \mathcal{M}_2^{K-1}, \mathcal{M}^K, d)$ implies that $z_1 = \zeta(x_1, \mathcal{M}_1^1, ..., \mathcal{M}_1^{K-1}, \mathcal{M}^K, d) \succeq z_2 = \zeta(x_2, \mathcal{M}_2^1, ..., \mathcal{M}_2^{K-1}, \mathcal{M}^K, d)$.[8] Repeating this two-step process multiple times generates a dataset consisting of opportunity sets (available mental state bundles) and choices from those sets. One can then use standard revealed preference techniques to infer the preferences $\succeq$. Our experimental design, summarized in Section 3, illustrates how one might elicit mental state bundles and associated choices. Our empirical framework, summarized in Section 4, exemplifies the use of the resulting data to estimate $\succeq$.

This approach avoids the Non-Comparability Problem by borrowing from the SWB paradigm: it evaluates welfare based on self-reports that encompass subjective reactions not only to the outcome, but also to the experience of choosing. At the same time, it avoids the Aggregation Problem by borrowing from the choice-based paradigm: it aggregates over the dimensions of subjective experience based on choice.

We close this section by addressing two potential concerns, one conceptual, the other practical. The conceptual concern is that, if preferences over objects are menu-dependent, then perhaps preferences over mental states are also menu-dependent. Fortunately, we can accommodate this possibility without reintroducing the Non-Comparability Problem. Within a mental-statist framework, the

[8]Note that, within any K-stage decision problem, $\mathcal{M}^K$ and d are fixed.

menu of mental state bundles, ξ, can matter to the individual only to the extent it affects mental states. In other words, ξ enters as an additional argument of the mental state mapping ζ. The mental-state menu induced by any decision problem D is then a fixed point of the mapping

$$\xi\left(D,\xi'\right) \equiv \{\zeta(x,\mathcal{M}^1,...,\mathcal{M}^K,d,\xi')\}_{(x,M^1,...,\mathcal{M}^K)\in\mathbb{M}(D)}$$

With this formulation,[9] each z already incorporates any preference-relevant aspects of menu-dependent reactions. Consequently, it is still appropriate to define preferences over z, rather than over pairs (z,ξ), and to recover those preferences from choices in the manner we propose.

The practical concern is that it may prove difficult to define and measure distinct mental states. Our solution is to measure proxies for various composites of mental states, which we call *Categorical Subjective Assessments* (CSAs). These CSAs include aggregates such as happiness and satisfaction, as well as narrower concepts such as pride and guilt. As long as (i) each CSA is a stable composite of underlying mental states (i.e., there is some function relating each CSA to those states), and (ii) the CSAs collectively span the pertinent mental states, we can think of people as having stable preferences over CSAs, and proceed as if the CSAs are measures of z. Section 5.5 proposes and implements several tests of the assumption that the spanning condition is satisfied in our application.

3 Experimental Design

The experiment consisted of three modules: (1) Dictator Games (DGs), (2) Computer Choices (CCs), and (3) Opt-out Games (OOs). In the DGs, participants were asked to choose one of two allocations of money between themselves and a randomly assigned partner. In the CCs, a computer exogenously determined the allocations. In the OOs, participants chose whether to participate in a DG, in which case they made a subsequent choice between the DG options, or to "quietly" opt-out. Following Dana et al. (2006a) and Lazear et al. (2012), the decision to opt out guaranteed that the other participant, who would otherwise have been on the receiving end of the DG, did not learn about that foregone possibility. Immediately following a choice by either a participant or the computer, participants were asked to report seven CSAs: guilt, pride, financial satisfaction, a sense of fairness, a sense of unfairness, happiness, and overall satisfaction. Appendix XXX contains the full study instructions.

Dictator Games. Participants selected allocations of payouts between themselves and a randomly assigned partner in seven DGs. Each DG appeared on a separate screen. Participants were told that their partners would learn both the choice set and their decision if the allocation was randomly selected for implementation. The choice sets for the seven DGs, which appeared in random order (rather than the order shown below), were as follows:

[9]The formulation assumes that such fixed points exist.

DG 1: (You: \$2.00; Partner: \$0.50) or (You: \$4.00; Partner: \$0.00)

DG 2: (You: \$2.00; Partner: \$1.00) or (You: \$4.00; Partner: \$0.00)

DG 3: (You: \$2.00; Partner: \$1.50) or (You: \$4.00; Partner: \$0.00)

DG 4: (You: \$2.00; Partner: \$2.00) or (You: \$4.00; Partner: \$0.00)

DG 5: (You: \$2.00; Partner: \$2.00) or (You: \$3.50; Partner: \$0.00)

DG 6: (You: \$2.00; Partner: \$2.00) or (You: \$3.00; Partner: \$0.00)

DG 7: (You: \$2.00; Partner: \$2.00) or (You: \$2.50; Partner: \$0.00)

In this manuscript (but not in the experiment), we refer to the option on the left in each row of the preceding list as "more equitable," and to the one on the right as "less equitable." In DG 1, switching from the less equitable to more equitable option increases the partner's payoff by only \$0.50 at a cost of \$2.00 to the decision maker; hence, we expected most participants to choose the less equitable option. Moving from DG 1 to DG 4, behaving more equitably becomes steadily more attractive because the partner's gain increases. Moving from DG 4 to DG 7, behaving more equitably becomes steadily more attractive because the cost to the decision maker decreases. In DG 7, switching from the less equitable to more equitable option increases the partner's payoff by \$2.00 at a cost of only \$0.50 to the decision maker; hence, we expected most participants to choose the more generous option. The differences in the attractiveness of the more versus less equitable option across the seven DGs thus created significant variation in participants' choices and CSAs.

Immediately after making a decision, participants reported the seven CSAs for both of the options available in that DG. We elicited the CSAs for each DG on a separate screen, which also displayed the two options and the participant's selection. We discuss this elicitation in more detail toward the end of this section.

Computer Choices. Participants were presented with eight different allocations chosen by a computer. These computer-determined allocations all appeared on separate screens, in random order. If one of these allocations was randomly selected for implementation, the partner learned that the participant was not responsible for the outcome.

We explored two distinct types of CC allocations. For the *main* variant, we left the unchosen options unspecified; for the *alternative* variant, we described the computer's choice set. Depending on circumstances, either version could correspond to a setting in which a planner makes a choice on behalf of some individual. We examine both to determine whether the presence of explicit alternatives affects participants' sense of well-being even when they do not control the selection.

For the CC variants that leave unchosen alternatives unspecified, the allocations were the eight possible outcomes associated with DG 1 through DG 7:

CC 1: (You: \$2.00; Partner: \$0.50)

CC 2: (You: \$2.00; Partner: \$1.00)

CC 3: (You: \$2.00; Partner: \$1.50)

CC 4: (You: \$2.00; Partner: \$2.00)

CC 5: (You: \$4.00; Partner: \$0.00)

CC 6: (You: \$3.50; Partner: \$0.00)

CC 7: (You: \$3.00; Partner: \$0.00)

CC 8: (You: \$2.50; Partner: \$0.00)

For the CC variants that specify the alternatives, the choice sets were the same as for the DGs.

Immediately after learning a CC decision, the participant reported the seven CSAs for the chosen allocation on a separate screen. For the CC variants that specify the alternatives, the screen displayed the computer's choice set.

Opt-Out Games. Following Dana et al. (2006a), Broberg et al. (2007), and Lazear et al. (2012), we designed games in which participants could "opt-out" of DGs. In effect, these games required our participants to make *meta-choice*s (choices over choice sets). Prior work has used meta-choices to evaluate the utility people derive from making decisions such as choosing an allocation in the DG. Such evaluations invoke two implicit assumptions: (i) the utility from opting out depends only on the opt-out payment and not on the decision the chooser thereby avoids, and (ii) the presence of the opt-out option does not alter the utility a chooser derives from either of the two options in the DG.

Our study presented participants with four OO decisions, shown on different screens. For each participant, all such decisions involved the same DG (DG 3, DG 4, or DG 5, assigned at random); only the opt-out payment differed The instructions told the participants that, if they opted out, "*the individual who you would have been paired with [for the DG] will not be told anything about this allocation task and will simply be paid the HIT fee without a bonus.*" Participants also learned that, if they opted in, then "*the participant you are paired with [for the DG] will be informed of the decision you made [in the DG].*" The four OO scenarios, which appeared in random order, were as follows:

OO 1: Opt into the DG or Opt out for \$5.00

OO 2: Opt into the DG or Opt out for \$4.00

OO 3: Opt into the DG or Opt out for \$3.50

OO 4: Opt into the DG or Opt out for $3.00

Each screen displayed the opt-in options first followed by the opt-out option. Immediately after making each decision, participants reported associated CSAs. We first elicited CSAs for the participant's chosen option. Then we elicited CSAs for the alternatives, starting with the more equitable option in the DG subgame, then the less equitable option, then the opt-out option (in each case, if unchosen). Each set of CSA elicitations appeared on a separate screen, which showed the alternatives and the participant's choice.

To ensure that potential partners would in fact be ignorant of decisions to opt out, only (roughly) half of the participants, selected at random, made OO decisions; the other half served as the OO partners. The instructions informed the latter group only that they may "*receive a bonus that is determined by a choice of another participant or randomly by the computer*," but by chance "*may not be eligible to receive a bonus in one of these two parts*." We provide additional details on randomization into treatments below.

CSA Elicitations We elicited the following CSAs: (1) guilt, (2) pride, (3) financial satisfaction, (4) a sense of fairness, (5) a sense of unfairness, (6) happiness, and (7) overall satisfaction with the study experience. We arrived at this list by adapting the social values orientation framework of Van Dijk (2015), which draws on a rich body of work concerning the psychology of prosocial behavior (e.g., Ketelaar and Tung Au, 2003; Tracy and Robins, 2004; Van Lange et al., 2007; Nelissen et al., 2007; Batson, 2011; Wubben et al., 2012). We included two overall assessments, happiness and satisfaction, to minimize the risk that the narrower CSAs do not encompass some important mental state. These measures also allowed us to assess the importance of the Aggregation Problem. Appendix B elaborates on the pertinent psychology literature and our reasons for choosing the CSAs used in our study.

We elicited the seven CSAs on five-point Likert scales. We randomized their order across participants but, to streamline participation, always presented them in the same order for a given participant. The CSA questions encompassed mental states both during and after the experiment. For DGs and OOs, we asked: *"We'd now like to know how you feel [about your chosen option/ if you had chosen the other option]... Considering both how you feel now and how you might feel in the future when this study is over, please indicate to what extent [your decision/this decision] led, or will lead, you to experience the following, on a scale of 1 (not at all) to 5 (very much)*." We used the phrase "your chosen option" and "your decision" for assessments pertaining to the chosen option, and the phrase "if you had chosen the other option" and "this decision" for those pertaining to the unchosen options. For CCs, we posed a slightly altered version of this question: "*Considering both how you feel now and how you might feel in the future when this study is over, please indicate to what extent the randomly determined outcome led, or will lead, you to experience the following, on a scale of 1 (not at all) to 5 (very much)*." To address the possibility that the Aggregation Problem might infect these pan-temporal assessments, we also conducted robustness analysis based on elicitations

that separately asked participants to report their present CSAs and predict their future CSAs; see below.

We also included an attention check that resembled the CSA elicitations, but that included the following preamble: "*This next question is not a question that needs to be answered. Rather, the goal of this question is to check to make sure that you are reading everything. To indicate this, please click the continue button without filling in any of the options below. You must click the continue button without filling anything below to have your HIT approved.*" We classified those who ignored the preamble as inattentive. The attention check appeared after the DG, CC, and OO modules, but before a closing battery of demographic questions.

Variations on CSA elicitations. To evaluate the robustness of our methods, we explored two alternative approaches to CSA measurement; see Section 5.6 for analyses of the resulting data.

For the first alternative approach, we elicited CSAs for the DG and OO games *before* participants made their decisions, rather than after. The motivation for this alternative is that, with our primary method, participants might skew their reported CSAs to rationalize the selections to which they have already committed. Elicitation questions took the following form for each of the two possible options: *"Considering both how you feel now and how you might feel in the future after this study is over, please indicate to what extent choosing the below option (in dark blue) would lead you to experience the following, on a scale of 1 (not at all) to 5 (very much)."*

For the second alternative approach, we separately elicited current CSAs, which reflect how the participant feels about an outcome immediately, and predicted future CSAs, which reflect how the participant expects to feel about that outcome in the future. The motivation for this alternative is that our primary method elicited a temporal aggregate, and was therefore potentially susceptible to the Aggregation Problem. Elicitation questions took the following form: *"(i) to what extent this decision [led/would lead] you to experience the following [seven CSAs] now, and (ii) how much you think this decision [will lead/would lead] you to experience the following [seven CSAs] in the future. Relative to how you would feel now, your experiences might be more intense if you keep thinking about them, or less intense if you quickly forget."* Due to concerns about the study's length, we omitted the OO module when using this alternative approach.

Procedures and incentives Upon joining the study, participants viewed a one-page consent form. We informed them that *"all information provided in this study is truthful and accurate. The decisions you make in this study are real and you will be paid in accordance with the instructions provided in the following pages.*"

The instructions explained that participants would receive $2 for taking part in the study, and possibly an additional bonus. Participants received reminders about the range of possible bonuses at the start of each of the three modules. For the DG and CC modules, the instructions described the study payments as follows: "*In this part, bonuses will range from $2 to $4 for you and $0 to*

$2 for your partner. Recall that overall, participants in this study will receive bonuses ranging from $0 to $5." For the OO module, we modified this explanation as follows: "*In this part, bonuses will range from $2 to $5 for you and $0 to $2 for the other participant. Recall that overall, participants in this study will receive bonuses ranging from $0 to $5.*" The purpose of these reminders was to deter participants from renormalizing the CSA Likert scales as they proceeded through the DG, CC, and OO modules.

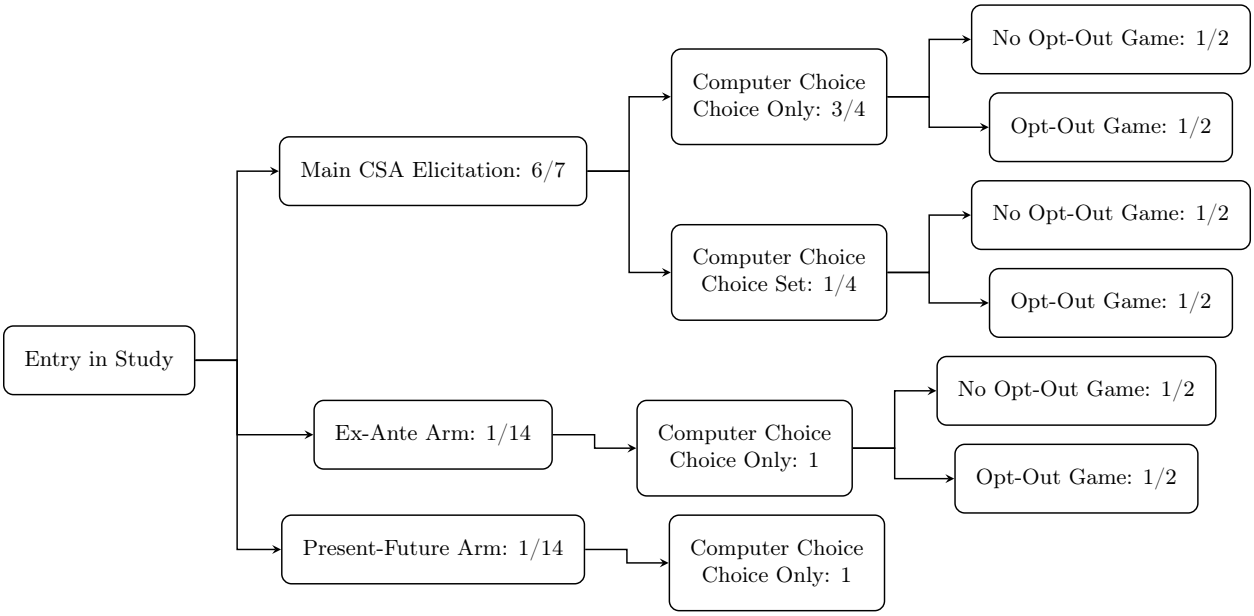
We used our primary CSA elicitation format for 6/7-ths of the participants. Of those, three-quarters viewed CC allocations with unspecified alternatives, while the remaining one-quarter viewed CC allocations with explicitly specified alternatives. We assigned the majority of participants to the first group because we use those data to determine the mapping from CSAs to equivalent variations (a money metric), as detailed in Section 4.

In each of the DG, CC, and OO modules, a participant could occupy either of two roles, Decider or Receiver. We use this terminology loosely for the CC module, where the Decider is the participant who received the (weakly) larger amount in each option and reports their CSAs, while the Receiver is the one who received the (weakly) smaller amount in each option.To conserve on costs, participants played both roles for the DG and CC modules. However, to ensure that Receivers in the OO games would infer nothing if Deciders opted out, half of the participants did not take on the role of OO Decider (or learn anything about that module). Those participants were matched with participants who viewed all three modules (DG, CC, and OO). Thus, approximately half of the participants served as Deciders in DGs, CCs, and OOs, and as Receivers in DGs and CCs, while the other half served as Deciders in DGs or CCs, or as Receivers in DGs, CCs, and OOs. For each participant, each of the five modules had a 20% chance of being the one that counted, and within each module each of the games' outcomes had an equal probability of being the one that is selected for the actual payout. This procedure ensured that all choices by participants in the Decider role were incentive compatible.

We communicated this randomization process to participants as follows. At start of each module in which the participant served as a Decider, we explained that *"there is a 20% chance that one of the scenarios from this part will be selected to determine the bonuses of you and your [randomly assigned] partner [in this study].*" At the end of the survey instrument, we informed participants who served as Deciders in all three modules (DGs, CCs, and OOs), and who were thus ineligible to be Receivers in the OOs, that "*your bonus will be determined by a choice of another participant or randomly by the computer*" with 40% probability. We informed participants who served as Deciders only in the DGs and CCs, and who were thus eligible to be Receivers in the OOs, that with 60% probability, "*you may receive a bonus that is determined by a choice of another participant or randomly by the computer. There is also a chance that you may not be eligible to receive a bonus.*"

We used the alternative CSA elicitation formats for the remaining 1/7-th of participants. Half reported CSAs prior to each decision in the DG and OO modules; the other half separately reported

Figure 1: Experimental mapping



Note: This figure reports the probabilities of being randomized into different treatment groups in the study. The first fork shows that 6/7-ths of participants were shown the main CSA elicitation format, while the remaining 1/7-th of participants were presented with a different format for CSA elicitations. In all CSA elicitation formats except in the present-future arm, participants had a 50% chance of being randomized into the OO game in their respective formats.

present and future CSAs (and did not view the OO module to keep the study length manageable). We assigned all of these participants to the main CC module, which leaves alternatives unspecified. Splitting them between the two versions of the CC module would have compromised the statistical precision of comparisons with the group using the main CSA elicitation format.

Figure 1 summarizes the randomization into treatment arms.

Participants. We recruited participants for the study though Amazon Mechanical Turk (MTurk) from late August 2021 through early September 2021. In total, 2,800 individuals completed the study. We dropped 60 from the analysis for the following reasons: 24 failed the attention check; 5 completed the study in less than 4 minutes; 6 inputted the incorrect completion code; 5 re-entered the study; and 20 encountered a technical error in submitting responses. Appendix Table A1 summarizes the sample's demographic composition: 53% identified as female, 83% were between the ages of 25 and 60, 69% stated that they held a Bachelor's or advanced degree, and 56% reported household income between $20k and $80k.

Unless otherwise stated, all of the following analyses pertain to the participants who used the main CSA elicitation format (6/7-ths of the total sample).

4 Empirical Framework for Analysis

4.1 Preferences and Choice

To implement the conceptual framework of Section 2, we require an empirical model that links CSAs to choice. Each Decider i obtains a vector X_{ijc} of CSAs from outcome j in environment c. The index c encodes the type of environment (DG, a CC, or OO), as well as the specific allocation problem the participants face. In the DG environments, there are two alternatives j: the more equitable and less equitable options. By design, the utility derived from these alternatives differs across the DG scenarios because the payouts associated with the more equitable and less equitable options vary. In the OO environments, there are three alternatives j: opting-out and the two DG alternatives.

We model decisions using standard discrete choice techniques. A Decider chooses alternative j if it maximizes

$$U = v(X_{ijc}) + \varepsilon_{ijc}. \tag{6}$$

The term $v(X_{ijc})$ is the deterministic component of utility. The term ε_{ijc} is an idiosyncratic realization that is distributed independently and identically for each pair (j, c). This term encompasses the determinants of utility not captured by our elicited CSAs (i.e., other mental states not spanned by the CSAs), and/or decision-making "noise" arising from changing beliefs about the value of alternatives (Block and Marschak, 1960; Woodford, 2019), or from other sources of "trembles" that often surface in experiments (McKelvey and Palfrey, 1995). We assume that the variation in utility across the scenarios in our experiment is small enough to justify using a first-order approximation. In other words, we take v to be locally linear and additive in the CSAs, so that $v(X_{ijc}) = X_{ijc}\beta$, where β is a vector of coefficients. We also assume that ε_{ijc} has a type I extreme value distribution (with scale parameter 1), and is drawn independently for each tuple of (i, j, c). Thus, the probability of choosing alternative j in environment c with alternatives $l = 1, \ldots J$ is

$$P_{jc} = \frac{e^{v(X_{ijc})}}{\sum_l e^{v(X_{ijc})}} \tag{7}$$

4.2 Translating Utility to Money Metrics

Our approach allows us to conduct welfare analysis using money-metric measures of consumer surplus. Specifically, we map variation in $v(X_{ijc})$ to dollar-denominated equivalent variations. This method generalizes earlier approaches that use only a single measure of happiness or satisfaction (e.g., Clark and Oswald, 2002; Finkelstein et al., 2013; see Benjamin et al., 2023, for a review). Translating utility into dollars makes our welfare measures economically interpretable. It also shows that our approach is compatible with the standard welfare paradigm, which typically expresses efficiency losses using money-metric scales. Deriving such measures involves two steps.

The first step is to specify a "benchmark" domain within which the individual's payoff varies while

other determinants of utility remain fixed. For our calculations, this domain consists of scenarios in which the individual receives an exogenously assigned payment, others receive nothing, and no alternatives are mentioned. These are the scenarios our participants encounter in the main CC module. In principle, one could use other benchmark domains, such as the scenarios encountered in our alternative CC module, wherein the individual receives an exogenously assigned payment selected from a specified menu, and others receive nothing. That procedure would also be economically interpretable, but the scale would be slightly different.

The second step is to estimate the relationship between utility and money within the benchmark domain. Formally, let u_{ijc} denote the money-metric measure of the deterministic component of utility person i obtains from alternative j in environment c (relative to being assigned $(0, 0)$ in the benchmark domain). Let m denote the marginal utility of money, which we assume does not vary within the relatively small range of payouts in our experiment. Let option j_y for environment c_0 denote the CC payment allocation $(y, 0)$ with no alternatives specified. Then by definition,

$$v(X_{ijc}) - v(X_{ij_0c_0}) = mu_{ijc}.$$

In the benchmark domain, where $u_{ij_yc_0} = y$, we therefore have

$$v(X_{ij_yc_0}) = my + v(X_{ij_0c_0}).$$

This approximation allows us to recover the marginal utility of money, m, by estimating a linear regression of $\widehat{v}(X_{ijc})$ (constructed from estimates of the choice model described in the preceding subsection) on the participant's payoff y using observations from the main CC module—specifically, all CC allocations for which the participant's partner receives no payout: $\{(4, 0), (3.5, 0), (3, 0), (2.5, 0)\}$. That is, we estimate the linear model $\mathbb{E}_i[\hat{v}(X_{ij_yc_0})] = my + \alpha$, where $\alpha = \mathbb{E}_i[v(X_{ij_0c_0})]$.

Equipped with an estimate of m, we can then write the average deterministic component of dollar-denominated utility for each alternative j in environment c as

$$\bar{u}_{jc} = \mathbb{E}_i[v(X_{ijc})/m] - \mathbb{E}_i[v(X_{ij_0c_0})/m], \tag{8}$$

where $\mathbb{E}_i$ denotes the expectation over individuals i. This quantity is interpretable as the equivalent variation for the deterministic portion of preferences associated with replacing the benchmark outcome (j_0, c_0) with (j, c).[10]

[10]To clarify, $\bar{u}_{jc}$ does not correspond to a measure of average equivalent variation in cases where we average over *chosen* alternatives, *and* the idiosyncratic term ε_{ijc} in (6) partially reflects actual preferences (rather than just noise). In such cases, one must account for the fact that the mean of ε_{ijc} conditional on any given choice is non-zero. In the special case where ε_{ijc} reflects only preferences and we consider utility of chosen alternatives, there are standard formulas for computing equivalent and compensation variation metrics for logit models (Dagsvik and Karlstrom, 2005) . We report results for $\bar{u}_{jc}$ rather than the logit formulas for equivalent variation because the assumption that ε_{ijc} reflects only preferences is probably unrealistic. Moreover, because ε_{ijc} is mean zero, the logit surplus formula

Technically, we cannot directly implement the preceding formula because the allocation $(0, 0)$ does not appear in the CC module. However, because we have assumed that utility is (approximately) linear in money within the benchmark domain, we can equivalently compute $\bar{u}_{jc}$ as

$$\bar{u}_{jc} = \mathbb{E}_i[v(X_{ijc})/m] - \mathbb{E}_i[v(X_{ij_4c_0})/m] + 4, \tag{9}$$

where (j_4, c_0) corresponds to the allocation $(4, 0)$ in the main CC module.[11]

In practice the marginal utility of money may vary across individuals: $v(X_{ijc}) = m_i u_{ijc}$, where m_i is participant i's marginal utility. In Appendix G, we find that m_i has a mean of $\bar{m} = 0.46$ and a variance of 0.09. Under the assumption that $m_i \perp u_{ijc}$, we have $v(X_{ijc}) - v(X_{ij_0c_0}) = \bar{m}u_{ijc}$, and thus

$$\bar{u}_{jc} = \mathbb{E}_i[v(X_{ijc})/\bar{m}] - \mathbb{E}_i[v(X_{ij_4c_0})/\bar{m}] + 4. \tag{10}$$

In other words, as long as the marginal utility of money is unrelated to variations in money-metric utility, our interpretation of $\bar{u}_{jc}$ is unchanged.[12] When the orthogonality assumption does not hold, $\bar{u}_{jc}$ is still interpretable as the average utility incremental, rescaled by the average marginal utility of money.

