A Representation of Preference over Preferences and Choices

Jun Hyun Ji*

August 18, 2026

Click here for the most recent version

Abstract

I develop an axiomatic framework for preferences over choices, where an *evaluator* values each observed choice for its outcome and for the quality of the choice itself, which reflects her own assessment of how well the choice aligns with *good preferences*. I define "second-order preference" through an axiom capturing how choice quality is relative to available alternatives. Within the expected-utility framework, my axioms yield a unique representation that separately identifies the evaluator's preference for outcomes and her second-order preference. As a special case, the model accommodates temptation and self-control as determinants of choice quality, generating a trade-off between achieving the best outcome and making the best choice. The framework provides a revealed-preference characterization of preferences over preferences, with implications for menu choice and welfare evaluation.

Keywords: Preference over preferences; second-order preference; ideal preference; choice quality

*Department of Economics, University of Pittsburgh. Email: juj25@pitt.edu. I am deeply indebted to Luca Rigotti for his guidance, support, and encouragement throughout this project. I would like to thank Jawwad Noor, Evan Piermont, Philip Reny, and Mauricio Ribeiro for valuable comments. I am also grateful for the helpful questions and suggestions from David Ahn, Cuimin Ba, Sven Neth, Gerelt Tserenjigmid, Kevin Zollman, and the participants at the Durham Economic Theory Conference, the Workshop on Choice and Welfare at Queen Mary University of London, the BSE Summer Forum, the Asian School in Economic Theory, and the PETCO poster session. I gratefully acknowledge financial support from the 2023 Tamara Horowitz Memorial Fund.

1. Introduction

The same outcome can come from a good choice or a bad one, depending on what was forgone in choosing it. Consider a judge who sentences three defendants convicted of the same crime. The first acted under threat, with no way out but to commit the crime. The second faced no such pressure and could easily have refrained. The third committed the crime *only* to prevent a worse consequence. Though their choices led to the same outcome, the judge would sentence the second most harshly and the third most leniently, responding to what each defendant willingly gave up in making that choice. She is, in effect, ranking the quality of the defendants' choices. Across many institutional contexts—courts, classrooms, insurance claims, politics, recruiting, and social settings—people tend to favor choices that align with their ideals, virtues, or *good preferences*, independently of their outcomes. Interestingly, the quality of acting on a preference depends on the menu of options available to the decision-maker (DM).

How do people assess the quality of choices? Can the preferences underlying these assessments be identified and characterized from observable comparisons among choices? These questions have not been explicitly addressed in economic theory, yet they pose challenges for both standard welfare judgments and our understanding of choice behavior. I develop an axiomatic framework for preferences over choices. Suppose an *evaluator* (she) observes and ranks choices made by a DM (he) who faces exogenously given menus. I define a choice by the DM as an outcome-menu pair (x, A). The evaluator has a preference $\succsim$ that assesses each choice (x, A) along two dimensions: the outcome x itself and the revelation that the DM *behaved as if* he preferred x to all else in A. The second component captures her assessment of choice quality, which I refer to as the *second-order preference*.

The preference $\succsim$ reflects mainly two ideas. First, the outcome itself does not determine the choice quality: when outcomes are controlled, any two choices from singleton menus are indifferent because no alternative was forgone. They differ only in their outcomes. To control for the outcome, I extend standard expected utility (EU) theory to preferences over choices: $\succsim$ is complete, transitive, continuous, and independent. When each outcome x is a lottery over some sure alternatives and menus are sets of lotteries, a preference relation $\succsim^Q$ can be induced from $\succsim$ such that $\succsim^Q$ ranks choices as if every choice yields the same outcome; hence, $(x, \{x\}) \sim^Q (y, \{y\})$ for all x, y. Essentially, the preference $\succsim^Q$ strips out the outcome component from $\succsim$

and becomes the second-order preference once the second idea is imposed.

Second, the quality of a choice is relative to its alternatives: every menu offers some choices that are relatively good and others that are relatively bad. The axiom called *Relativity* formalizes this idea. It states that each menu A contains some x, y such that *"choosing x willingly"* is at least as good as *"consuming it unwillingly,"* which is identified with $(x, A) \succsim (x, \{x\})$; conversely, *"choosing y willingly"* is at least as bad as *"being forced to consume it"*—i.e., $(y, \{y\}) \succsim (y, A)$. As in the opening example, if committing the crime is the best (worst) option available, the choice itself cannot be worse (better) than being forced into it. A direct violation is $(x, \{x\}) \succ (x, A)$ for all $x \in A$. One might recall *Sophie's Choice*: the evaluator may prefer that the DM be forced into any outcome rather than have to choose willingly among horrifying alternatives. Such a preference concerns the mental cost of having to choose, not the quality of choices. More broadly, Relativity rules out cases in which the menu itself determines the ranking of choice quality: e.g., $(x, A) \succ^Q (y, B)$ for all $x \in A$ and $y \in B$. Violations thus imply that the evaluator is interested in factors other than outcomes or choice quality (e.g., the DM's freedom or flexibility).[1]

My main theorems (Theorems 1-2) state that the preference over choices satisfies the axioms of the EU theory and Relativity if and only if it admits a unique representation U of the form

$$U(x, A) = u(x) + v(x) - v(r(A)) \tag{1}$$

where u, v are von Neumann-Morgenstern (vNM) utility functions over lotteries, and r is a function that maps menu A to a mixture $r(A)$ of lotteries in A. The function u describes a ranking of choices from singleton menus, representing the evaluator's outcome preference. The second-order preference $\succsim^Q$ is represented by the term $v(x) - v(r(A))$, where v represents *the ideal outcome preference*—the one that describes the evaluator's "ideal DM," whose choice is always of the highest quality given any menu. For example, $u(x) > u(y)$ implies that the evaluator prefers the outcome x to y regardless of which alternatives are given up in choosing each, whereas $v(x) > v(y)$ means that, whenever x and y are both available and outcomes are held fixed, the evaluator prefers the DM to give up y for x rather than x for y. Similarly,

[1] The evaluator also does not distinguish among choices according to motives or experiences that are not captured by the choice itself—e.g., whether the DM chooses strategically or out of desire.

