NBER WORKING PAPER SERIES

WELFARE AND THE ACT OF CHOOSING

B. Douglas Bernheim
Kristy Kim
Dmitry Taubinsky

Working Paper 32200
http://www.nber.org/papers/w32200

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2024, Revised 2025

We thank Dan Benjamin and Ori Heffetz, as well as seminar participants at UC Berkeley, Chicago, Harvard, Michigan, UC Santa Barbara, and Zurich 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). The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Welfare and the Act of Choosing
B. Douglas Bernheim, Kristy Kim, and Dmitry Taubinsky
NBER Working Paper No. 32200
March 2024, Revised 2025
JEL No. D60, D91

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 through an approach that redefines consumption bundles in terms of the sensations they produce, and measures welfare by blending methods involving choice and subjective well-being. In a laboratory experiment on redistributive preferences, and in surveys concerning consequential economic decisions, we find that revealed-preference methods, including those that exploit choices over menus, mismeasure welfare because preferences depend on choice sets, while happiness and satisfaction are not sufficient statistics for welfare.

B. Douglas Bernheim
Stanford University
Department of Economics
and NBER
bernheim@stanford.edu

Dmitry Taubinsky
University of California, Berkeley
Department of Economics
and NBER
dmitry.taubinsky@berkeley.edu

Kristy Kim
University of California, Berkeley
kristykim@berkeley.edu

A online appendix is available at http://www.nber.org/data-appendix/w32200

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, or because making complex decisions is cognitively taxing. 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 is not 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 or analyst guided by Norma's choices would thus incorrectly conclude that eliminating her opportunity to share equally would leave her worse off. An important research agenda that seeks to infer preferences over redistributive policies from incentivized choices (e.g., Fleurbaey et al., 2009; Saez and Stantcheva, 2016; Almås et al., 2020; Capozza and Srinivasan, 2024) ignores the fundamental challenges associated with such possibilities.

The Non-Comparability Problem potentially arises in many other consequential contexts. For example, countless studies use choice-based methods to evaluate the welfare effects of policies that limit people's opportunity sets. However, there is good reason to think that people care about autonomy. Because every choice involves the exercise of autonomy, people cannot choose to have no autonomy. Even if someone chooses to limit their subsequent opportunities, the existence of that choice implies that they have broader opportunities. Consequently, choices cannot reveal how people feel about some other party, such as the government, curtailing their options. As another example, there is also good reason to believe that making decisions within certain domains, such as financial budgeting, healthcare, and career planning, induces anxiety and that people make different choices when they are apprehensive. Consequently, the options they choose for themselves may differ from those that would satisfy their desires most effectively if someone else reduced their anxiety by choosing for them.

In this paper, we provide a general account of the Non-Comparability Problem (NCP) and offer a conceptual solution, which we summarize below. 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 methods and subjective well-being (SWB) measures. These findings suggest that our approach is potentially applicable to other policy-relevant domains.

Prior work has relied on two strategies for avoiding the NCP. 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 preferences like Norma's. 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., 2006; 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 opportunities to expand menus by including less desired options. 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 NCP 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 often 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 construct 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 NCP 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.

To illustrate our approach more concretely, consider the problem of determining Norma's preferences. In that context, we might elicit the financial satisfaction she derives from each alternative, the amount of guilt she experiences, and other emotional reactions. Now suppose we find that (i) Norma's preferences over mental state bundles place negative weight on guilt, and (ii) Norma experiences guilt only when choosing allocations herself. If the negative weight on guilt is sufficiently large in magnitude, we might then conclude that her recovered preferences favor selfish outcomes assigned by third parties, even though she never chooses them herself.

We implement our approach by studying three classes of two-person allocation problems in an online experiment. In "computer choices" (CCs), a computer selects an allocation exogenously, acting (in effect) as a social planner. According to standard welfare economics, one evaluates the planner's selections by asking what the participants would have chosen for themselves. Dictator Games (DGs) provide that information. We also examine opt-out games (OOs): participants chose between playing DGs as dictator and "quietly" opting out. Choices in OOs allow us to apply the meta-choice method.

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 and accounting for 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 (without revealing the alternative) 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 temptation and guilt depending on whether they treat their partner fairly or unfairly. We also find that the mere awareness of a financially advantageous alternative 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 NCP is important in standard choice problems.

For opt-out tasks, we find that the option to play the DG reduces the desirability of the opt-out alternative, and the existence of the opt-out alternative reduces the desirability of playing the DG. These patterns are again consistent with responses involving emotions such as guilt and temptation. 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 this approach typically understates the benefits people derive from opportunities to share. This conclusion is also a practical manifestation of the NCP.

The final part of the paper documents the relevance of our analysis for consequential decisions in domains such as health, financial budgeting, career choice, and charitable giving. Based on supplemental surveys, we reach four conclusions: (1) emotions associated with the act of choosing are important relative to other consequences even in major economic decisions; (2) such emotions are also significant for consequential meta-choices; (3) as a result of the first point, the NCP arises in practice; and (4) as a result of the second point, meta-choices do not resolve the NCP. Specifically, people report that negative emotions cause them to avoid spending time on important decisions about their finances, health, and careers (point (1)), that they experience similar emotions when thinking about setting aside time for making the same decisions (point (2)), and that they would make better financial decisions if they devoted more time to planning (which implies that the emotions distort decision making, consistent with points (3) and (4)). When asked to consider measures designed to improve aspects of their behavior,[1] they anticipate experiencing more desirable mental states and deriving greater overall benefit if their participation is voluntary rather than mandatory (point (3)). Finally, people expect to experience less desirable mental states when avoiding charitable solicitation if the avoidance is a consequence of their own action (a meta-choice) rather than attributable to circumstances (point (4)).

Our analysis draws on, extends, and connects several important branches of the literature. 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 measures, 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

[1] We ask about two types of measures: reducing avoidance of important decisions through regular reminders, and promoting self-control by increasing penalties on early withdrawal from retirement savings plans.

act prosocially (Dana et al., 2006; List, 2007; Broberg et al., 2007; Lazear et al., 2012; DellaVigna et al., 2012; Exley, 2015).

The paper is organized as follows. Section 2 describes 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. Section 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 presents survey evidence documenting relevance to consequential decisions. Section 8 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). After explaining why the problem is not solvable through the use of either meta-choices (Section 2.2) or SWB methods (Section 2.3), we describe our proposed solution (Section 2.4).

2.1 The Non-Comparability Problem

The central objective of welfare economics is to guide a benevolent non-paternalistic authority (the *Planner*, "he") who makes decisions on behalf of others. For the sake of simplicity, we focus on how the Planner can infer the welfare of a single individual (the *Agent*, "she"). We assume the Planner assesses the Agent's well-being according to the Agent's desires.[2] A core premise of standard welfare economics is that the Planner can learn the Agent's desires by observing or inferring her choices. Consequently, if the Agent would select choice situation A over choice situation B, a Planner charged with determining whether the Agent will be better off in choice situation A or B would conclude that A is better for the Agent.

For a practical illustration, suppose the Planner must divide a fixed sum of money between the Agent and a third party. As in recent work on the empirical foundations of generalized social marginal welfare weights (e.g., Saez and Stantcheva, 2016; Capozza and Srinivasan, 2024), the Planner wants this allocation to reflect, in part, the Agent's distributional preferences. To learn about those preferences, the Planner tries to determine how the Agent would allocate resources between herself and the third party. Upon discovering that the Agent consistently selects allocations benefiting the third party at her own expense, the Planner would conclude that he can fulfill the Agent's desires by doing the same.

This conclusion would be justified if the Agent's sole motivation were either altruism or fairness. However, the Agent's choice may also reflect hedonic sensations experienced as a result of exercising

[2] In other words, the Planner subscribes to Desire Satisfaction Theory, a philosophical school of thought holding that "well-being consists in having one's desires satisfied" (Kagan (1997); see also Parfit, 1984; Heathwood, 2016).

agency. For example, she may select a fair outcome because doing so makes her proud or because she would feel guilty about acting unfairly. In contrast, if the Planner simply assigns everything to the Agent and nothing to the third party, the Agent may feel little or no guilt. Likewise, if the Planner assigns a fair allocation, the Agent may feel little or no pride. As a result, even though the Agent chooses to redistribute when given the opportunity, she may nevertheless prefer the Planner to redistribute nothing at all. In that case—contrary to central premise of recent empirical studies (e.g., Saez and Stantcheva, 2016; Capozza and Srinivasan, 2024)—the Planner cannot learn about the Agent's preferences over redistributive policies by observing the agent's choices.

In short, standard welfare economics assumes that the objects of choice are the same regardless of whether the chooser is the Planner or the Agent. However, those objects may be bundles that include hedonic sensations arising from either the exercise or absence of agency. In that case, the objects of choice in the Agent's decision problem are fundamentally non-comparable to the objects of choice in the Planner's decision problem.

To formalize this point, suppose the Agent has preferences over bundles of the form (X, x), where X is the constraint set and x is the selected option. Suppose the Agent would choose $x^*(X)$ from the constraint set X, and that public policy determines whether the constraint set is X' or X''. In that case, the Planner needs to determine whether the Agent prefers $(X', x^*(X'))$ or $(X'', x^*(X''))$. However, when the Agent makes a choice involving any given constraint set X, we can conclude only that $(X, x^*(X)) \succeq (X, x)$ for all $x \in X$. Consequently, standard choices provide no basis for determining whether a policy that changes the constraint set helps or hurts the Agent. This observation is the essence of the NCP.

Another way to make the same point is to ask whether the Agent's utility function is recoverable from her choices (as in Koszegi and Rabin, 2008). Suppose the individual acts as if her choices maximize the utility function $U(X, x)$. In other words, $U(X, x^*(X)) \geq U(X, x)$ for all $x \in X$. Consider the alternative utility function $\tilde{U}(X, x) = U(X, x) + W(X)$, where $W(X)$ is an arbitrary function. Then it is also the case that $\tilde{U}(X, x^*(X)) \geq \tilde{U}(X, x)$ for all $x \in X$, which means that U and $\tilde{U}$ rationalize precisely the same behavior. It follows that the difference between the Agent's utility from $(X', x^*(X'))$ and $(X'', x^*(X''))$ is completely unidentified.

2.2 Meta-choices

One branch of the literature attempts to measure welfare-relevant sensations experienced during the act of choosing, and thereby resolve the Non-Comparability Problem, by studying *meta-choices* (that is, choices over choices). The most common designs involve *opt-out* (or *avoidance*) tasks. For instance, if someone is indifferent between accepting \$2 and performing a decision task that would reliably yield a payoff of \$3, one might infer that the money-metric cost of decision-making is \$1. Studies that have employed this method include Dana et al. (2006); Lazear et al. (2012); DellaVigna et al. (2012); Bartling et al. (2014); Allcott and Kessler (2019); Butera et al. (2022). Typically, these papers interpret the break-even opt-out transfer as a measure of the extent to which people are attracted to or repelled by the pertinent decision task.

To understand what can go wrong with the meta-choice method, recall the example from the introduction in which Norma divides a sum of money between herself and a friend. Because Norma is averse to bearing responsibility for an unequal split no matter how the decision is framed, meta-choices cannot reveal her fervent desire for someone to take the decision out of her hands and give her the entire prize.

Meta-choices fail to resolve the NCP in Norma's case because she has preferences over bundles of the form (X, x), just as in the previous section. When she chooses directly between an equitable option e and an inequitable option s, her constraint set is $X = \{e, s\}$; her utility is $U(e, \{e, s\})$ if she selects e and $U(s, \{e, s\})$ if she selects s. Suppose we allow her to delegate this choice to a party who will definitely select s. That option creates a meta-choice between the menus $\{s\}$ and $\{e, s\}$. Because the overall constraint set is still $X = \{e, s\}$, her utility is still $U(e, \{e, s\})$ if she ends up with e and $U(s, \{e, s\})$ if she ends up with s, regardless of whether she chooses s herself or delegates. Consequently, her meta-choice reveals no additional information concerning her preferences.

Implicitly, the meta-choice method invokes a different assumption about the structure of preferences. The underlying premise is that, instead of having a single emotional reaction based on the constraint set for the entire decision problem, the individual has different emotional reactions to each component choice of a sequenced decision, and the intensity of those reactions dissipates as the choice becomes more distant from the outcome. We next provide a general formulation for preferences that encompasses this possibility and prove that it also encounters the Non-Comparability Problem.

We consider settings in which an Agent 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 Agent selects $\mathcal{M}^1$ from a collection of level-1 menus $\mathcal{M}^2$, which we call a *level-2 menu.* Recursively, in a J-stage decision problem, the first stage involves the choice of $\mathcal{M}^{J-1}$ from a collection of level-$(J-1)$ menus $\mathcal{M}^J$, which we call a *level-J menu.*[3]

To illustrate, suppose the Agent makes selections from the constraint set $\{e, s\}$ in two steps: first she decides whether each of the options should be available; then she chooses 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, we might attempt to resolve the Non-Comparability Problem while allowing for the possibility that the act of choosing generates welfare-relevant sensations by assuming that the Agent's preferences depend only on the items she receives and the level-1 menus she faces. Under this assumption, choosing $\mathcal{M}^1_2 = \{e\}$ or $\mathcal{M}^1_3 = \{e, s\}$ over $\mathcal{M}^1_1 = \{s\}$ implies that, in a setting where a Planner assigns the level-1 menu, she would indeed prefer to have e on that menu. Furthermore, Agents who choose $\mathcal{M}^1_2 = \{e\}$ over $\mathcal{M}^1_1 = \{s\}$ prefer an e-mandate over an s-mandate,

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

and hence are better off with an e-mandate.

There is, however, little if any psychological foundation for the general assumption that only the final stage of a multi-stage decision process is emotionally consequential. For example, if choosing an entree from a restaurant's menu has hedonic consequences, then the same is plausibly true for the choice of a restaurant, or even the choice of a cuisine. And yet, without this assumption, the Non-Comparability Problem resurfaces. In the example above, an Agent's choice of $\{e\}$ from $\mathcal{M}^2 = \{\{s\}, \{e\}, \{e, s\}\}$ simply reveals that she prefers the triplet $\{e, \{e\}, \mathcal{M}^2\}$ (object, level-1 menu, and level-2 menu) over $\{s, \{s\}, \mathcal{M}^2\}$, $\{e, \{e, s\}, \mathcal{M}^2\}$, and $\{s, \{e\}, \mathcal{M}^2\}$. It does not follow, for example, that she prefers an e-mandate, which corresponds to the bundle $\{e, \{e\}, \{\{e\}\}\}$, over an s-mandate, which corresponds to the bundle $\{s, \{s\}, \{\{s\}\}\}$.

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

$$V(x, \mathcal{M}^1, ..., \mathcal{M}^J) = u(x) + \pi^1(x, \mathcal{M}^1) + \sum_{j=2}^{J} \pi^j(\mathcal{M}^{j-1}, \mathcal{M}^j) \qquad (1)$$

where

$$\pi^j(x, \{x\}) = 0 \text{ for all } x, \text{ and } \pi^j\left(\mathcal{M}^{j-1}, \left\{\mathcal{M}^{j-1}\right\}\right) = 0 \text{ for } j = 2, ..., J \text{ and all } \mathcal{M}^{j-1}. \qquad (2)$$

In other words, the individual receives intrinsic utility $u(x)$ from the item x, and also derives utility or disutility $\pi^j(\mathcal{M}^{j-1}, \mathcal{M}^j)$ when selecting from a level j menu. When there is no level-j choice (i.e., the level-j menu is degenerate in the sense that either $\mathcal{M}^1 = \{x\}$ or $\mathcal{M}^j = \left\{\mathcal{M}^{j-1}\right\}$ for some $j \in \{2, ..., J\}$), the individual experiences no level-j utility.

Knowledge of u is plainly crucial for welfare analysis. Indeed, the individual's preferences over x-mandates depend *only* on u. However, it is impossible to identify u using the hypothesized data (see Appendix A.1 for the proof):

Proposition 1. *For some fixed J, consider any utility function V with component functions $\left(u, \pi^1, ..., \pi^J\right)$ from the class described by equations (1) and (2). For all functions $\tilde{u} : X \to \mathbb{R}$, there exists a utility function $\tilde{V}$ from the same class with component functions $\left(\tilde{u}, \tilde{\pi}^1, ..., \tilde{\pi}^J\right)$ such that, for all $j \leq J$, V and $\tilde{V}$ are observationally equivalent with respect to all stage-j choices.*

Although u becomes identified if one *assumes* level-J menu independence, there is no way to *test* level-J menu independence with level-J data. As an alternative, when the choice data pertain to levels 1 through J, one might hope to test level-j menu independence for some $j < J$ (such as $j = 2$, which would allow for level-1 menu dependence, while validating opt-out designs). Failing to reject that hypothesis, one might then try to identify u from choice data pertaining to levels 1 through j. Unfortunately, that strategy does not work because $\tilde{\pi}^j$ is also unidentified (as shown in Appendix A.1), which means level-j menu independence is also untestable.

Most policies are not x-mandates. Instead, they impose constraints on choices (i.e., the menu structure). Even though u does not play the same decisive role in evaluations of those policies, the same identification problem nevertheless arises; see Appendix A.3 for an illustration.

Finally, we note that 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. We formalize this point in Appendix A.4.

2.3 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).

We begin by asking what a response to a question about happiness or satisfaction represents. One possibility is that the question prompts *transcription*: the respondent may simply "read a hedonic register" and report the reading. Alternatively, the question may prompt *aggregation*: the respondent attempts to construct a single summary index for a variety of past, present, and anticipated sensations such as amusement, boredom, sympathy, and anxiety. Considering the richness of hedonic experience as well as the sensitivity of SWB responses to wording (e.g., "happiness" versus "satisfaction"), we consider the first possibility less plausible, but will address it after discussing the second.

To illustrate the Aggregation Problem, suppose 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. 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 twice 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 twice as much weight on fatigue as on hunger. Thus, our use of natural language amounts to an instruction concerning the weights we want Norman to use when aggregating these sensations; it is simply less clear than a directive to employ specific numerical weights. 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 for the moment) 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?

An alternative defense is to insist that people answer SWB questions by performing transcription tasks rather than aggregation tasks. Plausibility aside, this defense is problematic because transcription is a special case of aggregation wherein the respondent places all weight on a single hedonic register. There is no foundation for assuming the existence of a hedonic sensation that serves as a sufficient statistic for overall well-being, or for assuming that the use of specific words and phrases reliably induces people to read the right register.

Critically, the available evidence is inconsistent with the hypothesis that people's responses reliably combine the various dimensions of hedonic experience in ways that reliably capture their desires. A central finding of Benjamin et al. (2014) is that people do not seek to maximize self-assessed happiness or satisfaction when making choices. Our own results, reported in Section 6.4, corroborate that conclusion by showing that single-index elicitations of happiness or satisfaction are not sufficient statistics for welfare.

2.4 The Proposed Solution

The central premise of our approach is that people are *mental statists*, in the sense that their primitive (as opposed to derivative) desires pertain to their internal hedonic sensations.[4] In keeping with this premise, we assume that, when someone makes a decision, they associate each option with an anticipated mental state bundle and then select their favorite bundle from the resulting menu. From the perspective of Desire Satisfaction Theory (which provides the philosophical foundation for standard welfare economics), a policy improves their welfare if and only if it provides them with a better mental state bundle according to those preferences.

Briefly, our method works as follows. First, we assess (proxies for) the mental state bundles the Agent expects to follow from each option in a collection of choice problems. Second, we use these data to estimate preferences over these anticipated mental state bundles using standard discrete choice techniques. Third, for any assigned choice situation (e.g., outcomes or constraint sets chosen by the Planner), we elicit the same (proxies for) mental state bundles.[5] The preferences estimated in step 2 then allow us to determine the Agent's preferences over the choice situations to which

[4] The mental statist view is commonly associated with John Stuart Mill (Mill, 2012). To be clear, this perspective does not presuppose selfishness; it subsumes the possibility that mental states depend on outcomes for others. 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.

[5] When the choice situation is non-degenerate, the relevant mental state bundles are the ones that correspond to Agent's selections.

some other party, such as a Planner, might assign her. For a formal general description of these steps, see Appendix A.5.

This approach avoids the Non-Comparability Problem that afflicts conventional choice-based methods 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 that afflicts the SWB approach by borrowing from the choice-based paradigm: it aggregates over the dimensions of subjective experience based on desires as captured by choices.