4.3 A Summary of Assumptions and Potential Confounds

In addition to employing the linear approximations mentioned above, our empirical model and approach to quantifying welfare require three assumptions. The first is the independence assumption $\varepsilon_{ijc} \perp X_{ijc}$. Violations of this assumption would take one of two forms. First, ε_{ijc} might subsume random "trembles," and participants might report CSAs to rationalize the resulting choices. Second, our CSAs might not span mental states that are correlated with those they do span. We provide evidence against both of these potential violations in Sections 5.5 and 5.6. The validity of our welfare estimates does not require the estimated relationship to satisfy any additional notion of "causality."[13]

The second assumption is that v is stable across the domains to which we apply our methods. In Section 5.4, we provide a test of the joint hypothesis that v is stable and correctly specified.

The third assumption is that measurement error in CSAs does not bias our estimate of β. Gillen et al. (2019) show how noisy measurement can bias coefficient estimates either upward or downward. We provide evidence against such bias in Section 5.3.

---

corresponds to $\bar{u}_{jc}$ in cases where $\bar{u}_{jc}$ is a full-sample mean for a fixed allocation.

[11] Under the linearity assumption, $\mathbb{E}_i[v(X_{ij_4c_0})/m] - \mathbb{E}_i[v(X_{ij_0c_0})/m] = 4$, so equations (8) and (9) are equivalent.

[12] Note that a linear regression of $\widehat{v}(X_{ijc})$ on the participant's payoff y in the main CC module correctly recovers the average marginal utility of money $\bar{m}$ in the case of heterogeneity in m_i because the distribution of payoffs y appearing in the regression is the same for every participant.

[13] For example, our approach remains valid if there are other CSAs that covary with choice, but that are fully spanned by the CSAs that we elicit. What matters is that we correctly predict the deterministic component of people's utility v_{ijc}, and that variation in this component is independent of ε_{ijc}.

5 The Relationship Between Choices and CSAs

In this section, we study the positive relationship between choices and CSAs in the Dictator and Opt-Out games. We begin by describing patterns of choices across variants of the games, and how the CSAs vary over the available alternatives in those environments. Then we estimate our empirical model linking choices to CSAs. Finally, we provide evidence for two key premises of our conceptual framework: that the estimated model is stable across classes of environments (DGs versus OOs), and that our CSAs adequately span the pertinent space of mental states.

5.1 Choices in DGs and OOs

Figure 2 summarizes Deciders' choices in the DG and OO games, and explores their relationships across these environments.

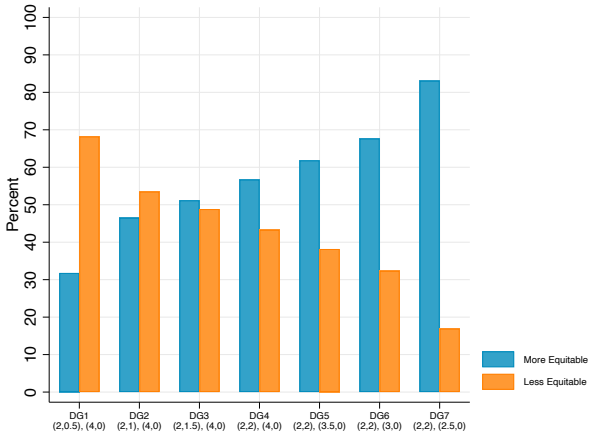
Panel (a) shows that the fraction of Deciders choosing the more equitable allocation varies considerably across the DGs, from 32 percent in DG 1, where it is least attractive, to 83 percent in DG 7, where it is most attractive. As expected, average generosity rises monotonically across games as the Receiver's benefit rises (DG 1 through DG 4) and as the Decider's cost shrinks (DG 4 through DG 7).

Panel (b) of Figure 2 shows that choice also varies significantly with the size of the opt-out payment in the OOs. The frequencies displayed in this panel reflect averages across the three opt-in DG subgames. Appendix Figure A1 provides a complete breakdown by both opt-out option and subgame. In OO 1 and OO 2, where the Decider receives a (weakly) larger payment by opting out than by choosing the less equitable option in the opt-in subgame, Deciders opt out well over half the time, and almost never opt in to choose the less equitable DG option. However, even when opting out is less financially advantageous than entering and choosing the less equitable option, as in OO 3 and OO 4, many participants still choose the opt-out option. These patterns of choice replicate, at larger scale, the results of Lazear et al. (2012), which those authors have interpreted as implying that people often act equitably in DGs to avoid the appearance of selfishness. As expected, the fraction opting out rises monotonically with the opt-out payment.

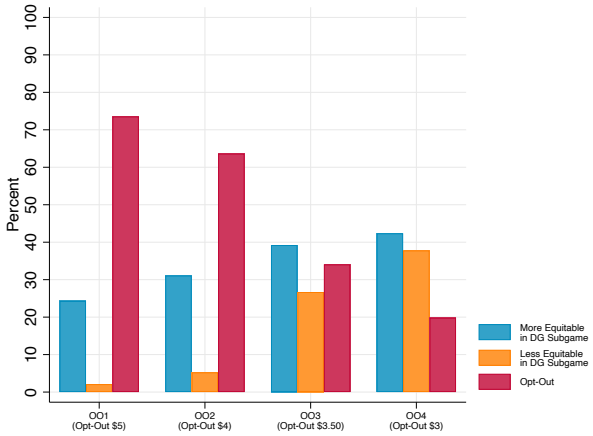
Panel (c) of Figure 2 exhibits intuitive relationships between choices in the DGs and OOs. Participants who behaved more equitably in a given DG are more likely to choose the more equitable alternative in the corresponding OOs. These participants either opt out or choose the more equitable allocation; they rarely opt in and choose the less equitable allocation. Participants who behaved less equitably in a given DG are more likely to maximize their payments in the OOs. They are likely to opt out when the associated payment is \$4 or \$5, and thus (weakly) greater than the highest opt-in payment. Conversely, they are likely to opt-in and choose the less equitable option when that strategy yields higher payouts than opting out. These participants almost never opt in and choose the more equitable option.

Figure 2: Distribution of choices in the dictator and opt-out games

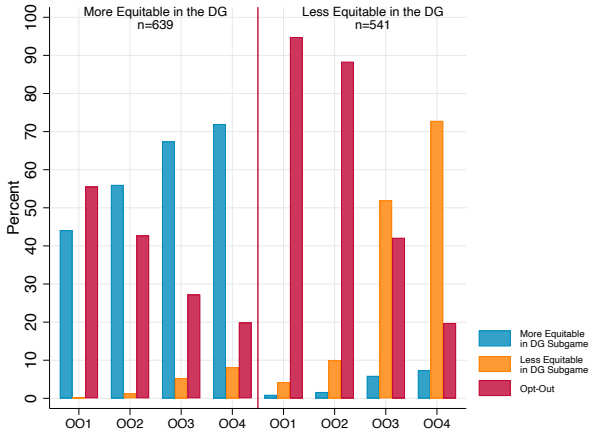
(a) Choice in dictator games



(b) Choice in opt-out games



(c) Choice in opt-out games, by choice in dictator games



Note: This figure reports the distribution of choices made in the DG and OO games for the main CSA elicitation format. Panel (a) reports the fraction of participants who chose the more equitable or less equitable option in each choice set in the DG. Panel (b) reports the fraction of participants who chose to (1) opt-out, (2) opt-in and choose the more equitable option, or (3) opt-in and choose the less equitable option. Panel (c) reports a similar plot to Panel (b), but split by whether or not the participant chose more equitably or less equitably in the DG choice set corresponding to the DG subgame in the OO game. The bars to the left of the red line report the distribution of OO choices made by participants who chose the more equitable option in the DG choice set corresponding to the DG subgame they viewed in the OO. The bars to the right of the red line report the distribution of OO choices made by participants who chose the less equitable option in the DG choice set corresponding to the DG subgame they viewed in the OO.

5.2 Patterns of CSAs

Figure 3 shows how CSAs for the more and less equitable options vary across the DG, CC (both the primary version and the one with an explicit choice set), and OO modules. Throughout our analysis, we divide the raw CSA responses by five (the highest rating on the Likert scale); this normalization ensures that our CSA measures lie between 0 and 1. To facilitate direct comparisons between utility in DGs and CCs, the figure matches each CC allocation with the DG menu that contains it. Thus, for example, although we did not present the allocations $(2, 0.5)$ and $(4, 0)$ together as a menu in the main CC module, the figure plots the average CSAs for both above the label for DG 1. For the OOs, we report CSAs just for the DG subgames.[14] The figure shows Deciders' CSAs for all available options (not just the ones they selected). Thus, differences between average CSAs across options for a given DGs, and across DGs for a given option, are not driven by differential sample selection.

The figures reveal several key patterns. First, CSAs vary between the more and less equitable allocations in predictable and intuitive ways. Pride and fairness are higher for the more equitable allocations, while guilt and unfairness are higher for less equitable allocations.

Second, holding the allocation constant, the CSAs vary strongly with the context. For example, for the less equitable allocations, respondents experience significantly more guilt and unfairness when they choose those allocations themselves rather than when the computer choses them. Similarly, pride is much higher for the more equitable allocations when the respondent, rather than when the computer, chooses them.

Third, pride and fairness are increasing in the Receiver's payout for more equitable options, and are relatively insensitive to payouts for less equitable options. Guilt and unfairness are decreasing in the Receiver's payout for the more equitable options, and are relatively insensitive to the Decider's payout for less equitable options. The financial satisfaction CSA primarily reflects payouts: it is increasing in the Receiver's payout for more equitable options, and increasing in the Decider's payout for less equitable options. Finally, the aggregate CSAs—happiness and satisfaction—appear to exhibit a blend of the patterns for all the other CSAs.

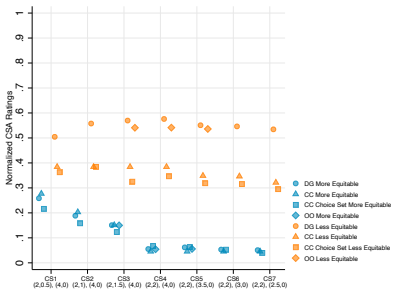
5.3 Model Estimates and the Limitations of Happiness and Satisfaction

Table 1 reports estimates of model (7) based on the DG choices and reported CSAs. With two possible outcomes, the model reduces to a standard logistic regression (without a constant term). Specifically, the dependent variable is an indicator, which equals one when the Decider chooses the more equitable alternative and zero otherwise. The covariate vector is the difference in reported CSA vectors between the more equitable and less equitable options, $\Delta_{ic} = X_{ij_1c} - X_{ij_2c}$, where

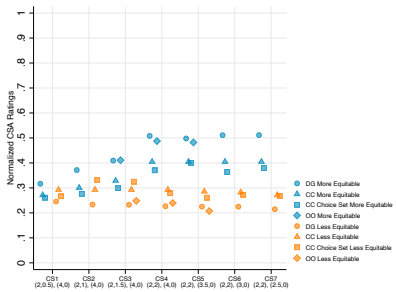
[14] CSAs for the opt-out options in the OO module appear in Figure [X] of Appendix [Y].

Figure 3: Average CSAs

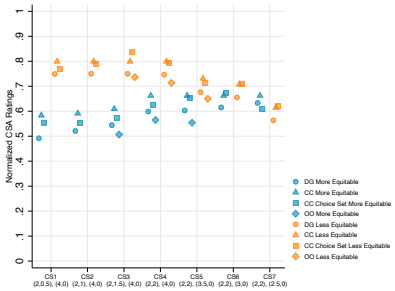
(a) Guilt



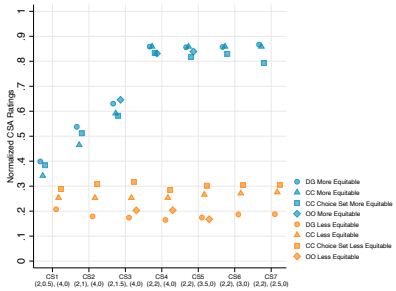
(b) Pride



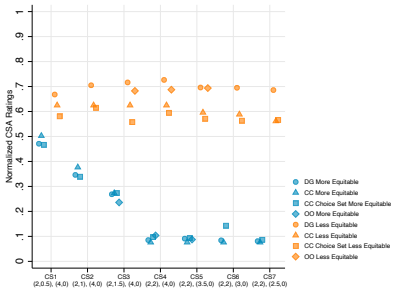
(c) Financial Satisfaction



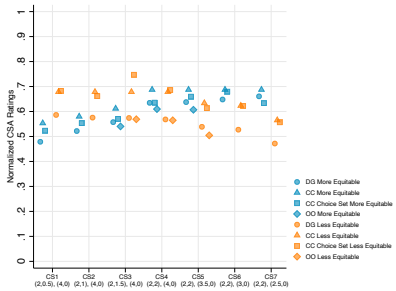
(d) Fairness



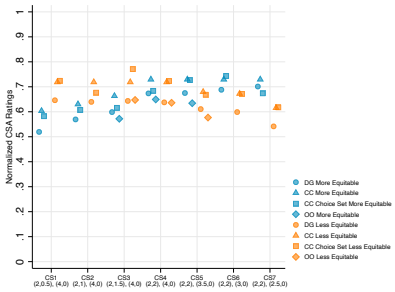
(e) Unfairness



(f) Happiness



(g) Satisfaction



Note: This figure reports the average CSA ratings for each option of each choice set in the DG, CC (main and alternative variant), and OO. We divide the reported CSA responses by five (the highest rating on the Likert scale so that the CSA measures lie between 0 and 1. The figure includes CSA ratings for both the actual and counterfactual options, so that the sample does not change across the different games considered. To facilitate direct comparisons between utility in DGs and CCs, the figure matches each CC allocation with the DG menu that contains it. Thus, for example, although we did not present the allocations $(2, 0.5)$ and $(4, 0)$ together as a menu in the main CC module, the figure plots the CSAs both above the label for DG 1. For the OOs, we report CSAs just for the DG subgames.

j_1 and j_2 index the more equitable and less equitable alternatives, respectively, in each scenario c. To facilitate interpretation of the coefficients, we report average marginal effects. That is, each coefficient represents the change in the likelihood of choosing more equitably, averaged across all participants in the DG, when the CSA corresponding to the pertinent option changes from a minimum value of 0 to a maximum value of 1, holding all other CSAs fixed at their means. Throughout, we cluster standard errors at the participant level.

Column (1) of Table 1 contains the baseline regression. As expected, Deciders are more likely to choose allocations with higher values of positive sensations like financial satisfaction, happiness, and satisfaction, and are less likely to choose allocations with higher values of negative sensations, like guilt and unfairness. The relationships between choice and the two broad CSAs, happiness and satisfaction, are especially strong. However, even when we condition on the values of those broad CSAs, we still find strong relationships between choices and several of the narrower CSAs (guilt, financial satisfaction, and unfairness). This finding, which is consistent with the results of Benjamin et al. (2014), suggests that happiness and satisfaction are not sufficient statistics for participants' desires (and hence welfare), as expressed through their choices.

In Column (2), we examine the robustness of these findings to corrections for measurement error in stated CSAs. In multivariate regressions, measurement error can attenuate some coefficients and amplify others (see, e.g., Gillen, Snowberg, and Yariv, 2019). For example, the coefficients of CSAs other than happiness in Column (1) may be significant, even though choice depends only on happiness, simply because measures of happiness are noisy. To address this potential confound, we restrict the sample to DGs $c \geq 2$, and instrument Δ_{ic} in DG c with Δ_{ic-1} from DG $c-1$. Under the assumption that measurement error is independent across the DGs, this strategy recovers the "true" coefficients on each CSA (see, e.g., Gillen, Snowberg, and Yariv, 2019). As Column (2) shows, this IV procedure leaves the CSA coefficients essentially unchanged. We obtain this result in part because the first stage coefficients are high (ranging from 0.69 to 0.77). These findings, including the narrow range within which the first-stage coefficients fall, are consistent with the degree of measurement error being similar across the CSAs.

The assumption that measurement error is independent across DGs is potentially objectionable. In Appendix D.3, we deploy an alternative statistical approach that does not require this assumption. That test also rejects the null hypothesis that choice, and hence welfare, vary only with the underlying levels of "true" happiness and satisfaction. The intuition for this test is as follows. Under the null hypothesis, the other five CSAs by definition form valid instruments for happiness and satisfaction: they are strongly correlated with happiness and satisfaction (relevance) and, under the null hypothesis, are independent of choice conditional on the true values of happiness or satisfaction (exclusion restriction). Thus, the null hypothesis implies that, in an IV regression of choice on happiness (or satisfaction) and one other CSA k, with happiness instrumented by the remaining four

Table 1: Association between deciders' choices and CSAs

	(1) Logit Choosing More Equitably	(2) IV Logit Choosing More Equitably	(3) Logit Choosing More Equitably
Δ Guilt	-0.13***	-0.14***	-0.19***
	(0.02)	(0.03)	(0.02)
Δ Pride	0.01	0.01	0.10***
	(0.02)	(0.03)	(0.02)
Δ Finan. Satis.	0.28***	0.27***	0.65***
	(0.02)	(0.04)	(0.02)
Δ Fairness	0.02	-0.01	0.03
	(0.02)	(0.03)	(0.02)
Δ Unfairness	-0.10***	-0.11***	-0.12***
	(0.02)	(0.04)	(0.02)
Δ Happiness	0.31***	0.37***	
	(0.02)	(0.04)	
Δ Satisfaction	0.39***	0.48***	
	(0.02)	(0.04)	
N. Participants: 2365			

Note: This table reports estimates of model (7) based on the DG choices and reported CSAs. Specifically, the table reports estimates of a logit regression, where the dependent variable is whether a more equitable option is selected, and the covariate vector is the difference in reported CSA vectors between the more equitable and less equitable options, $\Delta_{ic} = X_{ij_1c} - X_{ij_2c}$, where j_1 and j_2 index the more equitable and less equitable alternatives, respectively, in each scenario c. Coefficients are reported as average marginal effects, which are computed by averaging the change in the predicted probability of choosing the more equitable option when a CSA's normalized rating changes from 0 to 1, and all other CSAs are held constant for each participant. Column (1) reports the average marginal effects from our main model, which includes all CSAs. Column (2) reports estimates covariates are instrumented by the lagged CSA difference Δ_{ic-1}. These estimates are taken from a two-step control function approach (Terza et al., 2008): we estimate the residuals in a first stage linear regression of each Δ_{ic} on lagged Δ_{ic-1} . Then we include the residuals from the first stage in the main logit model. Column (3) reports the coefficients using the same specification as in Column (1) but with happiness and satisfaction omitted. Standard errors are reported in the parentheses, and calculated using bootstrap with resampling clusters at the participant level. * $p < 0.1$, ** $p < 0.05$, ***$p < 0.001$.

CSAs, the coefficient of CSA k should be zero, and the model should pass over-identification tests. Instead, when we implement this procedure for each CSA other than happiness or satisfaction, the t-statistic on the included CSA ranges from 10 to 20 in absolute value, and the model fails Hansen's J over-identification test dramatically, with the χ^2 statistic ranging from 300 to 450 (see Hansen, 1982).

Because happiness and satisfaction likely aggregate the five narrower CSAs to some degree, Column (3) of Table 1 explores the importance of the narrower CSAs in a regression that omits the two broad measures. The largest changes between Columns (1) and (3) are for the coefficients of pride and financial satisfaction. The coefficient of pride, which is essentially zero in Column (1), becomes moderately positive and statistically significant in Column (3). The coefficient of financial satisfaction almost doubles. Changes in the other coefficients are smaller. This finding points to a potential explanation for our finding that happiness and satisfaction are not sufficient statistics for welfare: these broad measures appear to encapsulate positive sensations (pride and financial satisfaction) to a greater extent than negative sensations (guilt and unfairness). Appendix Tables A7 and A8 report regressions of happiness and satisfaction on the other five CSAs. Table A8 shows directly that, relative to choice, assessments of happiness and satisfaction place much less weight on guilt and unfairness, and much greater weight on pride.

5.4 Stability Across Domains and Accuracy of Out-of-Sample Prediction

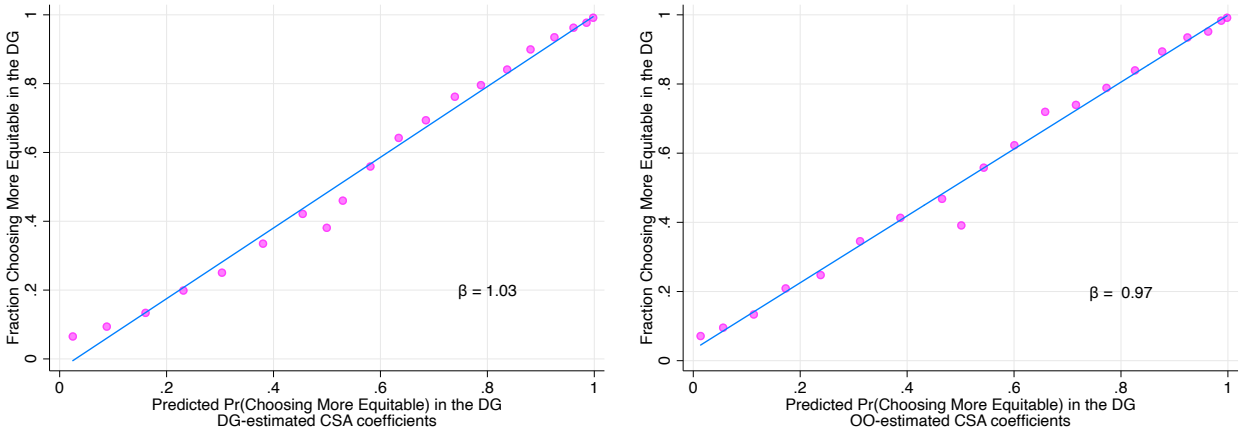
A key premise of our framework is that the empirical model relating choices to CSAs is stable across the domains to which we apply it. Appendix Table A6 formally tests this hypothesis by estimating the model of choice described in Section 4.1 using pooled data from all of the DGs and OOs. To allow for instability across domains, we include interactions between each CSA and an indicator for OO games. This procedure establishes a high bar, in that it entails a joint test of the assumption that the coefficients of the CSA vector are stable *and* that the model of choice in Section 4.1 is correctly specified. For example, even if the hypothesis of stability is valid, the data could fail this test if the ε_{ijc} for the OO games has a nested structure rather than a fully independent structure, or if the assumption of additively separable utility is not satisfied. Even so, Appendix Table A6 shows that there are no significant differences between the CSA coefficients for the two domains, with the single exception that financial satisfaction appears to gain more prominence in OO games.

Figure 4 assesses stability by comparing within-sample fit to cross-domain (and hence out-of-sample) predictive fit. Panel (a) depicts our model's in-sample fit for the DGs. We estimate the model using DG data only, use it to predict the probability of choosing the more equitable allocation for each tuple of CSAs corresponding to a (person, DG) pair, and then compare that predicted likelihood to the empirical likelihood. We divide the predicted likelihoods into twenty equally-sized bins, and for each bin plot the empirical frequency of choosing more equitably. For a model with perfect in-sample fit, all points would lie on the 45-degree line. The plotted points are in fact close to that line.

Figure 4: Observed versus fitted probabilities of choosing more equitably in the DGs

(a) Model estimated using DG data

(b) Model estimated using OO data



Note: This figure compares the model's predicted likelihood and the empirical likelihood of choosing the more equitable option in the DG. The predicted probabilities are divided into 20 bins, and the figure reports the average the empirical likelihood of choosing the more equitable option for each of those bins. Panel (a) estimates the model using the DG CSAs and reports the in-sample fit of the model. Panel (b) estimates the model using the OO CSAs and reports the out-of-sample fit. The blue line represents the 45 degree line, wherein all points would lie if the model was a perfect fit.

Panel (b) of Figure 4 depicts the model's cross-domain (out-of-sample) fit. Its construction is almost identical to that of panel (a), with one crucial difference: we estimate the empirical model using OO data instead of DG data. In other words, the coefficients of the CSA vector used to construct the predictions reflect choice patterns from OO scenarios, but the values of the CSAs to which the predictions apply reflect DG scenarios. Remarkably, panel (b) shows that the frequency pairs still line up along the 45-degree line. This finding establishes that it is possible to predict choice—and thus welfare—in another domain based on CSAs simply by estimating our model in the first domain and fitting it to CSAs from the second.

Appendix D.4 provides analogs of Figure 4 for choices in the OO games. The in-sample fit is slightly worse than for the DGs, potentially due to mis-specification of the error distribution, but there is again no evidence that the cross-domain (out-of-sample) fit is significantly worse than the in-sample fit.