$v(x) - v(r(A)) > v(y) - v(r(B))$ means that she prefers the choice (x, A) to (y, B) if both choices yield the same outcome. Intuitively, it reads as "*preferring x to all else in A* is preferred to *preferring y to all else in B*."[2]

The lottery $r(A)$ serves as a reference against which the quality of choosing x from A is assessed. In this sense, $r(\cdot)$ is *the reference function*, which can vary across individuals. That is, evaluators differ in their *attitudes toward preferences* even when they share the same ideal ranking v. I define a comparative measure of *evaluative strictness*, which captures how close the reference lies to the most ideal rather than the least ideal option on each menu (Theorem 3).

The representation above accommodates perhaps the broadest range of attitudes within the EU framework, imposing minimal structure on the reference function. For instance, it satisfies $(r(A), A) \sim (r(A), \{r(A)\})$ for every A that contains $r(A)$, but it does not explicitly specify what determines whether the quality of a choice is worth more or less than *"not choosing."* I address this by introducing a third idea as a special case in which the reference function has an explicit form.

The third idea is that the quality of a choice involves more than simply choosing a good option over a bad one. Preferences are valuable only when they induce "hard choices" (e.g., an alcoholic choosing coffee over beer; a dictator sacrificing his own interest for others), whereas easy choices (e.g., a dieter choosing fresh salad over rotten chocolate; a dictator choosing the Pareto-optimal allocation over inferior ones) are merely obvious. What counts as obvious, however, depends on the evaluator's taste—an idea central to this paper. The special-case representation provides a benchmark for capturing these evaluative differences. Building on Gul and Pesendorfer (2001)'s model of temptation and self-control, I formalize this idea through the axiom *Choice Betweenness*. In the resulting representation (Theorem 4), the evaluator assigns higher quality to choices that either exceed her expectation or require self-control; it thereby identifies her expectations about the DM's choices, temptations, and self-control—factors that jointly determine the reference of each menu. The novelty of this special case is that menus that offer good outcomes fundamentally differ from menus that offer good choices. This creates a trade-off between the best outcome and the best choice, a tension that is absent in the literature on menu

[2] The functions u and v need not be aligned. For example, the evaluator may prefer that the DM receive a particular outcome while regarding the choice of that outcome as bad. In a variety of contexts, u and v may interchangeably reflect the DM's immediate pleasure, normative goals, etc.

preferences in deterministic contexts (Gul and Pesendorfer, 2001).

I provide two applications of this special case in the context of menu choice. The first considers menu design by the DM. A dictator (DM) who anticipates the counterpart's (evaluator) reciprocation may secretly exclude the Pareto-optimal allocation from his menu in order to appear selfless, effectively giving up the best outcome for the best choice. The second considers menu design by the evaluator, who chooses the DM's menu based on the choices she expects or has observed. This reverses the usual menu-choice dynamics: menu choice is conditioned on outcome choice. A parent, for example, may restrict a child's menu after observing his poor choices often enough.

The remainder of this paper is organized as follows. In the rest of this section, I discuss the related literature and my contributions. In Section 2, I introduce the model, discuss the implications of my axioms, and characterize the second-order preference. Section 3 presents the representation, uniqueness result, and comparative measures. Sections 4 and 5 present the special case and its applications to menu choice, respectively. Section 6 discusses empirical challenges and possible extensions. Section 7 concludes. All omitted proofs are in the Appendix.

1.1. Related Literature and Contributions

Economists have long recognized that two choices leading to the same outcome may be evaluated differently depending on the menus from which they were made.[3] Several strands of research explore related ideas, but none study preference over the quality of choices directly.

The most closely related are the menu preference models of temptation and self-control that build on Gul and Pesendorfer (2001). In the standard two-period setting, the agent in period 1 chooses a menu from which his future self chooses an alternative in period 2. These models axiomatically characterize the choices in the first period when the agent anticipates a mental cost of forgoing an alternative in period 2.[4] Through the lens of my framework, suppose the DM chooses menus "as if" his utility function is $M(A) \coloneqq \max_{x \in A} U(x, A)$ where $U(x, A)$ is the representation in (1). The evaluator can be regarded as the agent's future self and the DM as the agent in period

[3] Regret (Bell, 1982; Loomes and Sugden, 1982) is a classic example. Contextual features (e.g., whether an option is the last one available in the market) can also be influential (see Dietrich and List, 2016; Shafir et al., 1993).

[4] The mental cost usually refers to self-control. Sarver (2008) models regret, Dillenberger and Sadowski (2012) shame, and Saito (2015) pride and shame.

1 who wants his future self to make better choices. In this sense, the function U represents the preference over his own period-2 choices. The function M nests several specifications of menu preferences in prior studies as special cases where the only variable is the reference function.[5] However, the standard dynamics of self-control problems, in which menu choices precede outcome choices, implicitly assume that the future self either mistakenly perceives the menu he chose as exogenously given or simply regards the quality of the menu choice as irrelevant to period-2 utility—an assumption that is unverifiable from menu-choice data (see Bernheim et al., 2024, who formally characterized this unverifiability).[6] My approach permits the evaluator and DM to be two distinct individuals, and thus the assumption is irrelevant.[7]

The closest precedent for using outcome-menu pairs as primitives in an axiomatic approach is Suzumura and Xu (2009), who study preferences for freedom using sure outcomes (see also Sen, 1988; Pattanaik and Xu, 1990, 1998). To my knowledge, my model is the first to employ the structure of lotteries—which is essential for my results—to study outcome-menu pairs in a similar style. The motivation is also different: e.g., the agent in their model may prefer (x, A) to (y, B) whenever A contains more alternatives than B, whereas the DM's autonomy itself has no direct value in my model; it is valuable only indirectly, depending on the menu.