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. Our method accommodates this possibility without reintroducing the Non-Comparability Problem. The motivating premise for this study is that the act of choosing has hedonic consequences—in other words, that menus matter because they affect mental states. Under that premise, the menu of mental state bundles, Z, can likewise matter to the individual only through mental states. If we elicit the anticipated mental state bundle the individual associates with each option when making their choice, the elicitation will subsume the effects of any such menu dependence. While there may be feedback between z and Z, in the end, only z matters. For a more formal explanation, see Appendix A.6.

The practical concern is that it may prove difficult to define and accurately 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. We can think of people as having stable preferences over CSAs and proceed as if the CSAs are, in effect, measures of z as long as (i) the CSAs collectively span the pertinent mental states, and (ii) each CSA is a stable composite of underlying mental states (i.e., there is some function relating each reported CSA to those states).[6] For our application, we test the stability and spanning conditions in Sections 5.4 and 5.5, respectively.

3 Experimental Design

The experiment involved three classes of mechanisms for dividing a fixed sum of money between a participant and a randomly assigned partner: (1) Computer Choices (CCs), (2) Dictator Games (DGs), and (3) Opt-out Games (OOs). In a CC, a computer played the role of a social planner in that it selected the allocation exogenously. Standard welfare economics evaluates the planner's selections by asking what the participants would have chosen for themselves. The DGs provide

[6] Because our goals differ from those of Benjamin et al. (2014), our approach does not require us to identify individual mental states or assume they are reported accurately. In contrast, Benjamin et al. (2014) focus on estimating the marginal utility of individual mental states, which requires that their measures of mental states are not only comprehensive, but also non-overlapping, maximally disaggregated, and correctly reported. For our purposes, CSAs can be overlapping composites and even deviate systematically from the truth as long as the spanning and stability conditions are satisfied.

that information. The OOs were meta-choices: participants chose whether to participate in a DG, in which case they made a subsequent choice between the DG options, or "quietly" opted out. Following Dana et al. (2006) 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 K contains the full study instructions.

Computer Choices. Participants were presented with eight allocations chosen by a computer, which 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 a participant's sense of well-being even when they do not control the selection.

We label the computer-selected allocations CC 1 through CC 8. For CC 1, the computer assigned (You: $2.00; Partner: $0.50). Moving from CC 1 to CC 4, the partner's payoff increased in increments of $0.50. For CC 5, the computer assigned (You: $4.00; Partner: $0.00). Moving from CC 5 to CC 8, the participant's payoff decreased in increments of $0.50. Henceforth, we refer to CC 1 through CC 4 as the *more equitable options* and to CC 5 through CC 8 as the *less equitable options.* To be clear, we did not use those labels in the experiment. For the CC variants that specify the alternatives, the choice sets were the same as for the DGs described below.

Dictator Games. Each of seven DGs 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. For DG 1, the participant chose between a more equitable option, (You: $2.00; Partner: $0.50), and a less equitable option, (You: $4.00; Partner: $0.00). Moving from DG 1 to DG 4, the partner's payoff for the more equitable option increased in increments of $0.50. Moving from DG 4 to DG 7, the participant's payoff for the less equitable option decreased in increments of $0.50. Thus, the options for DG 7 were (You: $2.00; Partner: $2.00) or (You: $2.50; Partner: $0.00). Critically, every DG offered two CC allocations and every CC allocation was an option in at least one DG.

Opt-Out Games. Following Dana et al. (2006), Broberg et al. (2007), and Lazear et al. (2012), we designed meta-choices that allowed participants to "opt-out" of DGs. There were four OO decisions shown on separate 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 the individual with whom they were paired would learn their decision if they opted in but would not know that any decision had been taken if they opted out. The four OO scenarios, which appeared in random order, involved opt-out payments of \$5.00, \$4.00, \$3.50, and \$3.00.

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.[7]

We elicited the seven CSAs on five-point Likert scales. In each case, the response encompassed mental states both during and after the experiment. We randomized the order of the questions across participants but always presented them in the same order for a given participant.[8] In the DGs and OOs, we first elicited CSAs for the participant's chosen option, then we elicited CSAs for the alternatives. For the OO's, we started 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 also showed the alternatives and the participant's choice. In the main CC module, we first elicited CSAs for the computer's chosen option, and then for the alternative.

Variations on CSA elicitations To evaluate the robustness of our methods, we explored two alternative approaches to CSA measurement

For the first alternative approach, we elicited CSAs for the DG and OO games *before* participants made their decisions, rather than after. We call this the *ex-ante arm.* 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.[9]

[7] To be clear, we selected these CSAs because they reflect critical motivations within the social preferences domain. For applications involving other domains, other CSAs may be more appropriate.

[8] For DGs and OOs, we asked: *"We'd now like to know how you [feel / think you would 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 D5 (very much).*"

[9] Elicitation questions took the following form for each of the two possible options: *"Considering both how you feel*

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. We call this condition the *present-future arm*. The motivation for this alternative is that our primary method elicits a temporal aggregate and is therefore potentially susceptible to the Aggregation Problem.[10] Due to concerns about the study's length, we omitted the OO module when using this alternative approach.

Procedures and incentives 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.[11] To conserve on costs, participants played both roles for the DG and CC modules, with one exception. 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 and CCs, and 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 selected for the actual payout. This procedure ensured that all choices by participants in the Decider role were incentive compatible. To deter participants from renormalizing the CSA Likert scales as they proceeded through the DG, CC, and OO modules, we reiterated the payout range for the entire study at the beginning of each module.

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 present and future CSAs. We assigned all 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.

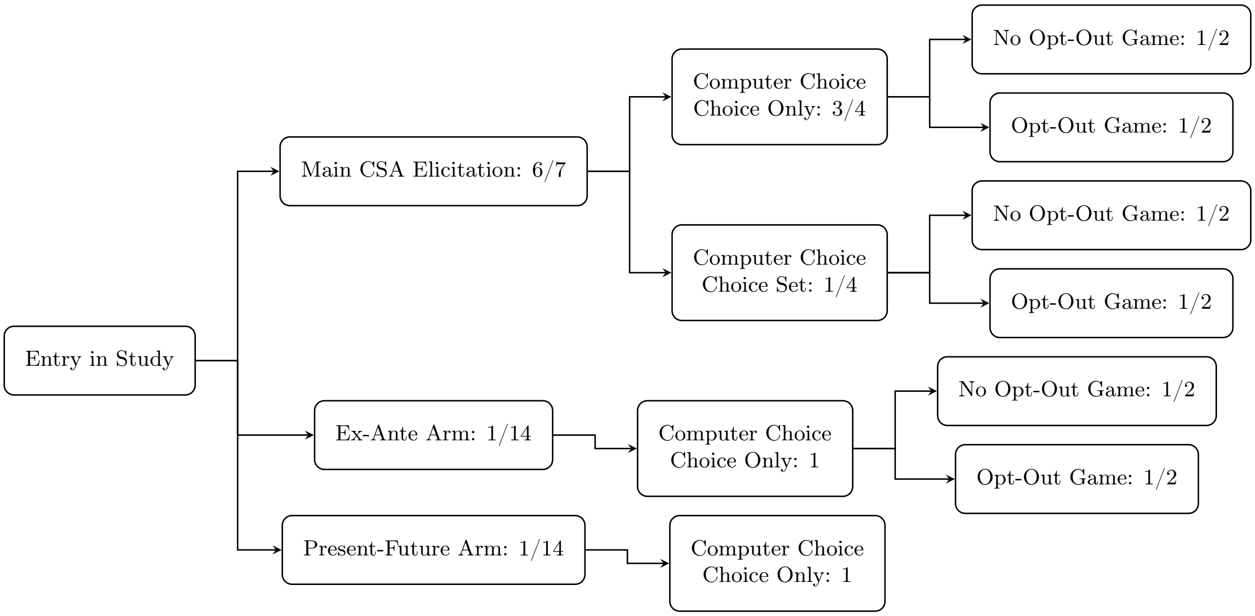
We also included an attention check that resembled the CSA elicitations, but that instructed

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)."

[10] 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."*

[11] 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 reported their CSAs, while the Receiver is the one who received the (weakly) smaller amount in each option.

Figure 1: Experimental conditions and randomization



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.

participants to simply click "continue" without answering.[12]

Figure 1 summarizes the randomization into treatment arms.

Participants. We recruited participants 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).

[12] The text was: "*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 instructions as inattentive. The attention check appeared after the DG, CC, and OO modules, but before a closing battery of demographic questions.

4 Framework for Empirical Analysis

4.1 Preferences and Choice

To implement the method proposed in Section 2.4, 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

$$U = v(X_{ijc}) + \varepsilon_{ijc}. \tag{3}$$

The term $v(X_{ijc})$ is the deterministic component of utility. The term ε_{ijc} is an idiosyncratic realization that 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. The quasi-linearity assumption is not strictly necessary but simplifies the analysis. 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

$$P_{jc} = \frac{e^{v(X_{ijc})}}{\sum_l e^{v(X_{ijc})}} \tag{4}$$

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., 2024, for a review). Translating utility into dollars makes our welfare measures economically interpretable. 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 among 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 c_y denote the CC setting that imposes the allocation $(y, 0)$ with no alternatives, and let j_y be the index for option $(y, 0)$ in that setting. Then by definition,

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

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

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

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: $\{(4, 0), (3.5, 0), (3, 0), (2.5, 0)\}$.

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

$$\bar{u}_{jc} = \mathbb{E}_i[v(X_{ijc})/m] - \mathbb{E}_i[v(X_{ij_0c_0})/m], \tag{5}$$

where $\mathbb{E}_i$ denotes the expectation over individuals i.[13] 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).[14]

In practice, the marginal utility of money may vary across individuals. However, as long as this parameter is unrelated to variations in money-metric utility, our interpretation of $\bar{u}_{jc}$ is unchanged;

[13] 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 $\bar{u}_{jc} = \mathbb{E}_i[v(X_{ijc})/m] - \mathbb{E}_i[v(X_{ij_4c_4})/m] + 4$, where (j_4, c_4) corresponds to the allocation $(4, 0)$ in the main CC module. Under the linearity assumption, $\mathbb{E}_i[v(X_{ij_4c_4})/m] - \mathbb{E}_i[v(X_{ij_0c_0})/m] = 4$, so the preceding expression is equivalent to equation (5).

[14] 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 (3) 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 the 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 corresponds to $\bar{u}_{jc}$ in cases where $\bar{u}_{jc}$ is a full-sample mean for a fixed allocation.

see Appendix G for details.

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 five 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, the 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."[15]

The second assumption is that v is stable across the domains to which we apply our methods. Even if preferences over mental states are stable, reduced-form preferences over CSAs could be unstable if the spanning condition fails or the relationship between CSAs and mental states is changeable. 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.[16] Gillen et al. (2019) show how noisy measurement can bias coefficient estimates either upward or downward. Section 5.3 provides evidence that such biases are not consequential in our setting.

The fourth assumption is that there is sufficient variation in CSAs within the estimation sample to identify preferences over the CSA bundles induced by the choice settings one seeks to evaluate. A problem would arise, for example, if a CSA varied across the latter but not within the former, or if two CSAs were perfectly collinear within the former but not across the latter.[17] For example, people might experience high financial satisfaction and low guilt when assigned inequitable allocations, but guilt and financial satisfaction might go hand-in-hand when they make choices. However, it is straightforward to check this assumption, and we do so for our data set.

The fifth assumption concerns misreporting of CSAs. Misreporting per se is not a problem if it preserves a stable relationship between reports and mental states: our method effectively inverts the mapping from mental states to (mis)reported CSAs when estimating how people's reports correspond to their choices. Problems would arise only if misreporting involves pooling (e.g., at the boundaries of the Likert scale) or if different strategic motives for misreporting come into play when the elicitation pertains to the Planner's options rather than to the choice tasks used for preference recovery. The issue of strategic misreporting arises generally in the literature on subjective well-

[15] For example, our approach remains valid if there are other CSAs that covary with choice but that are fully spanned by the CSAs 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}.

[16] Notice that we can rewrite equation (5) as $\bar{u}_{jc} = (\bar{X}_{ijc} - \bar{X}_{ij_0c_0})\beta/m$. With a reasonably large sample, $\bar{X}_{ijc}$ and $\bar{X}_{ij_0c_0}$ are measured with little or no error. Therefore, our method requires an estimate of β that is free from bias associated with measurement error.

[17] To be clear, no problem arises if two CSAs are perfectly collinear in both the estimation sample and the choice settings one seeks to evaluate.

being; see Benjamin et al. (2013; 2014) for a discussion and potential remedies.

Critically, our method does *not* assume that people have accurate expectations concerning the mental states that would follow from their options. Because choices depend on the outcomes people anticipate, errors in forecasting mental states would not confound our estimates of preferences over mental states. Our assumption is simply that, when someone chooses an option they think is A over an option they think is B, they desire A more than B, even if those beliefs are wrong. However, once one has recovered preferences, it is important to evaluate the Planner's alternatives by applying those preferences to the mental state bundle each Planner-assigned design problem actually induces rather than to the mental state bundle people expect it to induce. That is why, for every decision problem a Planner could potentially assign, we ask people after making their selection to report CSAs for the option they chose. Our method therefore allows us to measure welfare for any assigned decision problem without assuming that, prior to choice, people correctly anticipate the resulting mental states. Accurate expectations are required only when evaluating the welfare consequences of counterfactual choices. In subsequent sections, we sometimes conduct such evaluations to hold the sample constant when investigating certain forms of context dependence. Because we find no evidence of forecasting errors in our setting (see Section 5.6), these evaluations are also reliable. Notably, in domains where there is evidence of hedonic forecasting errors, one could use our method to evaluate the resulting welfare losses.

5 The Relationship Between Choices and CSAs

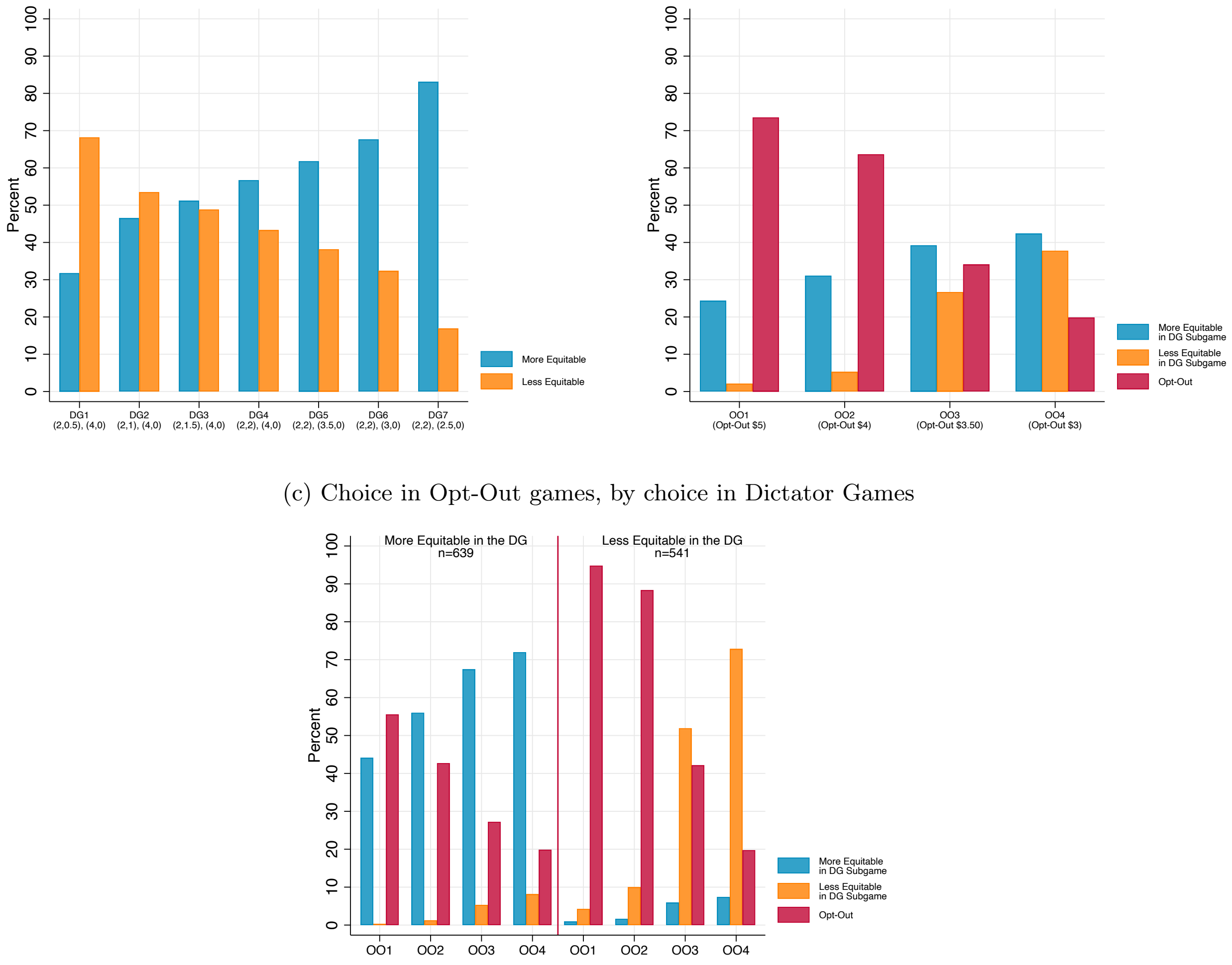
In this section, we study the relationship between choices and CSAs in the Dictator and Opt-Out Games. We begin by describing patterns of choices across variants of the games and how the CSAs vary over the available alternatives in those environments. Then we estimate our empirical model linking choices to CSAs. Finally, we provide evidence for key premises of our conceptual framework: that the estimated model is stable across classes of environments (DGs versus OOs), that the CSAs adequately span the pertinent space of mental states, that the independence assumption $\varepsilon_{ijc} \perp X_{ijc}$ holds, and that potential measurement error does not bias our conclusions.

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).

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.

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 A2 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.

5.2 Patterns of CSAs

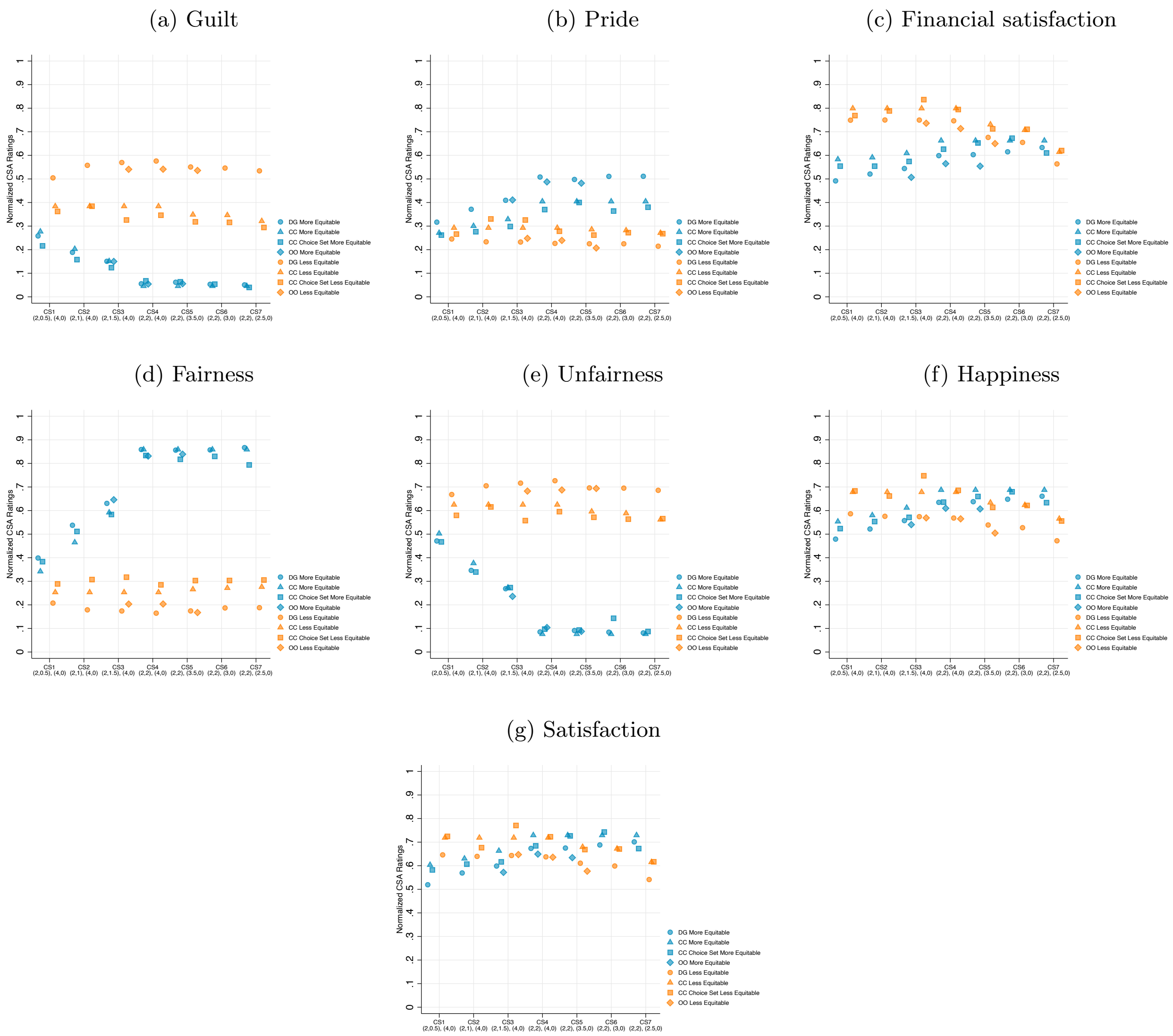
Figure 3 shows how CSAs for the more and less equitable options vary across the DG, CC (both the primary version and the one with an explicit choice set), and OO modules. Throughout our analysis, we divide normalize the raw CSAs responses by subtracting 1 and dividing by 4; this normalization ensures 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.[18] 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

[18] CSAs for the opt-out options in the OO module appear in Figure A1 of Appendix C.