5.5 Potentially-Omitted Mental States

A key assumption of our empirical model is that any pertinent mental states not spanned by our CSAs conform to our assumptions about the ε_{ijc} term—in other words, they must be orthogonal to the dimensions our CSAs do span. This orthogonality assumption implies that the average value of ε_{ijc} should not depend on the alternative j. Consequently, if we modify the logistic regressions

of Section 5.3 by including a constant term, (i) the estimated constant should be 0, and (ii) the CSA coefficients should be unchanged. In contrast, if a pertinent mental state that our CSAs do not fully span differs systematically between the more and less equitable options, we would expect to find a non-zero constant, as well as changes in the coefficients of CSAs that are correlated with that state. Appendix Table A2 replicates Table 1, but includes a constant term. Consistent with our assumptions, the constant term is close to zero, and the coefficients on the other CSAs are essentially unaltered.

As mentioned in Section 3, we drew extensively on a large literature from psychology concerning the types of motivations that influence choices involving equity. It is therefore reasonable to assume that our measures cover the most important motivations. To assess whether additional CSAs are likely to add important new dimensions, we conduct a principal component analysis (PCA) of the CSAs we have. Appendix Table A3 presents the loadings of the seven factors on the seven CSAs, as well as the fraction of variation each factor explains. The first two factors explain 73 percent of the variation, and the least significant factor explains only 3 percent of the variation. Moreover, Appendix Figure A2 shows that the choice predictions generated by the first two factors are nearly identical to the choice predictions obtained when we use all seven of our CSAs. Thus, we see no evidence that adding information concerning factors beyond the first two is consequential. These results cast doubt on the importance of any additional mental states that our CSAs do not span.

5.6 Additional Robustness Checks

Because our elicited CSAs encompass both immediate and subsequent mental states, they are potentially susceptible to the Aggregation Problem. To gauge the importance of this concern, we recruited another group of 200 participants (12 of whom we dropped from the analysis for reasons explained in Section 3) for a supplementary "temporal disaggregation" arm of the experiment. In this arm, we separately elicited CSAs associated with immediate and subsequent mental states. As detailed in Appendix F.2, we find that (i) the average reported values of the present and future CSAs for each option in each game are nearly identical, (ii) the correlations between present and future CSAs are nearly perfect, and (iii) an analog of the logit regression in Column (1) of Table 1 produces nearly identical coefficients for present and future CSAs.

Because we elicited CSAs after participants made their choices, a possible concern is that they may have distorted reported CSAs in ways that rationalized the options to which they previously committed. This tendency would lead to a violation of the assumption that $\varepsilon_{ijc} \perp X_{ijc}$. To gauge the importance of this concern, we recruited 200 participants (13 of whom we dropped from the analysis for reasons explained in Section 3) for a supplementary "ex-ante" arm of the experiment. We then elicited CSAs for all available options prior to choices. Appendix F.1 shows that this alternative elicitation format does not alter the reported CSAs or their relationships to choice.

Appendix F.3 provides a final test of the orthogonality assumption that $\varepsilon_{ijc} \perp X_{ijc}$. This test

assumes, plausibly, that variation across the DG choice sets is exogenous to the various components of the ε term, including trembles, random noise in perception, and idiosyncratic taste variation. In principle, this assumption allows one to use choice set dummies as instruments for X_{ijc}, and to test for violations of the orthogonality assumption by comparing our original estimates to the resulting IV estimates. In practice, there is insufficient independent variation in CSAs across the choice sets to instrument adequately for all seven CSAs. Instead, our approach is to modify our empirical model so that it does not rely on variation across choice sets for identification, and to check the stability of the CSA coefficients. If the data do not satisfy our orthogonality assumption, then purging our estimates of CSA variation that *does* satisfy the orthogonality assumption should alter the coefficients.

Concretely, we implement this test by adding choice set fixed effects to the logit regression that appears in Column (1) of Table 1. The modified regression, which we report in Appendix Table A13, shows that including choice set fixed effects has negligible effects on the pseudo R-squared and the CSA coefficients. This stability is notable because Appendix Table A12 shows that variation across the choice sets does account for a meaningful share of the variation in $\Delta_{ic}\hat{\beta}$, the predicted difference in average utility between the more and less equitable DG alternatives. Specifically, a regression of $\Delta_{ic}\hat{\beta}$ on choice set indicators has an adjusted R-squared of 0.12.[15] Thus, there is no evidence that choice responds differently to variation in Δ_{ic} that comes from plausibly exogenous differences in choice sets, and variation that comes from other sources. Moreover, stability of the pseudo R-squared statistic in Table A13 is consistent with our earlier finding (in Section 5.5) that, aside from the factors spanned by our CSAs, choice-relevant mental states do not vary systematically across the choice sets.

6 Estimating Welfare with CSAs

6.1 Motivations for Prosocial Behavior: Welfare in Dictator Games

Figure 5 plots $\bar{u}_{jc}$, the average deterministic component of Deciders' utility translated to dollar units (see Section 4.2), for each alternative in each DG. Panel (a) reports utility estimates for every participant and every alternative, both chosen and unchosen. Thus, the sample is the same for all options and DGs. In contrast, Panel (b) reports utility estimates only for the chosen alternatives. For all figures displaying welfare calculations, we bootstrap 95-percent confidence intervals, shown as vertical lines, blocking at the participant level.[16]

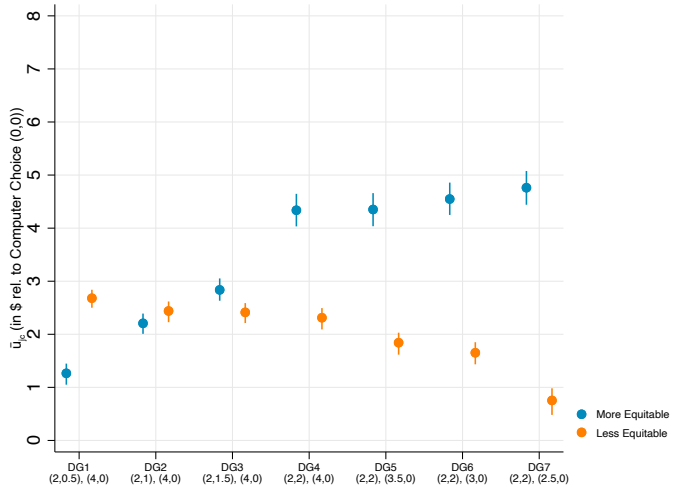
Several patterns in Panel (a) of Figure 5 illuminate participants' motivations for prosocial behavior.

[15] As expected, the coefficients increase steadily as one moves from DG 1 to DG 7.

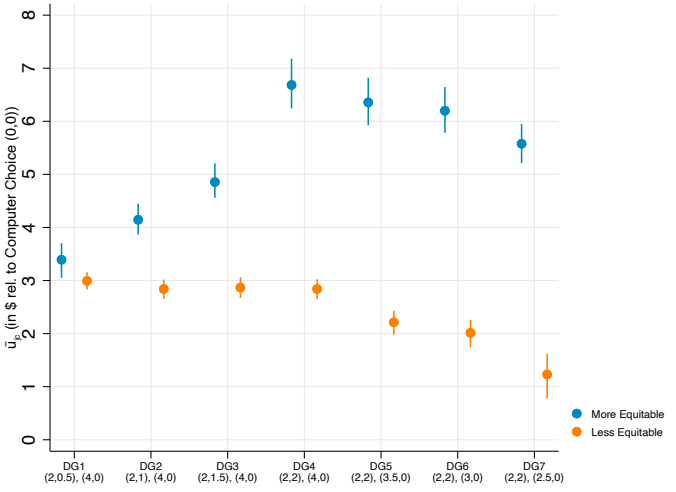
[16] Analytic standard errors for the money-metric welfare estimates are not readily available because they combine estimates from different steps, as detailed in Section 4.2.

Figure 5: Deciders' average utility in dictator games

(a) Full-sample results: utility from chosen and counterfactual alternatives



(b) Restricting only to the chosen options



Note: This figure reports the average money metric utility, $\bar{u}_{jc}$, for each option of each choice set in the DG. Panel (a) reports the average utility for both actual and counterfactual choices. Thus, each point in the figure contains the whole sample. Panel (b) reports the average money metric utility only for the chosen options. Thus, the sample of participants changes across the data points. The 95 percent confidence intervals are reported as the vertical bars, and calculated using bootstrap with 1,000 resampling clusters at the participant level.

First, the Deciders' average money-metric utility from choosing the less equitable allocation is always lower than the monetary payout that allocation provides. Recall that money-metric utility for any option is the dollar value, y, such that replacing $(0, 0)$ with $(y, 0)$ when both are exogenously assigned by the computer (in the main module), and with the option of interest when the individual selects it, both yield the same deterministic utility increment. The fact that participants derive lower utility from the less equitable allocation *when they have to choose it themselves than when the computer chooses it for them* is consistent with Deciders experiencing negative sensations, such as guilt, when *acting* selfishly. This finding corroborates the importance of negative sensations for motivating prosocial behavior (e.g., Rabin, 1995; Andreoni, 1995; Tangney and Dearing, 2002; Charness and Dufwenberg, 2006).

Second, Deciders' average money-metric utility from choosing the more equitable allocations generally exceeds the Deciders' monetary payout. This finding corroborates the importance of positive sensations for motivating prosocial behavior (e.g., Andreoni, 1989, 1990). Interestingly, the utility bonus from behaving equitably is particularly large when the more equitable option involves an equal split, as in DGs 4-7: for the $(2, 2)$ allocations, average money-metric utility exceeds the partners' combined payout. The special status of the 50-50 norm has been emphasized by Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Andreoni and Bernheim (2009) and others.

Third, holding the sample constant (as in Panel (a)), as we improve the Receiver's payout from the more equitable option (i.e., moving stepwise from DG 1 to DG 4), the average money-metric utility from that option increases substantially (as expected), while the average money-metric utility from the less equitable option declines (from 2.68 in DG 1 to 2.31 in DG 4; 95% CI for difference $= [0.28, 0.47]$).[17] The latter finding is consistent with the hypothesis that the typical Decider experiences more intense negative sensations such as guilt when the Receiver's payout in a foregone more-equitable allocation is greater.

Fourth, holding the sample constant, as we reduce the Decider's payout from the less equitable option (i.e., moving stepwise from DG 4 to DG 7), the average money-metric utility from that option declines significantly (as expected), while the average money-metric utility from the more equitable option increases (from 4.34 in DG 4 to 4.76 in DG 7; 95% CI for difference $= [-0.51, -0.34]$).[18] The latter finding is consistent with the hypothesis that Deciders experience negative sensations less intensely from comparison effects when their own payout in a foregone less-equitable outcome is greater.

Comparing Panel (a), which uses data for both chosen and unchosen options, and Panel (b), which only uses data for chosen options, we see that those who choose the less equitable option receive

[17] A regression of money-metric utility from the less equitable option on the DG number j, for $j \in \{1, 2, 3, 4\}$, yields a coefficient of -0.11, with a 95% CI of $[-0.14, -0.09]$. The confidence intervals for this and related analyses are computed by bootstrap, sampling at the participant level, and taking into account statistical uncertainty in our estimate of the marginal utility of money.

[18] A regression of money-metric utility from the less equitable option on the DG number j, for $j \in \{4, 5, 6, 7\}$, yields a coefficient of 0.15, with a 95% CI of $[0.12, 0.18]$.

greater-than-average utility from that option. The same is true for the more equitable option. These patterns reflects preference heterogeneity and self-selection: people choose the option they consider best for them.

In summary, our findings are consistent with optimization over mental-state bundles that encompass both negative and positive sensations. While some theories attribute generosity to the avoidance of negative sensations while others attribute it to pursuit of positive sensations, our results suggest that both positive and negative sensations motivate giving.

6.2 Welfare and the Act of Choosing: Dictator Versus Computer Choice

An important implication of Figure 5 is that the average money-metric utility a Decider obtains from the less equitable allocation in each DG is significantly lower than the Decider's monetary payout. Given our choice of benchmark domains (see Section 4.2), this payout is definitionally equivalent to the money-metric utility the Decider would receive if the computer assigned the same less equitable allocation exogenously. Thus, the act of choosing that allocation reduces the utility the Decider derives from it. With that example in mind, we now provide a more comprehensive analysis of utility differences between choosing an allocation and receiving it as an exogenous assignment.

Panel (a) of Figure 6 plots $\bar{u}_{jc}$ for each possible allocation using CSAs measured in the main CC module (where alternatives are unspecified). To facilitate direct comparisons between utility in DGs and CCs, the figure matches each CC allocation with the DG menu that contains it. Thus, although we did not present the allocations (2,0.5) and (4,0) together as a menu in the main CC module, the figure plots the money-metric utility for both above the label for DG 1, the scenario that offers those options.

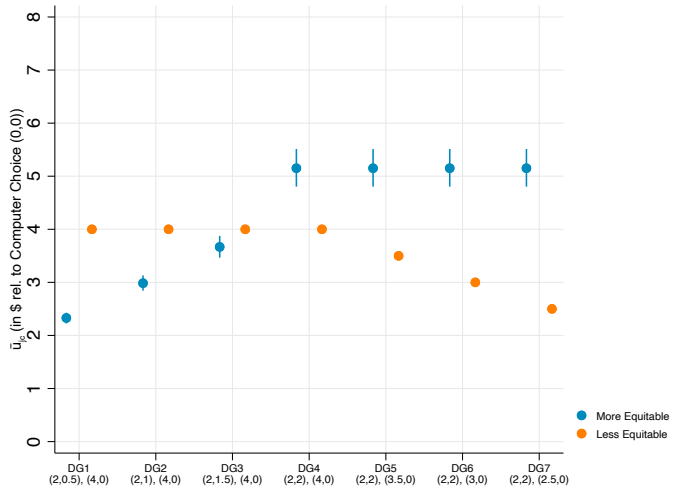
A comparison between the (a) panels of Figures 6 and 5 reveals that, as with the less equitable allocations, Deciders derive more utility from the more equitable allocations when the computer assigns them exogenously than when participants choose them. This finding is inconsistent with the hypothesis that the act of choosing an equitable option primarily engenders positive sensations such as pride.

Our findings therefore imply that the utility Deciders derive from any fixed outcome, whether less equitable or more equitable, depends on the process used to select that outcome. Moreover, in contrast to existing studies that claim to measure the value of autonomy using meta-choice methods (e.g., Fehr et al., 2013; Bartling et al., 2014), we find that Deciders are better off with each type of option if someone else chooses it for them than if they choose it for themselves.

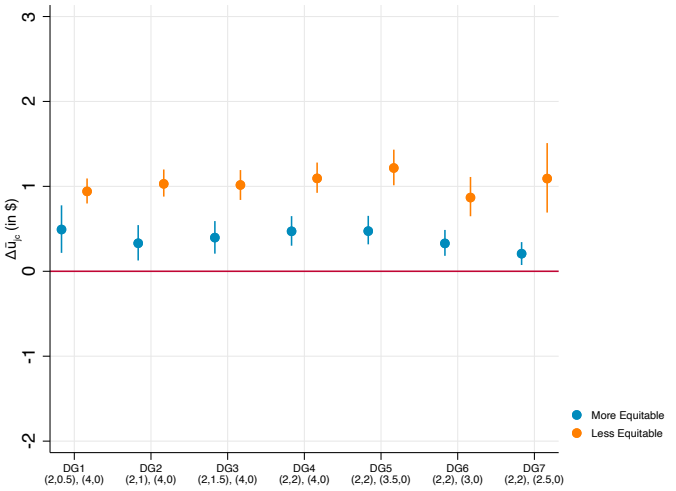
To appreciate the significance of these findings, consider a Planner charged with selecting an allocation from one of our seven DG menus, with the objective of maximizing the Decider's welfare. Standard welfare economics instructs the Planner to select the same option the Decider would choose

Figure 6: Deciders' average utility in the main computer choice model

(a) Average utility when the computer determines the allocation



(b) Welfare gain from having the computer choose instead of the Decider



Note: Panel (a) reports the average utilities from the main CC module (where alternatives are unspecified). To facilitate direct comparisons between utility in DGs and CCs, the figure matches each CC allocation with the DG menu that contains it. Thus, for example, although we did not present the allocations (2,0.5) and (4,0) together as a menu in the main CC module, the figure plots the money-metric utility for both above the label for DG. The sample is held constant in this panel because the money metric utility is reported for both actual and counterfactual DG options. Panel (b) reports average utility gains when the computer—instead of the Decider—chooses the allocation that the Decider chose in the DG. The 95 percent confidence intervals are reported as the vertical bars, and calculated using bootstrap with 1,000 resampling clusters at the participant level.

for herself. But we have just demonstrated that the mental-state bundle associated with each of the available options changes systematically when the Planner takes over the selection process, and that these changes are welfare-relevant. As a consequence, a Planner who tries to mimic the Decider's choices over DG allocations will not end up maximizing the Decider's welfare (a consequence of the Non-Comparability Problem). Furthermore, as we explain next, our methods allow us to quantify the error associated with that strategy.

Panel (b) of Figure 6 shows, for each allocation j in each DG c, how Deciders' average deterministic money-metric utility would change if, instead of choosing the allocation themselves, the computer chose it for them. Consistent with the patterns we noted in our discussion of Figures 5 and 6, we see that, on average across the seven DGs, the money-metric utility derived from a fixed allocation is higher when the computer is responsible for the selection by \$1.03 (95% CI [0.88,1.19]) in the case of less-equitable allocations, versus \$0.37 in the case of more-equitable allocations (95% CI [0.24,0.50]). Thus, standard welfare methods bias the Planner toward acting too generously on behalf of the Decider.

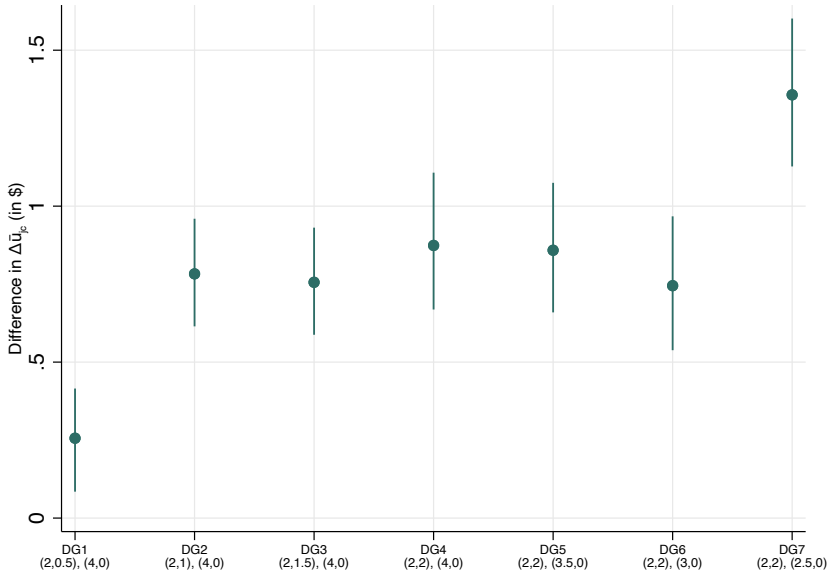
Panel (a) of Figure 7 exhibits this bias. It shows that, on average across the seven menus, the money-metric utility difference between the more equitable and the less equitable allocation is \$0.80 higher in the DGs than in the main CC module. A Planner who applies standard welfare methods will therefore be willing to overpay for opportunities to replace less equitable allocations with more equitable alternatives. Moreover, the magnitude of the potential overpayment—\$0.80—represents a large fraction of the stakes in this experiment.

Alternatively, imagine that the Planner uses standard welfare methods to compute the likelihood that the less-equitable allocation is optimal for the typical Decider. Panel (b) of Figure 7 quantifies the magnitude of the resulting error. Specifically, for each DG menu, we use the CSAs from the the DG module to determine the fraction of participants who would *appear* to be better off with the more equitable option based on those CSAs. Then we use the CSAs from the main CC module to determine the fraction who are *actually* better off with the more equitable option when someone else selects it for them. Taking the difference between these fractions and then dividing by the second, we obtain a measure of the degree to which standard welfare methods would lead the Planner to exaggerate the fraction of participants who are better off when someone else assigns the less equitable allocation. For all but one of the menus, the degree of exaggeration exceeds 10%.[19]

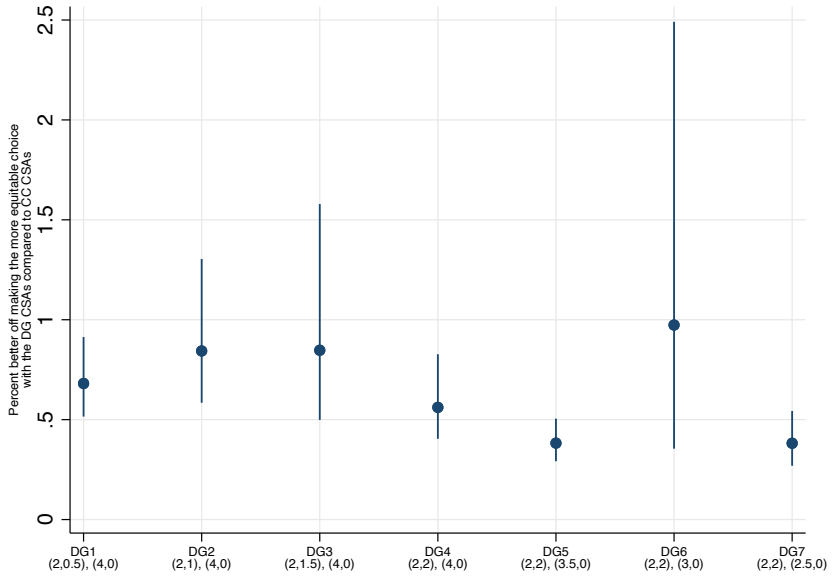
So far, we have focused on the main CC format, which does not identify alternative allocations. Next we ask whether our welfare conclusions change when the individual (here, the Decider) knows the menu from which a Planner (here, the computer) selects. Prior work in psychology (e.g., Tversky

[19]We rely on DG CSAs rather than actual DG choice ensure that the statistics in the DG and CC module are computed in a fully comparable manner. This is important if, e.g., the model linking CSAs to choice has any degree of mis-specification, or if the reported CSAs have some degree of noise to them. Our Section 5 results suggest that neither issue is particularly important, but also not completely absent.

Figure 7: Deciders' utility gain from replacing the less equitable option with the more equitable option: DG versus CC



(a) Difference in utility gain between DG and CC



(b) Difference in predicted probabilities of the less equitable option being optimal: DG vs. CC

Note: Panel (a) reports the average utility gains from switching from the more equitable allocation to the less equitable allocation in the DGs versus CCs. Panel (b) reports the percent change in those preferring the more equitable option when switching from the CC to the DG. The 95 percent confidence intervals are reported as the vertical bars, and calculated using bootstrap with 1,000 resampling clusters at the participant level.

and Shafir, 1992; Iyengar and Lepper, 2000) has suggested that contrasts with other options may lead to negative experiences, possibly even in situations where the outcome is exogenously determined. To address this issue, we replicate our analysis using data from the alternative CC module, in which we explicitly specify unchosen options.

First, we construct Figure 8, which is analogous to Figure 6 but uses data from the alternative CC module. Comparing Panel (a) from the respective figures, we see that Deciders derive less utility from the more equitable allocations when they know the unchosen alternative (the alternative CC module) than when they do not (the main CC module): on average across the seven menus, the difference is \$0.38 (95% CI of [0.07,0.72]). In contrast, for the less equitable allocations, the corresponding difference is roughly zero and insignificant (-0.08, with a 95% CI of [-0.41,0.23]).

With these differences in mind, we revisit the Planner's task considered above—i.e., selecting an allocation from one of our seven DG menus, with the objective of maximizing the Dictator's welfare. Panel (b) of Figure 8 replicates Panel (b) of 6 using data from the alternative CC module. On average across the seven menus, there is little difference between the money-metric utility Deciders derive from the more equitable allocation when the computer selects it (with the alternative specified) and when they choose it themselves (\$0.17 with 95% CI of [-0.06,0.42]). In contrast, the corresponding difference remains substantial for the less equitable allocation (\$0.76 with a 95% CI of [0.51, 1.03]). Consequently, even when people know the unchosen option, a Planner who applies standard welfare methods remains willing to overpay for opportunities to replace less equitable allocations with more equitable alternatives.

6.3 Welfare and Avoidance Opportunities

Next we analyze preferences and welfare in opt-out games. Prior work (e.g., Broberg et al., 2007; Lazear et al., 2012; DellaVigna et al., 2012) has (implicitly) assumed that the utility participants derive from opting in and from opting out does not depend on whether the participant or some other party makes the opt-out choice. We formalize this increasingly common assumption as follows:

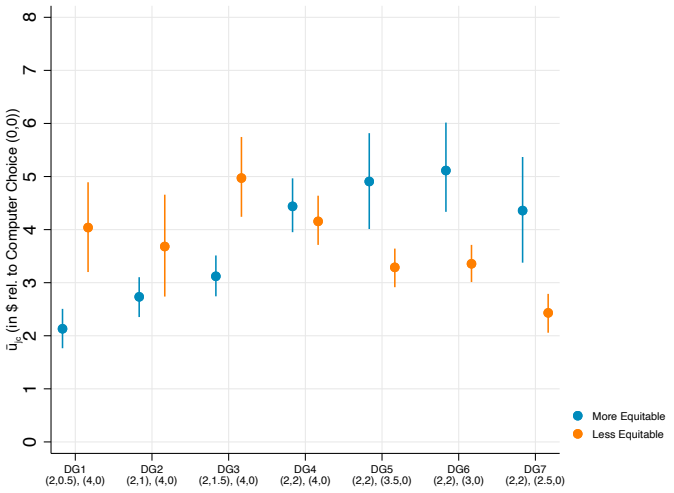
Comparability Hypothesis: A person chooses to opt-out for \$$X$ if and only if they would prefer an *exogenous* allocation of $(X, 0)$ over an *exogenous* assignment to the DG.