To my knowledge, this is the first paper to propose a characterization of a higher-order preference through a revealed-preference argument built directly on outcome–menu primitives, rather than focusing on the logic of higher-order preferences. There

[5] For example, Gul and Pesendorfer (2001)'s preference for commitment is represented by the function $M_{GP}(A) \coloneqq \max_{x\in A} u(x) + v(x) - \max_{y\in A} v(y)$. We can easily see that $M = M_{GP}$ when the function r satisfies $v(r(A)) = \max_{y\in A} v(y)$ for each A. As another example, the function M also nests Kopylov (2012)'s representation for a perfectionist who benefits from having normative goals on his menu. The representation takes the form $M_K(A) \coloneqq \bar{u}(x) + \bar{v}(x) - [\max_{y\in A} \bar{v}(y) - \kappa \max_{z\in A} \bar{u}(z)]$ where $\kappa \geq 0$. Let $u(x) = (1+\kappa)\bar{u}(x)$, $v(x) = \bar{v}(x) - \kappa\bar{u}(x)$, and $v(r(A)) = \max_{y\in A} \bar{v}(y) - \kappa \max_{z\in A} \bar{u}(z)$. Then $M = M_K$. There are representations in the literature that the function M fails to nest. Dekel et al. (2009) assume the context of uncertainty: e.g., the agent does not know which option will be tempting (see also Stovall, 2010; Dekel and Lipman, 2012). Noor and Takeoka (2015) relax the EU axioms to explore non-linear mental costs. Others expand the dynamics and study menu choices in three or more periods (see, for example, Stovall, 2018).

[6] Some later models even take higher-order menus (menus of menus of outcomes, and so on) as primitives (see Noor, 2011; Noor and Ren, 2023). Such constructions capture higher-order self-control but still require the unverifiable assumption.

[7] Even in the self-evaluation case, it is plausible that individuals do not always eliminate temptations ex ante (e.g., by refusing to own a television). Rather, they first observe their own behavior and, upon recognizing repeated failure to resist temptation, consider restricting their future menus (e.g., "I have been using my smartphone too much in the library; perhaps I should leave it behind when I go to study").

is a vast literature in philosophy and economics on metapreferences and freedom of will, which generally takes a prescriptive stance.[8] These approaches treat higher-order preferences as objects of choice, interpreted as one's higher-order desires. By contrast, my framework takes a descriptive, choice-theoretic stance: I do not directly model preference over *preference relations* interpreted as one's desires, but over actual choices. I show that *if* choices over choices are observable, the underlying second-order preference can be identified.[9] In a similar sense, this paper is the first to exploit menu variation to study higher-order preferences.[10]

This paper also contributes to the study of welfare when the act of choosing is itself welfare-relevant (Kőszegi and Rabin, 2008; Bernheim et al., 2024). When the evaluator is regarded as the DM himself, my representation can be interpreted as a welfare measure that depends not only on outcomes but also on the quality of his own choices. Bernheim et al. (2024) assume an arbitrary welfare function of choices for the DM, without structurally articulating how the set of forgone alternatives shapes that evaluation. Instead, they rely on the DM's self-reported well-being data to estimate it. I offer a preference-based utility framework for the welfare associated with choices. Alternatively, the evaluator can be interpreted as a social planner who has a preference over the DM's choices—e.g., a parent's paternalistic decisions based on children's choices or a policymaker's decisions based on people's choices. This differs from the standard welfare approach, which typically infers the DM's preference from his choices.

[8] Frankfurt (1971) described *freedom of the will* as the capacity to have consistent higher-order desires—wanting x, wanting to want x, and so on (see also Nehring, 2006; Pivato, 2024; Hayashi, 2024). Pivato (2025) formalizes this idea, asking how the freedom of will can be mathematically characterized.

[9] This raises an obvious difficulty in experimentation, as incentivizing choices over choices in laboratory settings may be unnatural (see Section 6 for more discussion). The contribution of the present analysis is instead to characterize the structure of second-order preference and what can be identified once sufficiently rich comparisons among choices are available.

[10] Previous studies have largely focused on a fixed choice problem (e.g., the heavy smoker of Jeffrey (1974) choosing from $\{\text{smoke}, \text{abstain}\}$), fixing an arbitrary menu. However, my model suggests that second-order preferences cannot be identified based on choices over choices from a fixed menu.

2. Model

2.1. Choices in Lottery Space

Let Z be the finite set of alternatives, and X be the set of lotteries on Z, endowed with a metric d generating the standard weak topology. Elements $x, y, z \in X$ are called lotteries, *outcomes*, or *options*. Let $\mathbb{M}$ denote the set of nonempty compact subsets of X. Its elements, denoted $A, B \in \mathbb{M}$, are called *menus*.[11] Let $\text{conv}(A)$ denote *the convex hull of* A. Naturally, we have $\text{conv}(A) \in \mathbb{M}$ for all $A \in \mathbb{M}$, which can be interpreted as the menu that offers every randomization of the options in A.

A *choice* is a pair (x, A) such that $x \in A \in \mathbb{M}$. Henceforth, I write (x, A) only when $x \in A$. *The set of choices* is the set $\mathbb{C} = \{(x, A) : x \in A \in \mathbb{M}\}$. Any element of the set $\{(x, \{x\}) : x \in X\}$ represents the exogenous consumption of a single lottery, and therefore is called a *vacuous choice*. The choice $(x, \{x\})$ indicates that the DM *vacuously chooses* x.[12] For $A, B \in \mathbb{M}$ and $\alpha \in [0, 1]$, define $\alpha A + (1-\alpha) B \coloneqq \{\alpha x + (1-\alpha) y : x \in A,\ y \in B\}$ where $\alpha x + (1-\alpha)\, y \in X$ denotes the convex combination of lotteries x and y. Accordingly, define $\alpha(x, A) + (1-\alpha)(y, B) \coloneqq (\alpha x + (1-\alpha) y, \alpha A + (1-\alpha) B)$ for any $\alpha \in [0, 1]$ and $(x, A), (y, B) \in \mathbb{C}$.

I assume that the evaluator does not distinguish between a lottery over choices and its reduced outcome-menu form (see Samuelson, 1952). This reduction is built into the domain $\mathbb{C}$.[13] In addition to the interpretation based on the reduction of compound lotteries, there are three other ways one might interpret convex combinations of choices in the lottery space. Consider the choice $(\alpha x + (1-\alpha)\, y, \alpha A + (1-\alpha)\, B)$ where $x \in A$, $y \in B$, and $\alpha \in (0, 1)$. First, α may represent a contingency plan: the DM faces menu A with probability α and B with probability $1 - \alpha$, committing to "choose x if A, and y if B." Second, α may represent relative frequency: menu

[11] I endow $\mathbb{M}$ with the Hausdorff metric $d_H(A, B) \coloneqq \max\{\max_{x \in A} \min_{y \in B} d(x, y),\ \max_{y \in B} \min_{x \in A} d(x, y)\}$.

[12] I use the word "vacuous" to emphasize that because a choice in general implies at least one forgone alternative, a singleton menu renders the notion of choice empty.