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 normalize the reported CSA responses subtracting 1 and dividing by 4. 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 for both above the label for DG 1. For the OOs, we report CSAs just for the DG subgames.

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 (4) 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, $\Delta_{ic} = X_{ijc} - X_{ij'c}$, where j and j' 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" coefficient for 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 for the instrument 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.

Table 1: Association between deciders' choices and CSAs

	(1)	(2)	(3)
	Logit	IV Logit	Logit
	Choosing	Choosing	Choosing
	More Equitably	More Equitably	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 (4) 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, clustered at the participant level, are reported in the parentheses. For the two-step estimator in columns 2 and 3, standard errors are calculated via bootstrap. * $p < 0.1$, ** $p < 0.05$, ***$p < 0.001$.

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 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).

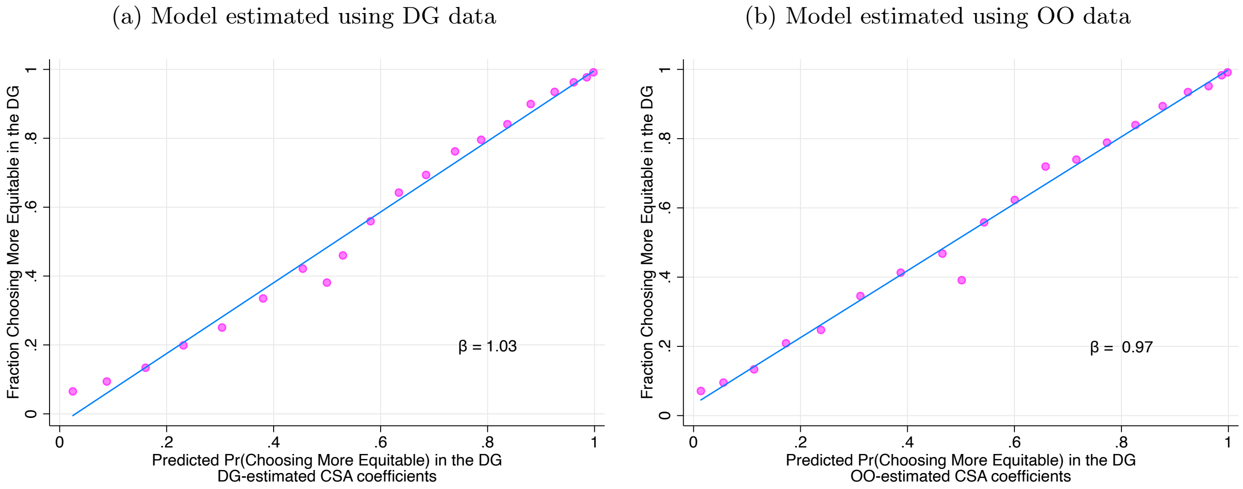
Because happiness and satisfaction likely aggregate the five narrower CSAs to some degree, Column (3) of Table 1 explores the importance of the narrower CSAs in a regression that omits the two broad measures. The largest changes between Columns (1) and (3) are for the coefficients of pride and financial satisfaction. This finding potentially explains why happiness and satisfaction are not sufficient statistics for welfare: relative to choice, these broad measures appear to encapsulate positive sensations (pride and financial satisfaction) to a greater extent than negative sensations (guilt and unfairness). Appendix Table A8, which reports regressions of happiness, satisfaction, and choice on the other five other CSAs, corroborates this inference.

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 with 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

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.

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.

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.[19]

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

[19] 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.

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 of 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 A3 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 the supplementary "temporal disaggregation" arm of the experiment. 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 their chosen options. 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 the supplementary "ex-ante" arm of the experiment. Appendix F.1 shows that eliciting CSAs before choices does not alter the reported CSAs or their relationships to choice. The results also imply that participants have no trouble reporting CSAs for unchosen alternatives.

Finally, the results for the ex-ante arm help rule out the possibility that eliciting CSAs amplifies emotional responses, as well as the related possibility that it alters the relationship between mental states and choices.[20] If these effects were present in our setting, the ex-ante elicitations would likely amplify them. But we find no evidence of such amplification.

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. As shown in Appendix Table A14, this modification has negligible effects on the pseudo R-squared and the CSA coefficients. This finding is consistent with absence of a meaningful distinction between variation in Δ_{ic} that comes from plausibly exogenous differences in choice sets and variation that comes from other sources. However, there is another potential explanation: choice set fixed effects may not account for much variation in $\Delta_{ic}\hat{\beta}$, the predicted difference in average utility between the more and less equitable DG alternatives. Appendix Table A13 rules out the second possibility by showing that a regression of $\Delta_{ic}\hat{\beta}$ on choice-set indicators has a non-trivial adjusted R-squared (0.12).[21] Moreover, stability of the pseudo R-squared statistic in Table A14 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.

Finally, we show in Appendix F.4 that our results are not influenced by our reliance on within-participant variation. For example, it is in principle possible that being exposed to both DG and CC modules influences the mental states that participants experience in the module they encounter second, or maybe even changes their preferences over mental states. Appendix F.4 rules out this possibility. At the same time, had we found that order effects altered participants' experiences, our method would have remained internally consistent, and would have permitted us to evaluate the impact of order on the welfare derived from each decision task.

6 Estimating Welfare with CSAs

Having recovered Deciders' preferences over CSA bundles, we now use those preferences to evaluate their welfare. Because the allocation problems we consider involve both a Decider and a Receiver,

[20] While such effects would not confound internal validity, they could limit external relevance for assessing welfare for policies that are not accompanied by CSA elicitations.

[21] As expected, the coefficients increase steadily as one moves from DG 1 to DG 7, because the more equitable allocation becomes more and more attractive.

two features of our analysis merit emphasis. First, as is common throughout the economic literature, our perspective on welfare is individualistic rather than communal. Second, while we focus on assessing the Decider's well-being, our methods are also suitable for drawing conclusions about the Receiver. In principle, we might find that Receivers care not only about their payoffs, but also about the process that generated those payoffs. For example, knowing that another party could have been more generous might reduce a Receiver's enjoyment of a given outcome.

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.[22]

Several patterns in Panel (a) of Figure 5 illuminate participants' prosocial motivations. 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 the option of interest, when the individual selects it, 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.

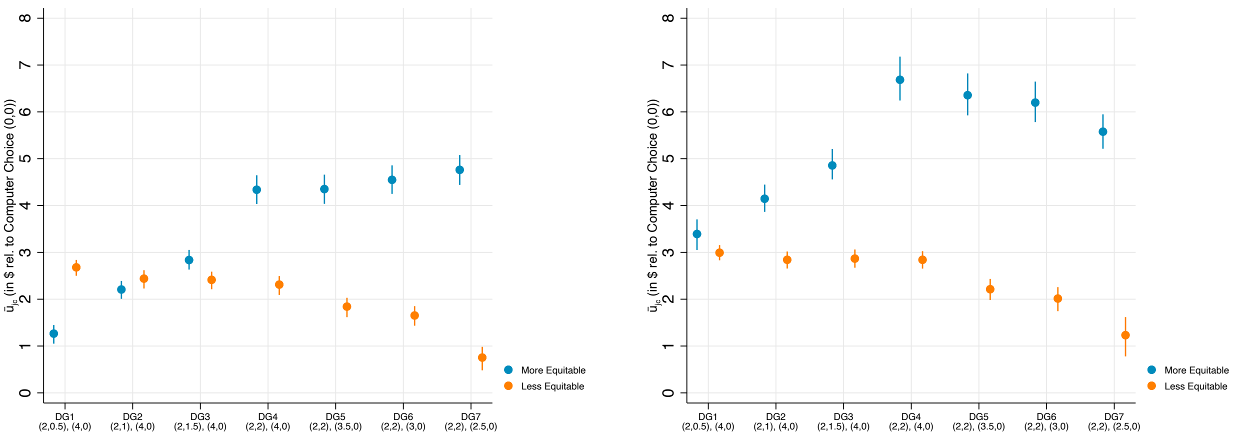
Third, holding the sample constant (as in Panel (a)), as we improve the Receiver's payout from the more equitable option (i.e., moving stepwise from DG 1 to DG 4), the average money-metric utility from that option increases substantially (as expected), while the average money-metric utility from the less equitable option declines (from 2.68 in DG 1 to 2.31 in DG 4; 95% CI for difference $= [0.28, 0.47]$).[23] The latter finding is consistent with the hypothesis that the typical Decider

[22] 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.

[23] A regression of money-metric utility from the less equitable option on the DG number j, for $j \in \{1, 2, 3, 4\}$,

experiences more intense negative sensations such as guilt when the Receiver's payout in a foregone more-equitable allocation is greater.

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 is based on 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.

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 = $[-0.51, -0.34]$).[24] The latter finding is consistent with Deciders experiencing negative sensations from comparison effects less intensely 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 each option receive greater-than-average utility from that option. This pattern is a natural consequence of preference heterogeneity and self-selection.

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.

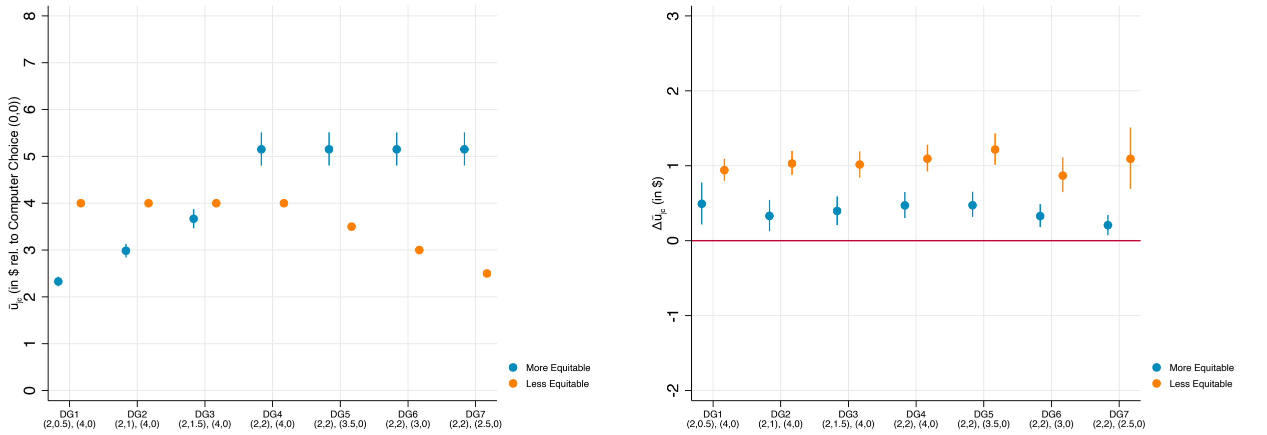
[24] A regression of money-metric utility from the less equitable option on the DG number j, for $j \in \{4, 5, 6, 7\}$, yields a coefficient of 0.15, with a 95% CI of $[0.12, 0.18]$.

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 finding 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.

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



(a) Average utility when the computer determines the allocation

(b) Welfare gain when computer chooses 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 across all choice sets 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 (actually or counterfactually) 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.

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. Quantitatively, 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 (95% CI [0.24,0.50]) in the case of more-equitable allocations.

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, suppose the Planner is concerned, at least in part, with the Decider's welfare. Standard welfare economics instructs the Planner that the option the Decider would choose for herself is also the one that is best for her when instead the Planner makes the choice. 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, for example, mimics 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 discrepancy.

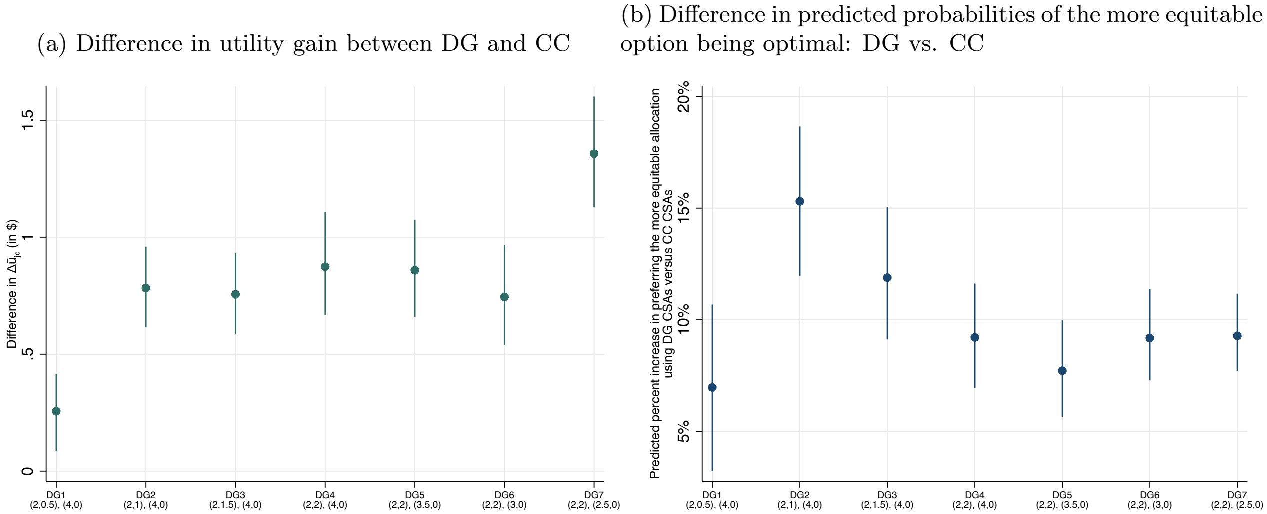
Panel (a) of Figure 7 shows that, on average across the seven menus, the money-metric utility difference between the more equitable and the less equitable allocation is \$0.80 higher in the DGs than in the main CC module.[25] A Planner who applies standard welfare methods will thus underestimate how much Deciders' welfare decreases when the Planner chooses more equitable allocations instead of less equitable ones.

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 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

[25] Note that Figure 7a is not directly comparable to Figure 6b because Figure 6b considers only the alternatives actually chosen by the Deciders (meaning that the sample changes across the different choice sets due to selection), whereas Figure 7a compares chosen and unchosen alternatives for every Decider in each choice set (meaning that there is no selection and the sample is constant across the choice sets).

exaggerate the fraction of Deciders who are better off when someone else assigns the less equitable allocation. The degree of exaggeration .ranges from 7% to more than 15%, depending on the decision problem.[26]

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 more equitable option being optimal: DG vs. CC

Note: Panel (a) reports the average utility gains from choosing the more equitable allocation instead of 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 sample is held constant across all choice sets in both panels. The 95 percent confidence intervals are reported as the vertical bars and calculated using bootstrap with 1,000 resampling clusters at the participant level.

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 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. Our methods allow us to study this possibility directly. We show in Appendix H that Deciders derive less utility from the more equitable allocations when they know the unchosen alternative ($0.38, with a 95% CI of [0.07,0.72]). However, we do not find a significant difference for the less equitable allocation.

Due to these patterns, when we perform essentially the same analysis for the alternative CC format, we conclude that, on average across the seven menus, there is little difference between

[26] We rely on DG CSAs rather than on actual DG choices to ensure that the statistics in the DG and CC modules are computed in a fully comparable manner. This distinction could be important if, e.g., the model linking CSAs to choice is mispecified, or if the reported CSAs are sufficiently noisy. The results in Section 5 suggest that neither issue is particularly important.

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]). Thus, even when people know the unchosen option, a Planner who applies standard welfare methods continues to underestimate how much Deciders' welfare decreases when the Planner replaces the less equitable allocations with more equitable alternatives. See Appendix H for details.

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 8, which plots participants' utility, $\bar{u}_{jc}$, for each alternative in each of the OO games. The figure provides three main insights.

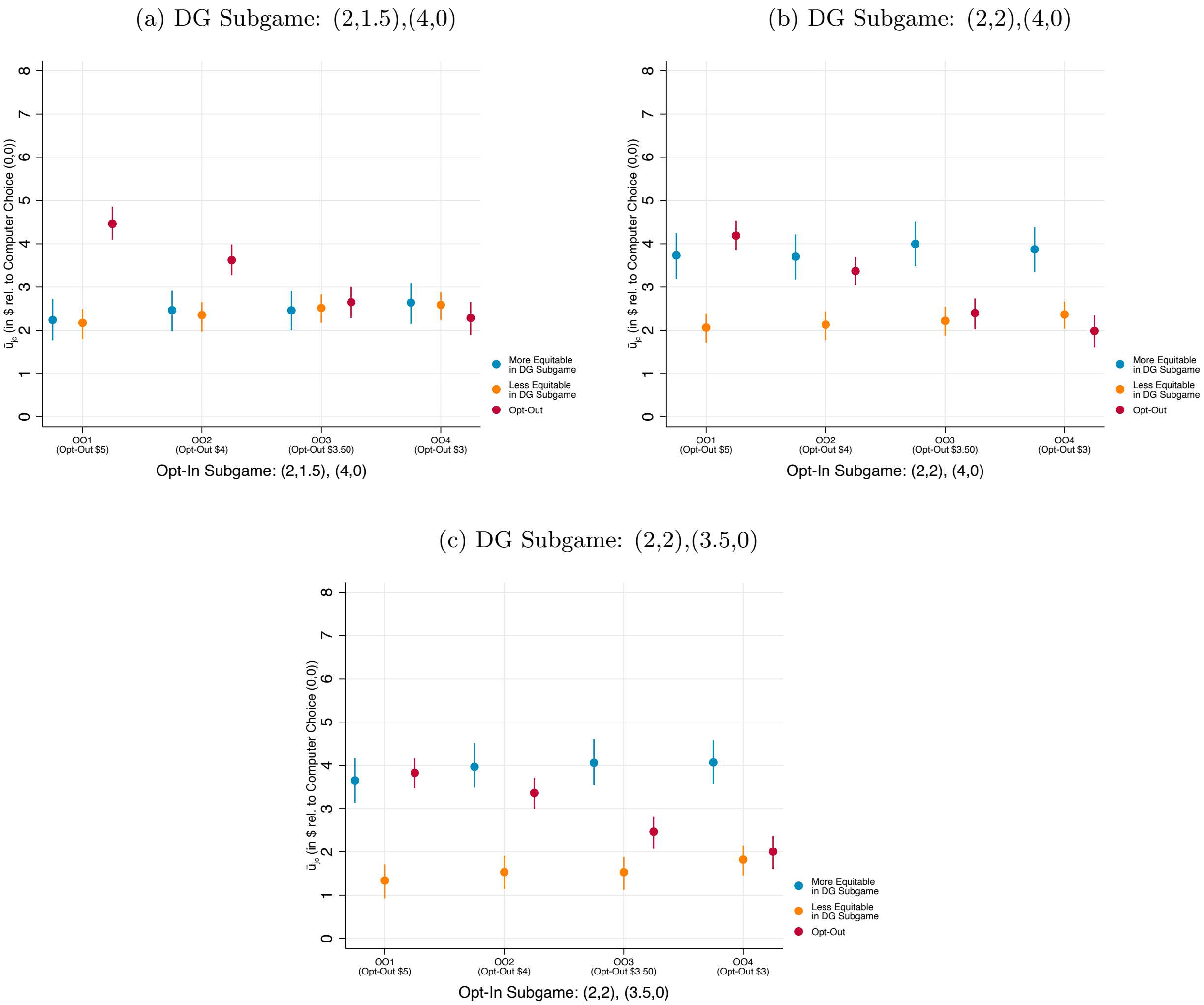
First, the money-metric utility Deciders would receive from choosing the opt-out option themselves is significantly lower than if the computer assigned that allocation exogenously (in the main CC module).[27] This finding provides direct evidence that the Non-Comparability Problem affects meta-choices: participants experience disutility from the experience of opting out.