Our methods allow us to test this hypothesis formally, and to provide welfare estimates that do not require it to hold. As a first step, we construct Figure 9, which plots participants' utility, $\bar{u}_{jc}$, for each alternative in each of the OO games. The figure provides three main insights.

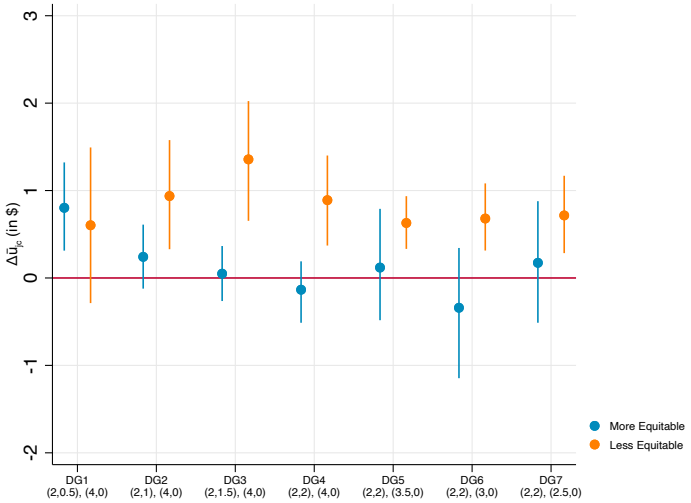
First, the money-metric utility Deciders would receive if the computer assigned the opt-out allocation exogenously is significantly below the opt-out payout. This finding provides direct evidence that the Non-Comparability Problem affects metachoices: participants experience disutility from the experience of choosing the opt-out option.

Figure 8: Deciders' average utility in the alternative computer choice module

(a) Average utility when the computer determines the allocation



(b) Welfare gain from having the computer choose instead of the Decider



Note: Panel (a) reports the average money metric utility in the alternative CC module (where alternatives are specified). To facilitate direct comparisons between utility in DGs and CCs, the figure matches each CC allocation with the DG menu that contains it. Thus, although we did not present the allocations (2,0.5) and (4,0) together as a menu in the main CC module, the figure plots the money-metric utility for both above the label for DG 1. The sample is held constant in this panel because the money metric utility is reported for both actual and counterfactual DG options. Panel (b) reports average utility gains when the computer—instead of the Decider—chooses the allocation that the Decider chose in the DG. The 95 percent confidence intervals are reported as the vertical bars, and calculated using bootstrap with 1,000 resampling clusters at the participant level.

Second, opting out nevertheless generates a less negative experience than opting into the DG subgame and choosing the less equitable option. Specifically, in panels (a) and (b), the utility estimates for OO 2 clearly show that Deciders receive much higher utility from opting out for \$4 rather than opting in and choosing $(4, 0)$. Similarly, in panel (c), Deciders receive higher utility from opting out for \$3.50 than from opting in and choosing $(3.50, 0)$ in OO 3. Averaging across the three DG subgames, these differences imply that Deciders would be willing to pay \$1.15 (95% CI [0.97,1.34]) to experience the mental-state bundle associated with opting out to $(y, 0)$ rather than the one associated with opting in and choosing $(y, 0)$ (thereby revealing the choice to their partners).

Third, a comparison to Panel (a) of Figure 5 reveals that Deciders' utility from choosing the more equitable allocation in the DG component of the OO game is generally lower than the utility they would obtain from choosing that allocation in the corresponding DG. On average, the money-metric difference is \$0.42 (95% CI [0.21,0.62]). This finding is consistent with the negative contrast effects observed in Panel (b) of Figure 6; i.e., that participants' derive less utility from an outcome when additional alternatives are available. Interestingly, these effects are more muted for the less equitable options in the DG subgames.

A secondary manifestation of contrast effects in Figure 9 is that utility from both the more equitable and less equitable options is slightly higher when the opt-out payout is lower.

The first and third insights reflect the Non-Comparability Problem: adding an opt-out option changes the environment and thus the experience of choosing, such that (i) opting out leads Deciders to obtain lower utility than they would if the computer exogenously assigned the same allocation exogenously, and (ii) the utility Deciders obtain from each option in the DG declines. The first effect makes Deciders less eager to opt out, while the second effect has the opposite effect. In principle, these two effects might cancel out, in which case the Comparability Hypothesis would hold, as prior studies have assumed. To evaluate this possibility, we proceed in three steps.

First, we compute the value of having the option to play in the DG (relative to an exogenous allocation of $(0, 0)$), assuming the Comparability Hypothesis holds. Under that assumption, we can depict the opt-out decision using the following logit model:

$$Pr(\text{opt in}) = \frac{exp\left(\frac{\bar{U}_c - \pi_o}{\sigma_c}\right)}{1 + exp\left(\frac{\bar{U}_c - \pi_o}{\sigma_c}\right)},$$

where $\bar{U}_c$ is the average money-metric utility Deciders derive from participating in the DG c, π_o is the opt-out payment, and σ_c is the (DG-specific) variance of the error term. Column (1) of Table 2 provides estimates of $\bar{U}_c$ for the three DG subgames, along with 95% confidence intervals. Consistent with prior research, participants' relatively high opt-out rates imply they do not, on average, place a great deal of value on the option to share. The average implied value of participating in a DG is below \$4 for the menu {(2,1.5), (4,0)}, slightly higher than \$4 for the menu {(2,2),(4,0)}, and

Table 2: Comparing welfare estimates from standard approaches vs. hybrid approach

	(1) Choice-based inference of playing the DG using the Opt-Out Game	(2) $\bar{u}_{jc}$ of playing in the DG using CSAs in the DG	(3) $\bar{u}_{jc}$ of playing in the DG using CSAs in the OO	(4) Difference (1)-(2)	(5) Difference (2)-(3)
Subgame: (2,1.5) vs. (4,0)	3.82	3.88	3.48	-0.06	0.40**
	[3.82, 3.92]	[3.71, 4.06]	[3.12, 3.85]	[-0.26, 0.14]	[0.08, 0.74]
Subgame: (2,2) vs. (4,0)	4.11	5.01	4.25	-0.91***	0.76***
	[4.00, 4.17]	[4.74, 5.33]	[3.86, 4.69]	[-1.21, -0.63]	[0.39, 1.14]
Subgame: (2,2) vs. (3.5,0)	3.87	4.77	3.76	-0.90***	1.01***
	[3.81, 3.93]	[4.49, 5.07]	[3.34, 4.23]	[-1.24, -0.60]	[0.58, 1.42]
N. Participants: 2365					

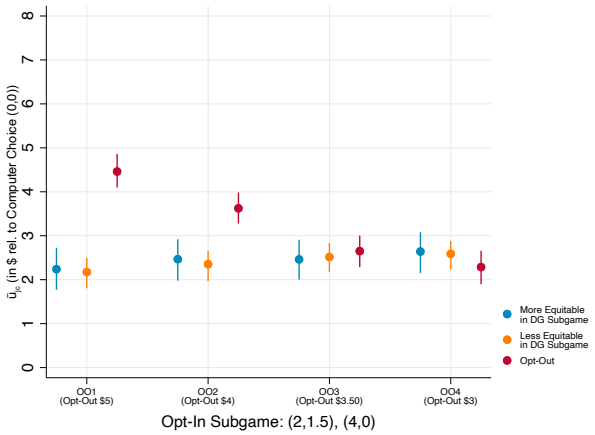
Note: This table reports Deciders' estimated utilities of playing in the DG using the approaches described in Section 6.3. Column (1) reports the money-metric utility estimates obtained from standard choice-based methods that assume Comparability Hypothesis for OO games. Column (2) reports the money-metric utility estimates obtained from our approach. Column (3) reports the average utility Deciders would derive from their DG choices if the corresponding CSAs were those they reported in the OOs, rather than in the DGs. Column (4) is the difference between the estimates in Columns (1) and (2), while Column (5) is the difference between the estimates in Columns (2) and (3). The 95 percent confidence intervals are reported in brackets, and calculated using bootstrap with 1,000 resampling clusters at the participant level. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.001$.

slightly higher than \$3.5 for the menu {(2,2),(3.5,0)}.

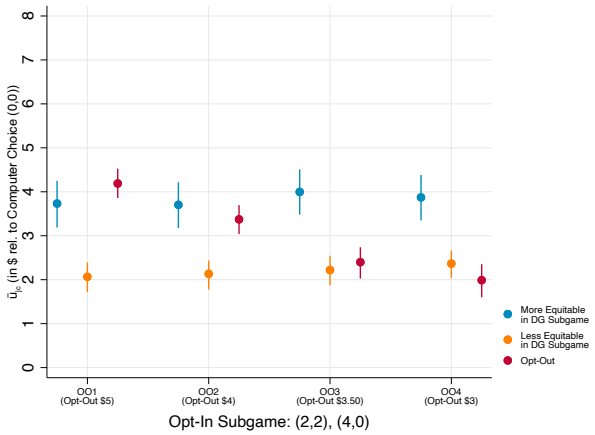
In the second step, we estimate the money-metric utility Deciders actually derive from playing exogenously assigned DGs; see Column (2). We calculate these measures by taking a weighted average of the estimated money-metric utility Deciders derive from their chosen option in assigned DGs (Panel (b) of Figure 5), with the weights equal to the fraction of participants choosing each option. Comparing Columns (1) and (2), we see that Deciders obtain significantly higher utility when playing in an assigned DG than their opt-out choices imply under the Comparability Hypothesis. Column (4) shows that we reject the implications of the Comparability Hypothesis for DG valuation ($p < 0.05$) for two out of the three DGs.

Our final step helps to clarify why the meta-choice approach (Column (1)) underestimates the utility Decider's derive from an assigned DG. In Column (3), we use our method to compute the average utility Deciders would derive from their DG choices if the corresponding CSAs were those they reported in the OOs, rather than in the DGs. Consistent with the contrast between Figures 9 and 5, the presence of an opt-out option significantly reduces the value of playing a DG; we report the differences in Column (5).

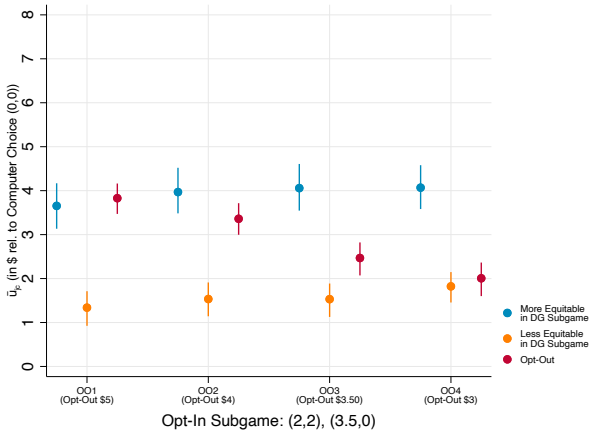
Figure 9: Deciders' average utility from different possible options in the opt-out games



(a) DG Subgame: (2,1.5),(4,0)



(b) DG Subgame: (2,2),(4,0)



(c) DG Subgame: (2,2),(3.5,0)

Note: This figure reports the average money-metric utility for each option of each choice set in the OO. We average across both chosen counterfactual options so that the sample is held constant throughout in all the estimates. Panel (a) reports the average utilities for the OO game where the opt-in subgame is (2,1.5) vs. (4,0); Panel (b) reports the average utilities for the OO game where the opt-in subgame is (2,2) vs. (4,0); and Panel (c) reports the average utilities for the OO game where the opt-in subgame is (2,2) vs. (3.5,0)}. The 95 percent confidence intervals are reported as the vertical bars, and calculated using bootstrap with 1,000 resampling clusters at the participant level.

6.4 Revisiting Happiness and Satisfaction as Proxies for Welfare

As Table 1 showed, neither happiness nor satisfaction are sufficient statistics for choice and welfare in our experiment. Appendix E.2 demonstrates explicitly that using happiness or satisfaction as a proxy for welfare leads to systematic bias. Specifically, we replicate the preceding analysis using models that incorporate only one CSA, either happiness or satisfaction. This procedure is analogous to prior work that converts happiness or satisfaction indices to money-metric scales (e.g., **??**). For the DGs, our main finding is that the resulting welfare measures slightly overestimate utility from choosing the less equitable option, and substantially underestimate utility from choosing the more equitable option. In the CCs, relying only on happiness or satisfaction also leads to underestimates of Deciders' utilities from the more equitable allocations. In the OOs, relying only on happiness or satisfaction does not bias estimates of welfare from the opt-out option, but leads to qualitatively and quantitatively similar biases for the alternatives in the DG subgames.

7 Concluding Remarks

This paper develops a conceptual and empirical framework that combines SWB and revealed preference methods to estimate welfare in situations where neither method alone is adequate. A key motivation for this framework is that, in settings where the act of choosing creates welfare-relevant emotions, standard choice-based methods suffer from the Non-Comparability Problem, which can render welfare unrecoverable from standard data. As proof of concept, we apply this framework to sharing and avoidance decisions, and obtain six insights. First, happiness and satisfaction are not sufficient statistics for welfare in our settings. Second, Deciders who are tasked with dividing money between themselves and a partner are better off if their choice is taken out of their hands, conditional on achieving the same outcome (regardless of whether it is the more or less equitable alternative). Third, Deciders act generously in part to avoid negative affect, and thus their choices over-state their preferences for exogenously-imposed equitable allocations. Fourth, "meta choices," such as opting out of a sharing opportunity, generate welfare-consequential bundles of mental states. In our setting, Deciders experience negative affect such as guilt from opting out. Fifth, we document a novel context effect whereby including opt-out opportunities decreases Deciders' utility from choosing one of the existing options. Sixth, these findings collectively imply the presence of serious Non-Comparability Problems in the setting we study. It follows that the prior literature's choice-based analyses of Deciders' opt-out decisions provides a misleading gauge of the extent to which Deciders benefit from exogenously-created sharing opportunities.

Prior studies have used metachoices to assess welfare effects associated with the act of choosing in a variety of other settings. Each provides a potentially fruitful context for applying our framework, in part because the Non-Comparability Problem may arise in those settings. One set of studies uses metachoices to assess the value of authority, autonomy, and control (e.g., Fehr et al., 2013; Owens et al., 2014; Bartling et al., 2014). The Non-Comparability Problem potentially arises in that

context for a variety of reasons. Someone might opt for control to avoid feeling guilty about shirking responsibility, rather than out of an inherent desire for authority. Alternatively, the experience of exercising control may depend on whether one takes control or is granted authority. Another set of studies explores how restrictions on future options (e.g., commitment contracts) can benefit time-inconsistent decision makers by promoting self-control (see, e.g., Carrera et al., 2022, for a recent summary). And yet, one might value self-imposed restrictions while resenting restrictions imposed by others. Additionally, the act of choosing a restriction may be directly welfare-relevant, for example because such choices signal positive personal attributes (either to the chooser or to others), or because it also depletes self control, thereby altering the benefit of the commitment. A third set of studies uses metachoices to estimate the direct welfare effects of leveraging peer comparisons and social image considerations (e.g., Allcott and Kessler, 2019; Butera et al., 2022). But those who feel accountable to their peers may consent to such comparisons out of a sense of social or moral obligation, despite wishing the option did not exist.

There are a variety of other welfare questions that standard choice-based methods cannot resolve, but that extensions of our approach may fruitfully inform. One example concerns non-standard choice patterns such as loss aversion or sunk cost effects. Are these patterns mistakes, or do they reflect welfare-relevant sensations? Through an extension of our approach, one could determine whether such patterns reflect anticipated CSAs (as opposed to some non-hedonic aspect of framing), whether those expectations are accurate, how people would choose if they held accurate expectations, and the magnitude of any associated welfare loss. Another example concerns the evaluation of welfare when people's identities and mindsets change endogenously over time (e.g., Akerlof and Kranton, 2000; Bernheim et al., 2021). If these phenomena reflect changes in mappings from consumption experiences (broadly construed) to mental states rather than changes in preferences over mental states, then our approach offers a path forward. The mental-states approach to behavioral welfare analysis may offer solutions to many important conceptual challeges such as these.

References

Akerlof, G. A. and R. E. Kranton (2000). Economics and Identity. *The Quarterly Journal of Economics 115*(3), 715–753. Publisher: Oxford University Press.

Allcott, H. and J. B. Kessler (2019). The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons. *American Economic Journal: Applied Economics 11*(1), 236–76.

Anderson, W. D. and M. L. Patterson (2008). Effects of social value orientations on fairness judgments. *The Journal of Social Psychology 148*(2), 223–246. Publisher: Taylor & Francis.

Andreoni, J. (1989). Giving with Impure Altruism: Applications to Charity and Ricardian Equivalance. *Journal of Political Economy 97*(6), 1447–1458.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The economic journal 100*(401), 464–477. Publisher: JSTOR.

Andreoni, J. (1995). Warm-Glow versus Cold-Prickle: The Effects of Positive and Negative Framing on Cooperation in Experiments. *Quarterly Journal of Economics 110*(1), 1–21.

Andreoni, J. and B. D. Bernheim (2009). Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects. *Econometrica 77*(5), 1607–1636.

Bartling, B., E. Fehr, and H. Herz (2014). The Intrinsic Value of Decision Rights. *Econometrica 82*(6), 2005–2039. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA11573.

Batson, C. D. (2011). *Altruism in humans*. Oxford University Press, USA.

Benjamin, D. J., K. Cooper, O. Heffetz, M. Kimball, and J. Zhou (2023). Adjusting for Scale-Use Heterogeneity in Self-Reported Well-Being. *working paper*.

Benjamin, D. J., O. Heffetz, M. S. Kimball, and N. Szembrot (2014, September). Beyond Happiness and Satisfaction: Toward Well-Being Indices Based on Stated Preference. *American Economic Review 104*(9), 2698–2735.

Benjamin, D. J., M. S. Kimball, O. Heffetz, and A. Rees-Jones (2012, August). What Do You Think Would Make You Happier? What Do You Think You Would Choose? *The American economic review 102*(5), 2083–2110.

Bernheim, B. D. (2016). The Good, the Bad, and the Ugly: A Unified Approach to Behavioral Welfare Economics. *Journal of Benefit-Cost Analysis 7*(1), 12–68.

Bernheim, B. D., L. Braghieri, A. Martínez-Marquina, and D. Zuckerman (2021, February). A Theory of Chosen Preferences. *American Economic Review 111*(2), 720–754.

Bernheim, B. D. and D. Taubinsky (2018). Behavioral Public Economics. In B. D. Bernheim, S. DellaVigna, and D. Laibson (Eds.), *The Handbook of Behavioral Economics*, Volume 1. New York: Elsevier.

Block, H. D. and J. Marschak (1960). Random Orderings and Stochastic Theories of Response. In I. Olkin (Ed.), *Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling*. Stanford University Press.

Bolton, G. E. and A. Ockenfels (2000). ERC: A theory of equity, reciprocity, and competition. *American economic review 90*(1), 166–193.

Broberg, T., T. Ellingsen, and M. Johannesson (2007, January). Is generosity involuntary? *Eco-*

nomics Letters 94(1), 32–37.

Butera, L., R. Metcalfe, W. Morrison, and D. Taubinsky (2022). Measuring the Welfare Effects of Shame and Pride. *American Economic Review 112*(1), 122–168.

Carrera, M., H. Royer, M. Stehr, J. Sydnor, and D. Taubinsky (2022). Who Chooses Commitment? Evidence and Welfare Implications. *Review of Economic Studies 89*(3), 1205–1244.

Charness, G. and M. Dufwenberg (2006). Promies and Partnership. *Econometrica 74*(6), 1579–1601.

Clark, A. and A. J. Oswald (2002). A simple statistical method for measuring how life events affect happiness. *International Journal of Epidemiology 31*(6), 1139–1144.

Dagsvik, J. K. and A. Karlstrom (2005, January). Compensating Variation and Hicksian Choice Probabilities in Random Utility Models that are Nonlinear in Income. *The Review of Economic Studies 72*(1), 57–76.

Dana, J., D. M. Cain, and R. M. Dawes (2006a, July). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes 100*(2), 193–201.

Dana, J., D. M. Cain, and R. M. Dawes (2006b, July). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes 100*(2), 193–201.

Deaton, A. (2018). What do self-reports of wellbeing say about life-cycle theory and policy? *Journal of Public Economics 162*, 18–25.

DellaVigna, S., J. A. List, and U. Malmendier (2012, February). Testing for Altruism and Social Pressure in Charitable Giving. *The Quarterly Journal of Economics 127*(1), 1–56.

Deutsch, M. (1960). The effect of motivational orientation upon trust and suspicion. *Human relations 13*(2), 123–139. Publisher: Sage Publications Sage UK: London, England.

DiTella, R., R. J. MacCulloch, and A. J. Oswald (2001, March). Preferences over Inflation and Unemployment: Evidence from Surveys of Happiness. *American Economic Review 91*(1), 335–341.

Exley, C. L. (2015, October). Excusing Selfishness in Charitable Giving: The Role of Risk. *The Review of Economic Studies 83*(2), 587–628. _eprint: https://academic.oup.com/restud/article-pdf/83/2/587/17417166/rdv051.pdf.

Fehr, E., H. Herz, and T. Wilkening (2013). The Lure of Authority: Motivation and Incentive Effects of Power. *American Economic Review 104*(4), 1325–1359.

Fehr, E. and K. M. Schmidt (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics 114*(3), 817–868. Publisher: MIT Press.

Finkelstein, A., E. F. P. Luttmer, and M. J. Notowidigdo (2013). What Good Is Wealth Without Health? The Effect of Health on the Satisfaction Derived from Consumption. *Journal of the European Economic Association 11*(1), 221–258.

Gillen, B., E. Snowberg, and L. Yariv (2019, August). Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study. *Journal of Political Economy 127*(4), 1826–1863. Publisher: The University of Chicago Press.

Gilovich, T. and V. H. Medvec (1995). The experience of regret: what, when, and why. *Psychological review 102*(2), 379. Publisher: American Psychological Association.

Gruber, J. and S. Mullainathan (2005). Do Cigarette Taxes Make Smokers Happier? *B.E. Journal of Economic Analysis and Policy 5*(1).

Gul, F. and W. Pesendorfer (2001). Temptation and Self-Control. *Econometrica 69*(6), 1403–1435. Publisher: [Wiley, Econometric Society].

Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica 50*(4), 1029–1054. Publisher: [Wiley, Econometric Society].

Heathwood, C. (2016). Desire-Fulfillment Theory. In G. Fletcher (Ed.), *The Routledge Handbook of Philosophy of Well-Being*. Routledge.

Iyengar, S. S. and M. R. Lepper (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology 79*(6), 995–1006. Place: US Publisher: American Psychological Association.

Josephs, R. A., R. P. Larrick, C. M. Steele, and R. E. Nisbett (1992). Protecting the self from the negative consequences of risky decisions. *Journal of personality and social psychology 62*(1), 26. Publisher: American Psychological Association.

Kelley, H. H. and J. W. Thibaut (1978). *Interpersonal relations: A theory of interdependence*. John Wiley & Sons.

Ketelaar, T. and W. Tung Au (2003). The effects of feelings of guilt on the behaviour of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction. *Cognition and emotion 17*(3), 429–453. Publisher: Taylor & Francis.

Koszegi, B. and M. Rabin (2008). Choices, situations, and happiness. *Journal of Public Economics 92*(8-9), 1821–1832.

Krusell, P., B. Kuruscu, and A. A. Smith, Jr. (2010). Temptation and Taxation. *Econometrica 78*(6), 2063–2084.

Lazear, E. P., U. Malmendier, and R. A. Weber (2012). Sorting in Experiments with Application to Social Preferences. *American Economic Journal: Applied Economics 4*(1), 136–163. Publisher: American Economic Association.

Lewis, Helen, B. (1971). Shame and guilt in neurosis. *Psychoanalytic review 58*(3), 419–438. Publisher: National Psychological Association for Psychoanalysis.

Lewis, M. (2008). Self-conscious emotions: Embarrassment, pride, shame, and guilt. Publisher: The Guilford Press.

Ludwig, J., G. J. Duncan, L. A. Gennetian, L. F. Katz, R. C. Kessler, J. R. Kling, and L. Sanbonmatsu (2012). Neighborhood effects on the long-term well-being of low-income adults. *Science 337*(6101), 1505–1510.

McKelvey, R. D. and T. R. Palfrey (1995, July). Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior 10*(1), 6–38.

Messick, D. M. and C. G. McClintock (1968). Motivational bases of choice in experimental games.

Journal of experimental social psychology 4(1), 1–25. Publisher: Elsevier.

Mill, J. S. (2012). *Utilitarianism*. Renaissance Classics.

Miller, R. S. and J. P. Tangney (1994). Differentiating embarrassment and shame. *Journal of Social and Clinical Psychology 13*(3), 273–287. Publisher: Guilford Press.

Nelissen, R. M., A. J. Dijker, and N. K. de Vries (2007). Emotions and goals: Assessing relations between values and emotions. *Cognition and Emotion 21*(4), 902–911. Publisher: Taylor & Francis.

Nozick, R. (1974). *Anarchy, State, and Utopia*. Basic Books.

Owens, D., Z. Grossman, and R. Fackler (2014). The Control Premium: A Preference for Payoff Autonomy. *American Economic Journal: Microeconomics 6*(4), 138–161.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Rabin, M. (1995). Moral Preferneces, Moral Constraints, and Self-Serving Biases. *working paper*.