[13] Alternatively, one could define the primitive domain as the space of simple lotteries over choices, $\Delta(\mathbb{C})$, and impose the assumption explicitly alongside the standard *Independence axiom*. In particular, consider the lottery that yields the choice (x, A) with probability α and the choice (y, B) with probability $1 - \alpha$, denoted $\alpha \circ (x, A) + (1 - \alpha) \circ (y, B)$. This lottery, which belongs to $\Delta(\mathbb{C})$, would be treated as equivalent to the reduced choice $(\alpha x + (1-\alpha) y, \alpha A + (1-\alpha) B) \in \mathbb{C}$. An analogous lottery-over-menus construction is used to motivate indifference to the timing of uncertainty resolution in Gul and Pesendorfer (2001) and Dekel et al. (2001, 2007).

A was given α times in which case, the DM chose x, and menu B the remaining $1-\alpha$ times, in which case he chose y.[14] Together, these interpretations extend Ok and Tserenjigmid (2022)'s interpretations of stochastic choice—originally defined for a single deterministic menu—to the space of lotteries where both choices and menus are stochastic. Lastly, the choice may be the evaluator's expectation about the DM's choice from the risky menu $\alpha A + (1-\alpha) B$.

2.2. Standard Axioms and Implications

The primitive of the model is a binary relation $\succsim$ on $\mathbb{C}$. I first impose natural extensions of the standard vNM axioms and discuss their implications.

Axiom 1 (Weak Order)**.** *$\succsim$ is complete and transitive.*

Axiom 2 (Independence)**.** *For all $\lambda \in (0,1)$, $(x,A) \succ (y,B)$ implies $\lambda(x,A) + (1-\lambda)(z,C) \succ \lambda(y,B) + (1-\lambda)(z,C)$.*

Axiom 3 (Continuity)**.** *$\{(x,A) : (x,A) \succsim (y,B)\}$ and $\{(x,A) : (y,B) \succsim (x,A)\}$ are closed.*

One of the most conventional implications of the EU theory in a standard setting is that randomization itself has no value. Likewise, if the evaluator's preference over choices conforms to the EU principles, then she finds no value in *giving up* randomization. Suppose the DM's choice is $(x, \{x,y\})$. Even though his menu does not explicitly include any nondegenerate randomization between x and y, the evaluator cannot prevent him from randomizing—e.g., flipping an imaginary coin, creating multiple states or personal contingencies. The standard axioms render $\succsim$ insensitive to whether such randomization was available. This is formally written as $(x, \{x,y\}) \sim (x, \mathrm{conv}(\{x,y\}))$ where $\mathrm{conv}(\{x,y\}) = \{\alpha x + (1-\alpha)\, y : \alpha \in [0,1]\}$.

The following result formalizes this implication.

Lemma 1 (Irrelevance of Forgone Randomization)**.** *Suppose $\succsim$ satisfies Axioms 1-3. Then for each $A, B \in \mathbb{M}$, $B \subseteq \mathrm{conv}(A)$ implies $(x,A) \sim (x, A \cup B)$.*

Proof. See Appendix. *Q.E.D.*

[14] The empirical interpretation abstracts from scale. When outcomes accumulate with the number of choices, meaningful comparison between two choices would require normalizing the total number of choices given to each DM.

As a proof sketch, let $A = \{x, y\}$ for simplicity, and let $A_2 \coloneqq \frac{1}{2}A + \frac{1}{2}A$. The DM, for instance, can condition his choice on good and bad weather, each occurring with equal probability. His menu effectively becomes A_2. Yet, if he ultimately chooses x from $\{x, y\}$ in both states, the vNM axioms imply that the evaluator is indifferent between the choice of x in this two-state environment and the original single-state choice. Suppose instead that $(x, A_2) \not\sim (x, A)$. By completeness and Axiom 2, $\frac{1}{2}(x, A) + \frac{1}{2}(x, \text{conv}(A)) \not\sim \frac{1}{2}(x, A_2) + \frac{1}{2}(x, \text{conv}(A))$. However, the two choices are both equal to $(x, \text{conv}(A))$ since $\frac{1}{2}A + \frac{1}{2}\text{conv}(A) = \frac{1}{2}A_2 + \frac{1}{2}\text{conv}(A) = \text{conv}(A)$, a contradiction. Hence, we have $(x, A) \sim (x, A_2)$.

By the same reasoning, the evaluator remains indifferent to any further decomposition of states, as long as the DM chooses x in each one:

$$\begin{aligned}(x, \{x, y\}) \sim & \tfrac{1}{2}(x, \{x, y\}) + \tfrac{1}{2}(x, \{x, y\}) \\ & \sim \tfrac{1}{3}(x, \{x, y\}) + \tfrac{1}{3}(x, \{x, y\}) + \tfrac{1}{3}(x, \{x, y\}) \sim \cdots .\end{aligned}$$

As the DM considers more finely divided contingencies, his choice approaches $(x, \text{conv}(\{x, y\}))$. By Axiom 3 and transitivity, the indifference is preserved in the limit: $(x, \{x, y\}) \sim (x, \text{conv}(\{x, y\}))$. Then, for any $B \subseteq \text{conv}(\{x, y\})$, we have $(x, \{x, y\}) \sim (x, \{x, y\} \cup B)$ since $\text{conv}(\{x, y\} \cup B) = \text{conv}(\{x, y\})$.[15] This completes the proof.

Another conventional consequence of the EU theory is an affine representation. I use the standard definitions: the function $U : \mathbb{C} \to \mathbb{R}$ represents $\succsim$ if $U(x, A) \geq U(y, B)$ is equivalent to $(x, A) \succsim (y, B)$, and it is affine if $U(\lambda(x, A) + (1 - \lambda)(y, B)) = \lambda U(x, A) + (1 - \lambda) U(y, B)$ for all $(x, A), (y, B) \in \mathbb{C}$ and $\lambda \in [0, 1]$. However, the mixture space theorem (Herstein and Milnor, 1953) cannot be used directly since $\mathbb{C}$ is not a mixture space.[16] Nevertheless, we can show that $\succsim$ still admits an affine representation using Lemma 1.