Second, opting out nevertheless generates a less negative experience than opting into the DG subgame and choosing the less equitable option. In Panels (a) and (b), Deciders receive higher utility in OO 2 from opting out for \$4 than from opting in and choosing $(4, 0)$. Similarly, in Panel (c), Deciders receive higher utility in OO 3 from opting out for \$3.50 than from opting in and choosing $(3.50, 0)$. 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 the difference between the main and alternative CC modules, as revealed in Appendix Figure 6. That is, participants derive less utility from an outcome when additional alternatives are available. These effects are more muted for the less equitable options in the DG subgames.

[27] Again, recall that by construction, the money-metric utility Deciders obtain from an exogenously-imposed allocation of $(\$X, 0)$ in the main CC module is simply $\$X$.

Figure 8: 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 and 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.

A secondary manifestation of the contrast effects in Figure 8 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, 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 does the opposite. 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.

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. $*p < 0.1$, $**p < 0.05$, $***p < 0.01$.

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 differences in money-metric utility Deciders obtain from the DG (subgame) options in Figures 8 and 5, the presence of an opt-out option significantly reduces the value of playing a DG; we report the differences in Column (5). This finding provides additional quantitative confirmation that adding options has negative contrast effects on the utility Deciders derive from existing alternatives.

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 elaborates on this point by replicating 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., Clark and Oswald, 2002; Finkelstein et al., 2013). 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 Relevance for Consequential Decisions

Having focused so far on small-stakes decisions in a laboratory experiment, we next document the relevance of our analysis for consequential decisions. The survey evidence presented in this section makes four points. First, emotions associated with the act of choosing are important relative to other consequences even in major economic decisions. Second, emotions associated with act of choosing are also significant for consequential meta-choices. Third, as a result of the first point, the NCP arises in practice. Fourth, as a result of the second point, meta-choices do not resolve the NCP in practice. We address the first two questions in Section 7.1 and the second two in Section 7.2. While our method can potentially resolve the NCP in the settings we examine, full applications are beyond the scope of the current study.

Our analysis is based on three pre-registered surveys which we fielded through Prolific Academic on March 31, 2025. Each survey involved 500 participants, of whom 482, 468, and 496 passed both our attention check (see survey instruction appendix L) and our AI detection check (see appendix

J).[28] The surveys took 11.3, 14.4, and 5.8 minutes, respectively.

7.1 The importance of emotions associated with the act of choosing

We begin by providing evidence that emotions associated with the act of choosing are important relative to other consequences in major economic decisions. Our strategy is to show that these emotions cause people to avoid spending time on decisions, even when this behavior impairs decision quality. The existing literature discusses a specific variant of such avoidance, the "ostrich effect," wherein the negative emotions are associated with information acquisition (Karlsson et al., 2009; Olafsson and Pagel, ming). Our elicitations are more general, and our results show that the contemplation of difficult choices regularly generates negative affect for other reasons.[29]

We opened Survey 1 by asking respondents if they avoid thinking about important decisions because doing creates negative emotions. We instructed them to identify those decisions and describe the associated emotions. Following the principles set out by Haaland et al. (2024), we used open-ended questions to avoid priming respondents with our own preconceptions. On average, respondents listed 3.4 decisions.

Next, we asked respondents to categorize the decisions they mentioned using a list of categories we developed based on open-ended responses from a pilot survey. This novel self-categorization approach avoids investigator bias as well as the potential arbitrariness of LLMs. Significantly, no participant selected "None of the Above," which indicates that our list was appropriate and comprehensive.

The frequency with which respondents reported avoidance was highest for financial decisions (58%). Reported frequencies were also high for decisions involving family (43%), social matters (40%), health (38%), and career (33%). Reported avoidance is less common for decisions involving personal development (28%), daily tasks (20%), education (6%), and legal matters (4%).

Survey 2 focused more narrowly on decisions involving financial matters, health, and career planning. Our main objective was to gauge the importance of negative emotions relative to more standard explanations for decision-avoidance such as lack of time, procrastination, and pessimism concerning the benefits of further deliberation. The survey first asked respondents if they believed they would have better financial, health, or career outcomes if they devoted more time to those decisions. 79.5, 84.4, and 71.4 percent, respectively, indicated that they would. We then asked those respondents to describe their reasons for limiting the time they spent on those decisions. We used the same self-categorization method as in Survey 1 to sort their open-ended responses into categories.[30]

Figure 9 presents the results. Three categories involve emotions associated with the act of choosing: "I experience negative emotions such as stress, anxiety, fear," "It makes me feel overwhelmed,"

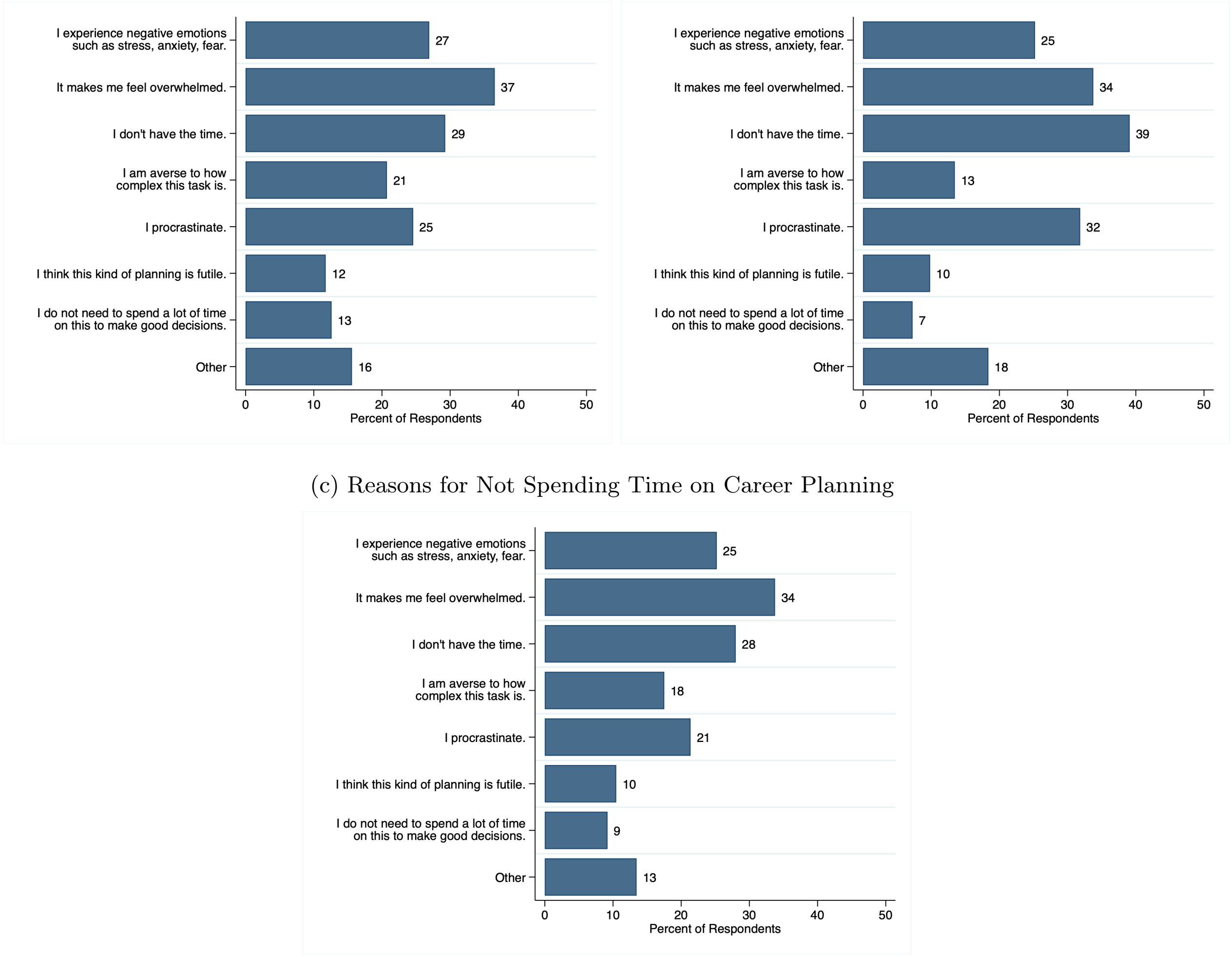
[28] The exclusion criteria for our failing attention checks or AI-use checks were fully pre-registered. See the OSF registration link here.

[29] In open responses, few participants mention that new information is what drives their negative emotions.

[30] Respondents could list multiple categories, and on average each respondent selected 1.8, 1.6, and 1.8 categories, respectively.

and “I am averse to how complex this task is.”

Figure 9: Reasons for not spending more time on financial, health, and career decisions



(a) Reasons for Not Spending Time on Financial Planning

(b) Reasons for Not Spending Time on Health Planning

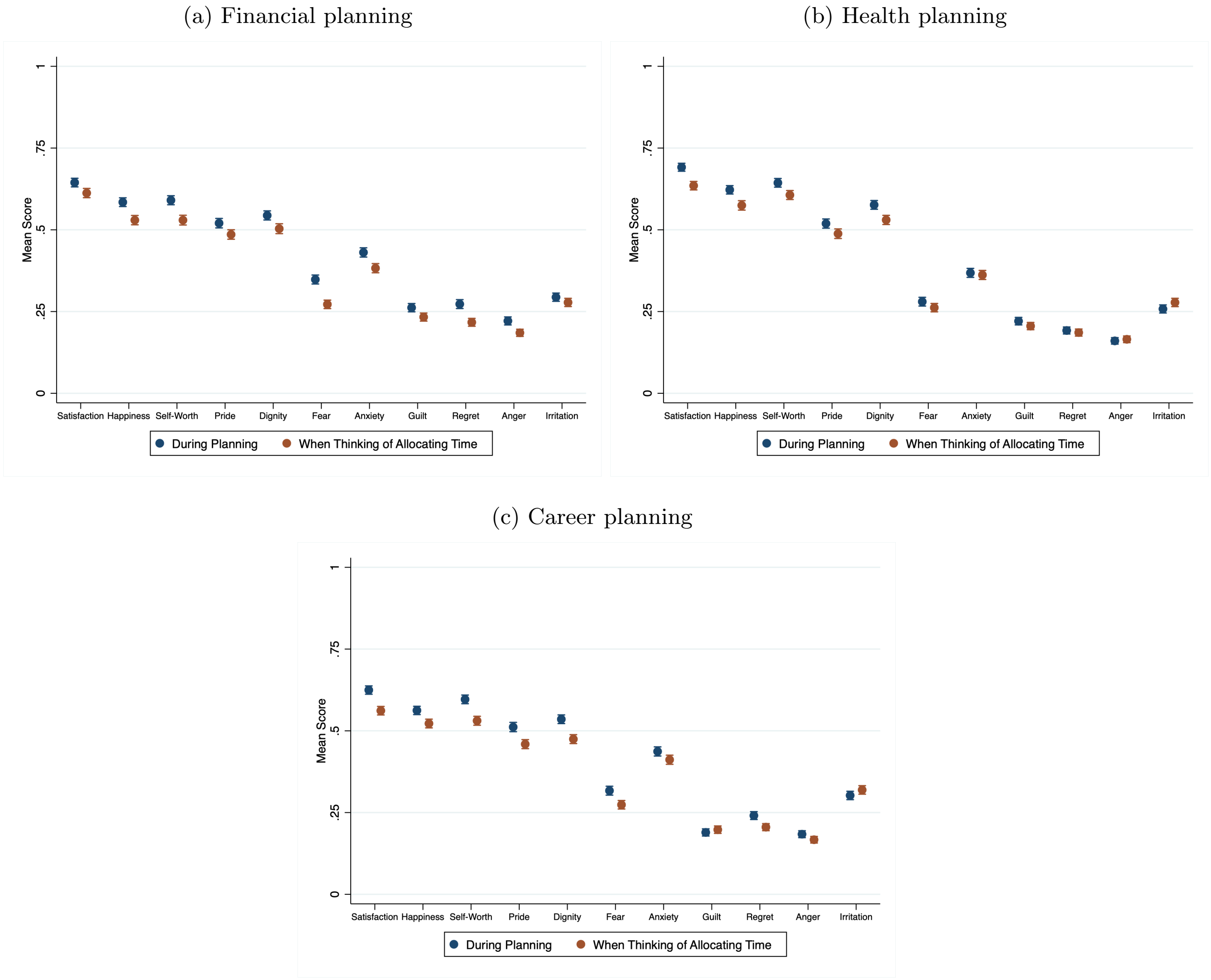
(c) Reasons for Not Spending Time on Career Planning

Note: This figure reports the percent of respondents citing reasons for not spending time on planning in the following categories: “I experience negative emotions such as stress, anxiety, fear,” “It makes me feel overwhelmed,” "I don’t have the time," “I am averse to how complex this task is," "I procrastinate," "I think this kind of planning is futile," "I do not need to spend a lot of time on this to make good decisions," and "Other." Panel (a) presents responses related to financial planning, Panel (b) to health planning, and Panel (c) to career planning. Respondents included in the figure answered "Yes" to a prior question asking whether they would achieve better financial, health, or career outcomes if they devoted more time to decision-making. Respondents then self-classified their reasons into the 8 categories. The sample consists of 372, 395, and 334 respondents, respectively.

Across the three categories of decisions, between 45% and 49% of respondents listed at least one of these reasons. Between 34% and 37% cited concerns about feeling overwhelmed; it was the most common reason given for avoiding two of the three types of decisions and the second most common reason for avoiding the third. In comparison, 28% to 39% cited limited time and 21% to 32% mentioned procrastination.

The second part of Survey 2 investigated whether meta-choices also involve emotions associated with the act of choosing. Specifically, we asked about the degree to which making plans involving finances, health, and career (choices) impacted specific emotional responses, and about the degree to which thinking about setting aside time for such planning (meta-choices) impacted those responses. To facilitate quantitative comparisons, we specified 11 CSAs which we assessed using Likert scales (as in our main experiment). Open-ended responses would have made such comparisons impractical.

Figure 10: CSAs during financial/health/career planning, and during setting aside time for planning


(a) Financial planning

(b) Health planning

(c) Career planning

Note: This figure reports participants' self reports of how the 11 CSAs would be impacted by (i) financial, health, and career planning, and (ii) thinking about setting aside time for financial, health, and career planning. Blue points correspond to answers about the planning phase, and red points correspond to answers about setting aside time for planning. Panel (a) reports average CSA scores for financial planning, Panel (b) for health planning, and Panel (c) for career planning. CSA scores, ranging from 1 to 5, were normalized by subtracting 1 and dividing by 4. Vertical bars represent the 95% confidence intervals. The sample consists of 468 respondents.

Results appear in Figure 10.[31] Participants report that both planning and thinking about setting aside time for planning impacts all CSAs. While the latter responses are typically somewhat attenuated versions of the former, the magnitudes are comparable and the patterns are similar.

7.2 The practical relevance of the NCP

Documenting the importance of emotions experienced during the act of choosing establishes a critical precondition for the NCP. To illustrate, suppose the ultimate consequences of a financial plan x_2 yield higher utility $u(x_2)$ than an alternative plan x_1, but that working out the details of x_2 requires more time and hence involves higher emotional costs, $e(x_2) > e(x_1)$. Suppose further that people do not experience those emotional costs when the Planner chooses on their behalf. In that case, people's decisions are non-comparable to those of the Planner. As a result, the Planner should select x_2 (because $u(x_2) > u(x_1)$) even when people choose x_1 (because $u(x_2) - e(x_2) < u(x_1) - e(x_1)$). The fact that people believe such emotions lead them to make lower quality decisions involving their finances, health, and careers is consistent with this possibility.

In this section, we provide more direct evidence of the NCP in three settings. Appendix I.1 reports additional survey results supporting our conclusions.

The first setting again involves the domains of finance, health, and career. In the second part of Survey 2, we asked participants how they would feel about a program that would provide them with regular reminders (through text messages, emails, push notifications, and occasional phone calls) to make time for planning. We described two scenarios. In one, the individual signs up for the program themselves. In the other, the government mandates participation. Appendix Figure A13 reports the results. For each of these scenarios, we elicited the same 11 CSAs described in the preceding section. In each of the three domains of finance, health, and career, the difference between the responses for the two scenarios is highly statistically significant ($p < 0.01$) for all CSAs other than guilt; i.e., for 10 out of the 11 CSAs. The largest difference in each domain is for anger, which is much lower when choosing the reminders oneself rather than having them mandated. Volunteering for reminders yields more satisfaction, happiness, self-worth, pride, and dignity, as well less fear, anxiety, regret, and especially irritation. The differences in average responses between the two scenarios for these 10 CSAs are substantial, exceeding 10% of the difference between the highest and lowest possible Likert rating for 9 CSAs in the context of finances, 9 in the context of health, and 7 in the context of career. A Planner who judges the desirability of mandated reminders based on standard "revealed preference" analyses would therefore substantially overstate the benefits of the mandate.

The second setting concerns retirement savings plans. In the second part of Survey 1, we asked respondents how they would feel about increasing the penalty on early withdrawals from 10% to 30%, which we describe as a strategy for helping people exercise self-control. Once again, we elicited participants' reactions to two scenarios, one in which the individual chooses the higher penalty, the other in which the government imposes it. In addition to eliciting CSAs as above, we also asked

[31] As before, we normalize Likert scale responses by subtracting 1 and dividing by 4.

participants whether, in each scenario, they would consider themselves better off with the policy than without it. Critically, 44.0 percent said the policy would make them better off if they chose it, but only 19.3 percent said it would make them better off if mandated. In addition, we find highly statistically significant differences between the two scenarios ($p < 0.01$) for 10 of the 11 CSAs (all except regret), as summarized in Appendix Figure A12. The largest differences are for anger and irritation, which are much higher with mandates. Choosing the higher penalty oneself also yields far more satisfaction, happiness, self-worth, pride, and dignity, as well as slightly more guilt, less fear, and less anxiety. The differences in average responses between the two scenarios for these 10 CSAs are also substantial, exceeding 10% of the difference between the highest and lowest possible Likert rating for 9 out of 11 CSAs, 20% for 4 CSAs, and 30% for two CSAs. A Planner who judges the desirability of higher early withdrawal penalties based on standard "revealed preference" analyses would therefore substantially overstate the benefits of the penalty.

Results for the first two settings suggest a future research agenda that leverages our methods to better understand the welfare effects of government intervention. They raise the possibility that—even in settings where externalities or consumer mistakes make people unambiguously worse off—government intervention to counteract these market failures may not increase social welfare.

The third setting, which involves charitable donations, is based on a well-known study by DellaVigna et al. (2012) that uses meta-choices to quantify aversion to solicitation. In Survey 3, we ask respondents to consider three scenarios. In all three, someone arrives at their door soliciting contributions to a non-profit children's hospital. For the first scenario, the respondent opens the door, discovers that the visitor is a solicitor, and decides whether to contribute. For the second scenario, they are aware that the person at the door is a solicitor before opening it (having received prior notice); they decide whether to answer the door and, if so, whether to contribute. For the third scenario—which is not part of the DellaVigna et al. (2012) study, but which we included to investigate the NCP—the person is not at home when the solicitor arrives and is therefore unable to answer the door, but they are nevertheless aware of the visit (having seen the solicitor on a doorbell camera). In each scenario, we ask respondents to report the same CSAs as in our main experiment for every option, as well as their likely selection.