Tangney, J. P. (1990). Assessing individual differences in proneness to shame and guilt: Development of the Self-Conscious Affect and Attribution Inventory. *Journal of personality and social psychology 59*(1), 102. Publisher: American Psychological Association.

Tangney, J. P. and R. L. Dearing (2002). *Shame and Guilt*. Guilford.

Terza, J. V., A. Basu, and P. J. Rathouz (2008, May). Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of Health Economics 27*(3), 531–543.

Tracy, J. L. and R. W. Robins (2004). "Putting the Self Into Self-Conscious Emotions: A Theoretical Model". *Psychological Inquiry 15*(2), 103–125. Publisher: Taylor & Francis.

Tversky, A. and E. Shafir (1992, November). Choice under Conflict: The Dynamics of Deferred Decision. *Psychological Science 3*(6), 358–361. Publisher: SAGE Publications Inc.

Van Dijk, E. (2015). The economics of prosocial behavior. *The Oxford handbook of prosocial behavior*, 86–99. Publisher: Oxford University Press Oxford.

Van Lange, P. A., D. De Cremer, E. Van Dijk, and M. Van Vugt (2007). Self-interest and beyond. *Social psychology: Handbook of basic principles*, 540–561. Publisher: Guilford press New York, NY.

Woodford, M. (2019). Modeling Imprecision in Perception, Valuation and Choice. *Annual Review of Economics 12*, 579–601.

Wubben, M. J., D. De Cremer, and E. van Dijk (2012). Is pride a prosocial emotion? Interpersonal effects of authentic and hubristic pride. *Cognition & Emotion 26*(6), 1084–1097. Publisher: Taylor & Francis.

Part

Online Appendix

Welfare and the Act of Choosing

B. Douglas Bernheim, Kristy Kim, and Dmitry Taubinsky

Table of Contents

A The Non-Comparability Problem with General Forms of Process Dependence

The Non-Comparability Problem is not limited to settings with constraint-set dependence and menu dependence. It also applies when preferences encompass other aspects of the decision process.

As a general matter, we can write any choice problem as a pair, (X, d), where X is the constraint set and d specifies all *details* concerning the process used for choosing an element of X. A detail is a characteristic of a decision problem, such as a decision tree or a feature of information presentation, that has nothing to do with the features of the available items. Within any such problem, there may be a variety of ways to select any given item.[20] We will use σ to denote a *trajectory*, defined as a particular combination of choices that leads to the selection of some item. For any decision problem (X, d) and item $x \in X$, there is a set of trajectories, $\Sigma_x(X, d)$ (*x-trajectories*) that yield the item x.[21]

We will assume throughout that an individual cares about the selected item, x. Her preferences are *constraint-dependent* if she also cares intrinsically about X, *detail-dependent* if she also cares intrinsically about d, and *trajectory-dependent* if she also cares intrinsically about σ. Each of these possibilities is an aspect of *process-dependent* preferences. Menu dependence, studied in the preceding section, is a form of trajectory dependence.[22]

To allow for arbitrary process dependence, we assume preferences are defined over objects of the form (X, d, σ, x), where $x \in X$ and $\sigma \in \Sigma_x(X, d)$. In other words, an individual potentially cares about the item she ends up with (x), the set of potential alternatives (X), the structure of the decision problem (d), and the particular combination of component choices that delivered the selected item, σ.

We can now state the general Non-Comparability Problem. If an individual chooses $x^*(X, d)$ and $\sigma^*(X, d) \in \Sigma_{x^*(X,d)}(X, d)$ when presented with the problem (X, d), we can conclude only that

$$(X, d, \sigma^*(X, d), x^*(X, d)) \succeq (X, d, \sigma, x)$$

for all $x \in X$ and $\sigma_x \in \Sigma_x(X, d)$. It follows that for two distinct decision problems, (X, d) and (X', d'), an individual's choices provide us with no basis for determining whether she is better off with $(X, d, \sigma^*(X, d), x^*(X, d))$ or $(X', d', \sigma^*(X', d'), x^*(X', d'))$. Consequently, we can never say whether a policy that changes the decision problem facing an individual helps or hurts her. This conclusion follows even if the domain of preferences does not encompass X or σ (so that they exhibit neither constraint dependence nor menu dependence), provided people care about other details of choice encapsulated in d. As before, metachoices do not resolve this problem. A metachoice between (X, d) and (X', d') is simply a new choice problem of the form (X'', d''), where $X'' = X \cup X'$ and

[20] Recall that, for the example in which Norma chooses between salad and pizza, there are three distinct ways to select salad and three distinct ways to select pizza.

[21] For the sake of simplicity, we abstract from decision processes that include random elements. Such processes may be of interest, and accomodating them requires a more general frameworrk.

[22] A trajectory corresponds to a sequence of menus $\mathcal{M}^{[k]}, \mathcal{M}^{[k-1]}, ..., M^{[0]}$. Once again consider the example described in footnote 20. One S-trajectory involves choosing S in the first period. In that case, $\mathcal{M}^{[1]}$ is $\{\{S\}\}$ and $\mathcal{M}^{[0]}$ is $\{S\}$. Another S-trajectory involves choosing S in the second period. In that case, $\mathcal{M}^{[1]}$ is $\{\{S\}, \{P\}, \{S, P\}\}$ and $\mathcal{M}^{[0]}$ is $\{S\}$. The final S-trajectory involves choosing S in the third period. In that case, $\mathcal{M}^{[1]}$ is $\{\{S\}, \{P\}, \{S, P\}\}$ and $\mathcal{M}^{[0]}$ is $\{S, P\}$. In a level-k decision problem, the level-k menu is a fixed feature of the problem, one that implies the constraint set, X. Thus, level-k menu dependence is most appropriately classified as constraint dependence and detail dependence, rather than trajectory dependence.

d'' captures the fact that the decision is now structured as a choice between two "continuation procedures." There is no opportunity for the choices in this new setting to reveal an individual's preferences between an *unchosen assignment* to one decision problem or the other.

B Psychological Literature Motivating our Choice of CSAs

Social Value Orientation and Emotional Motivations

Psychologists have long studied how social distributive preferences affect interdependent situations (Kelley and Thibaut, 1978; Ketelaar and Tung Au, 2003). The social values orientation (SVO) framework, defined as the preferences over distributional outcomes between the self and others (Van Dijk, 2015), provides a guide to determine what kinds of motivations, affects, and cognitive processes take place tradeoffs between prosocial and self-interested motivations. SVOs are operationally defined by maximizing welfare for the self and others (prosocial), maximizing one's own well-being (individualistic), and maximizing relative well-being (competitive) (Deutsch, 1960; Messick and McClintock, 1968). Batson (2011) breaks down the preferences of maximizing one's own well-being (individualistic) and others (prosocial) into four motivations: principlism, egoism, altruism, and collectivism. Principlism is the motive to uphold some moral principle, egoism is the motive to maximize one's own welfare, altruism is the motive to maximize other's welfare, and collectivism is the motive to maximize group welfare. Moral self-conscious emotions (e.g. guilt and shame) are intimately tied with egoism and principlism since their self-evaluative nature serve as inhibitors of selfish tendencies (Batson, 2011; Lewis, 2008). Using the aforementioned frameworks and theories, we provide justifications of our four measures below. We begin with individualistic and competitive social preferences.

Financial Satisfaction. In line with a pro-self orientation (both individualistic and competitive view) and the standard economic theory of rationality, we elicit the participant's level of financial satisfaction. Because our experiment involves a simple distribution of dollar amounts between the self and another unknown individual, the level of financial satisfaction serves as a proxy for objective self-interest. We argue that eliciting self-interest through "financial satisfaction" is relatively objective and minimizes possibly biases from loaded framing.

We now examine possible motives for prosocial preferences.

Fairness. A plethora of research has shown evidence of individuals' propensities towards egalitarian distributions. Van Lange et al. (2007) proposes an integrative framework of prosocial behavior via collectivism and principlism, based on the evidence that prosocial individuals choose outcomes in terms of maximization of joint outcomes and equality in outcomes. Furthermore, Anderson and Patterson (2008) illustrate that fairness judgments are utilized for both prosocial and pro-self orientations, though in different ways.

Guilt. Batson (2011) defines avoidance of self-punishment as an egotistical motivation to inhibit selfish tendencies; self-punishment often takes the form of moral/self-conscious emotions which are self-evaluative emotions induced from a specific event (Lewis, 2008). Guilt is often used as a mechanism to ensure prosocial behavior (Ketelaar and Tung Au, 2003; Nelissen et al., 2007). Specifically, the anticipation of guilt often serves as a functional emotion that induces avoidance of self-interest actions (Ketelaar and Tung Au, 2003; Nelissen et al., 2007). The anticipation of guilt, a consequential emotion, is intimately tied with mixed-motive social dilemmas, hence is a first-order emotion for all SVOs.

Pride. Lastly, we incorporate a positive, self-conscious moral emotion: pride. Pride can be distinguished into two categories: hubristic pride and achievement-oriented pride (Lewis, 2008; Tangney,

1990; Tracy and Robins, 2004), both of which affect decisions in interpersonal situations. Hubristic pride (synonymous to "conceit" or "arrogance") is associated with antisocial or aggressive behavior and relates to individualistic orientations. Achievement-oriented pride (synonymous to "confident" or "successful") is associated with prosocial behavior (Wubben et al., 2012). Evidence points to these two categories being "semantically and experientially distinct" (Tracy and Robins, 2004). We argue that the component of pride which is relevant in our experiment is achievement-oriented pride because it is tied with the interpersonal actions taken. Measures of pride stemming from the decision in our experiment does not include hubristic pride, given our sample is random and the level of hubris within-participant is constant. Hence, we argue that an elicitation of "pride" will capture achievement-oriented pride and not hubristic pride.

Other Possible Measures

Provisionally, we do not include three widely studied behavioral responses in the interpersonal literature: regret, empathy and shame.

Regret is a self-conscious emotional response to discovering "better" alternatives and realizing the degree of control over a "negative" outcome (Gilovich and Medvec, 1995). Anticipated regret is a key feature in decisions which incorporate risk (Josephs et al., 1992). Given that our experiment gives participants full control over deterministic outcomes, regret is not a first-order emotion in this setting. Even if it is still possible to feel regret for one's choices, we assume any dissatisfaction from a simple, non-repeated, anonymous game would relate to choosing a division that was more "fair" or that provided more "financial satisfaction."

Empathy has also been widely studied as an important component of altruism and prosocial behavior (Batson, 2011). In a similar vein of why we exclude hubristic pride, the level of empathy one has should not change for each iteration of the game, as there is nothing in our experiment which evokes changes in levels of social-connectedness.

Shame is a common self-conscious moral emotion that is studied along with guilt and pride. Though shame and guilt may be induced by similar events, shame is often distinguished as being more public in nature (Miller and Tangney, 1994) and as more of a transgression of self rather than of behavior (Lewis, 1971). As such, shame is not a primary emotional response to our experiment. Furthermore, shame and guilt are colloquially related, so if some shame was implicated during this experiment, we suspect measures of guilt and shame to be nearly-equivalent from the perspective of our participants.

C Supplementary Descriptive Results

C.1 Sample Demographics

Table A1: Study sample demographics

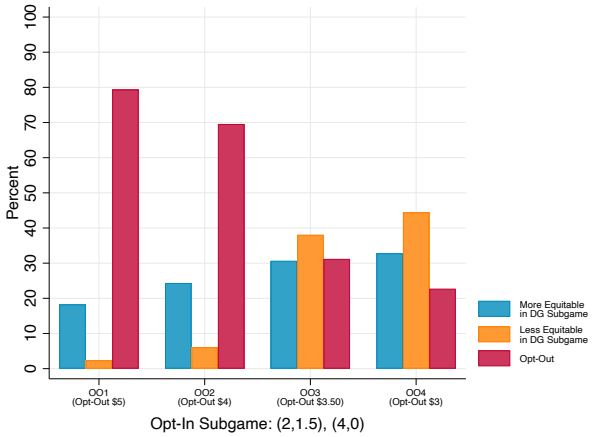
	Count	Percent
Gender		
Female	1,453	53
Male	1,264	46
Other	13	0
Decline to State	10	0
Age		
18-24	166	6
25-39	1,249	46
40-60	1,021	37
60+	296	11
Decline to state	8	0
Education		
High school graduate	10	0
Some college	243	9
Vocational / trade / technical school	114	4
Bachelor's degree	733	27
Advanced degree	1,166	43
Decline to state	463	17
Less than high school	11	0
Household Income		
$0 - $19,999	265	10
$20,000 - $39,999	530	19
$40,000 - $59,999	564	21
$60,000 - $79,999	452	16
$80,000 - $99,999	342	12
$100,000 - $119,999	207	8
$120,000 or more	323	12
Decline to state	57	2
Total	2,740	100

Note: This table reports the demographic characteristics of the study sample. The count column reports the number of participants while the percent column reports the percent of participants that fall under each demographic category.

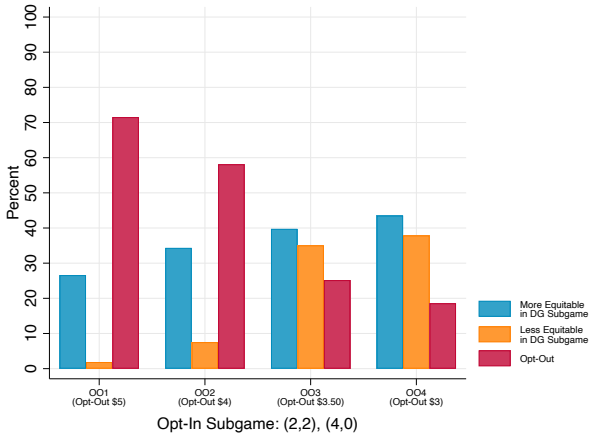
C.2 Additional Results About Choice in the OO games

Figure A1: Distribution of choices in the opt-out games by subgame

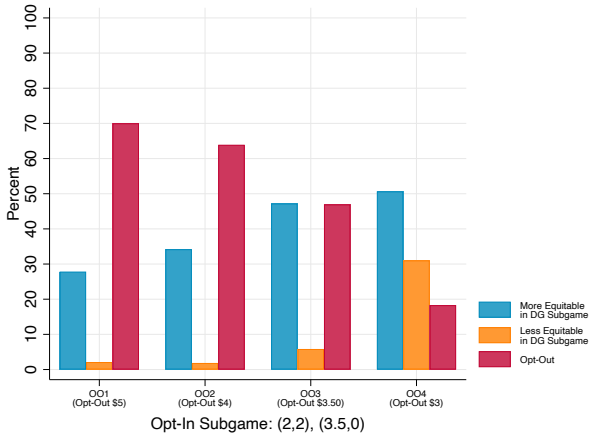
(a) Subgame DG 3: (2,1.5) vs. (4,0)



(b) Subgame DG 4: (2,2) vs. (4,0)



(c) Subgame DG 5: (2,2) vs. (3.5,0)



Note: This figure shows the distribution of choices made in each of the choice sets in the OO games. Panel (a) presents the likelihood with which each option was chosen in each OO game where DG 3 was the opt-in subgame. Panels (b) and (c) present analogous results OO games where DG 4 and DG 5, respectively, were the opt-in subgames.

D Supplementary Results on the Relationship Between Choice and CSAs

D.1 Including a Constant Term in Table 1 Regressions

To test the the orthogonality assumption of our model as outlined in Section 4.1, we can examine how a constant term in our logit regression may affect the coefficients of the CSA vector. The intuition is that if there are omitted CSAs that are relevant to choices but not encompassed by our CSAs, including a constant term should result in (1) a large constant coefficient and (2) changes to the coefficients of other CSAs. We recreate Table 1 in Columns (1), (3), and (5) in Appendix Table A2 and estimate a similar regression in Columns (2), (4), and (6), respectively, with a constant term included. We find small changes in the coefficients and...

Table A2: Association between choices and CSAs

	(1) Logit (no cons) Choosing More Equitably	(2) Logit (w cons) Choosing More Equitably	(3) IV Logit (no cons) Choosing More Equitably	(4) IV Logit (w cons) Choosing More Equitably	(5) Logit (no cons) Choosing More Equitably	(6) Logit (w cons) Choosing More Equitably
Δ Guilt	-0.85***	-0.96***	-0.73***	-0.90***	-1.11***	-1.24***
	(0.12)	(0.12)	(0.26)	(0.26)	(0.12)	(0.12)
Δ Pride	0.08	0.06	0.04	-0.12	0.58***	0.55***
	(0.12)	(0.12)	(0.22)	(0.22)	(0.12)	(0.12)
Δ Finan. Satis.	1.85***	1.66***	1.74***	1.42***	3.82***	3.56***
	(0.14)	(0.15)	(0.32)	(0.33)	(0.14)	(0.14)
Δ Fairness	0.13	0.33**	-0.28	0.15	0.20	0.44***
	(0.13)	(0.13)	(0.37)	(0.37)	(0.13)	(0.12)
Δ Unfairness	-0.66***	-0.72***	-1.18***	-1.08**	-0.71***	-0.80***
	(0.13)	(0.13)	(0.44)	(0.43)	(0.13)	(0.13)
Δ Happiness	2.05***	2.03***	2.32***	2.38***		
	(0.16)	(0.16)	(0.44)	(0.44)		
Δ Satisfaction	2.60***	2.57***	3.38***	3.44***		
	(0.17)	(0.17)	(0.45)	(0.45)		
Constant				-0.32***		
				(0.09)		
N. Participants: 2365						

Note: This table estimates logit regressions analogous to Table 1. Columns (1), (3), and (5) contain the regressions from Columns (1), (2), and (3) of table 1, except that the coefficients are reported as log odds. Columns (2), (4), and (6) report the regressions from Columns (1), (3), and (5), respectively, but include a constant term. Standard errors are reported in the parentheses, and calculated using bootstrap with 1,000 resampling clusters at the participant level. $^*p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.001$.

D.2 Principal Components Analysis of the CSAs

We assess how the principal components of the CSAs explain variation in the CSA ratings that we elicited and choices that we observe in the data. Table A3 summarizes the principle component analysis, and shows that Component 1 and Component 2 explain a large proportion of the variance in CSAs.

We can then apply the methodology from Section 4 to the first two factors from Table A3. Figure A2 shows that the estimates that we obtain for Decider's money-metric utility are similar to the

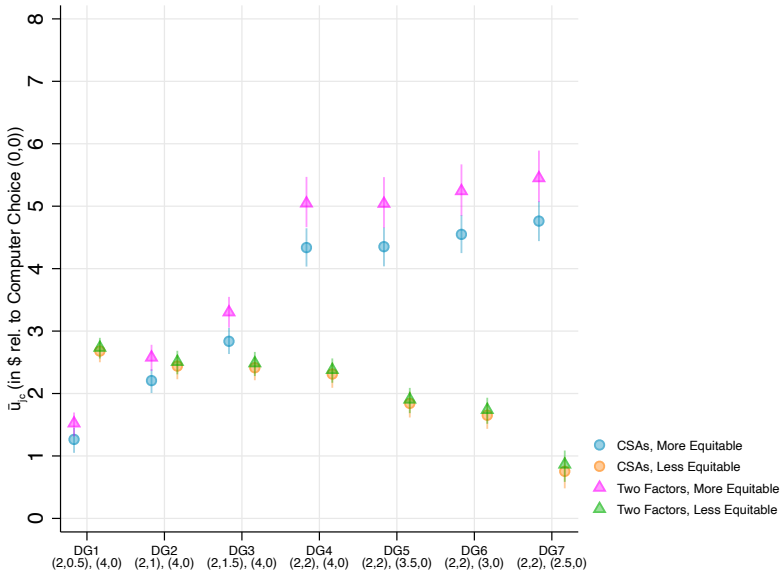
ones we obtain when using all seven CSAs. This again suggests that the first two factors appear to span much of the relevant space of mental states.

Table A3: Principal Components of the 7 CSAs

	Eigenvalue	Proportion of Variance	Cumulative
Component 1	3.19	0.46	0.46
Component 2	1.93	0.28	0.73
Component 3	0.75	0.11	0.84
Component 4	0.38	0.05	0.89
Component 5	0.32	0.05	0.94
Component 6	0.23	0.03	0.97
Component 7	0.20	0.03	1.00

Note: This table reports the eigenvalues and explanatory variance from the principal component analysis of the seven CSAs. The first column reports the eigenvalues of the correlation matrix, ordered by size. The second column reports the percent variation of the CSAs explained by each component. The last column reports the cumulative explanatory variance.

Figure A2: Deciders' average utility using CSAs vs principal components in utility



Note: This figure reports the average money-metric utility, $\bar{u}_{jc}$, for each option of each choice set in the DG. Each point in the figure contains the same sample composition. The blue and orange (rounded) points represent the average utilities obtained from applying the methodology in Section 4 to all 7 CSAs. The pink and green (non-rounded) points represent the average utilities estimated from applying the methodology in Section 4 to the first two factors from Table A3 only. The 95 percent confidence intervals are reported as the vertical bars, and calculated using bootstrap with 1,000 resampling clusters at the participant level.

Table A5: Testing the significance of additional CSAs with satisfaction

	(1) IV Logit Choosing More Equitably on Satisfaction and Guilt	(2) IV Logit Choosing More Equitably on Satisfaction and Pride	(3) IV Logit Choosing More Equitably on Satisfaction and Finan.	(4) IV Logit Choosing More Equitably on Satisfaction and Fair.	(5) IV Logit Choosing More Equitably on Satisfaction and Unfair.
T-Stat for Additional CSA	-18.21	13.02	15.90	18.88	-19.94
Hansen's J chi2	306.7	434.0	387.5	368.1	292.2
Hansen's J p-value	3.96e-65	1.25e-92	1.41e-82	2.16e-78	5.22e-62

Note: This table reports tests of whether the disaggregated CSA listed in the column title is relevant for predicting choice when there is a (measurement error free) measure of overall satisfaction.

D.3 Correlated Measurement Error

Tables A4 and A5 report on an alternative strategy to support the evidence that happiness and satisfaction alone do not encompass all welfare-relevant CSAs. If it were true that that happiness and satisfaction were the only welfare-relevant CSAs, then the other five CSAs by definition form valid instruments for happiness and satisfaction: they are strongly correlated with happiness and satisfaction (relevance) and, under the null hypothesis, are independent of choice conditional on the true values of happiness or satisfaction (exclusion restriction). Thus, the null hypothesis implies that, in an IV regression of choice on happiness (or satisfaction) and one other CSA k, with happiness instrumented by the remaining four CSAs, the coefficient of CSA k should be zero, and the model should pass over-identification tests.

Each column of Table A4 reports results from a Logit regression where the dependent variable is choosing more equitably and the covariates are (i) the difference in reported happiness between the more and less equitable allocation and (ii) Δ_{ikc}, where Δ_{ikc} is the kth CSA from the vector Δ_{ic}, for k corresponding to one of the five disaggregated CSAs. The difference in reported happiness is instrumented with the other four disaggregated CSA differences, $\Delta_{ik'c}$, where k' indexes the other four disaggregated CSAs. Because there are multiple instruments, we can report an overidentification test (Hansen's J), which is listed in each column for each of the five Logit regressions. In each column we also list the t-statistic for the null hypothesis that the coefficient of Δ_{ikc} is zero. Under the null hypothesis that CSA k is irrelevant once happiness is controlled for, the model would pass the overidentification test and the null hypothesis that the coefficient of Δ_{ikc} is zero would not be rejected. Instead, the table shows dramatic rejections of the null hypothesis and dramatic failures of the overidentification test. Table A5 reports on analogous analysis for overall satisfaction.

Table A4: Testing the significance of additional CSAs with happiness

	(1) IV Logit Choosing More Equitably on Happiness and Guilt	(2) IV Logit Choosing More Equitably on Happiness and Pride	(3) IV Logit Choosing More Equitably on Happiness and Finan.	(4) IV Logit Choosing More Equitably on Happiness and Fair.	(5) IV Logit Choosing More Equitably on Happiness and Unfair.
T-Stat for Additional CSA	-16.80	10.94	16.56	18.15	-18.81
Hansen's J chi2	342.7	449.9	351.5	351.1	317.8
Hansen's J p-value	6.53e-73	4.56e-96	8.40e-75	1.03e-74	1.56e-67

Note: This table reports tests of whether the disaggregated CSA listed in the column title is relevant for predicting choice when there is a (measurement error free) measure of happiness.

D.4 Stability of Mapping from CSAs to Choice

Table A6 reports estimates a multinomial logit regression of choices on the associated coefficients of the CSA vector and on interaction variables between them and an indicator for OO. This regression pools the choices and CSAs reported in the DG and OO games. We find that, with the exception of financial satisfaction, the OO setting has no significant bearing on the CSA coefficients in the model that is specified in Section 4.1. (Recall also that we show in Section 5.4 that the OO-estimated CSA coefficients perform as well as DG-estimated CSA coefficients in predicting the likelihood of a participant choosing more equitably in the DG.)

Figures A3 and A4 are analogous to Figure 4, but concern choice in the OO games. The figures plot the actual versus predicted likelihood of choosing the more equitable allocation in the opt-in subgame (figure A3) and of choosing to opt out (figure A4). Panel (a) in each of the figures forms predictions by estimating our empirical model on DG data only, and thus constitutes a test of out-of-sample fit. Panel (b) in both figures forms predictions by estimating our empirical model on OO data only, and thus constitutes a test of in-sample fit.