Proposition 1. *$\succsim$ satisfies Axioms 1-3 if and only if there exists a continuous affine function $U : \mathbb{C} \to \mathbb{R}$ representing $\succsim$.*

[15] The conditioning device need not be weather per se. The DM may condition his choice on any external state that leaves the menu unchanged; under the vNM axioms, such conditioning is irrelevant for the evaluator's assessment. In contrast, the evaluator will pay attention to states that alter feasibility—e.g., the DM's illness that eliminates x but not y from his menu.

[16] $\mathbb{C}$ under the operation used in this paper is not a mixture space (*one-sided distributivity* fails) because, for a nonconvex menu A and $\lambda \in (0, 1)$, $\lambda A + (1 - \lambda)A = A$ does not hold in general.

Proof. Restrict $\succsim$ to $\mathbb{C}^c := \{(x, A) \in \mathbb{C} : A \text{ is convex}\}$. Note that $\mathbb{C}^c$ is a mixture space (I leave the proof in the appendix). Since $\succsim$ satisfies Axioms 1-3 on $\mathbb{C}$, its restriction to $\mathbb{C}^c$ also satisfies them. By the mixture space theorem, there exists a continuous affine function $\widehat{U} : \mathbb{C}^c \to \mathbb{R}$ representing the restriction of $\succsim$ to $\mathbb{C}^c$. Now define $U : \mathbb{C} \to \mathbb{R}$ by $U(x, A) := \widehat{U}(x, \text{conv}(A))$. By Lemma 1, we have $(x, A) \sim (x, \text{conv}(A))$ for every $(x, A) \in \mathbb{C}$, and thus $(x, A) \succsim (y, B)$ if and only if $(x, \text{conv}(A)) \succsim (y, \text{conv}(B))$. Since $\widehat{U}$ represents $\succsim$ on $\mathbb{C}^c$, it follows that $(x, A) \succsim (y, B)$ if and only if $U(x, A) \geq U(y, B)$. Hence, U represents $\succsim$ on $\mathbb{C}$. It is straightforward to show that U is affine on $\mathbb{C}$. *Q.E.D.*

In addition to the affine representation, the vNM axioms also grant separability: the evaluator's utility function must be additively separable in its arguments—the outcomes and the menus. This is formally shown as follows.

Proposition 2 (Separability)**.** *Suppose $\succsim$ satisfies Axioms 1-3. Then, for all (x, A), $(y, A), (x, B), (y, B) \in \mathbb{C}$,*

(a) $(x, A) \succsim (y, A)$ if and only if $(x, B) \succsim (y, B)$, and

(b) $(x, A) \succsim (x, B)$ if and only if $(y, A) \succsim (y, B)$.

Proof. For (a), suppose toward a contradiction that $(x, A) \succsim (y, A)$ and $(y, B) \succ (x, B)$ for some $x, y \in A \cap B$. By Proposition 1, we have $\frac{1}{2}(x, A) + \frac{1}{2}(y, B) \succ \frac{1}{2}(y, A) + \frac{1}{2}(x, B)$. However, these two choices are identical, since $\frac{1}{2}(x, A) + \frac{1}{2}(y, B) = \left(\frac{1}{2}x + \frac{1}{2}y, \frac{1}{2}A + \frac{1}{2}B\right) = \frac{1}{2}(y, A) + \frac{1}{2}(x, B)$. This is a contradiction. Similarly, $(x, B) \succsim (y, B)$ also implies $(x, A) \succsim (y, A)$. For (b), suppose $(x, A) \succsim (x, B)$ but $(y, B) \succ (y, A)$. There are two cases to consider: (i) $(x, A) \succsim (y, A)$ and (ii) $(y, A) \succ (x, A)$. By (a), (i) implies $(x, B) \succsim (y, B)$. Then, we have $(x, A) \succsim (x, B) \succsim (y, B) \succ (y, A)$, which implies $\frac{1}{2}(x, A) + \frac{1}{2}(y, B) \succ \frac{1}{2}(x, B) + \frac{1}{2}(y, A)$ by Proposition 1. It follows that $(\frac{1}{2}x + \frac{1}{2}y, \frac{1}{2}A + \frac{1}{2}B) \succ (\frac{1}{2}x + \frac{1}{2}y, \frac{1}{2}B + \frac{1}{2}A)$, which is a contradiction because the two choices are identical. A similar contradiction is reached in case (ii). *Q.E.D.*

Proposition 2(a) states that the evaluator's preference induces a consistent ranking of outcomes conditional on menus: she prefers "choosing x from A" to "choosing y from A" if and only if "choosing x from any other menu B" is preferred to "choosing y from B."[17] Note that this does not determine the ranking *across* menus: i.e.,

[17] Proposition 2 also implies that the outcome-choice rule induced by $\succsim$, defined by $c(A; \succsim) := \{x \in A : (x, A) \succsim (y, A) \ \forall y \in A\}$, satisfies the weak axiom of revealed preference.

$(x, A) \succsim (y, A)$ does not guarantee $(x, A) \succsim (y, B)$. This fundamentally differs from the behavior of individuals who care only about outcomes because they would prefer (x, A) to (y, B) whenever x yields a better outcome than y. Proposition 2(b) implies that if *"giving up A for x"* is preferred to *"giving up B for x,"* then *"giving up A"* is preferred to *"giving up B"* for any other common option.

2.3. Decomposition of Preference over Choices

I decompose the preference $\succsim$ over choices into two parts. The first part is the evaluator's preference over outcomes, which captures how much the evaluator values the consequence of each choice, regardless of how it is chosen. This means it can be characterized by a ranking of lotteries alone. In that sense, the outcome preference is also referred to as the evaluator's first-order preference, which is defined as follows:

Definition 1. The binary relation $\succcurlyeq_1^*$ on X is *the first-order preference* if $x \succcurlyeq_1^* y$ whenever $(x, \{x\}) \succsim (y, \{y\})$.

$x \succcurlyeq_1^* y$ denotes that the outcome of x is preferred to that of y, which holds whenever vacuously choosing x is preferred to vacuously choosing y.