In the spirit of DellaVigna et al. (2012), one might attempt to measure the welfare effect for the respondent of a ban on solicitation by assessing their willingness-to-accept (or willingness-to-pay) to answer the door in Scenario 2 (a meta-choice). That method implicitly invokes the comparability hypothesis. Specifically, it assumes that not opening the door in Scenario 2 provides the same hedonic experience as Scenario 3 (wherein, as with a ban, the respondent is not responsible for avoiding solicitation). It also assumes that, conditional on opening the door in Scenario 2, each option (contributing and not contributing) provides the same hedonic experience as in Scenario 1 (wherein, as without a ban, the respondent cannot avoid solicitation).

Our survey results imply that the comparability hypothesis is violated in this setting. We find highly statistically significant differences between not opening the door in Scenarios 2 and 3 ($p < 0.01$) for 6 of the 7 CSAs (all except financial satisfaction). When restricting to the pre-registered sub-sample of those with revealed prosocial inclinations, Appendix Figure A14 shows that when choosing not to open the door themselves, people experience more guilt and unfairness, and less happiness, satisfaction, pride, and fairness. The differences in average responses between the two scenarios are at least 10% of the difference between the highest and lowest possible Likert rating for happiness, satisfaction, and fairness, more than 20% for unfairness, and more than 30% for guilt. Notably, these differences are similar to and in some cases larger than those we find in our main experiment when we compare the opt-out option to receiving the same allocation in the CC module. These findings call the validity of the meta-choice method into question.

8 Concluding Remarks

This paper develops a conceptual and empirical framework that combines SWB and revealed preference methods to measure 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, which meta-choices do not resolve. We also use a series of surveys to illustrate the relevance of the Non-Comparability Problem for consequential economic decisions and associated government interventions.

Prior studies have used meta-choices to assess welfare effects associated with the act of choosing in a variety of settings. Each provides a potentially fruitful context for applying our framework. One set of studies uses meta-choices 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, to achieve a sense of pride, or out of an inherent desire for authority. The experience of exercising control might also 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. Alternatively, the act of choosing a restriction may signal positive personal attributes (either to the chooser or to others) or deplete self-control. A third set of studies uses meta-choices 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. One could also use our method to investigate the intriguing possibility that welfare-relevant sensations associated with act of choosing die off for sufficiently high-level menus, in which case appropriately designed meta-choices might also overcome the NCP. However, it would be important to determine the reasons those sensations die off. If the explanation is that people have trouble seeing through complex multi-stage problems, this approach would not offer a valid solution.

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 reference-dependence or sunk cost effects. Do these patterns reflect welfare-relevant sensations? For example, if one's choice of alternative a is implemented with probability p, while an exogenous alternative b is implemented with probability $1 - p$, how does p influence people's anticipated and realized welfare when b is realized? Through an extension of our approach, one could determine whether people correctly anticipate their CSAs when exhibiting patterns such as loss aversion, a question that is central to models such as those of Kőszegi and Rabin (2006). 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 challenges 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.

Almås, I., A. W. Cappelen, and B. Tungodden (2020). Cutthroat capitalism versus cuddly socialism: Are americans more meritocratic and efficiency-seeking than scandinavians? *Journal of Political Economy 128*(5), 1753–1788.

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., G. Carroll, O. Heffetz, and M. S. Kimball (2014, October). Aggregating local preferences to guide policy. Working paper.

Benjamin, D. J., K. Cooper, O. Heffetz, and M. Kimball (2024). From happiness data to economic conclusions. *Annual Review of Economics 16*, 359–391.

Benjamin, D. J., O. Heffetz, M. S. Kimball, and N. Szembrot (2013). Aggregating local preferences to guide marginal policy adjustments. *American Economic Review 103*(3), 605–610.

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? *Economics 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.

Capozza, F. and K. Srinivasan (2024, April). Who should get money? estimating welfare weights in the u.s. CESifo Working Paper 11086, CESifo. URPP Equality of Opportunity Discussion Paper Series No. 50, University of Zurich.

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 (2006, 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.

Fleurbaey, M., E. Schokkaert, and K. Decancq (2009). An experimental study on individual choice, social welfare, and fairness. *European Economic Review 53*(4), 385–400.

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).

Haaland, I., C. Roth, S. Stantcheva, and J. Wohlfart (2024). Measuring what is top of mind. *working paper*.

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.

Kagan, S. (1997). *Normative ethics*. Routledge.

Karlsson, N., G. Loewenstein, and D. Seppi (2009). The ostrich effect: Selective attention to information. *Journal of Risk and Uncertainty 2*(38), 95–115.

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.

Kőszegi, B. and M. Rabin (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics 121*(4), 1133–1165.

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

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.

List, J. A. (2007, June). On the interpretation of giving in dictator games. *Journal of Political Economy 115*(3), 482–493.

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.

Olafsson, A. and M. Pagel (forthcoming). The ostrich in us: Selective attention to personal finances. *The Review of Economics and Statistics*.

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*.

Saez, E. and S. Stantcheva (2016). Generalized social marginal welfare weights for optimal tax theory. *American Economic Review 106*(1), 24–45.

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 Theory

A.1 Proof of Proposition 1

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 $j = 2, ..., J$, likewise define

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

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

With this construction, we have

$$\begin{aligned} \tilde{V}(x, \mathcal{M}^1, ..., \mathcal{M}^J) &= \tilde{u}(x) + \tilde{\pi}^1(x, \mathcal{M}^1) + \sum_{j=2}^{J} \tilde{\pi}^j(\mathcal{M}^{j-1}, \mathcal{M}^j) \\ &= V(x, \mathcal{M}^1, ..., \mathcal{M}^J) + f^J(\mathcal{M}^J) \end{aligned} \tag{6}$$

Furthermore, the $\tilde{\pi}^j$ functions have the same property as the π^j function: when there is no level-j choice, the individual experiences no level-j utility (condition (2)). Consequently, $\tilde{V}$ belongs to the same class of utility representations as V.

Importantly, if V rationalizes choices for levels 1 through J, so does $\tilde{V}$. The reason is that such choices do not involve the selection of $\mathcal{M}^J$, and thus adding $f^J(\mathcal{M}^J)$ to V changes none of its implications for behavior. The conclusion is that u is fundamentally unidentified because the construction produces an observationally equivalent utility function for *any* $\tilde{u}$.

A.2 Problems with testing level-j menu independence for $j < J$

The construction in Section A.1 shows that it is impossible to test level-j menu independence for $j < J$ using choice data for levels 1 through J. For if V satisfies level-j menu independence, then $\tilde{\pi}^j(\mathcal{M}^{j-1}, \mathcal{M}^j) = -f^{j-1}(\mathcal{M}^{j-1}) + f^j(\mathcal{M}^j)$, which means we can choose f^j so that $\tilde{V}$ does not, and we have already established that V and $\tilde{V}$ are observationally equivalent for the hypothesized data.

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 $(\mathcal{M}^2, ..., \mathcal{M}^{K-1})$.

A.3 Welfare analysis for policies other than x-mandates

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 $J \geq s$. 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 obtain $\tilde{\pi}^j(\mathcal{M}^{j-1}, \mathcal{M}^j)$ for $j > s$. Reversing the recursive construction, we obtain $\tilde{\pi}^j(\mathcal{M}^{j-1}, \mathcal{M}^j)$ for $j < s$, and ultimately

$\tilde{u}$.[32]Putting these components together, we obtain a utility function $\tilde{V}$ within the pertinent class that is related to V, as before, by equation (6). 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), the welfare impacts implied by $\tilde{V}$ versus V differ by $f^s(\mathcal{N}^s) - f^s(\mathcal{M}^s)$. Because we are free to select any function f^s, the impact is ambiguous.

A.4 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.[33] 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.[34]

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.[35]

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

[32] 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 $\tilde{u}(x) = u(x) + f^0(x)$.

[33] 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.

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

[35] 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 33. 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.

(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 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.

A.5 Formal description of the method

Let z denote a vector (or bundle) of mental states, and let Z be the set of feasible mental-state bundles. We assume that an individual's preferences correspond to a well-behaved binary relation, $\succsim$, over such vectors. To allow for the full range of considerations that can give rise to the Non-Comparability Problem, we assume we can write the anticipated 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 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 anticipated 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 $z_1 = \zeta(x_1, \mathcal{M}_1^1, ..., \mathcal{M}_1^{K-1}, \mathcal{M}^K, d) \succeq z_2 = \zeta(x_2, \mathcal{M}_2^1, ..., \mathcal{M}_2^{K-1}, \mathcal{M}^K, d)$.[36] Repeating this two-step process multiple times generates a dataset consisting of perceived 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 anticipated mental state bundles and associated choices. Our empirical framework, summarized in Section 4, exemplifies the use of the resulting data to estimate $\succeq$.

A.6 Menu dependence in preferences over mental states

Suppose the individual evaluates the options in decision problem D while imagining that the menu of mental state bundles is $Z = \xi^0(D)$. A possible concern is that, when contemplating any given

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

option $(x, \mathcal{M}^1, ..., \mathcal{M}^K)$ for D, the anticipated mental state bundle z might depend on $\xi^0(D)$, which means we can write it as $z = \zeta(x, \mathcal{M}^1, ..., \mathcal{M}^K, d, \xi^0(D))$.

Despite this menu dependence, the mental-statism hypothesis implies that the preference relation $\succeq$ still applies to z, rather than to pairs (z, Z); each z already incorporates any preference-relevant aspects of menu-dependent reactions. It follows that the individual chooses the $\succeq$-maximal element of the following set:

$$\xi\left(D, \xi^0(D)\right) \equiv \{\zeta(x, \mathcal{M}^1, ..., \mathcal{M}^K, d, \xi^0(D))\}_{(x, M^1, ..., \mathcal{M}^K) \in \mathbb{M}(D)}$$

If we elicit direct measures of the mental states associated with each option, $\zeta(x, \mathcal{M}^1, ..., \mathcal{M}^K, d, \xi^0(D))$, we can therefore recover $\succeq$, just as we would if there were no mental-state-menu dependence.[37] Likewise, we can use the recovered preferences to evaluate well-being in any other setting D' for which we have similar data.

[37] An individual who notices that $\xi\left(D, \xi^0(D)\right) \neq \xi^0(D)$ might reevaluate the options imagining that the mental state bundle is $\xi\left(D, \xi^0(D)\right)$ rather than $\xi^0(D)$. Indeed, they might do so repeatedly until the process converged to a fixed point, or they might locate a fixed point through some other intuitive process. Interpreting $\xi^0(D)$ as a fixed point of the mapping $\zeta(x, \mathcal{M}^1, ..., \mathcal{M}^K, d, \xi')$ does not alter our conclusions. Our reasoning likewise encompasses any other process governing the generation of $\xi^0(D)$, irrespective of whether it yields a fixed point.

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 trade-offs 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

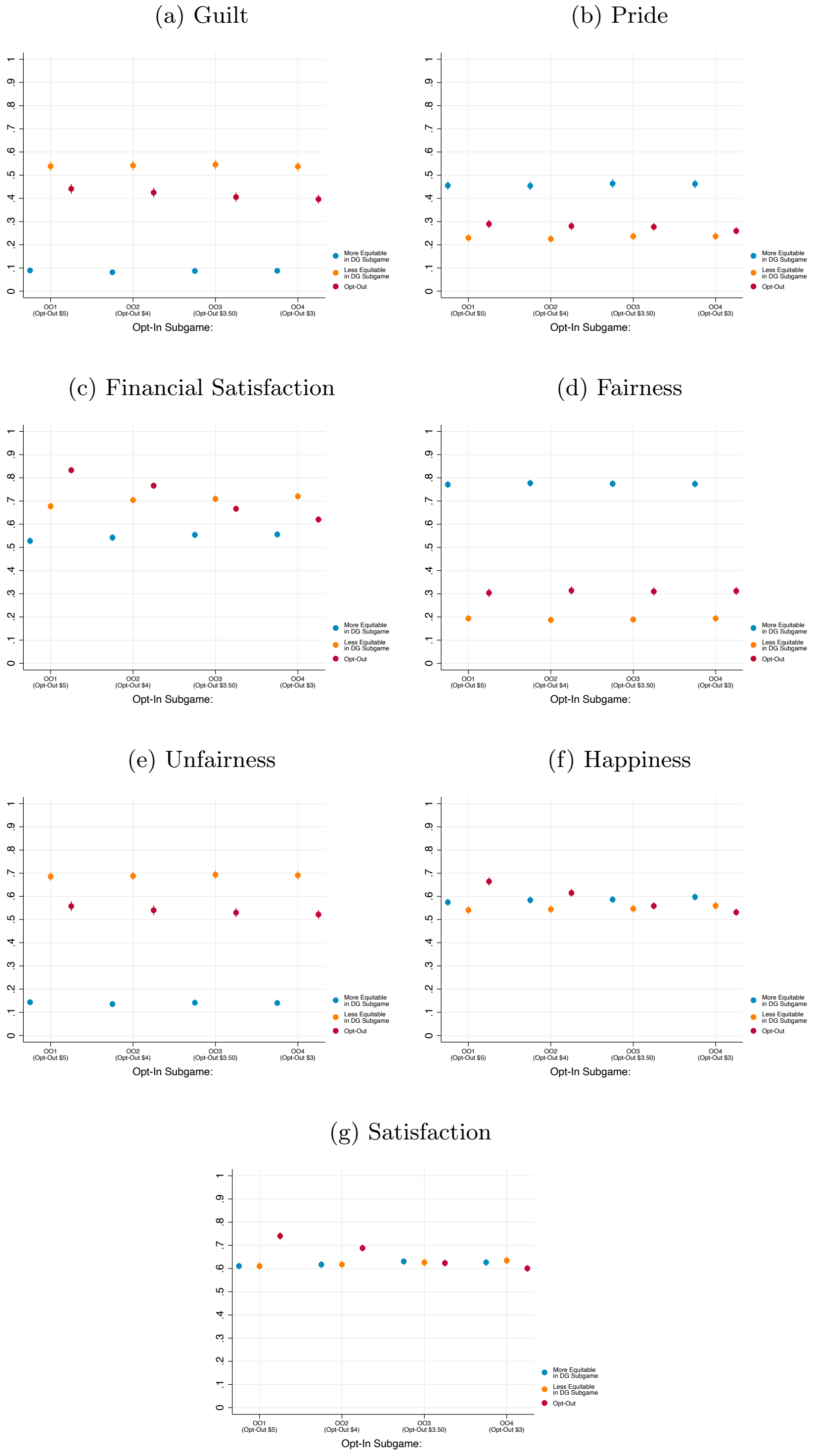
C.1 Sample Demographics

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.

Figure A1: Average CSAs in the OO



(a) Guilt

(b) Pride

(c) Financial Satisfaction

(d) Fairness

(e) Unfairness

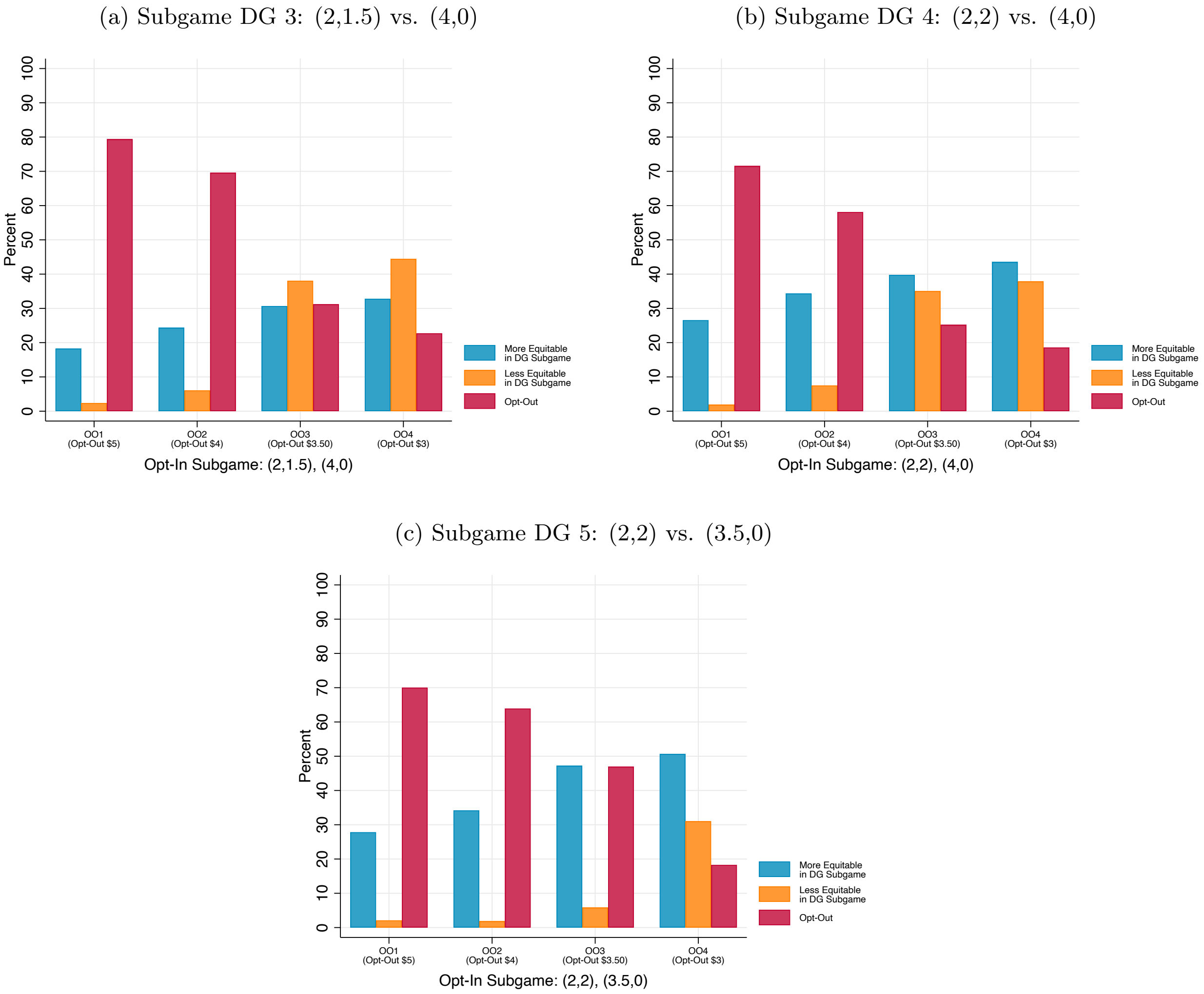
(f) Happiness

(g) Satisfaction

Note: This figure reports the average CSA ratings for each option of each choice set in the OO. We normalize the CSA responses, which range from 1 to 5, by subtracting one and diving by four, such 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. We report CSAs across all DG subgames. The 95 confidence interval is shown in the vertical bars.

C.2 Additional Results About CSA and Choice in the OO games

Figure A2: 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 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 report coefficients as odds ratios to make the constant term interpretable. We find small changes in the coefficients and a small (though statistically significant) constant term.

Table A2: Association between choices and CSAs

	(1)	(2)	(3)	(4)	(5)	(6)
	Logit (no cons)	Logit (w cons)	IV Logit (no cons)	IV Logit (w cons)	Logit (no cons)	Logit (w cons)
	Choosing	Choosing	Choosing	Choosing	Choosing	Choosing
	More Equitably	More Equitably	More Equitably	More Equitably	More Equitably	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		-0.29***		-0.32***		-0.36***
		(0.06)		(0.08)		(0.06)
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. $*p < 0.1$, $**p < 0.05$, $***p < 0.01$.

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 principal component analysis, and shows that Component 1 and Component 2 explain a large proportion of the variance in CSAs.