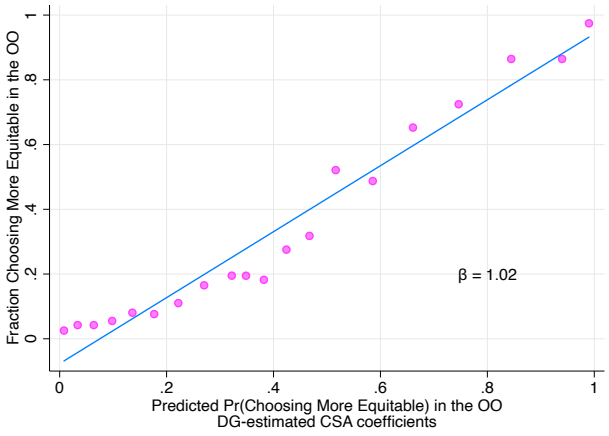
Table A6: Coefficients of the CSA vector in DG and OO games

Dependent Var.	(1) Mult. Logit Choice
Guilt	-0.849***
	(0.118)
Pride	0.081
	(0.123)
Finan. Satis.	1.852***
	(0.144)
Fairness	0.134
	(0.129)
Unfairness	-0.661***
	(0.129)
Happiness	2.051***
	(0.161)
Satis.	2.597***
	(0.176)
Guilt × (Opt-Out Game)	-0.073
	(0.168)
Pride × (Opt-Out Game)	0.250
	(0.178)
Finan. × (Opt-Out Game)	1.088***
	(0.229)
Fairness × (Opt-Out Game)	-0.334
	(0.176)
Unfairness × (Opt-Out Game)	-0.173
	(0.184)
Happiness × (Opt-Out Game)	-0.246
	(0.263)
Satis. × (Opt-Out Game)	-0.021
	(0.290)
N. Participants	2365

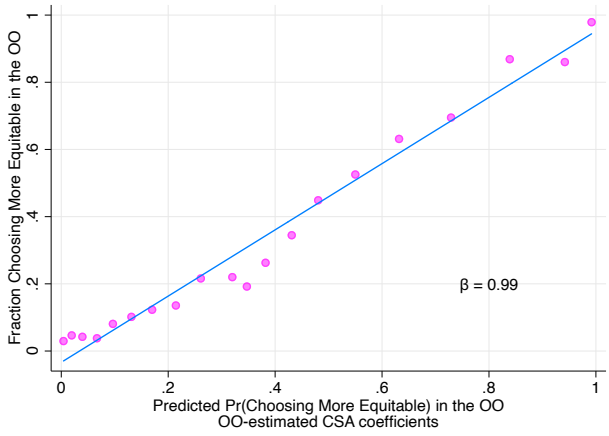
Note: This table reports coefficients of the CSAs in the OO game vs the DG. The table reports the average marginal effects from an alternative-specific conditional logit of choosing a more equitable option on the seven CSAs and on interactions of the seven CSAS with an indicator for whether or not they were reported for the OO game. The regression pools the choice and CSA data from the DG and OO game in the main CSA elicitation module. Standard errors are reported in the parentheses, and calculated using bootstrap with 1,000 resampling clusters at the participant level. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.001$.

Figure A3: Observed versus fitted probabilities of choosing more equitably in the OOs

(a) Model estimated using DG data



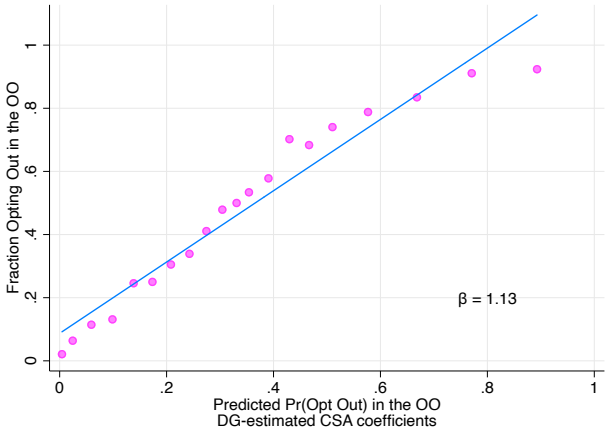
(b) Model estimated using OO data



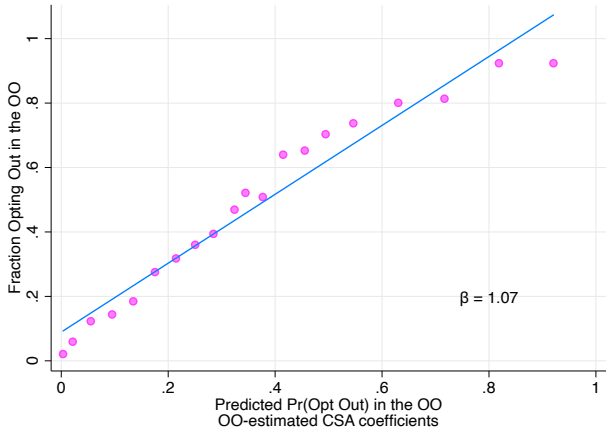
Note: This figure compares the model's predicted likelihood and the empirical likelihood of choosing the more equitable option (from the opt-in subgame) in the OO games. The predicted probabilities are divided into 20 bins, and the figure reports the average the empirical likelihood of choosing the more equitable option for each of those bins. Panel (a) estimates the model using the DG CSAs and reports the out-of-sample fit of the model. Panel (b) estimates the model using the OO CSAs and thus reports the in-sample fit. The blue line represents the 45 degree line, wherein all points would lie if the model was a perfect fit.

Figure A4: Observed versus fitted probabilities of opting out in the OOs

(a) Model estimated using DG data



(b) Model estimated using OO data



Note: This figure compares the model's predicted likelihood and the empirical likelihood of choosing the opt-out option in the OO games. The predicted probabilities are divided into 20 bins, and the figure reports the average the empirical likelihood of option out for each of those bins. Panel (a) estimates the model using the DG CSAs and reports the out-of-sample fit of the model. Panel (b) estimates the model using the OO CSAs and thus reports the in-sample fit. The blue line represents the 45 degree line, wherein all points would lie if the model was a perfect fit.

E Additional Results on Using Happiness and Satisfaction as Proxies for Welfare

E.1 Relationship of Happiness and Satisfaction to the Other CSAs

Table A7 reports how happiness and satisfaction may aggregate guilt, pride, financial satisfaction, fairness, and unfairness. We find that happiness and satisfaction similarly aggregate the other five CSAs, with positive CSAs, such as financial satisfaction, being heavily weighted. We compare this with how the same five CSAs are associated with choices in the DG in Table A8. This table reports the same two columns found in Appendix Table A7 and a third column which reports the weights of the five CSAs on choice in the DG. The coefficients in each column are normalized such that the financial satisfaction coefficient is 1. Again, we find that happiness and satisfaction have similar ways of aggregating the other five CSAs; however, those weights are different from how the five CSAs are weighted in choices. Guilt and unfairness matter relatively more in determining choices than happiness and satisfaction, while pride and fairness are relatively less important for choice than for happiness and satisfaction

Table A7: Regressions of happiness and satisfaction on the disaggregated CSAs

	(1) OLS	(2) OLS
Dependent Var.	Δ Happiness	Δ Satisfaction
Δ Guilt	-0.11***	-0.09***
	(0.01)	(0.01)
Δ Pride	0.19***	0.12***
	(0.01)	(0.01)
Δ Finan. Satis.	0.59***	0.59***
	(0.01)	(0.01)
Δ Fairness	0.04***	0.03***
	(0.01)	(0.01)
Δ Unfairness	-0.04***	-0.03**
	(0.01)	(0.01)
N. Participants	2365	2365

Note: This table reports the associations of guilt, pride, financial satisfaction, fairness, and unfairness with the ratings of happiness and satisfaction, respectively, in the DG. Column (1) reports the OLS coefficients from a regression of relative happiness on the five relative CSAs, where relative refer to the ratings of the CSAs for the more equitable option relative to the less equitable option. Column (2) reports the OLS coefficients from a regression of relative satisfaction on on the five relative CSAs. Standard errors are reported in the parentheses, and calculated using bootstrap with 1,000 resampling clusters at the participant level. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.001$.

Table A8: Regressions of happiness, satisfaction, and choice on the disaggregated CSAs

Dependent Var.	(1) OLS Δ Happiness	(2) OLS Δ Satisfaction	(3) Logit Choosing More Equitably
Δ Guilt	-0.18***	-0.14***	-0.29***
	(0.02)	(0.02)	(0.03)
Δ Pride	0.32***	0.21***	0.15***
	(0.03)	(0.02)	(0.03)
Δ Finan. Satis.	1.00	1.00	1.00
	(.)	(.)	(.)
Δ Fairness	0.06***	0.06***	0.05
	(0.02)	(0.02)	(0.03)
Δ Unfairness	-0.06***	-0.05**	-0.19***
	(0.02)	(0.02)	(0.03)
N. Participants: 2365			

Note: This table reports the coefficients of the CSA vector on relative happiness (Column 1), relative satisfaction (Column 2), and choices made in the DG (Column 3). Columns (1) and (2) report OLS coefficients, normalized by its financial satisfaction coefficient; while Column (3) reports the logit coefficients (in log-odds), normalized by the financial its financial satisfaction coefficient. Standard errors are reported in the parentheses, and calculated using bootstrap with 1,000 resampling clusters at the participant level. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.001$.

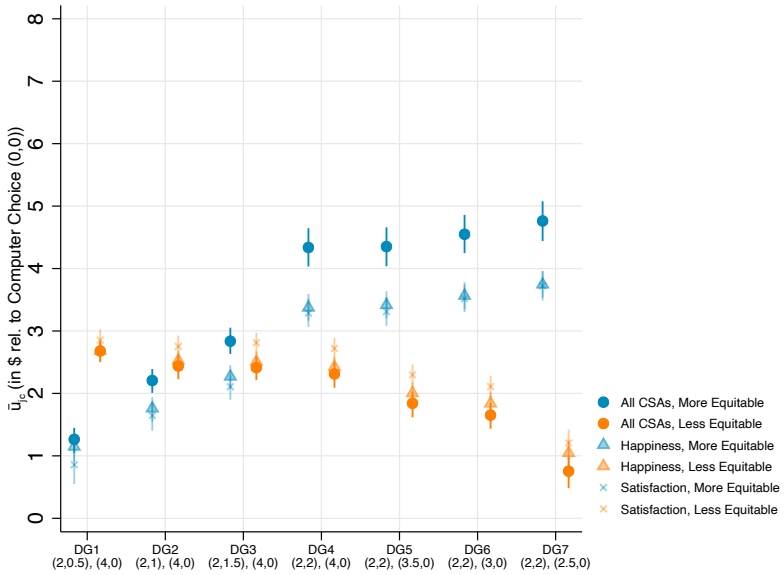
E.2 Bias from Using Happiness and Satisfaction to Estimate Deciders' Welfare

In this section, we compare welfare estimates using all CSAs to welfare estimates that assume that happiness or satisfaction are sufficient statistics for welfare. In the model, a Decider chooses alternative j if it maximizes

$$U = v(X_{ijc}) + \varepsilon_{ijc}.$$

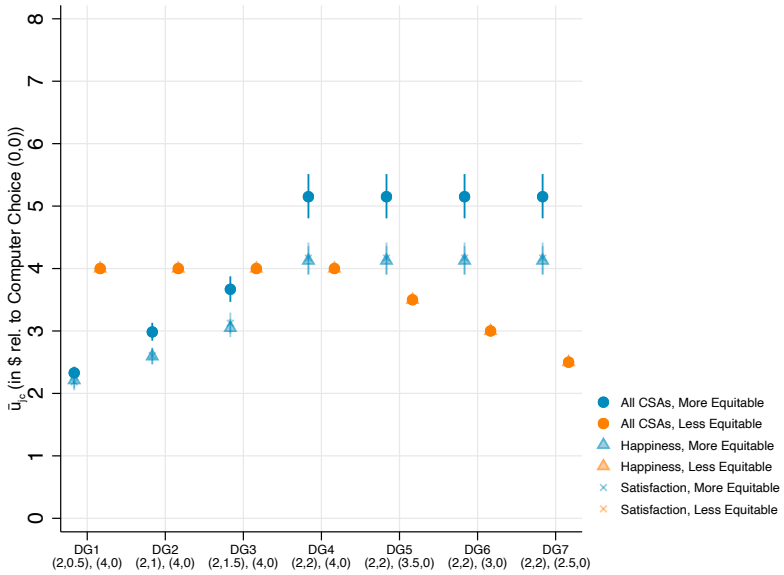
The term $v(X_{ijc})$ is the observable, deterministic component of utility and is locally linear and additive in the CSAs, such that $v(X_{ijc}) = X_{ijc}\beta$, where β is a vector of coefficients and X_{ijc} is a vector of all seven CSAs . When we assume happiness and satisfaction are sufficient statistics, then $v(X_{ijc}) = X_{ijc}\beta$, where X_{ijc} just denotes the normalized happiness or satisfaction rating. Thus, using this model, we can construct three versions of $\bar{u}_{jc}$: (1) one which includes the vector of the seven CSAs, (2) one which includes only happiness ratings, and (3) one which includes only satisfaction ratings. These three versions of $\bar{u}_{jc}$ are graphed together in Figures A5, A6, and A7 for the DG, OO, and CC games, respectively. Notably, welfare estimates including only happiness or satisfaction are very similar to each other. However, they deviate from our main estimates (which are baed on all seven CSAs) in a several important ways: (1) happiness and satisfaction underestimate welfare gains from choosing a more equitable option and (2) happiness and satisfaction overestimate the welfare gains from choosing a less equitable option. This can be explained by the fact that happiness and satisfaction measures overweigh positive CSAs relative to negative ones. See Appendix Table A8 for more details.

Figure A5: Comparing welfare estimates in the DG module using all CSAs versus just happiness or satisfaction



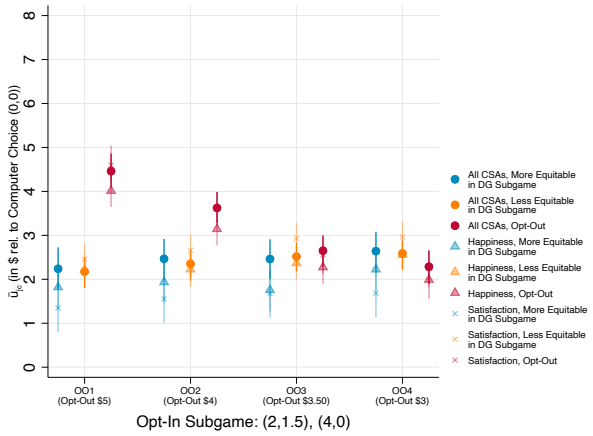
Note: This figure reports the average utility for each option of each choice set in the DG, using both actual and counterfactual choices. Thus, each point in the graph contains the utilities of all participants. The rounded points report the average utilities calculated by our hybrid method. The non-rounded points report the average utilities using ratings of happiness only (triangle points) and satisfaction only (x points). The 95 percent confidence intervals are reported as the vertical bars, and calculated using bootstrap with 1,000 resampling clusters at the participant level.

Figure A6: Comparing welfare estimates in the main CC module using all CSAs versus just happiness or satisfaction

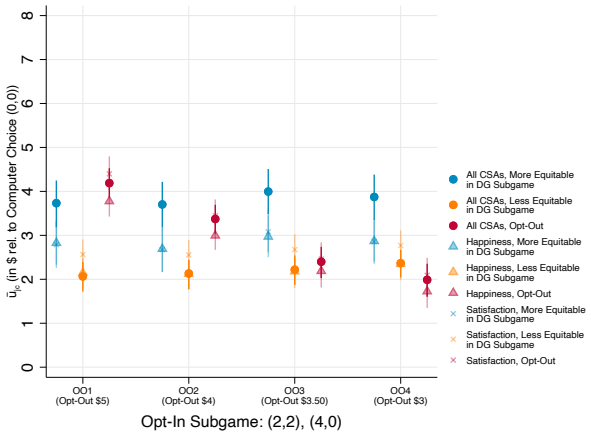


Note: This figure reports the average utility for each option of each choice set in the CC. The figure averages the utilities from the CC version where the computer shows a singleton choice. To make the graphs comparable, we place them on a DG axis; however, the alternative option is not shown to the participant. Hence, the utilities for the more equitable option in DG4-DG7 are the same, and the utilities for the less equitable option in DG1-DG4 are the same. The rounded points report the average utilities calculated using our hybrid method. The non-rounded points report the average utilities using ratings of happiness only (triangle points) and satisfaction only (x points). The 95 percent confidence intervals are reported as the vertical bars, and calculated using bootstrap with 1,000 resampling clusters at the participant level.

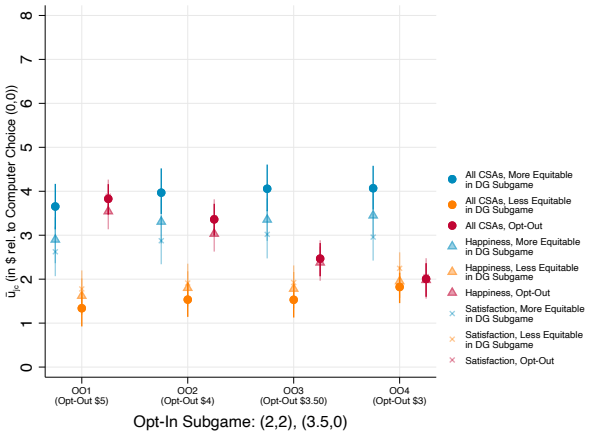
Figure A7: Comparing welfare measures using the hybrid method vs. happiness/satisfaction in the OO



(a) Subgame: (2,1.5) vs (4,0)



(b) Subgame: (2,2) vs (4,0)



(c) Subgame: (2,2) vs (3.5,0)

Note: This figure reports the average utility for each option of each choice set in the OO. The figures are separated by the opt-in subgame that the participant saw: DG3, DG4, or DG5. The rounded points report the average utilities calculated using our hybrid method. The non-rounded points report the average utilities using ratings of happiness only (triangle points) and satisfaction only (x points). The 95 percent confidence intervals are reported as the vertical bars, and calculated using bootstrap with 1,000 resampling clusters at the participant level.

F Additional Robustness Checks

F.1 Anticipation of Experienced CSAs

Table A9 examines whether eliciting CSA ratings prior to the decision alters the coefficients of the CSA vector in our model. The table presents estimates of the model in Table, 1, but with interactions with an indicator for whether or not the participant was in the "ex-ante" module. We find that this alternative CSA elicitation format does not meaningfully change the weights on any one CSA.

Table A9: Effect of the ex-ante module on coefficients of the CSAs

Dependent Var.	(1) Logit Choosing More Equitably	(2) Logit Choosing More Equitably
Δ Guilt	-0.13***	-0.13***
	(0.02)	(0.02)
Δ Pride	0.02	0.01
	(0.02)	(0.02)
Δ Finan. Satis.	0.29***	0.28***
	(0.02)	(0.02)
Δ Fairness	0.02	0.02
	(0.02)	(0.02)
Δ Unfairness	-0.10***	-0.10***
	(0.02)	(0.02)
Δ Happiness	0.32***	0.31***
	(0.02)	(0.02)
Δ Satisfaction	0.39***	0.39***
	(0.02)	(0.02)
Δ Guilt × Ex-Ante Arm		-0.01
		(0.06)
Δ Pride × Ex-Ante Arm		0.04
		(0.07)
Δ Finan. Satis. × Ex-Ante Arm		0.09
		(0.08)
Δ Fairness × Ex-Ante Arm		0.02
		(0.07)
Δ Unfairness × Ex-Ante Arm		0.06
		(0.07)
Δ Happiness × Ex-Ante Arm		0.10
		(0.10)
Δ Satis. × Ex-Ante Arm		-0.00
		(0.10)
Observations	17864	17864
N. Participants	2552	2552

Note: Coefficients are reported as average marginal effects, which are computed by averaging the change in the predicted probability of choosing the more equitable option when a CSA's normalized rating changes from 0 to 1, and all other CSAs are held constant for each participant. The sample includes the main elicitation module and the ex-ante module. Standard errors are reported in the parentheses, and calculated using bootstrap with 1,000 resampling clusters at the participant level. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.001$.

Table A10: Correlation between present and future CSAs

	Present-Future Correlation
Guilt	0.88
Pride	0.94
Financial Satis	0.84
Fairness	0.93
Unfairness	0.90
Happiness	0.84
Satisfaction	0.84
Observations	7896
N. Participants	188

Note: This table displays correlation between present and future CSAs for all reported CSA in the DG, CC, and OO. The sample constitutes those that saw an alternative CSA elicitation where the participants were asked to disaggregate their ratings between the present and future.

F.2 Aggregation Across Present and Future CSAs

Below we examine how inter-temporally disaggregating CSAs may affect our analysis. Table A10 shows that CSAs elicited for the present and CSAs elicited for the future are highly correlated with one another (ranging from 0.84 to 0.94). Further, we find that when estimating the coefficients of the CSA vector separately using present and future CSAs, the coefficients are nearly identical as shown in Table A11.

Table A11: Present and Future coefficient of the CSAs

Dependent Var.	(1) Logit Choosing More Equitably (Present)	(2) Logit Choosing More Equitably (Future)
Δ Guilt	-0.18***	-0.16**
	(0.05)	(0.05)
Δ Pride	0.06	0.09
	(0.06)	(0.07)
Δ Finan. Satis.	0.26***	0.25***
	(0.06)	(0.07)
Δ Fairness	0.09*	0.12*
	(0.05)	(0.05)
Δ Unfairness	-0.03	-0.04
	(0.05)	(0.05)
Δ Happiness	0.16*	0.12
	(0.07)	(0.07)
Δ Satisfaction	0.46***	0.34***
	(0.07)	(0.07)
Choice Set FE	No	Yes
Observations	1316	1316
N. Participants	188	188

Note: This table reports a version of Column (1) of Table 1, but using data only from the subset of the participants in the present-future CSA elicitation arm of our experiment. Column (1) reports estimates when using the present CSAs, while Column (2) reports estimates when using the future CSAs. Standard errors are reported in the parentheses, and clustered at the participant level. $^*p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.001$.

F.3 Further Tests of the Orthogonality Assumption

Table A12: Choice Set Effect on Relative CSAs

	(1) Reg $\Delta_{ijc}\hat{\beta}$.
Choice Set 2	0.55***
	(0.03)
Choice Set 3	0.85***
	(0.03)
Choice Set 4	1.59***
	(0.04)
Choice Set 5	1.82***
	(0.04)
Choice Set 6	2.00***
	(0.04)
Choice Set 7	2.51***
	(0.05)
Constant	-0.66***
	(0.04)
Adj. R-Squared	0.123
Observations	16555
N. Participants	2365

Note: This table reports the choice set fixed effects on our estimate of the utility difference between choosing the more versus less equitable allocation in the DGs, $\Delta_{ijc}\hat{\beta}$. Standard errors are reported in the parentheses, and clustered at the participant level. *$p < 0.1$, **$p < 0.05$, ***$p < 0.001$.

Table A13: Association between Deciders' choices and CSAs

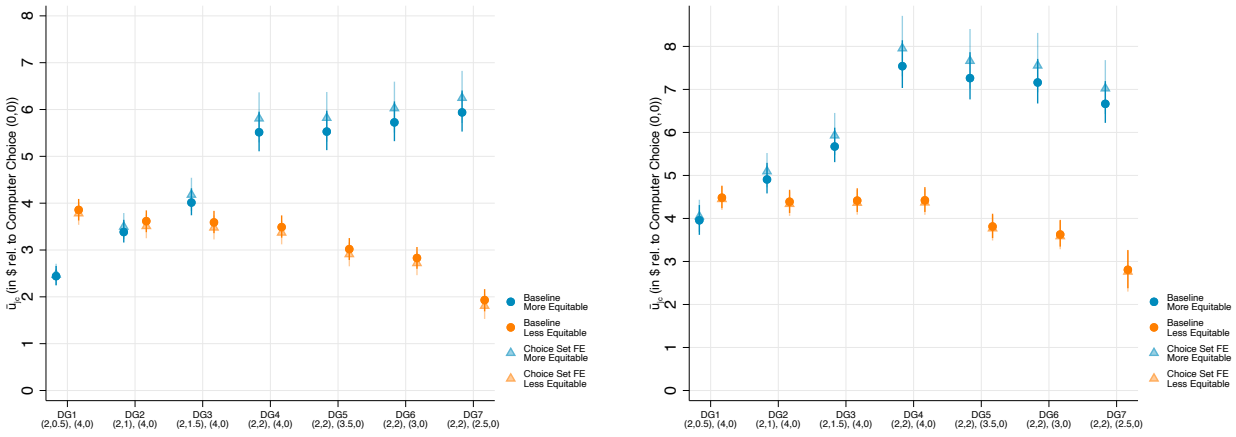
	(1) Logit Choosing More Equitably	(2) FE Logit Choosing More Equitably
Δ Guilt	-0.13***	-0.13***
	(0.02)	(0.02)
Δ Pride	0.01	0.01
	(0.02)	(0.02)
Δ Finan. Satis.	0.28***	0.24***
	(0.02)	(0.02)
Δ Fairness	0.02	0.02
	(0.02)	(0.02)
Δ Unfairness	-0.10***	-0.11***
	(0.02)	(0.02)
Δ Happiness	0.31***	0.30***
	(0.02)	(0.02)
Δ Satisfaction	0.39***	0.38***
	(0.02)	(0.02)
Choice Set FE	No	Yes
Pseudo R-Squared	0.33	0.34
Observations	33110	33110
N. Participants	2365	2365

Note: Column (1) is identical to Column (1) of Table 1. Column (2) is a variation of Column (1) that include choice set fixed effects. Standard errors are reported in the parentheses, and clustered at the participant level. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.001$.