The second part is the evaluator's preference that is unrelated to the outcomes, which is called the outcome-irrelevant (OI) preference. It is characterized by fixing a stochastic choice environment in which each choice yields the same outcome but corresponds to a different preference. This is the stage of my model where exploiting the lottery space is crucial. For example, suppose two DMs consumed x and y equally often, but their choices were made differently. One DM chose y only when x was not available, while the other chose x only when y was not available. Whenever they were free to choose between x and y, the first DM gave up y, and the second DM x. If the evaluator cares only about outcomes and conforms to the EU principles, she must be indifferent between the two DMs' choices because overall they both yielded the same distribution of alternatives. However, the two choices are aligned with different preferences: if the evaluator wants a DM who prefers x to y, she would favor the first DM's choice. The OI-preference is the evaluator's preference restricted to this environment.

Formally, the OI-preference is defined as follows:

Definition 2. The binary relation $\succsim^Q$ on $\mathbb{C}$ is called *the OI-preference over choices* if $(x, A) \succsim^Q (y, B)$ whenever $\left(\frac{1}{2}x + \frac{1}{2}y, \frac{1}{2}A + \frac{1}{2}\{y\}\right) \succsim \left(\frac{1}{2}y + \frac{1}{2}x, \frac{1}{2}B + \frac{1}{2}\{x\}\right)$.

The OI-preference can be interpreted in two ways. First, we can consider the evaluator comparing the choices of two DMs: DM_1 and DM_2 choose from menus A and B, respectively. To isolate the evaluator's outcome-irrelevant concern, an experimenter could stipulate that, regardless of what each DM selects, each will receive his own chosen outcome with probability 0.5 and the other DM's chosen outcome otherwise. Alternatively, we can consider the two DMs in an interdependent choice situation, each influencing the other's outcome with probability $\frac{1}{2}$, which is described by the menus $\frac{1}{2}A + \frac{1}{2}\{y\}$ and $\frac{1}{2}B + \frac{1}{2}\{x\}$.

In either interpretation, $\succsim^Q$ captures how $\succsim$ behaves when outcomes are irrelevant because the outcomes from the two menus are the same for all $x \in A$ and $y \in B$. The two menus differ only in how the DM's preferences are revealed. The choice of $\frac{1}{2}x + \frac{1}{2}y$ from menu $\frac{1}{2}A + \frac{1}{2}\{y\}$ reveals that x is preferred to everything else in A. Similarly, $(\frac{1}{2}y + \frac{1}{2}x, \frac{1}{2}B + \frac{1}{2}\{x\})$ reveals that y is preferred to all else in B. Hence, any strict comparison made by the OI-preference $\succsim^Q$ reflects the evaluator's outcome-irrelevant concern.

By construction, $\succsim^Q$ treats all vacuous choices as indifferent (see the following result; the proof is straightforward and omitted for brevity).

Proposition 3 (Indifference of Vacuous Choices)**.** *If $\succsim$ satisfies Axioms 1-3, then the induced OI-preference $\succsim^Q$ satisfies $(x, \{x\}) \sim^Q (y, \{y\})$ for all $x, y \in X$.*

The indifference of vacuous choices is an essential characteristic of preferences for the quality of choices because vacuous choices do not involve any willingly forgone opportunities.[18]

2.4. Second-order Preference and Menu-favoritism

The following axiom captures the general concept of preferences for the quality of choices.

Axiom 4 (Relativity)**.** *Each $A \in \mathbb{M}$ satisfies $(x, A) \succsim (x, \{x\})$ and $(y, \{y\}) \succsim (y, A)$ for some $x, y \in A$.*

[18] Suzumura and Xu (2009), in their study of preferences for freedom, impose the axiom *Indifference of No-Choice Situations*, which likewise requires any two choices from singleton menus to be indifferent because they provide no freedom, not because they do not involve willingly forgone alternatives.

Relativity captures the intuition that the quality of a choice is relative to constraints and therefore cannot be determined by outcomes or menus in isolation, nor independently of the choices themselves. Every menu comes with a relatively "good" option and a relatively "bad" option: the former is the one that the evaluator would like the DM to willingly choose, and the latter is what she would like him to give up. Let x and y denote the good and bad options on menu A. Then "choosing x willingly" must be at least as good as "consuming it unwillingly," which is identified with $(x, A) \succsim (x, \{x\})$; conversely, "choosing y willingly" must be at least as bad as "being forced to consume it"—i.e., $(y, \{y\}) \succsim (y, A)$.

Within the EU theory, I say the OI-preference $\succsim^Q$ is the evaluator's second-order preference if Relativity holds.

Definition 3. The OI-preference $\succsim^Q$ induced by $\succsim$ is called *the second-order preference over choices* if $\succsim$ satisfies Axioms 1-3 and Relativity.

Under Relativity, $(x, A) \succsim^Q (y, B)$ can be read intuitively as follows: *"preferring x to all else in A"* is preferred to *"preferring y to all else in B."* Naturally, $\succsim^Q$ also satisfies Axioms 1-3 and Relativity.[19]

The violation of Relativity is formally referred to as *menu-favoritism*, an evaluative bias toward the features of the menu itself, independent of outcomes or preferences.

Definition 4 (Menu-favoritism)**.** The preference $\succsim$ exhibits *menu-favoritism* if there are convex non-singletons $A, B \in \mathbb{M}$ such that $(x, A) \succ^Q (y, B)$ for all $x \in A$ and $y \in B$, where $\succsim^Q$ is the OI-preference induced by $\succsim$.

Suppose $(x, A) \succ^Q (y, B)$ for all $x \in A$ and $y \in B$. That is, *"preferring anything in A"* is strictly preferred to *"preferring anything in B,"* which implies that the quality of choices is independent of choices. The evaluator simply prefers a DM facing menu A to the one facing B.[20] The following result formally shows that violations of Relativity are equivalent to menu-favoritism.

Proposition 4. *Suppose $\succsim$ satisfies Axioms 1-3. Then, $\succsim$ violates Relativity if and only if $\succsim$ exhibits menu-favoritism.*

[19] Notice that Definition 3 is specific to the EU framework because the OI-preference relies on the EU principles. However, Relativity itself is theory-neutral: it is independent of the EU theory and remains well defined even for evaluators who violate Axioms 2-3.

[20] For instance, menu-favoritism arises when the evaluator is an extreme non-consequentialist as in Suzumura and Xu (2009), and has an extreme preference for the DM's freedom: i.e., $(x, A) \succ (y, B)$ simply because A has more options than B.