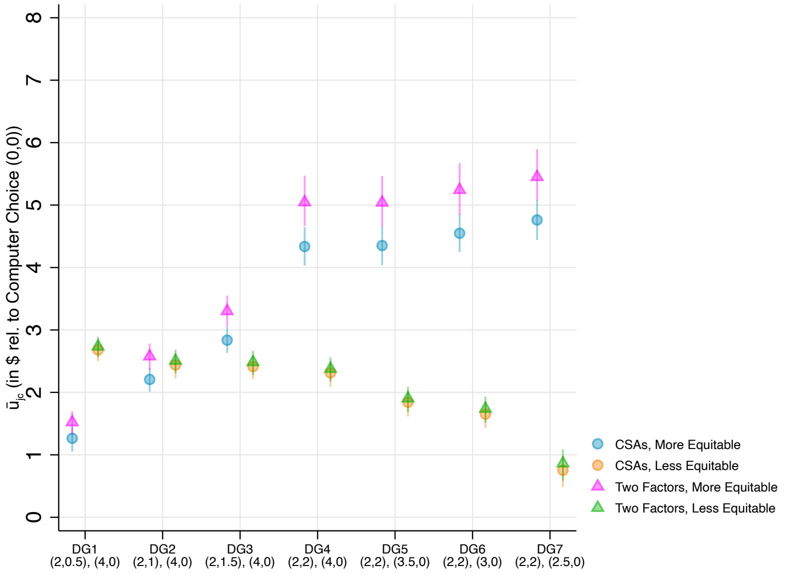
We can then apply the methodology from Section 4 to the first two factors from Table A3. Figure A3 shows that the estimates that we obtain for Decider's money-metric utility are similar to the ones we obtain when using all seven CSAs. This again suggests that the first two factors appear to span much 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 A3: 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 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.)

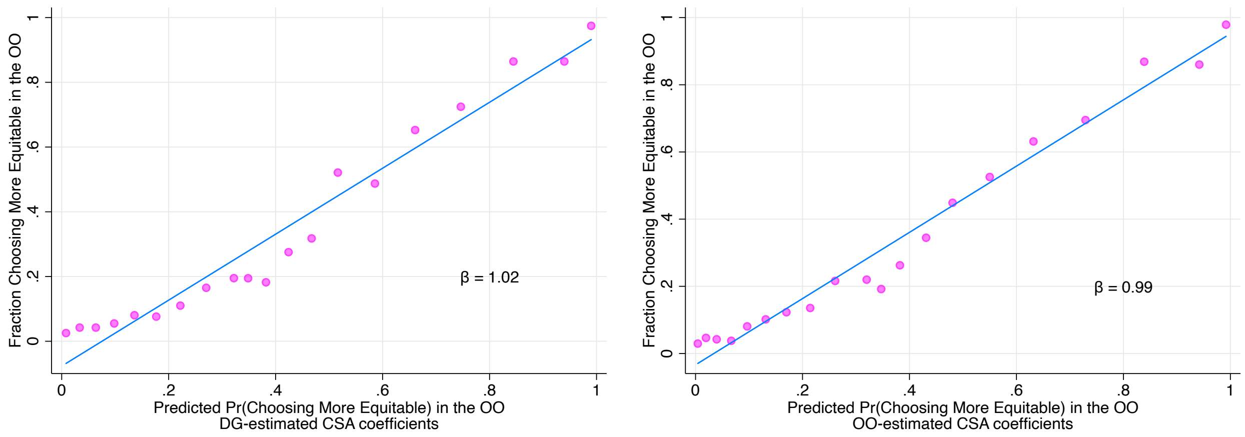
Figures A4 and A5 are analogous to Figure 4, but concern choice in the OO games. The figures plot the actual versus predicted likelihood of choosing the more equitable allocation in the opt-in subgame (figure A4) and of choosing to opt out (figure A5). Panel (a) in each of the figures forms predictions 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. *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$.

Figure A4: Observed versus fitted probabilities of choosing more equitably in the OOs

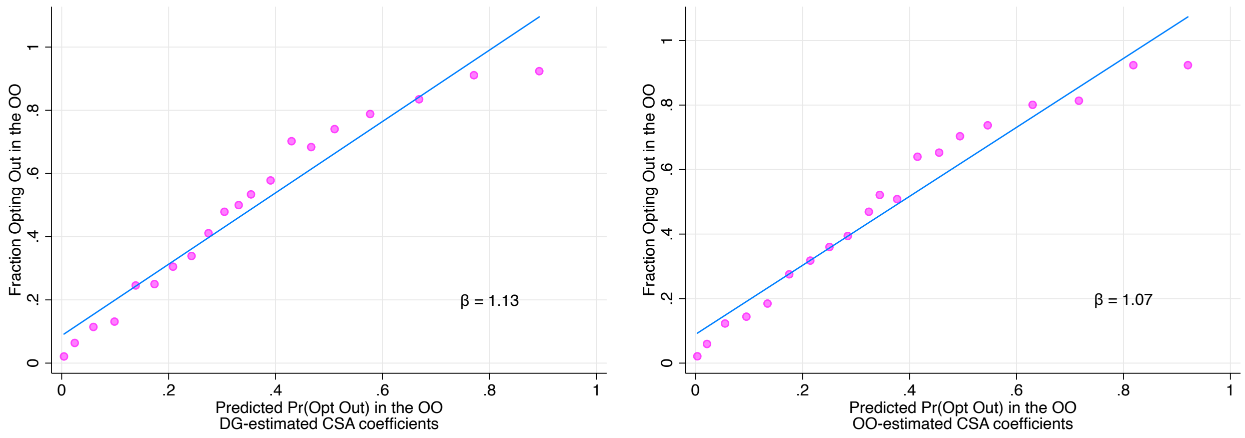


(a) Model estimated using DG data

(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 A5: 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)	(2)
	OLS	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. *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$.

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

	(1)	(2)	(3)
	OLS	OLS	Logit
Dependent Var.	Δ Happiness	Δ Satisfaction	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. *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$.

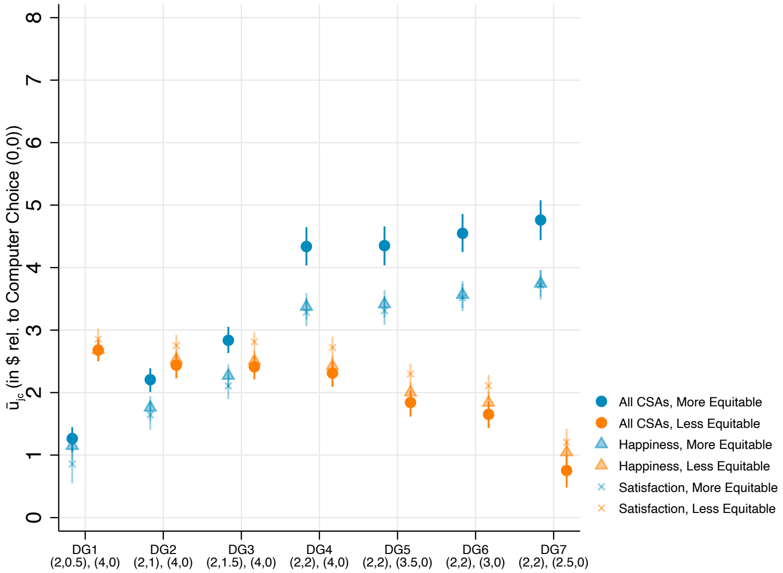
E.2 Bias from Using Happiness and Satisfaction to Estimate Deciders' Welfare

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

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

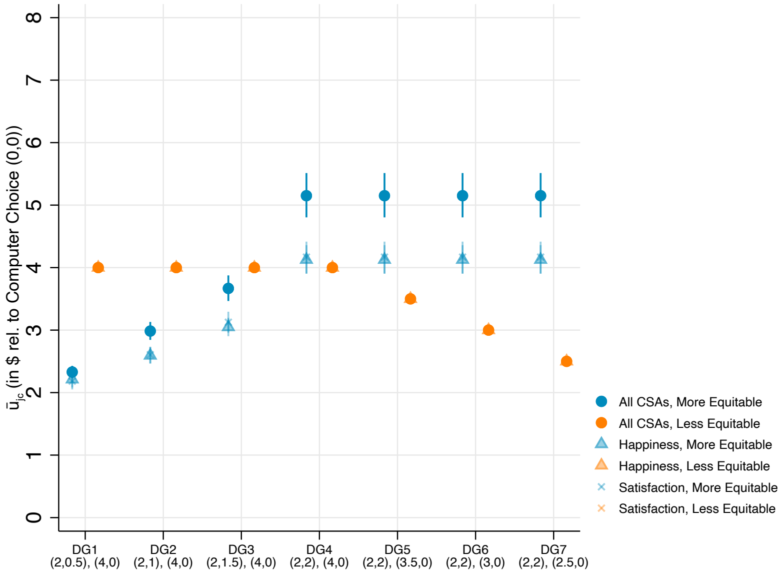
The term $v(X_{ijc})$ is the observable, deterministic component of utility and is locally linear and additive in the CSAs, such that $v(X_{ijc}) = X_{ijc}\beta$, where β is a vector of coefficients and X_{ijc} is a vector of all seven CSAs . When we assume happiness and satisfaction are sufficient statistics, 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 A6, A7, and A8 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 based on all seven CSAs) in 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 A6: Comparing welfare estimates in the DG module using all CSAs versus just happiness or satisfaction



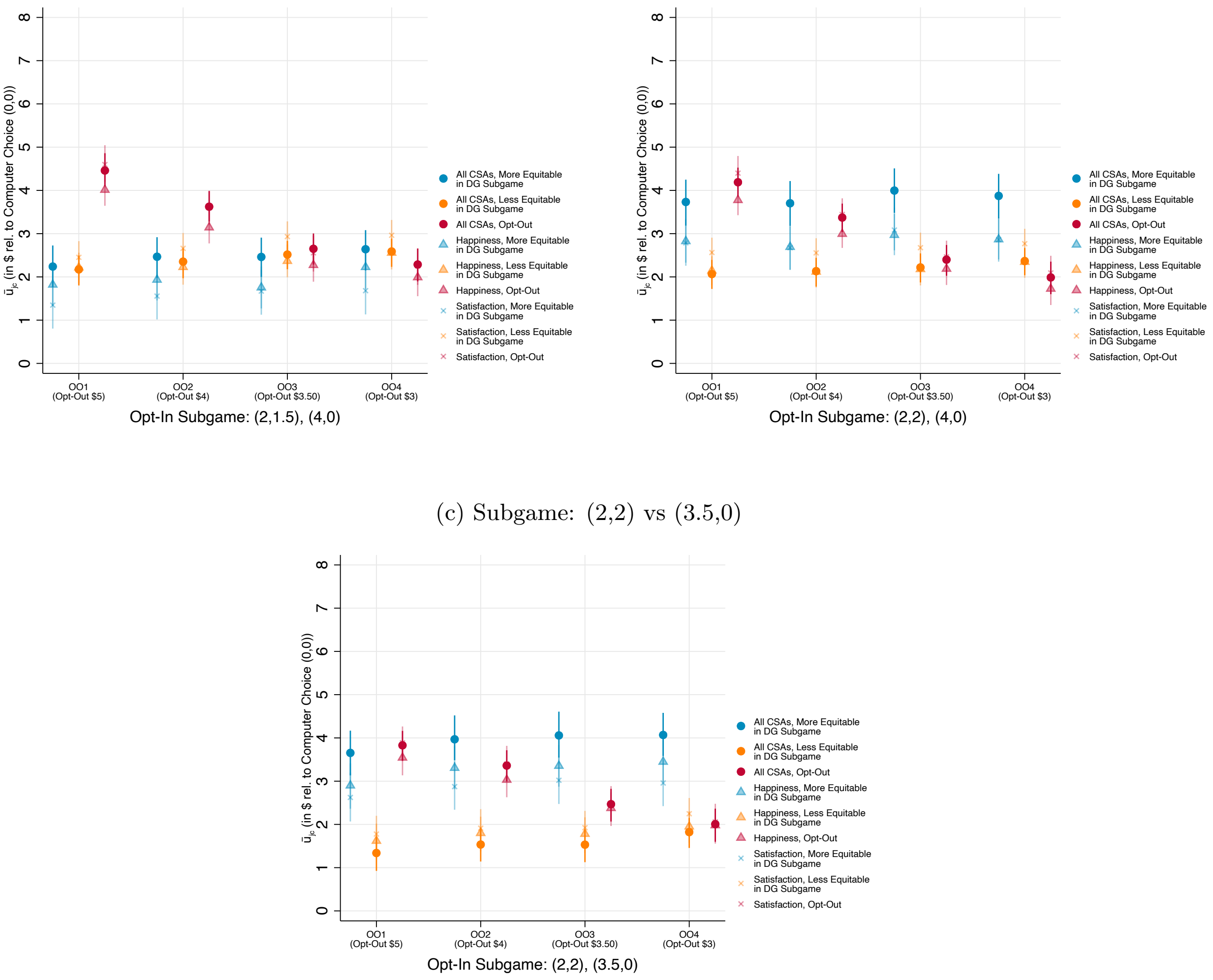
Note: This figure reports the average utility for each option of each choice set in the DG, using both actual and counterfactual choices. Thus, each point in the graph contains the utilities of all participants. The rounded points report the average utilities calculated by our hybrid method. The non-rounded 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 A7: 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 A8: 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.

Table A9: Differences in average utilities and CSA ratings between ex-ante and ex-post elicitations in the DG

Dep. Var.	(1) OLS $\bar{u}$	(2) OLS Guilt	(3) OLS Pride	(4) OLS Finan. Satis.	(5) OLS Fairness	(6) OLS Unfairness	(7) OLS Happiness
Ex-Ante	-0.010	0.012	0.032	0.006	0.002	0.023	0.024
	(0.026)	(0.020)	(0.018)	(0.019)	(0.023)	(0.022)	(0.021)
Constant	0.548***	0.229***	0.699***	0.182***	0.699***	0.549***	0.617***
	(0.007)	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)
Observations	17864	17864	17864	17864	17864	17864	17864
N. Participants	2552	2552	2552	2552	2552	2552	2552

Note: This table displays the average difference in DG utilities and CSA ratings when eliciting CSAs ex-ante. The coefficients are reported from an OLS regression of average utilities or CSA ratings (as named in the columns) in the DG on an indicator for whether the CSAs were elicited ex-ante (i.e. prior to the choice). Standard errors are reported in the parentheses and clustered at the participant level. $* * p < 0.1$, $** p < 0.05$, $*** p < 0.01$.

F Additional Robustness Checks

F.1 Anticipation of Experienced CSAs

Table A10 examines whether CSAs and average utility are different between when the CSAs are before versus after making a decision. We do not find a significant difference in any of the regressions. Table A10 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 A10: 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 clustered at the participant level. *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$.

Table A11: 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 A11 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 A12.

Table A12: Present and Future coefficient of the CSAs

	(1)	(2)
	Logit	Logit
	Choosing	Choosing
	More Equitably	More Equitably
Dependent Var.	(Present)	(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. *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$.

F.3 Further Tests of the Orthogonality Assumption

Table A13: 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. *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$.

Table A14: 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

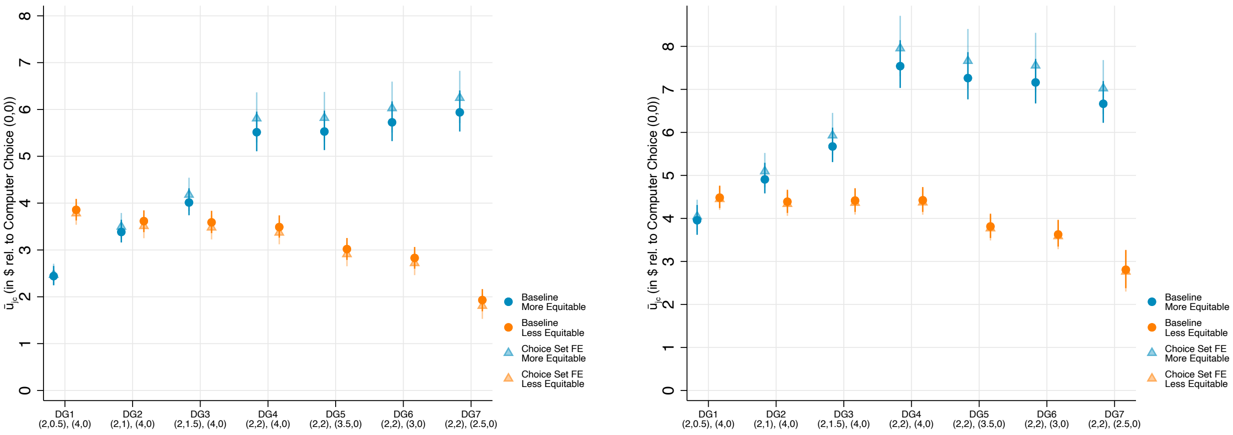
Note: Column (1) is identical to Column (1) of Table 1. Column (2) is a variation of Column (1) that include choice set fixed effects. Standard errors are reported in the parentheses, and clustered at the participant level. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

F.4 Order effects

An additional concern is that the order in which we present participants with their decisions alters their preferences over CSAs in the respective decisions. This would be an impediment to our methodology, because it would imply that preferences over CSAs are not stable and easily altered by order. Our results in Figures 4, A4, and A5 do not provide any evidence of instability. Here, we perform an additional stability check where we specifically investigate stability across identical choices presented in different orders.

Specifically, we split our sample into two groups, based on the order of the DG and CC module, and compare the predictive fit with the empirical probabilities. We report our results in Figure A10. Figure A10(a) is constructed as follows. In the first step, we estimate the model using DG data for the subsample of participants who participated in the DG module before the CC module. We then apply this estimated model to the subsample of participants who participated in the DG module after the CC module. Specifically, for these participants, we take their elicited CSAs, and form a prediction about the likelihood of choosing the more equitable allocation by using the estimated model. In the final step, we compare the predicted probabilities to the actual choices of

Figure A9: Deciders' average utility in dictator games



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

(b) Restricting only to the chosen options

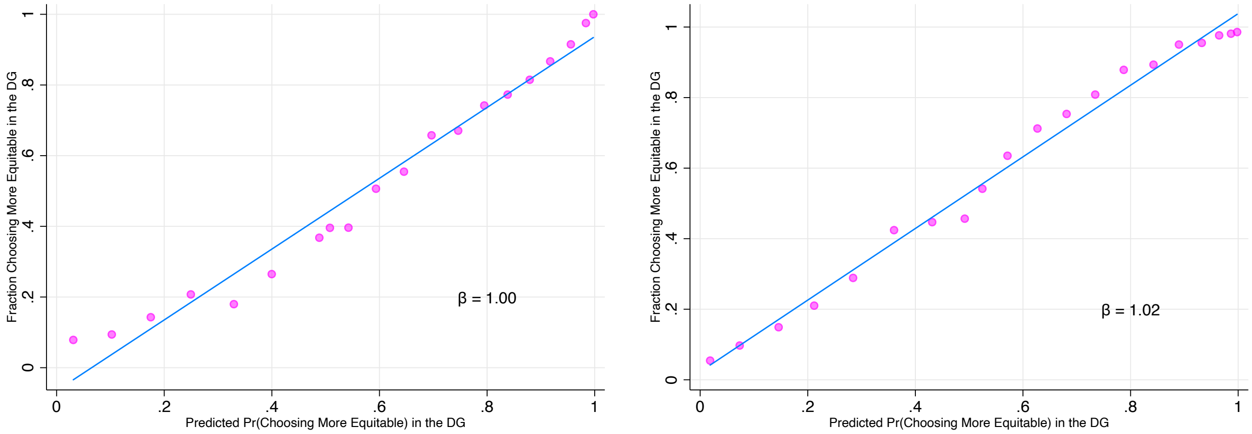
Note: This figure extends Figure 5 by also including estimate of money-metric utility based on the coefficients from Column (2) of Table A14. The original estimates are in circles, while the new estimate based on Column (2) are in triangles. The 95 percent confidence intervals are reported as the vertical bars, and calculated using bootstrap with 1,000 resampling clusters at the participant level.

the participants who participated in the DG module after the CC module. Figure A10(b) performs an analogous exercise, except the model is estimated using participants who participated in the DG module after the CC module, and we make predictions about the choices of participants who participated in the DG module before the CC module.

The two panels of Figure A10 show that the order in which participants' decisions are presented does not alter our conclusions about participants' preferences over the CSAs. For example, if seeing the DG module first made participants more sensitive to some types of emotions versus others, then we would not obtain the stability that we observe here.