Figure A8: Deciders' average utility in dictator games

(a) Full-sample results (chosen and counterfactual alternatives)

(b) Restricting only to the chosen options



Note: This figure extends Figure 5 by also including estimate of money-metric utility based on the coefficients from Column (2) of Table A13. The original estimates are in circles, while the new estimate based on Column (2) are in triangles. The 95 percent confidence intervals are reported as the vertical bars, and calculated using bootstrap with 1,000 resampling clusters at the participant level.

G Heterogeneity in the Marginal Utility of Money

By our assumptions, we have that

$$v(X_{ij_yc_0}) = m_i y + v(X_{ij_0c_0}).$$

To estimate the distribution of m_i, we estimate the mixed effects model

$$\hat{v}(X_{ij_xc_0}) = m_i y + \alpha_i + \varepsilon_{ijc}$$

where α_i captures individual-level variation in $v(X_{ij_0c_0})$ and ε_{ijc} captures idiosyncratic noise, such as that coming from trebles in people's reporting of the CSAs. We assume that $m_i \sim N(\bar{m}, \sigma_m^2)$ and $\alpha_i \sim N(\bar{\alpha}, \sigma_\alpha^2)$. We estimate this model via a standard maximum likelihood estimator to find that $\bar{m} = 0.46$ (95% CI $= [0.42, 0.51]$)), and $\sigma^2 = 0.09$ (95% CI $= [0.19, 0.30]$). By comparison, if we assume homogeneity, $m_i \equiv \bar{m}$, and apply a standard OLS estimator, we obtain $\bar{m} = 0.46$ (95% CI $= [0.42, 0.52]$).

H Study Instructions Appendix

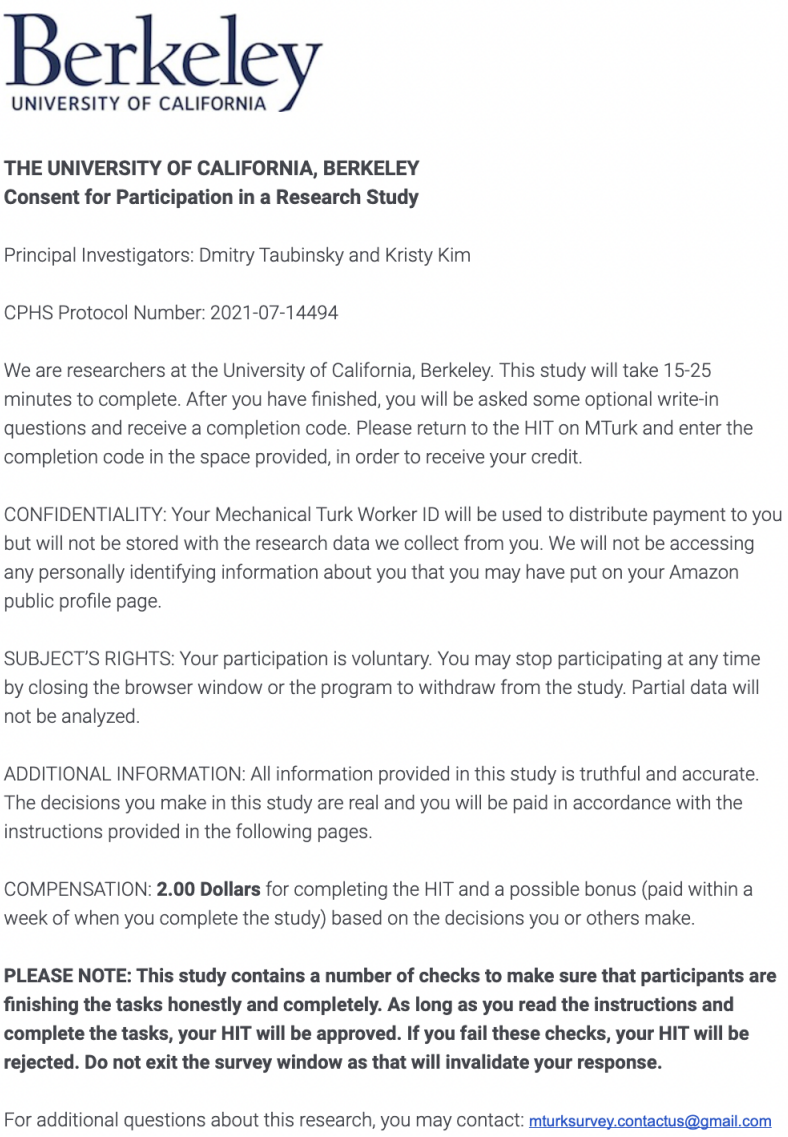
Welfare and the Act of Choosing

B. Douglas Bernheim, Kristy Kim, and Dmitry Taubinsky

Introduction and Consent

Study participants were recruited via Amazon Mechanical Turk (MTurk) for a "15-25 minute survey about preferences for money allocation to yourself and another person." Qualifications required being 18 years of age or older, having a high approval rating, completing more than 100 surveys, and residing in the United States. A web link to the study opened the Qualtrics survey with the following introduction and consent screen (Figure S1).

Figure S1: Consent and Introduction for the Study



Berkeley
UNIVERSITY OF CALIFORNIA

THE UNIVERSITY OF CALIFORNIA, BERKELEY
Consent for Participation in a Research Study

Principal Investigators: Dmitry Taubinsky and Kristy Kim

CPHS Protocol Number: 2021-07-14494

We are researchers at the University of California, Berkeley. This study will take 15-25 minutes to complete. After you have finished, you will be asked some optional write-in questions and receive a completion code. Please return to the HIT on MTurk and enter the completion code in the space provided, in order to receive your credit.

CONFIDENTIALITY: Your Mechanical Turk Worker ID will be used to distribute payment to you but will not be stored with the research data we collect from you. We will not be accessing any personally identifying information about you that you may have put on your Amazon public profile page.

SUBJECT'S RIGHTS: Your participation is voluntary. You may stop participating at any time by closing the browser window or the program to withdraw from the study. Partial data will not be analyzed.

ADDITIONAL INFORMATION: All information provided in this study is truthful and accurate. The decisions you make in this study are real and you will be paid in accordance with the instructions provided in the following pages.

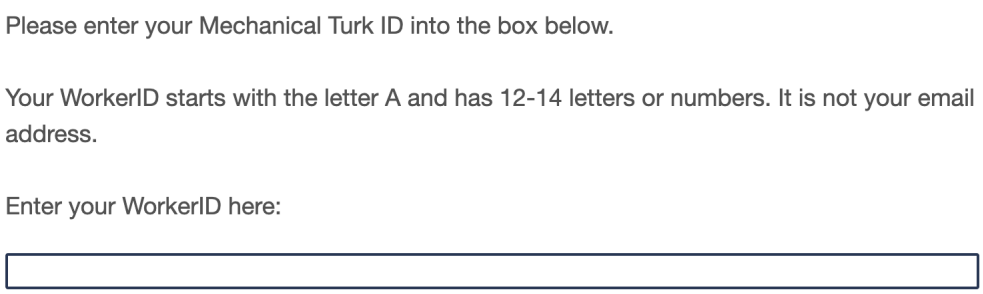
COMPENSATION: **2.00 Dollars** for completing the HIT and a possible bonus (paid within a week of when you complete the study) based on the decisions you or others make.

PLEASE NOTE: This study contains a number of checks to make sure that participants are finishing the tasks honestly and completely. As long as you read the instructions and complete the tasks, your HIT will be approved. If you fail these checks, your HIT will be rejected. Do not exit the survey window as that will invalidate your response.

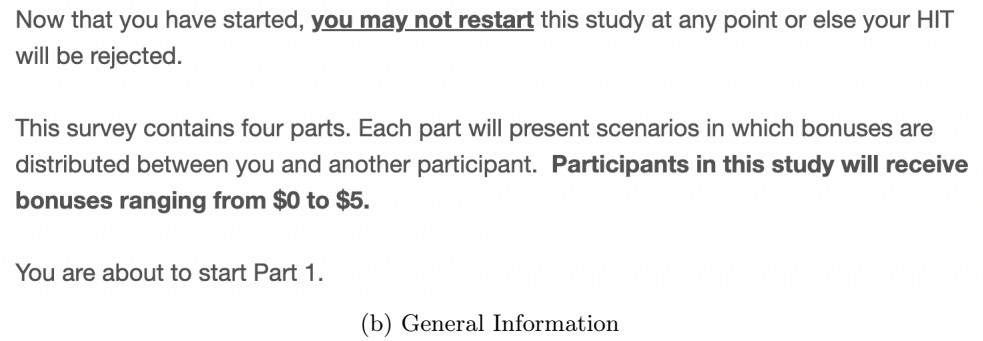
For additional questions about this research, you may contact: mturksurvey.contactus@gmail.com

Following a confirmation of age and consent, participants were asked to provide their MTurk ID and then shown some general information about the survey (Figure S2).

Figure S2: Introduction Continued



(a) MTurk ID Submission Box



(b) General Information

Main Study Arms

The Dictator Game

The instructions for the Dictator Game are shown in Figure S3 and a sample question is shown in Figure S4. Participants were asked to choose between two bonus allocations for themselves and a partner.

Figure S3: Dictator Game Instructions

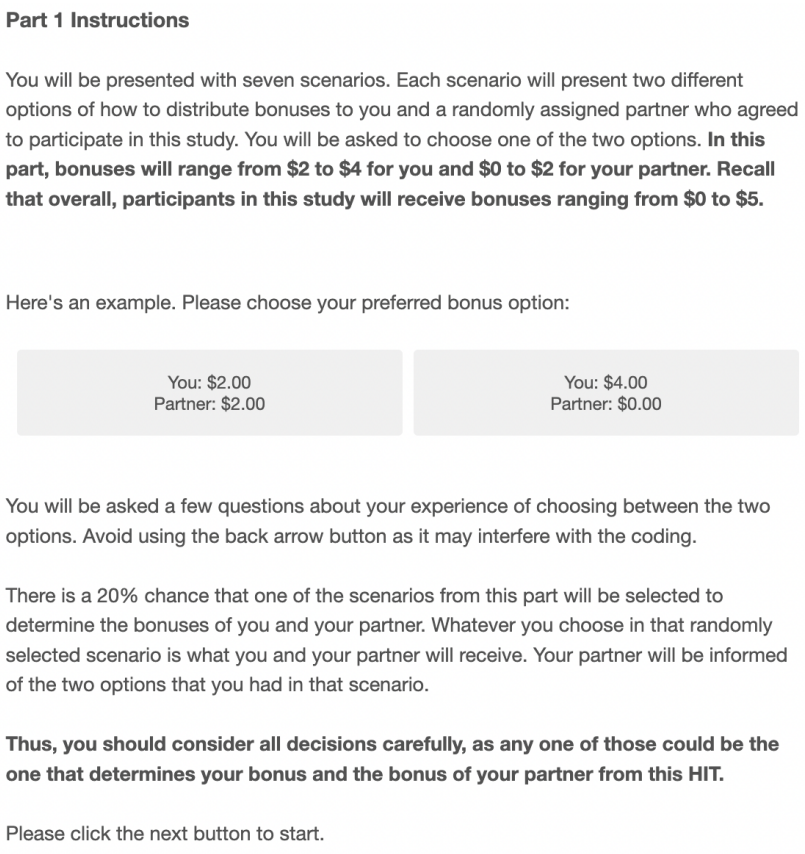
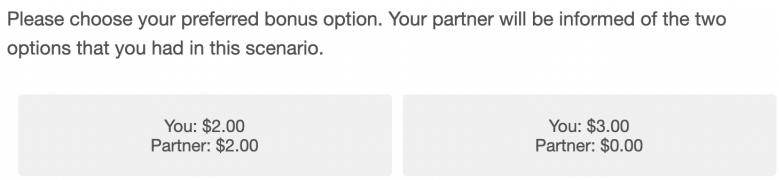
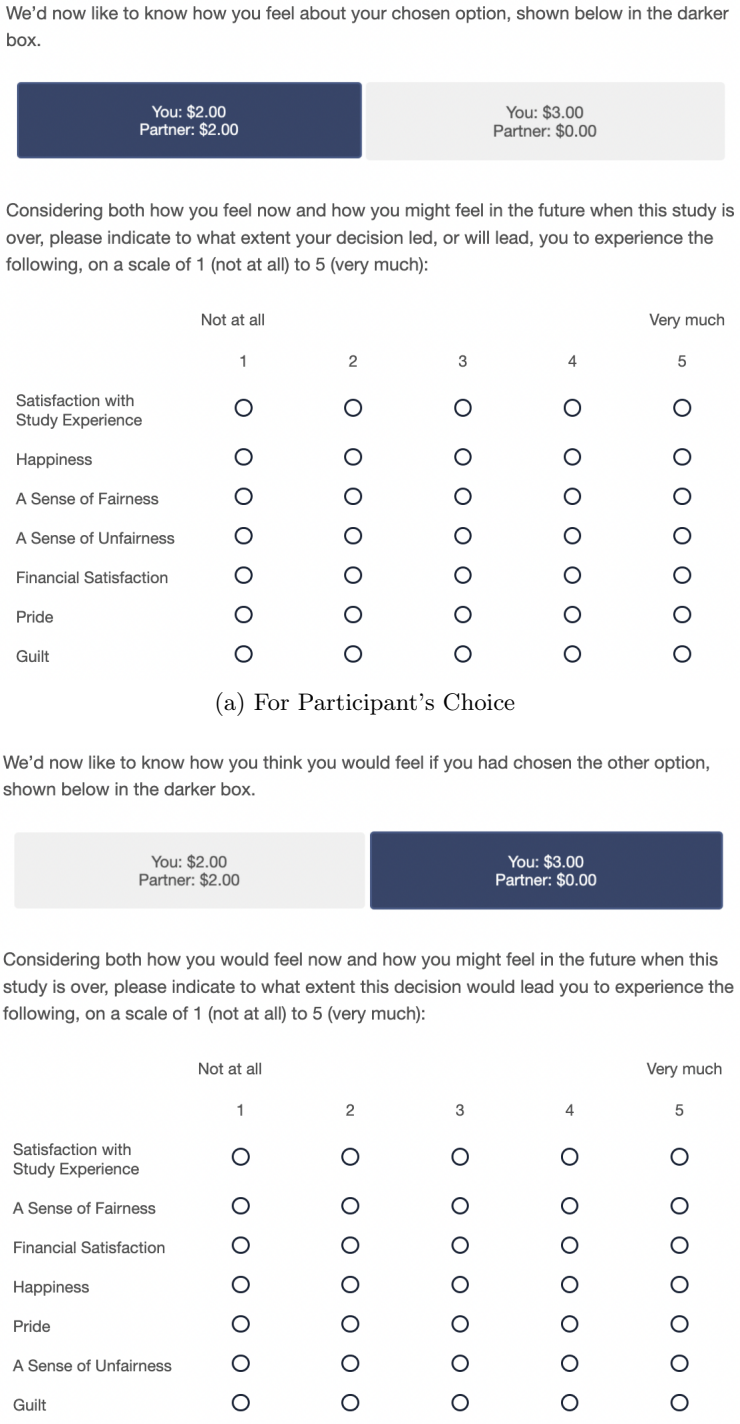


Figure S4: Dictator Game Sample Question



Following each choice, participants were prompted to report their levels of guilt, pride, financial satisfaction, sense of fairness, sense of unfairness, happiness, and satisfaction (henceforth referred to as "categorical subjective appraisals," or CSAs) with the study experience for (1) their choice and (2) the alternative had they chosen it ("counterfactual choice"), as shown in Figure S5 .

Figure S5: Dictator Game CSA Elicitations



(a) For Participant's Choice

(b) For Participant's Counterfactual Choice

Main Computer Choice Module

The instructions for the main Computer Choice module are shown in Figure S6. A sample question is shown in Figure S7.

Figure S6: Computer Choice (Displayed Choice) Instructions

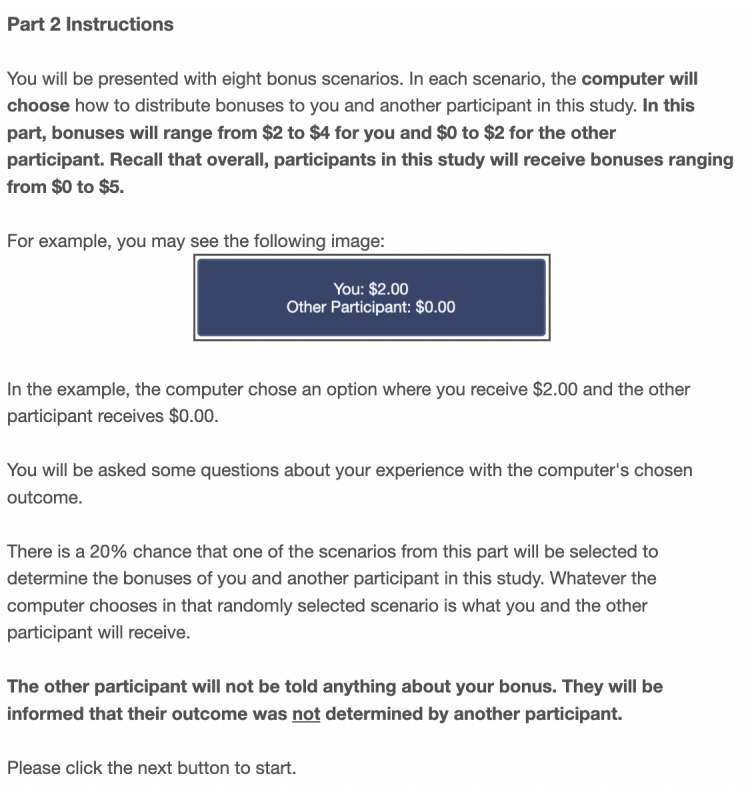
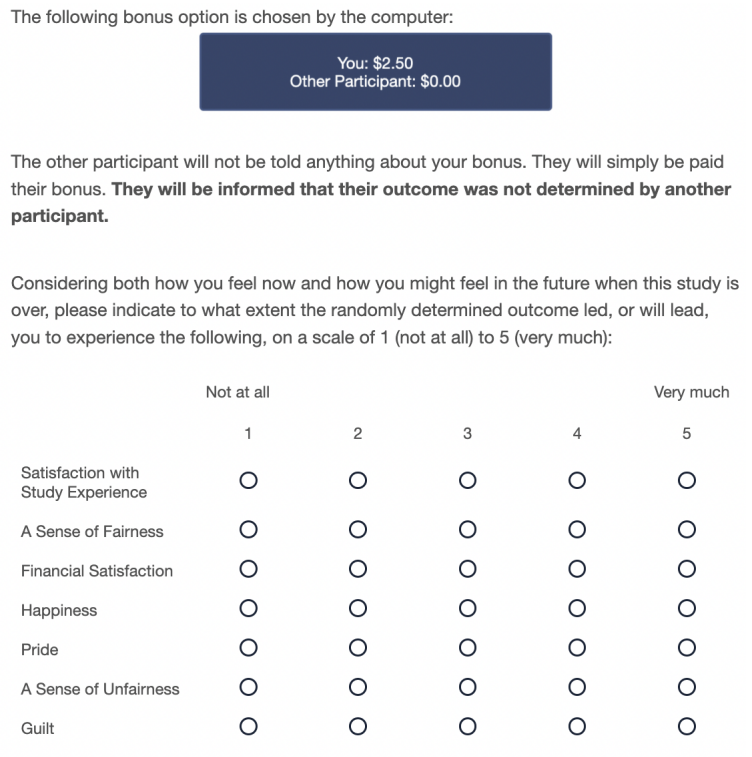


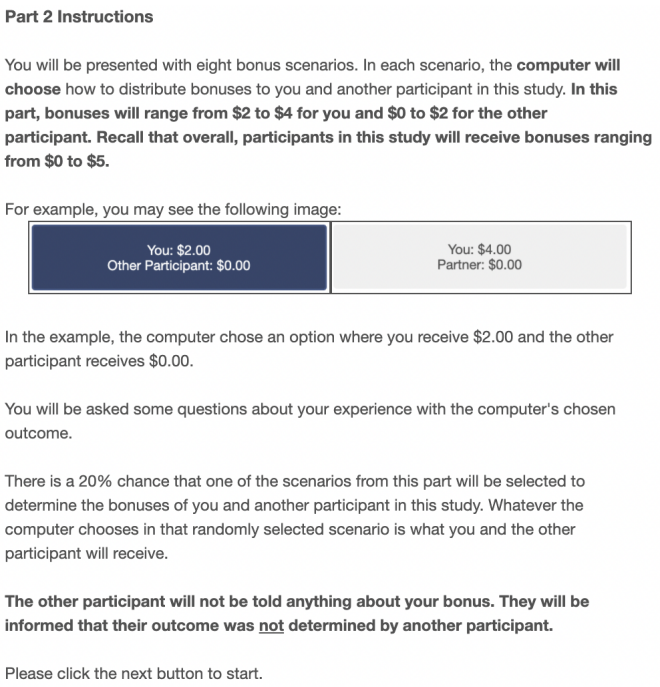
Figure S7: Computer Choice (Displayed Choice) CSA Elicitation



Alternative Computer Choice Module with a Displayed Choice Set

The instructions for the Computer Choice with a displayed choice are shown in Figure S8. A sample question is shown in Figure S9.

Figure S8: Computer Choice (Displayed Choice Set) Instructions



Part 2 Instructions

You will be presented with eight bonus scenarios. In each scenario, the **computer will choose** how to distribute bonuses to you and another participant in this study. **In this part, bonuses will range from $2 to $4 for you and $0 to $2 for the other participant. Recall that overall, participants in this study will receive bonuses ranging from $0 to $5.**

For example, you may see the following image:

You: $2.00 Other Participant: $0.00	You: $4.00 Partner: $0.00

In the example, the computer chose an option where you receive $2.00 and the other participant receives $0.00.

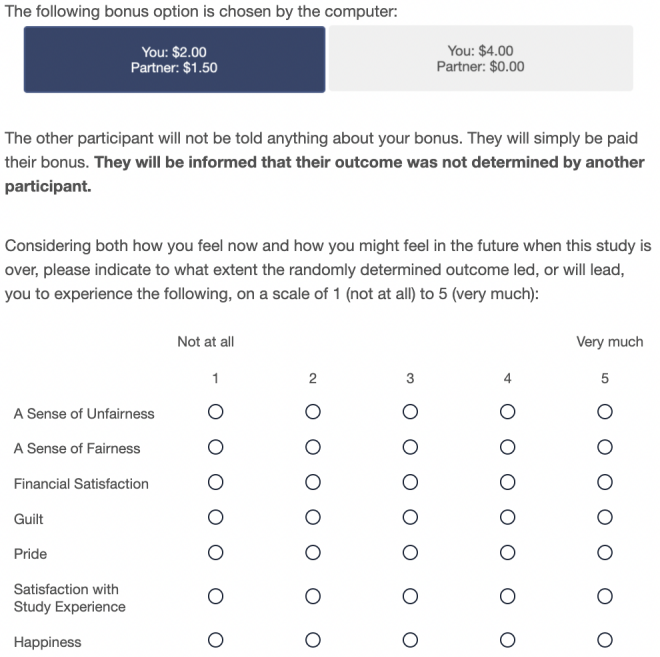
You will be asked some questions about your experience with the computer's chosen outcome.

There is a 20% chance that one of the scenarios from this part will be selected to determine the bonuses of you and another participant in this study. Whatever the computer chooses in that randomly selected scenario is what you and the other participant will receive.

The other participant will not be told anything about your bonus. They will be informed that their outcome was not determined by another participant.

Please click the next button to start.

Figure S9: Computer Choice (Displayed Choice Set) CSA Elicitation



The following bonus option is chosen by the computer:

You: $2.00 Partner: $1.50	You: $4.00 Partner: $0.00

The other participant will not be told anything about your bonus. They will simply be paid their bonus. **They will be informed that their outcome was not determined by another participant.**

Considering both how you feel now and how you might feel in the future when this study is over, please indicate to what extent the randomly determined outcome led, or will lead, you to experience the following, on a scale of 1 (not at all) to 5 (very much):

	Not at all				Very much
	1	2	3	4	5
A Sense of Unfairness	○	○	○	○	○
A Sense of Fairness	○	○	○	○	○
Financial Satisfaction	○	○	○	○	○
Guilt	○	○	○	○	○
Pride	○	○	○	○	○
Satisfaction with Study Experience	○	○	○	○	○
Happiness	○	○	○	○	○

Each participant was shown seven computer choices in random order. The choice sets were identical to the seven shown in the Dictator Game, but the choices that the computer makes were randomized in a way

such that each unique choice had an equal probability of being seen, as shown in Table S1. This maintained comparability to the main CC module.

Table S1: Probabilites for Each Option in the Computer Choice With a Displayed Choice Set

Choice Set	Pr(Prosocial)	Pr(Profit-maximizing)
1: ($2.00, $0.50) or ($4.00, $0.00)	7/8	1/8
2: ($2.00, $1.00) or ($4.00, $0.00)	7/8	1/8
3: ($2.00, $1.50) or ($4.00, $0.00)	7/8	1/8
4: ($2.00, $2.00) or ($4.00, $0.00)	1/2	1/2
5: ($2.00, $2.00) or ($3.50, $0.00)	1/8	7/8
6: ($2.00, $2.00) or ($3.00, $0.00)	1/8	7/8
7: ($2.00, $2.00) or ($2.50, $0.00)	1/8	7/8

Opt-Out Games

The instructions for the Opt-Out Games are shown in Figure S10 and a sample question is shown in Figure S11.