Proof. For the "if" part, suppose there exist convex non-singletons $A, B \in \mathbb{M}$ such that $(x, A) \succ^Q (y, B)$ for all $x \in A$ and $y \in B$. Suppose, toward a contradiction, that Relativity holds. Since $\succsim^Q$ also satisfies Relativity, there exist $x \in A$ and $y \in B$ such that $(x, \{x\}) \succsim^Q (x, A)$ and $(y, B) \succsim^Q (y, \{y\})$. By Proposition 3, all vacuous choices are indifferent under $\succsim^Q$, and thus $(y, \{y\}) \sim^Q (x, \{x\})$. Hence, by transitivity, $(y, B) \succsim^Q (y, \{y\}) \sim^Q (x, \{x\}) \succsim^Q (x, A)$, contradicting $(x, A) \succ^Q (y, B)$.

For the "only if" part, suppose there is a non-singleton menu B that violates Relativity. (Relativity holds trivially for singleton menus as long as $\succsim$ is complete.) This implies that either $(y, B) \not\succsim (y, \{y\})$ for all $y \in B$, or $(y, \{y\}) \not\succsim (y, B)$ for all $y \in B$. By completeness, these two cases are mutually exclusive: i.e., either (i) $(y, \{y\}) \succ (y, B)$ for all $y \in B$, or (ii) $(y, B) \succ (y, \{y\})$ for all $y \in B$, *but not both*. I consider only case (i), because the proof for case (ii) follows by an analogous argument. By Lemma 1, we have $(y, \{y\}) \succ (y, \mathrm{conv}(B))$ for all $y \in \mathrm{conv}(B)$; hence, for brevity, I assume $B = \mathrm{conv}(B)$.

I now show that, using B, we can construct a stochastic menu such that, controlling for the outcomes, every choice from that menu is strictly preferred to every choice from menu B. Pick any $b \in B$. For $\lambda \in (0, 1)$, define $A_\lambda := \lambda\{b\} + (1 - \lambda) B$. Note that A_λ is a convex non-singleton. Consider any choice from A_λ: a pair (x_{y_1}, A_λ) where $x_{y_1} := \lambda b + (1 - \lambda) y_1$ for some $y_1 \in B$. It follows from the definition of $\succsim^Q$ that $(x_{y_1}, A_\lambda) \succ^Q (y_2, B)$ is equivalent to $(\frac{1}{2}x_{y_1} + \frac{1}{2}y_2, \frac{1}{2}A_\lambda + \frac{1}{2}\{y_2\}) \succ (\frac{1}{2}y_2 + \frac{1}{2}x_{y_1}, \frac{1}{2}B + \frac{1}{2}\{x_{y_1}\})$, which, by the vNM axioms, is also equivalent to

$$\begin{aligned} &\tfrac{1}{2}\Big[\lambda(b, \{b\}) + (1 - \lambda)\,(y_1, B)\Big] + \tfrac{1}{2}(y_2, \{y_2\}) \\ &\qquad\qquad \succ \tfrac{1}{2}(y_2, B) + \tfrac{1}{2}\Big[\lambda(b, \{b\}) + (1 - \lambda)\,(y_1, \{y_1\})\Big]. \end{aligned} \tag{2}$$

Note that the choice on each side yields the same outcome of $\frac{\lambda}{2}b + \frac{1-\lambda}{2}y_1 + \frac{1}{2}y_2$. Notice that $(b, \{b\})$, with a probability weight of $\frac{\lambda}{2}$ on both sides, is irrelevant, and thus (2) is equivalent to

$$\left[\frac{1-\lambda}{2-\lambda}\right](y_1, B) + \left[\frac{1}{2-\lambda}\right](y_2, \{y_2\}) \succ \left[\frac{1}{2-\lambda}\right](y_2, B) + \left[\frac{1-\lambda}{2-\lambda}\right](y_1, \{y_1\}). \tag{3}$$

The vNM axioms imply that if λ is close enough to 1, the strict preference in (3) holds for any $y_1, y_2 \in B$, since $(y, \{y\}) \succ (y, B)$ for all $y \in B$. Hence, by the definition of the OI-preference $\succsim^Q$, we have menu-favoritism: $(x, A_\lambda) \succ^Q (y, B)$ for all $x \in A_\lambda$ and

$y \in B$. This completes the proof. *Q.E.D.*

The role of λ in this argument is worth noting. Under the OI-preference $\succsim^Q$, once λ is close enough to 1, every choice from A_λ is strictly preferred to every choice from B, as in (3). Yet λ is only a feature of menu A_λ taken as exogenous. Under Relativity, menu features such as λ cannot by themselves determine the ranking of choices. We can choose $y_1, y_2 \in B$ such that $(y_1, \{y_1\}) \succsim (y_1, B)$ and $(y_2, B) \succsim (y_2, \{y_2\})$. The strict preference in (2) then fails for every $\lambda \in (0,1)$. We can also choose $y_1, y_2 \in B$ such that $(y_2, \{y_2\}) \succ (y_2, B)$ and $(y_1, B) \succ (y_1, \{y_1\})$, then (2) holds for every $\lambda \in (0,1)$. Thus, whether the comparison holds depends instead on which options are chosen.

2.4.1. Remarks on Relativity

What does Relativity rule out? A violation of Relativity is a sign that the evaluator is interested in external factors, beyond the outcomes and the quality of the DM's choices. For example, she may have a preference over the DM's freedom itself (Suzumura and Xu, 2009) or flexibility (Kreps, 1979). Or she may prefer that an option be enforced rather than willingly chosen—desiring to spare him the cognitive cost of processing large menus (see, for example, Iyengar and Lepper, 2000), valuing her control over his outcomes, or, as in *Sophie's Choice*, wishing that any particular selection be forced rather than made freely in an agonizing decision problem.[21] Relativity rules out all such motives.