A second question relating to order effects is whether our results about the non-comparability problem—particularly the novel result for meta-choices—are robust to between-subject comparisons. For example, it is possible that participating in the DGs or OOs alters the experience in one or both types of games. This would not create a confound in the sense of undermining the internal validity of our approach and results, since our framework is capable of addressing legitimate aspects of preferences like potential history-dependence. However, this may be important if the goal is to answer the following question: among participants whose first set of decisions is either the DG or the OO games, are the meta-choices still subject to the non-comparability problem? We address this question with Table A15, where we restrict the sample so that each column only includes participants who saw the referenced set of games first. Column (1) restricts to participants who saw the OO module before any other module, and is otherwise analogous to column (1) of Table 2. Column (2) restricts to participants who saw the DG module before any other module, and is otherwise analogous to column (2) of Table 2. Column (3) presents the difference, analogous to column (4) of Table 2. Overall, the estimates are similar to Table 2, with the difference between columns (1) and (2) slightly higher in Table A15. This demonstrates that our conclusions about

Figure A10: Observed versus fitted probabilities of choosing more equitably in the DGs, by viewing order



(a) Preference estimates from DG-before-CC sample, predictions for CC-before-DG sample

(b) Preference estimates from CC-before-DG sample, predictions for CC-before-DG sample

Note: This figure compares the model's predicted likelihood and the observed 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 for the subset of participants that viewed the DG module before the CC module and predicts the probability of choosing more equitably for the sample that viewed the CC module before the DG module. We then plot the predictions against the empirical probabilities. Panel (b) estimates the model using the DG CSAs for the subset of participants that viewed the CC module before the DG module and predicts the probability of choosing more equitably for the sample that viewed the DG module before the CC module.. The blue line represents the 45 degree line, wherein all points would lie if the model was a perfect fit.

the non-comparability problem in meta-choices are robust to both between- and within-subject comparisons.

Table A15: Comparing welfare estimates from standard approaches vs. hybrid approach, OO shown first

	(1) Choice-based inference of playing the DG using the Opt-Out Game OO first	(2) $\bar{u}_{jc}$ of playing in the DG using CSAs in the DG DG first	(3) Difference (1)-(2)
Subgame: (2,1.5) vs. (4,0)	3.78***	3.89***	-0.11
	[3.61, 3.97]	[3.64, 4.16]	[-0.42, 0.21]
Subgame: (2,2) vs. (4,0)	4.12***	5.15***	-1.03***
	[3.89, 4.36]	[4.79, 5.54]	[-1.46, -0.59]
Subgame: (2,2) vs. (3.5,0)	3.79***	4.99***	-1.20***
	[3.55, 4.04]	[4.62, 5.41]	[-1.68, -0.78]

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, using the sample of participants that viewed the OO module first. Column (2) reports the money-metric utility estimates obtained from our approach, using the sample of participants that viewed the DG module first. Column (3) is the difference between the estimates in Columns (1) and (2). The 95 percent confidence intervals are reported in brackets, and calculated using bootstrap with 1,000 resampling clusters at the participant level. *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$.

G Heterogeneity in the Marginal Utility of Money

We first show that heterogeneity in the marginal utility of money does not pose a problem for our approach. 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

$$\bar{u}_{jc} = \mathbb{E}_i[v(X_{ijc})/\bar{m}] - \mathbb{E}_i[v(X_{ij_4c_0})/\bar{m}] + 4. \tag{7}$$

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.[38] When the orthogonality assumption does not hold, $\bar{u}_{jc}$ is still interpretable as the average utility increment rescaled by the average marginal utility of money.

As a positive question, we next investigate whether there is indeed variation 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 = $[0.42, 0.52]$).

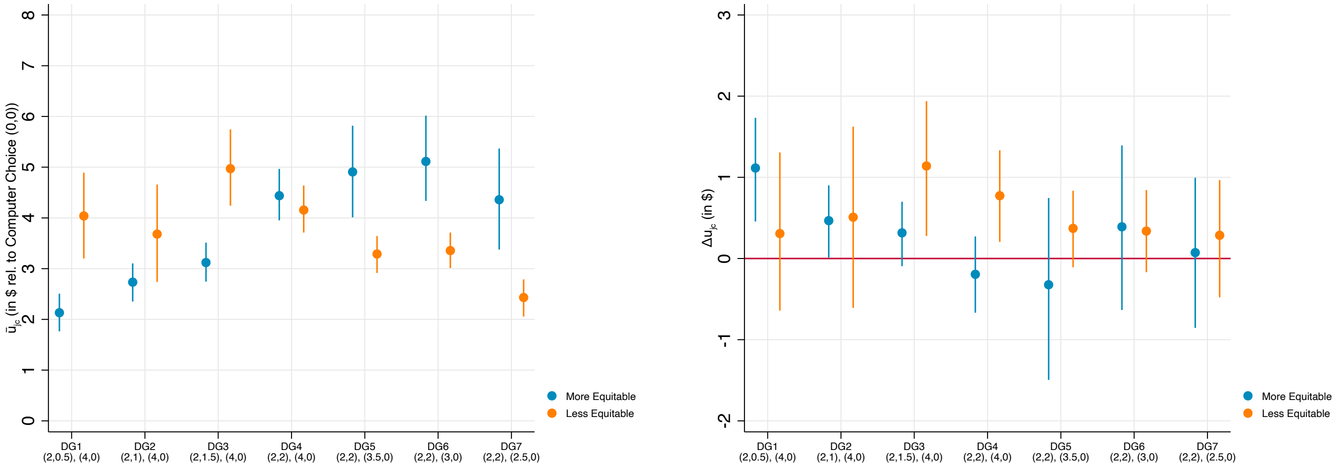
H Comparing the Main and Alternative CC Modules

We construct Figure A11, which is analogous to Figure 6 but uses data from the alternative CC module. Comparing Panels (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]). Thus, including contrasting allocations lowers the utility Deciders obtain from having the more-equitable allocations exogenously assigned, but not from having the less-equitable allocations exogenously assigned.

Panel (b) of Figure A11 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

[38] 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.

Figure A11: Deciders' average utility in the alternative computer choice module



(a) Utility when the computer chooses 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.

standard welfare methods continues to underestimate how much Deciders' welfare decreases when the Planner replaces the less equitable allocations with more equitable alternatives.

I Additional Survey Results

I.1 Preferences for Own Versus Government Choice

I.1.1 Retirement Savings and Early Withdrawal

Part 2 of Survey 1 started with the following prompt:

> *People sometimes make early withdrawals from their retirement savings accounts (such as employer-sponsored 401(k) plans) because of lapses of self control. That is, people have good intentions to accumulate funds for old age, but they don't always follow through on those intentions because they prioritize immediate "wants," such as spending on entertainment, hobbies, unnecessary luxuries, or dining out.*
>
> *Right now, the penalty for early withdrawals before age 59.5 is 10% of the amount withdrawn. Some economists think that a 30% penalty would be better. The idea is that a higher penalty might help people exercise self-control more effectively by discouraging them from making early withdrawals.*

Participants then answered questions about choosing a higher early-withdrawal penalty themselves, and about the government choosing a higher early withdrawal penalty. We randomized the order of whether participants first answered questions about voluntarily increasing the penalty, or whether they first answered questions about the government increasing the penalty.

In answer to the question of whether participants would be better off if they voluntarily increased the penalty, 44.0 percent said yes. By contrast, only 19.3 percent said that they would be better off if the government increased the penalty. Figure A12 presents the CSAs that participants anticipated would be associated with the government making choosing a higher penalty versus the participants themselves choosing a higher penalty. For either outcome, and for every CSA, participants report a different experience when they versus the government make a decision. The positive CSAs—happiness, satisfaction, pride, self-worth, dignity—are more intense when the participants choose the penalty themselves versus when the government chooses it. By contrast, fear, anxiety, anger, and irritation are all higher when the government implements the penalty. Of the negative CSAs, only guilt and regret are lower when the government implements the penalty, which makes sense because these are emotions tied to a sense of responsibility for the outcome.

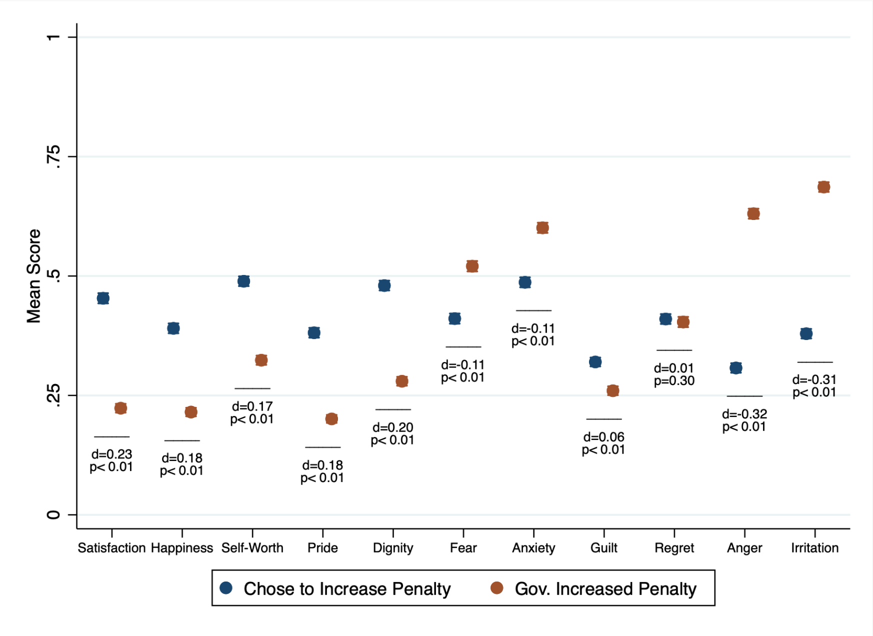
I.1.2 Reminders for Financial, Health and Career Planning

In the last part of Survey 2 we asked questions about financial, health, and career planning that started with the following prompt:

> *Imagine that you had the opportunity to sign yourself up for a program that provides regular reminders—through text messages, emails, push notifications, and occasional phone calls—to take more time for [financial/health/career] planning.*
>
> *Suppose [you signed yourself up for this program/the government mandated that everyone, including you, must be part of this program]. Please indicate how you think you would feel, both immediately and going forward.*

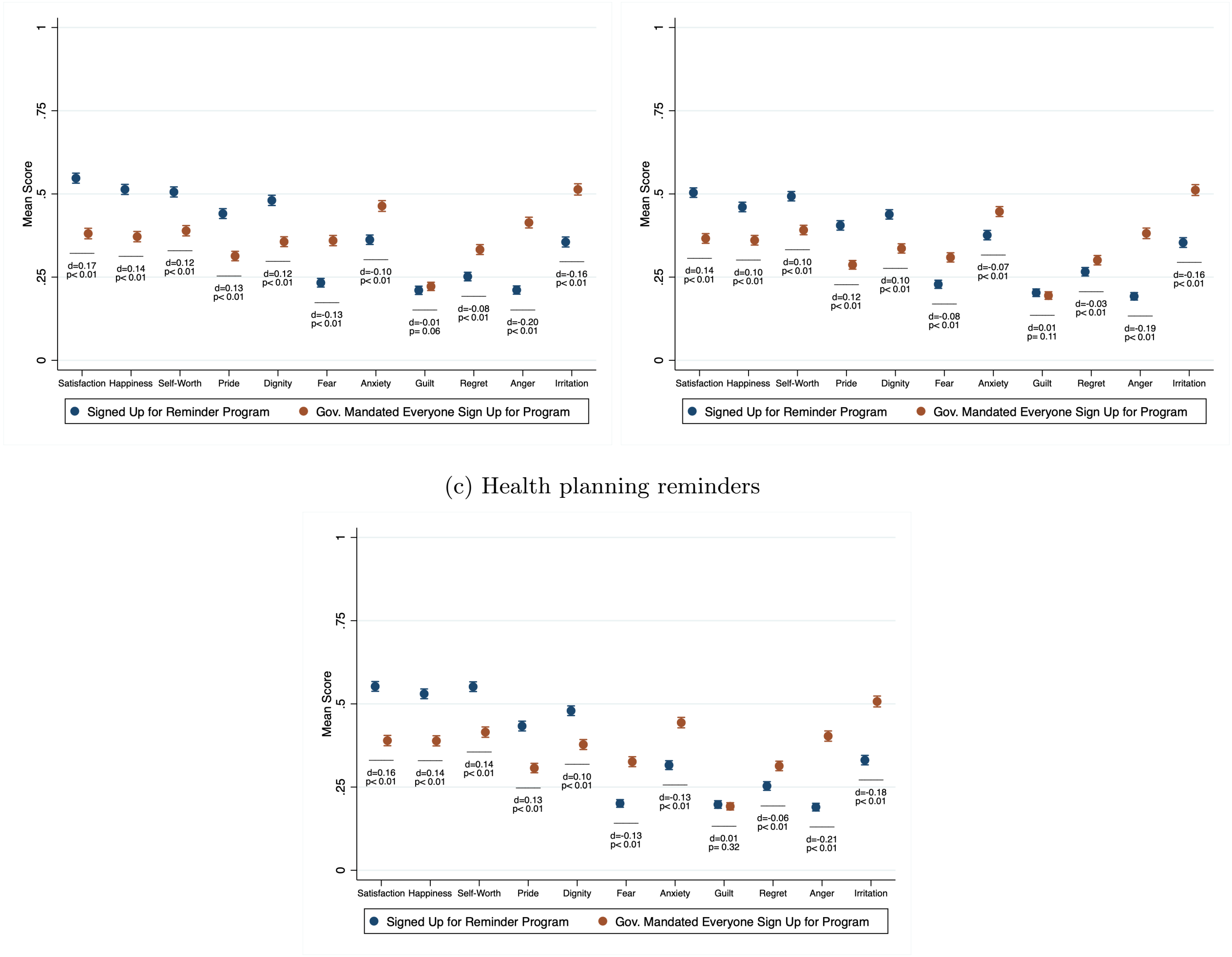
Figure A13 presents the results. Mirroring the results on paternalistic increases in the early withdrawal penalty, we again see that the positive CSAs—happiness, satisfaction, pride, self-worth, dignity—are all lower when the government implements the reminder program. By contrast, the negative CSAs—fear, anxiety, anger, irritation, regret—are higher when the government implements the reminder program.

Figure A12: CSAs when the participant or government makes a decision about the early withdrawal penalty



Note: This figure reports the mean self-reported CSAs in the scenarios where (i) the respondent chooses to increase the early retirement withdrawal penalty, and (ii) the government chooses to increase the early retirement withdrawal penalty. Blue points correspond to the question where the respondent chooses the penalty, and red points correspond to the case where the government imposes the penalty. CSA scores, ranging from 1 to 5, were normalized by subtracting 1 and dividing by 4. Differences in means between the self-imposed and government-imposed penalties are reported below each CSA, denoted as d, with corresponding p-values denoted as p. Vertical bars represent the 95% confidence interval. The sample includes 482 respondents.

Figure A13: CSAs in reminder program: Self versus government



(a) Financial planning reminders

(b) Career planning reminders

(c) Health planning reminders

Note: This figure reports the mean self-reported CSAs in the scenarios where (i) respondent signs up for a financial/career/health reminder program, and (ii) the government implements the program. Blue points correspond to the question where the respondent chooses the penalty, and red points correspond to the case where the government imposes the penalty. CSA scores, ranging from 1 to 5, were normalized by subtracting 1 and dividing by 4. Differences in means between the self-imposed and government-imposed penalties are reported below each CSA, denoted as d, with corresponding p-values denoted as p. Vertical bars represent the 95% confidence interval. The sample consists of 468 respondents.

I.2 Social Pressure and Charitable Giving

Survey 3 involved a stated choice experiment and CSA elicitation that was motivated by the door-to-door charitable giving experiment of DellaVigna et al. (2012). This experiment involved three scenarios, presented in random order.

Scenario 1 was described as follows:

> *Imagine that there is a person at your door soliciting charitable contributions for a non-profit children's hospital. This person previously left a flier at your door, so that's how you know their intention and the charity for which they are fundraising. There are three things you could do: 1. Open the door and make a donation, 2. Open the door and not make a donation, or 3. Not open the door, pretending to be away from home. What would you choose?*

Scenario 2 instead was described in the following way:

> *Imagine that someone rings your doorbell, and you open the door without knowing this person's motive. You then learn that this person is soliciting charitable contributions for a non-profit children's hospital. There are two things you could do: 1. Make a donation, or 2. Not make a donation. What would you choose?*

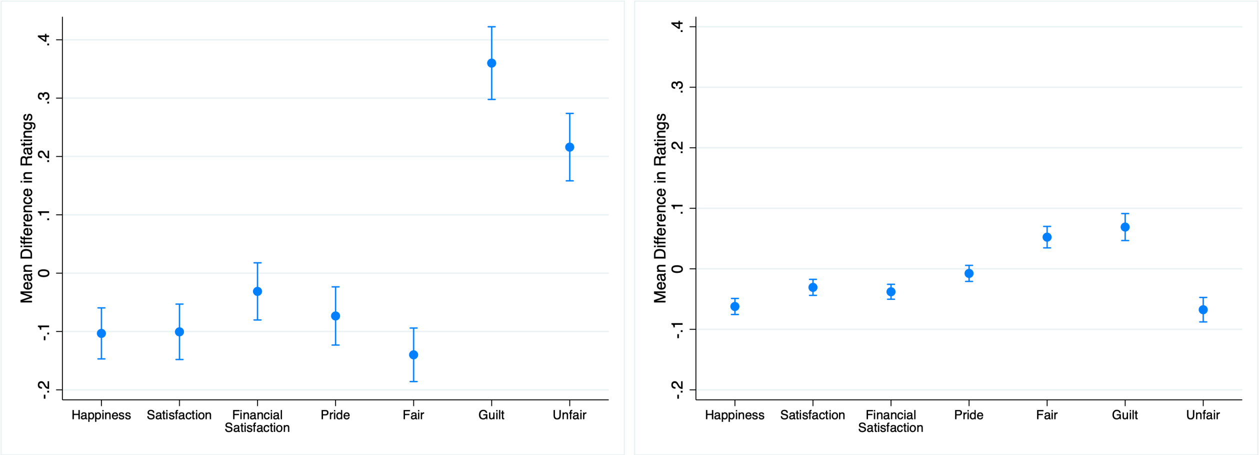
Both Scenarios 1 and 2 capture the main experimental conditions in the DellaVigna et al. (2012) field experiment. For the survey, we also designed a third scenario, where participants never had the ability to open the door:

> *Imagine that you were at work, and therefore not at home, when a person came by your house to solicit donations for a non-profit children's hospital. Therefore, you had no opportunity to open the door. You only found out that the solicitor came by because you saw them on your doorbell camera and recognized the hospital's name on their shirt. (If you don't have a doorbell camera, imagine that you do). What would you choose?*

Within each scenario we also elicited participants' CSAs for all possible outcomes in the respective scenario. The set of CSAs, and the elicitation, was analogous to our main experiment. In scenarios 1 and 2, the elicitation came after participants stated their preferences, and in scenario 3 the elicitation came after participants read the prompt.

Figure A16 presents participants' choices. Figures A14 and A15 depict the CSAs that participants report they would experience if they chose to avoid the giving opportunity, relative to a situation in which it is impossible for them to give. These figures compare the CSAs in our survey experiment (panel a) to the CSAs in analogous decisions in OOs versus CCs (panel b). For the charitable donations survey this is the difference between choosing not to answer the door versus not having the option to be at home to open the door. For our main experiment, this is the difference between opting out of a DG, versus having that same allocation instead exogenously assigned by the computer. Because there are multiple OOs in our main experiment, we simply take the average over all the variants. As outlined in the pre-registration plan, our focus is on Figure A14, which restricts to participants more sensitive to social motivations rather than narrow self-interest. For the charitable giving survey, these are the participants who give in Scenario 2. For our main experiment, these are the participants who chose the more equitable allocation in the DG game that corresponds to the subgame of the OO being analyzed.