Figure S10: Opt-Out Game Instructions

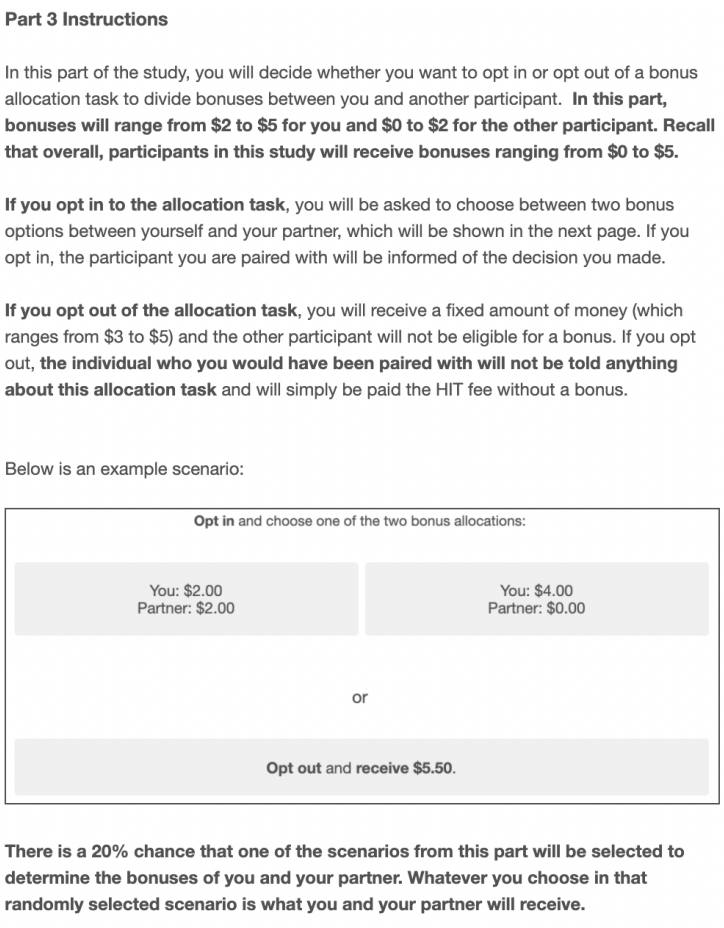
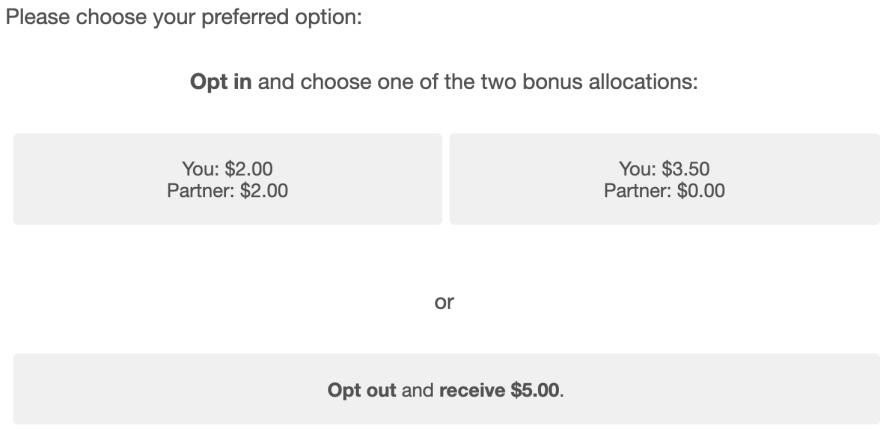


Figure S11: Opt-Out Game Example



Following each choice, participants were prompted to report their CSAs for (1) their choice and (2) the other two alternatives had they chosen it ("counterfactual choice"), as shown in Figure S13. The example of the second counterfactual question is not shown to avoid redundancy.

Figure S12: Opt-Out Game CSA Elicitation

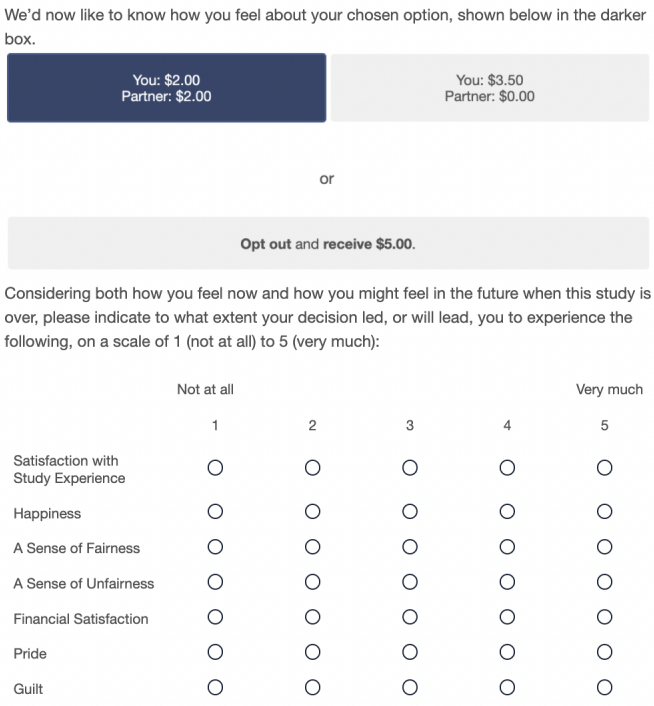
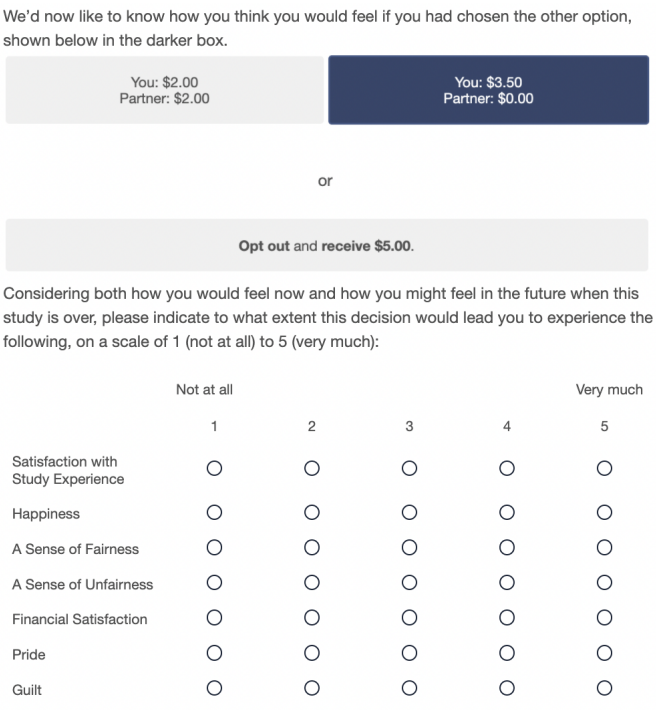
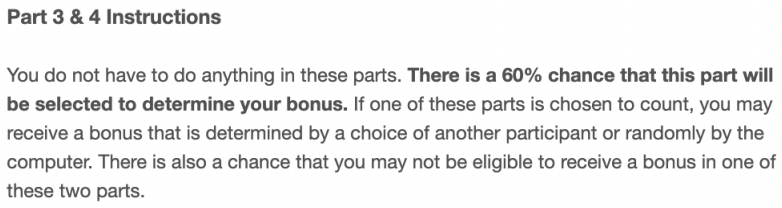


Figure S13: Opt-Out Game CSA Elicitation (Counterfactual Choice Example)



To ensure that the participant's potential partner would not be aware of the opt-out choice, each participant who played the Opt-Out Game was paired with another individual who did not play the Opt-Out Game. Those who did not see the Opt-Out Game were shown the screen in Figure S14.

Figure S14: Instructions for Participants Who Do Not View the Opt-Out Game

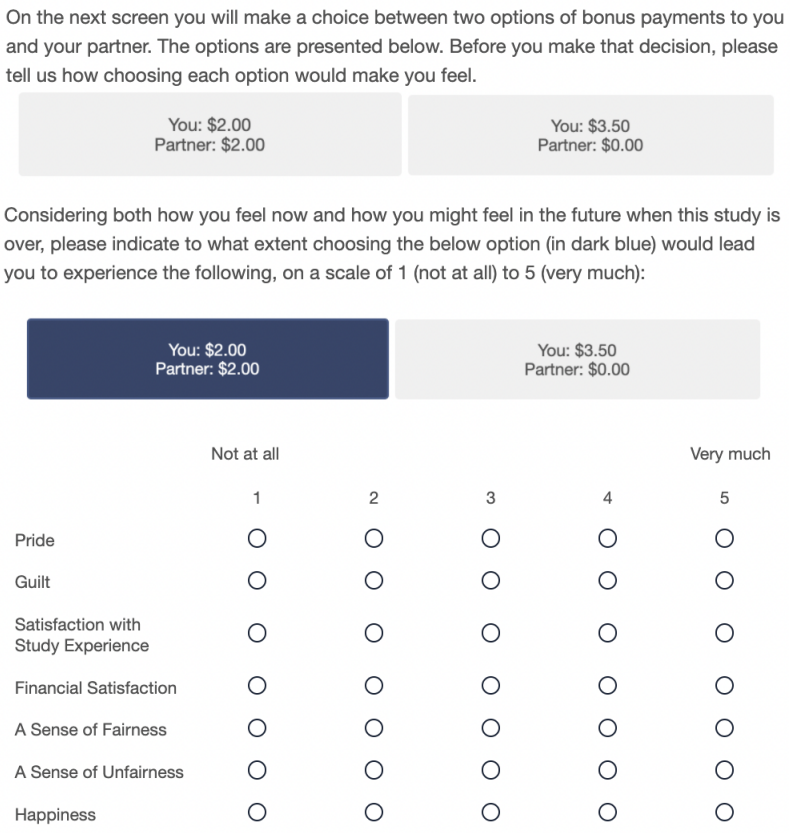


Robustness Check Arms

Ex-Ante Questions

Those who entered this arm were shown similar questions with identical choice sets to those found in the main survey; however, the difference was that these participants were asked to report their CSAs for each possible option prior to making a decision for each option in the Dictator Game and Opt-Out Game. The instructions for each part were the same as in the main survey. Examples of the Dictator Game and Opt-Out Game are shown in Figures S15 and S16. All participants in this arm were shown the main Computer Choice module.

Figure S15: Dictator Game in the Ex-Ante Arm



Present and Future CSAs

Those who entered this arm were shown similar questions with identical choice sets to those found in the main study arm except that these participants were asked to report their CSAs for the present and future separately. An example of the Dictator Game and Computer Choice CSA elicitations is shown in Figures S17 and S18. All participants in this arm were in the main Computer Choice module. Additionally, participants in this arm did not participate in the Opt-Out Games due to the increased length of the elicitations.

Figure S17: Dictator Game CSA Elicitations in the Present-Future Arm

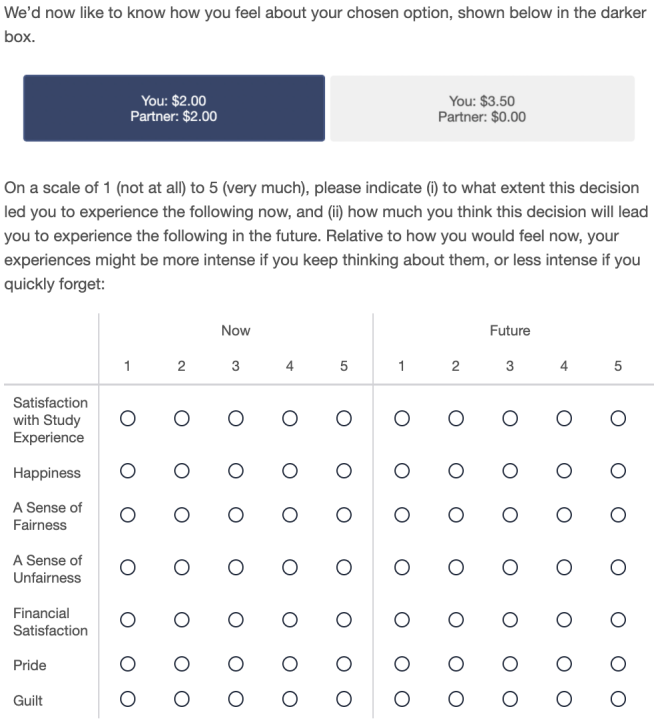


Figure S16: Opt-Out Game in the Ex-Ante Arm

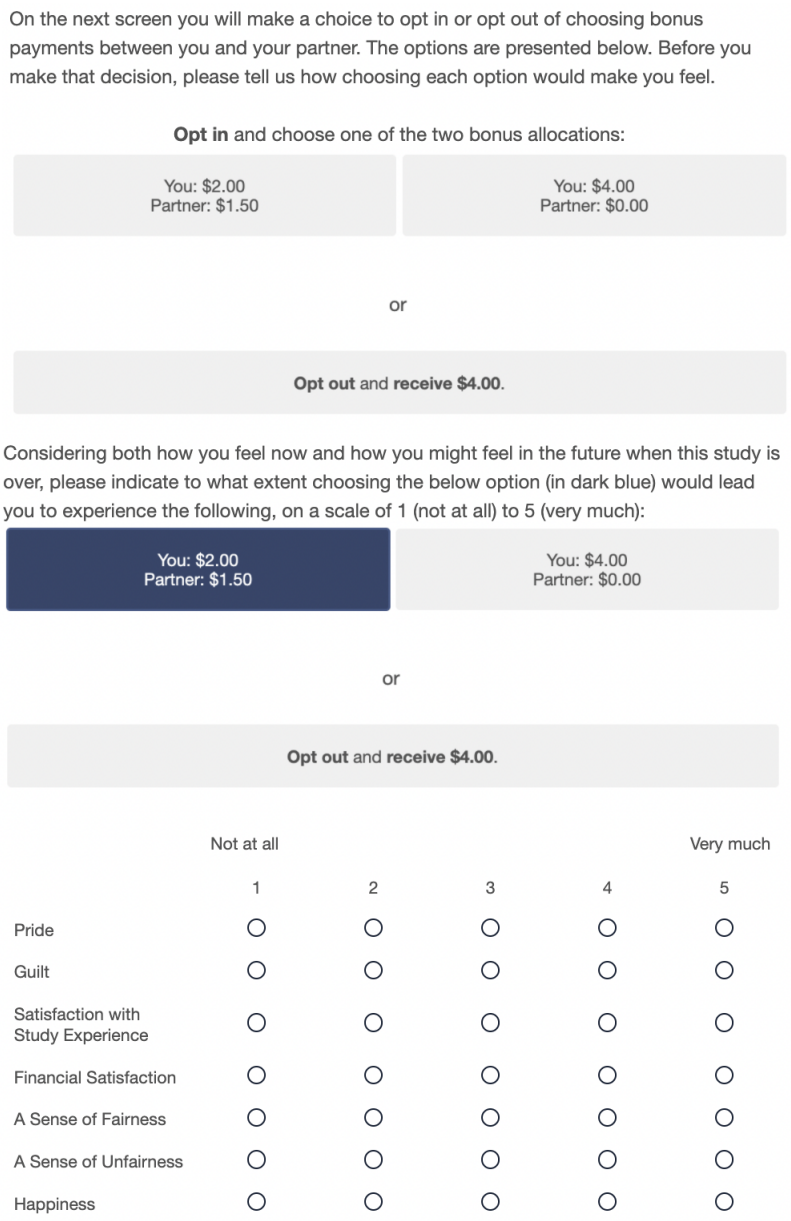
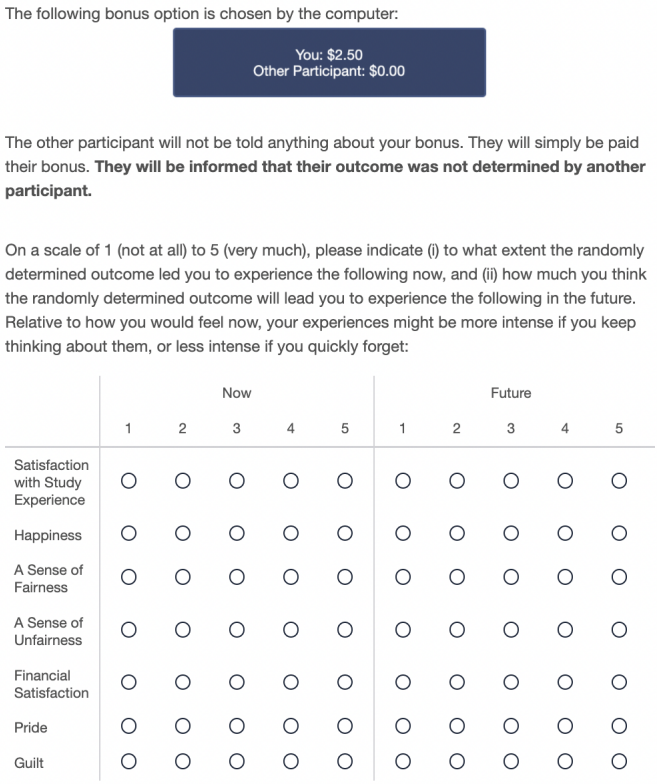


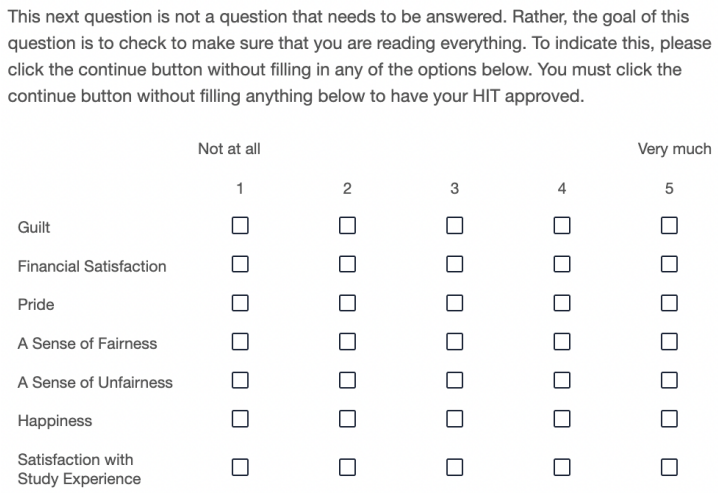
Figure S18: Computer Choice CSA Elicitations in the Present-Future Arm



Attention Check

After all games were completed, the participants were shown an attention check question. The attention check question asked the participant to leave the answer fields blank and simply click the continue button. The question is shown in Figure S19.

Figure S19: Attention Check Question



Demographic and Other Questions

After completing all the games and the attention check question, participants were shown a screen which read, "Thank you for completing this part of the study. We will let you know which part and scenario from the study was selected to count within two weeks, when we give you your bonus payment. We have only a few short questions left." This was followed by questions about the participant's age, gender, education level, and household income. A screenshot of these questions are shown in Figures S20-S23.

Figure S20: Demographic Question: Age

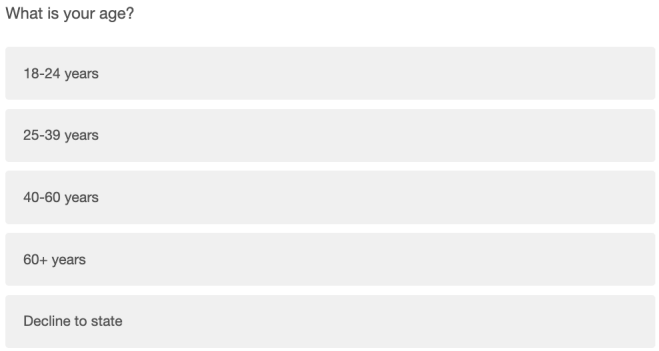


Figure S21: Demographic Question: Gender

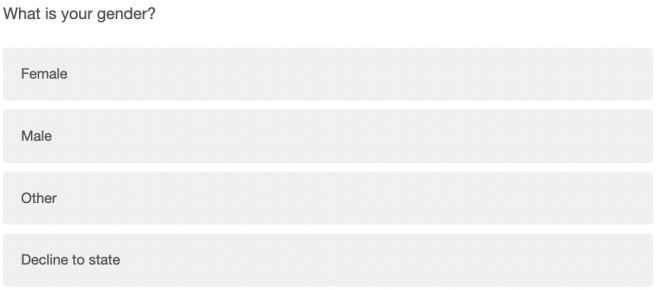


Figure S22: Demographic Question: Education

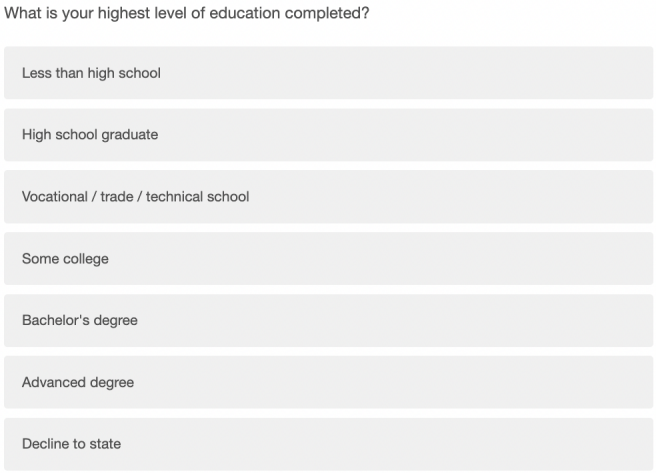
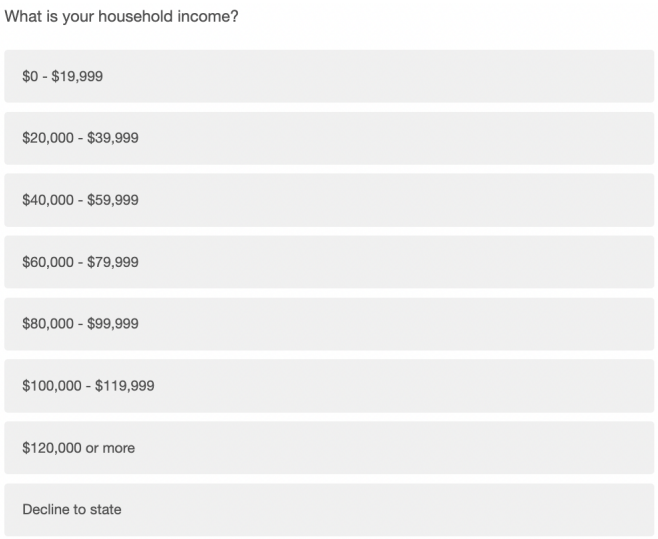
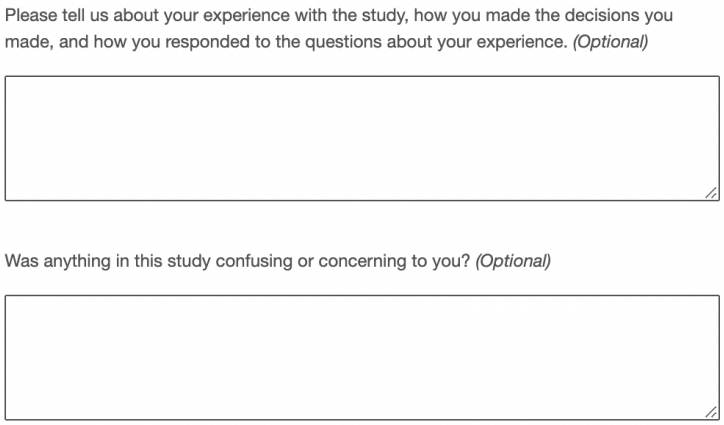


Figure S23: Demographic Question: Household Income



Last, participants were invited to comment about their experiences in the study in an optional question shown in Figure S24.

Figure S24: Comments and Concerns



Follow-Up Bonus Messages

Bonus payouts and messages were sent to all participants within two weeks of completing the survey. There were six versions of the bonus messages.

When the participant's own decision was chosen to determine the payouts, the messages were as follows:

- Dictator Game: "One of your survey choices was randomly selected for the bonus of you and another MTurker. You were given the scenario, **[(You: $X, Partner: $Y) or (You: $W, Partner: $Z)]** of which you chose the **[first/second]** option. Your bonus is $**[X or W].** Thank you for participating!"

- Computer Choice: "Your survey was randomly selected for the bonus of you and another participant. The computer chose the scenario, **[(You: $X, Partner: $Y)/ the [first/second] option from the scenario (You: $X,Partner: $Y) or (You: $W, Partner: $Z)].** Your bonus is $**[X/W]**. Thank you for participating!"
- Opt-Out Game: "One of your survey choices was randomly selected for the bonus of you and another MTurker. You were given the scenario, **[(You: $X; Partner: $Y) or (You: $W; Partner: $Z) or (Opt out of dividing money and take $V)],** of which you chose the first option. Your bonus is $**[X/W/V]**. Thank you for participating!"

When another participant's choice determined the outcome, the messages were as follows:

- Dictator Game: "You were randomly paired with another MTurker whose survey was randomly selected for the bonus. They were given the scenario, **[(Partner: $X; You: $Y) or (Partner: $W; You: $Z)]**, of which they chose the **[first/second]** option. Your bonus is $**[Y/Z]**. Thank you for participating!"
- Computer Choice: "You were randomly paired with another participant whose survey was randomly selected for the bonus. In their survey, the computer chose a bonus allocation where your bonus is $**[Y/Z]**. Thank you for participating!"
- Opt-Out Game where the partner opts out: No message
- Opt-Out Game where the partner does not opt out: "You were randomly paired with another MTurker whose survey was randomly selected for the bonus. They were given the scenario, **[(Partner: $X; You: $Y) or (Partner: $W; You: $Z) or (Opt out of dividing money and take $V)]**, of which they chose the **[first/second]** option. Your bonus is $**[Y/Z]**. Thank you for participating!"