Figure A14: Mean difference in CSAs from opting out of giving versus not being able to give (among those more likely to give)



(a) Not opening the door (among those giving in Scenario 1)

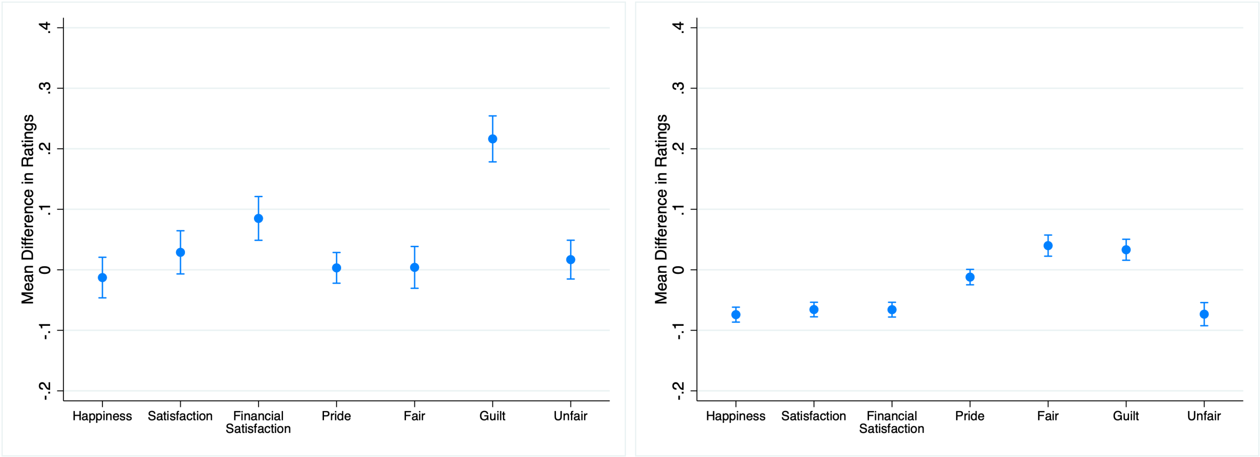
(b) Opting out of the OO (among those giving in that DG)

Note: This figure reports the average change in CSA scores that participants experience when choosing to avoid the giving opportunity, relative to a situation in which giving is impossible. Panel (a) focuses on the charitable donations survey experiment and reports the difference between the average CSA score for avoiding the solicitor in Scenario 1 and the average CSA score for being unable to answer the door in Scenario 3, among participants who chose to donate in Scenario 2. This sample consist of 184 respondents. Panel (b) uses the main experiment and reports the difference between opting out of a DG and having that same opt-out allocation instead exogenously assigned by the computer, among participants who chose the more equitable allocation in the DG game that corresponds to the subgame of the OO being analyzed. Because there are multiple OOs in our main experiment, we take the average over all the variants. This sample consist of 470 respondents. CSA scores, ranging from 1 to 5, were normalized by subtracting 1 and dividing by 4. Vertical bars represent 95% confidence intervals.

Figure A14a shows that participants experience not opening the door very differently between Scenario 1—where they choose to not open it—and Scenario 3—where they have no option to open the door. All of the CSAs—with the exception of financial satisfaction—differ significantly between Scenarios 1 and 3 for not opening the door. The differences are more muted for participants who chose not to give in Scenario 1 (Figure A15a), as these participants plausibly put less weight on the mental states that drive charitable giving. The second key result shown in Figure A14 is that the difference in CSAs—and thus the extent of the NCP—appears to be at least as large in the charitable giving context (panel a) as it is in our main experiment (panel b). This suggests that our lab-experimental conclusions about the NCP are likely to be relevant in related field settings. This manifestation of the NCP is inconsistent with the identifying assumptions of DellaVigna et al. (2012), but parallels our results on experimental opt-out games.

Figures A17, A18, and A19 present the CSAs for all other options in all other scenarios, and for three respective sets of participants: (i) all participants, (ii) those who donated in Scenario 2 and thus are more likely to be driven by guilt, fairness and other social motives for giving, and (iii) those who did not donate in Scenario 2 and thus are less driven by social motives for giving.

Figure A15: Mean difference in CSAs from opting out of giving versus not being able to give (among those not likely to give)

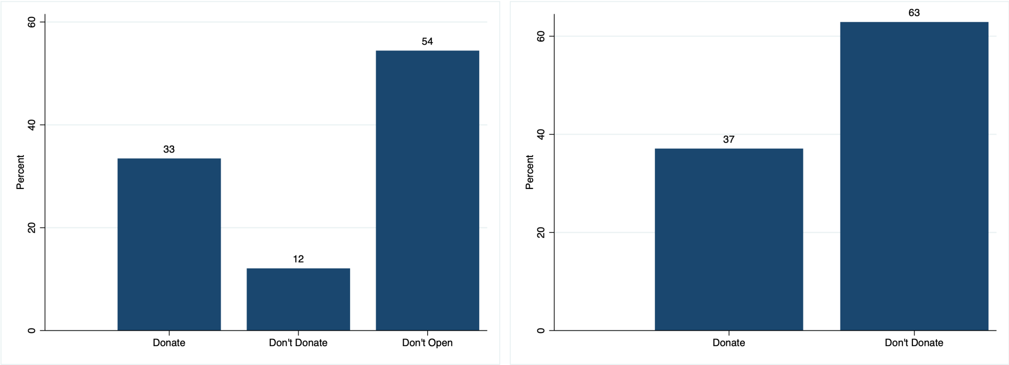


(a) Not opening the door (among those not giving in Scenario 2)

(b) Opting out of the OO (among those not giving in that DG)

Note: This figure reports the average change in CSA scores that participants experience when choosing to avoid the giving opportunity, relative to a situation in which giving is impossible. Panel (a) focuses on the charitable donations survey experiment and reports the difference between the average CSA score for avoiding the solicitor in Scenario 1 and the average CSA score for being unable to answer the door in Scenario 3, among participants who chose to *not* donate in Scenario 2. This sample consist of 312 respondents. Panel (b) uses the main experiment and reports the difference between opting out of a DG and having that same opt-out allocation instead exogenously assigned by the computer, among participants who did *not* choose the more equitable allocation in the DG game that corresponds to the subgame of the OO being analyzed. Because there are multiple OOs in our main experiment, we take the average over all the variants. This sample consist of 415 respondents. CSA scores, ranging from 1 to 5, were normalized by subtracting 1 and dividing by 4. Vertical bars represent 95% confidence intervals.

Figure A16: Choices in Scenarios 1 and 2A13

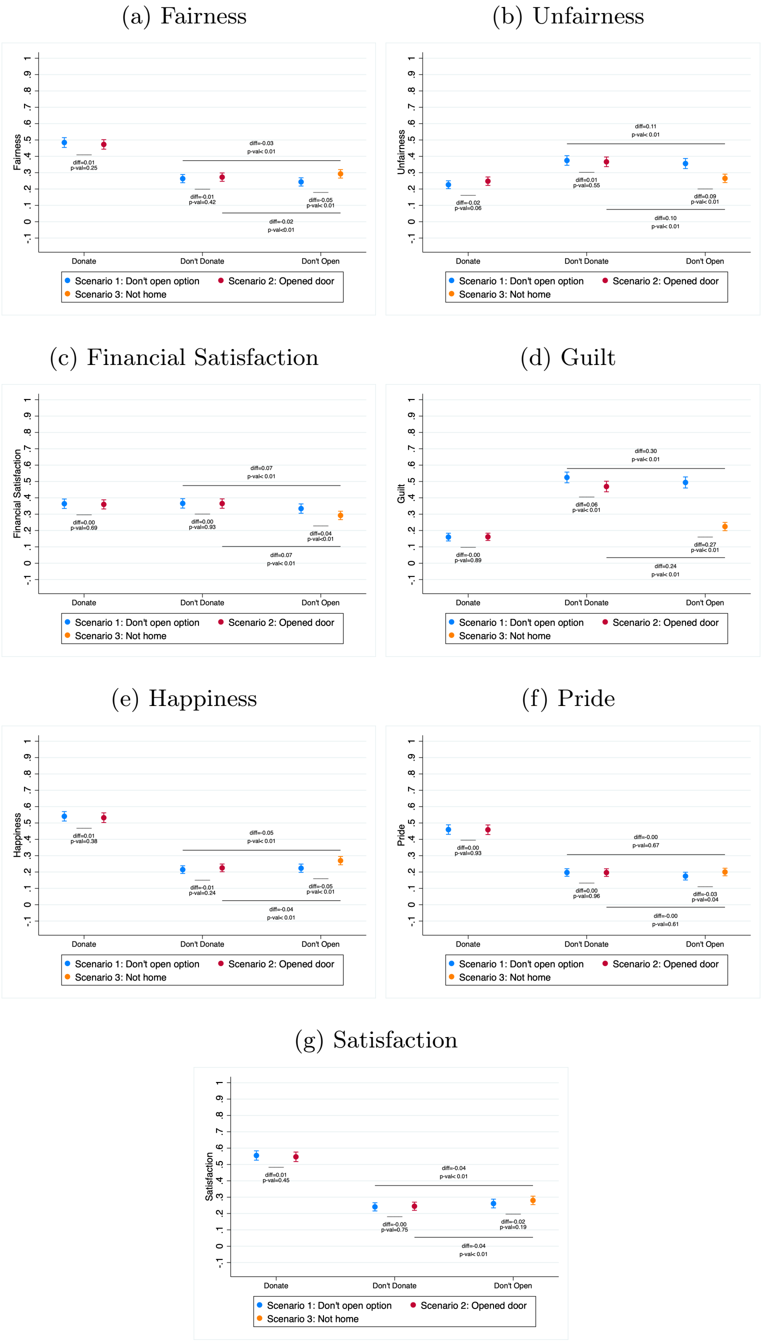


(a) Scenario 1 (option to not open door)

(b) Scenario 2 (no option not to open the door)

Notes: This figure reports the distribution of choices made in Scenarios 1 and 2 of the charity survey experiment. Panel (a) reports the percentage of respondents who chose to donate, not donate, or not open their door in Scenario 1. Panel (b) reports the percentage of respondents who chose to donate or not donate in Scenario 2. The sample consist of 496 individuals.

Figure A17: Reported CSAs for charity solicitations



(a) Fairness

(b) Unfairness

(c) Financial Satisfaction

(d) Guilt

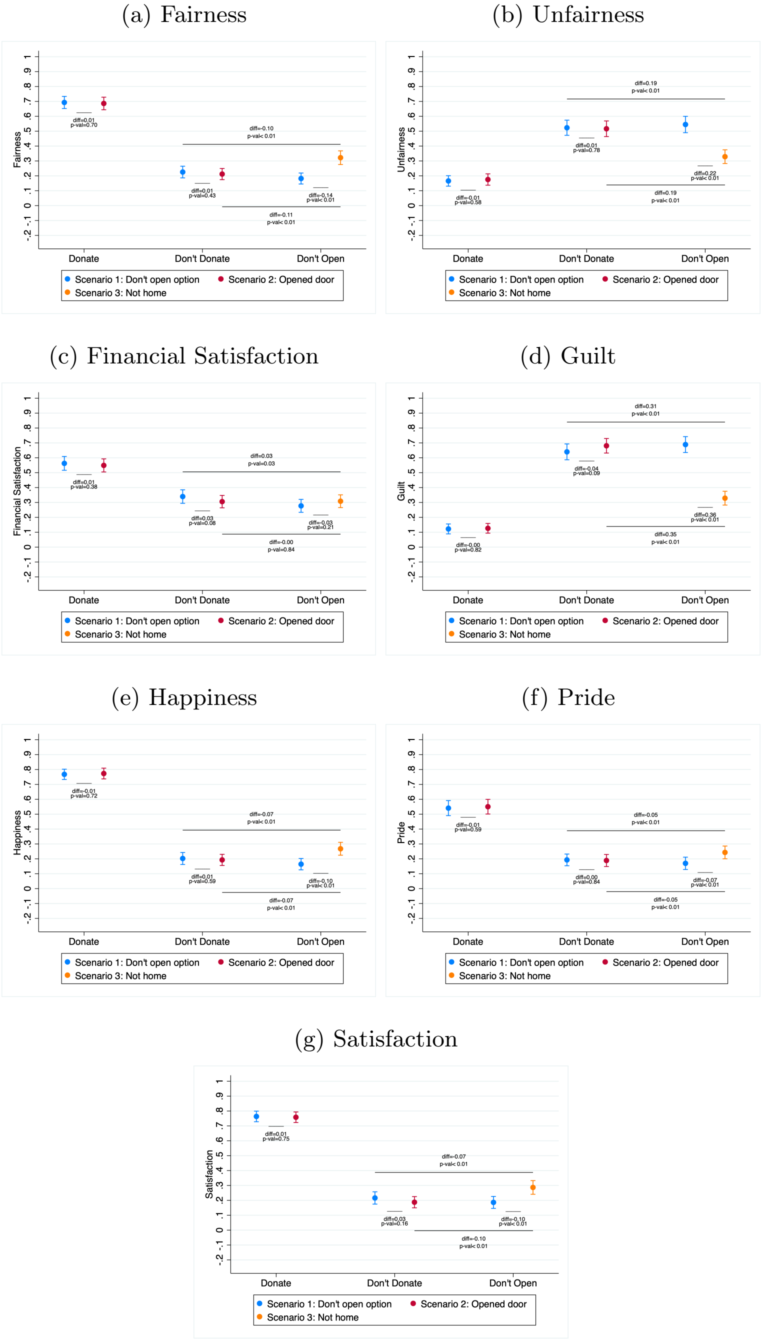
(e) Happiness

(f) Pride

(g) Satisfaction

Notes: This figure reports the average CSAs for each action that is available to the participant in the supplementary charity experiment. Each panel corresponds to a CSA, labeled in the title. Blue points represent the scenario in which the participant knows there is a solicitor at the door. Red points represent the scenario in which they open the door to a solicitor unknowingly. The yellow point represents the scenario in which they missed the solicitor who came by their door. CSA scores, ranging from 1 to 5, were normalized by subtracting 1 and dividing by 4. Differences in means between the same action in different scenarios are reported below each CSA denoted as d, with corresponding p-values denoted as p. Vertical bars represent the 95% confidence interval. This sample consist of 496 respondents.

Figure A18: Reported CSAs for charity solicitations, only those who donated Scenario 2



(a) Fairness

(b) Unfairness

(c) Financial Satisfaction

(d) Guilt

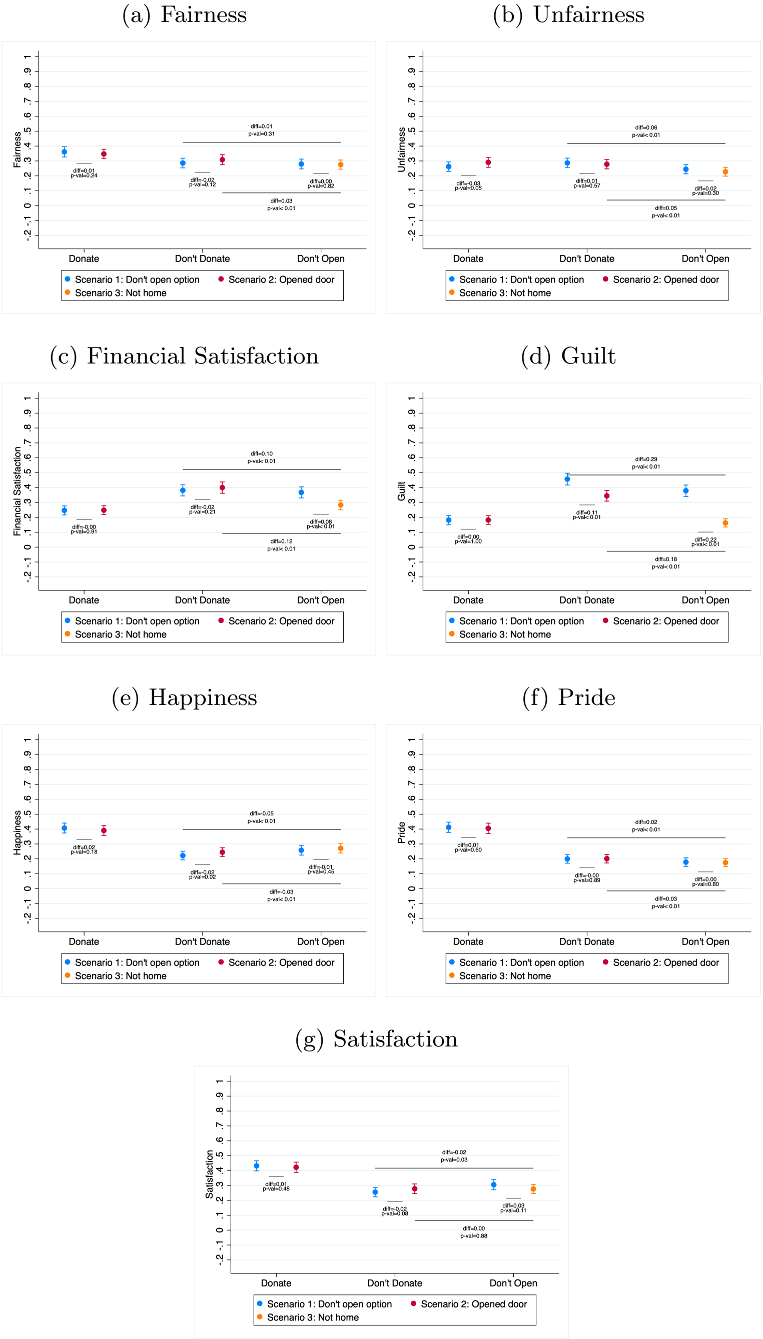
(e) Happiness

(f) Pride

(g) Satisfaction

Notes: This figure reports the average CSAs for each action that is available to the participant in the supplementary charity experiment. The sample is restricted to those who chose to donate in Scenario 2. Each panel corresponds to a CSA, labeled in the title. Blue points represent the scenario in which the participant knows there is a solicitor at the door. Red points represent the scenario in which they open the door to a solicitor unknowingly. The yellow point represents the scenario in which they missed the solicitor who came by their door. CSA scores, ranging from 1 to 5, were normalized by subtracting 1 and dividing by 4. Differences in means between the same action in different scenarios are reported below each CSA denoted as d, with corresponding p-values denoted as p. Vertical bars represent the 95% confidence interval. This sample consist of 184 respondents.

Figure A19: Reported CSAs for charity solicitations, only those who did not donate Scenario 2



(a) Fairness

(b) Unfairness

(c) Financial Satisfaction

(d) Guilt

(e) Happiness

(f) Pride

(g) Satisfaction

Notes: This figure reports the average CSAs for each action that is available to the participant when a charity solicitor appears at their door. The sample is restricted to those who chose not to donate in Scenario 2. Each panel corresponds to a CSA, labeled in the title. Blue points represent the scenario in which the participant knows there is a solicitor at the door. Red points represent the scenario in which they open the door to a solicitor unknowingly. The yellow point represents the scenario in which they missed the solicitor who came by their door. CSA scores, ranging from 1 to 5, were normalized by subtracting 1 and dividing by 4. Differences in means between the same action in different scenarios are reported below each CSA denoted as d, with corresponding p-values denoted as p. Vertical bars represent the 95% confidence interval. This sample consist of 312 respondents.

J Protocol for Detecting AI-Generated Responses

This section outlines the methods used to prevent and detect responses that may have been produced by generative AI tools. Below we describe our approach to ensure the authenticity and integrity of the data.

To minimize the risk of AI-assisted responses, we implemented two key preventive measures. First, participants were explicitly informed that AI-generated responses or responses from external sources would result in non-payment. The consent page included the following statement:

> *Please ensure that all responses are generated solely by you. If your responses are flagged as potentially generated by artificial intelligence software (such as ChatGPT) or copied from external sources, your payment may be delayed or withheld. We may use automated plagiarism detection tools and web page navigation tracking, among other methods, to identify such behavior. Surveys rejected due to AI-generated or copied responses will not be eligible for payment.*

This language established clear expectations for respondents before the study and served as a deterrent by introducing payment penalties for responses that were not generated by the participant. Second, we disabled the paste function for open-ended response fields. Removing this function made it more difficult for respondents to input pre-generated content from external sources, further deterring this action.

In addition to implementing preventive measures, we conducted an ex-post review—which we fully pre-registered—to identify AI-generated responses. We first flagged respondents based on criteria that we describe below, and then analyzed their responses through an online AI content detector. For this study, we utilized the AI detector Copyleaks. A respondent was flagged if they met two of the three following criteria:

1. Use of title case (i.e. when the first letter of each word is capitalized, like in a heading or title)—partially or fully–in any open-ended response;
2. Lack of first-person pronouns (i.e. “I,” “me,” “my,” “mine,” “we,” “us,” “our,” and “ours”) in all open-ended responses;
3. Use of list formatting in any open-ended response (e.g. the presence of “1.” or “1)”).

Additionally, flagged respondents had to meet a minimum total response length of 350 characters, the minimum required length for reliable AI detection.

For those identified as potential AI users, we submitted all their open-ended responses to the online AI detector, Copyleaks. If the detector classified the responses as “90% AI content found” or greater, the respondent was excluded from our analysis