Despite ruling out many interesting motives, Relativity amounts to a simple normative statement, highlighted by Propositions 3-4: *the value of the DM's willingness (choices) should not be evaluated independently of the choices themselves, based solely on their consequences or exogenously given abilities (menus).* Proposition 3 captures the irrelevance of outcomes, and Proposition 4 ensures that menus cannot by them-

[21] For example, let h and t denote homework and television, respectively, and consider a parent who not only wants her child to do h rather than watch t but also wants him to prefer h to t. In terms of her preference, we have $(h, \{h\}) \succ (t, \{t\})$ and $(h, \{h,t\}) \succ (t, \{h,t\})$. A typical violation of Relativity would be either $(t, \{h,t\}) \succ (t, \{t\})$ or $(h, \{h\}) \succ (h, \{h,t\})$. In the first case, observing a bad choice is strictly preferred to letting him watch television when no homework is assigned; in the second, enforcing homework time is strictly preferred to the child's decision to make the right choice. Both cases indicate that the parent's evaluation cannot be reduced to the child's preferences or outcomes alone; for instance, the first case may stem from a preference for the child's freedom, while the second may reflect a preference for exercising control.

selves determine choice quality.[22]

2.5. Ideal Preference

The second-order preference can be characterized with two components. The first is a ranking of lotteries, referred to as the ideal outcome preference. The second is a function called the reference function. The first component is defined as follows.

Definition 5. Let $\succsim^Q$ be the OI-preference induced by $\succsim$. The binary relation $\succcurlyeq^{\mathcal{I}}$ on X is *the ideal preference of* $\succsim$ if $x \succcurlyeq^{\mathcal{I}} y$ whenever $(x, \{x, y\}) \succsim^Q (y, \{x, y\})$.

Since the evaluator wants the DM to follow certain preferences in each decision environment, some notion of an "ideal DM"—whose choice is always of the highest quality given any menu—must underlie her assessment. Since $(x, \{x, y\}) \succsim^Q (y, \{x, y\})$ means "preferring x to y" is preferred to "preferring y to x," I denote this relation by $\succcurlyeq^{\mathcal{I}}$, written as $x \succcurlyeq^{\mathcal{I}} y$, and read it as "$x$ is *ideally (weakly) preferred* to y."

Yet the above definition specifies the ideal DM's behavior only for doubleton menus and is therefore an incomplete characterization. Consider the following rankings: $(x, \{x, y\}) \succ^Q (y, \{x, y\})$ and $(y, \{x, y, z\}) \succ^Q (x, \{x, y, z\})$. The former implies that when menu $\{x, y\}$ is involved, the evaluator's ideal DM would prefer x to y. The latter, however, implies the opposite when a third option z becomes available. This phenomenon can be referred to as second-order preference reversals. The vNM axioms rule out these reversals.

Proposition 5. *Suppose* $\succsim$ *satisfies Axioms 1-3. Let* $\succsim^Q$ *and* $\succcurlyeq^{\mathcal{I}}$ *be the induced OI-preference and the ideal preference of* $\succsim$*, respectively. Then,*

(a) $\succsim^Q$ *satisfies* $(x, A) \succsim^Q (y, A)$ *if and only if* $(x, B) \succsim^Q (y, B)$*,*

(b) $x \succcurlyeq^{\mathcal{I}} y$ *if and only if* $(x, A) \succsim^Q (y, A)$ *for all* $A \ni x, y$*,*

[22] When a menu is given to the DM rather than chosen by him, the options open to him are not of his own doing. Immanuel Kant in his book *Groundwork of the Metaphysics of Morals (1785)* writes that *"A good will is not good because of what it effects or accomplishes, because of its fitness to attain some proposed end, but only because of its volition"* (Kant, 1785; Gregor, 1998). Kant excludes two things from the goodness of a will: the outcome it produces, and its fitness to bring about some possible outcome. In my setting, this fitness corresponds to the DM's menu: it settles what he is able to accomplish before he chooses anything. Under menu-favoritism, the quality of a choice is determined by this given capacity; under Relativity it depends instead on the DM's choice.

(c) $\succcurlyeq^{\mathcal{I}}$ *is complete, transitive, continuous, and independent.*[23]

Proof. Part (a) is Proposition 2 extended to $\succsim^{Q}$. For (b), note that (a) implies

$$x \succcurlyeq^{\mathcal{I}} y \iff (x, \{x, y\}) \succsim^{Q} (y, \{x, y\}) \iff (x, A) \succsim^{Q} (y, A) \quad \forall A \ni x, y.$$

For (c), note that (b) implies $x \succcurlyeq^{\mathcal{I}} y$ if and only if $(x, X) \succsim^{Q} (y, X)$. Since $\succsim^{Q}$ satisfies Axioms 1-3, the restriction of $\succsim^{Q}$ to $\{(x, X) : x \in X\}$ also satisfies Axioms 1-3. Hence, (c) is true. *Q.E.D.*

2.6. Reference Function

I now characterize the second component of the second-order preference: the reference function. I use the following result.

Proposition 6. *Suppose* $\succsim$ *satisfies Axioms 1-3 and Relativity. Then, each* $A \in \mathbb{M}$ *satisfies* $(x, A \cup \{x\}) \sim (x, \{x\})$ *for some* $x \in \text{conv}(A)$*.*

Proof. By Relativity, each $A \in \mathbb{M}$ contains $b, w \in A$ such that $(b, A) \succsim (b, \{b\})$ and $(w, \{w\}) \succsim (w, A)$. Then, Axioms 1-3 ensure the existence of a $\lambda \in [0, 1]$ such that $\lambda(b, A) + (1 - \lambda)(w, A) \sim \lambda(b, \{b\}) + (1 - \lambda)(w, \{w\})$, which is equivalent to $(x, \lambda A + (1 - \lambda) A) \sim (x, \{x\})$ where $x = \lambda b + (1 - \lambda) w \in \text{conv}(A)$. Since the set $\lambda A + (1 - \lambda) A$ is equal to $(A \cup \{x\}) \cup B$ for some $B \subset \text{conv}(A)$, it follows from Lemma 1 that $(x, \lambda A + (1 - \lambda) A) \sim (x, A \cup \{x\})$. By transitivity, we have $(x, A \cup \{x\}) \sim (x, \{x\})$. *Q.E.D.*

According to Proposition 6, Relativity within the EU framework implies that if b and w are, respectively, the "good" and "bad" options to willingly choose from A, then there must be a mixture x of b and w such that choosing x from A is neither better nor worse than consuming it unwillingly. This allows us to define a choice function that selects such a mixture for each menu.

I say the function $c : \mathbb{M} \to X$ is a *(stochastic) choice function* if $c(A) \in \text{conv}(A)$ for all $A \in \mathbb{M}$. I define the evaluator's *reference function* as follows:

Definition 6. A choice function $r : \mathbb{M} \to X$ is *a reference function* induced by $\succsim$ if $r(A) \in \{x \in \text{conv}(A) : (x, A \cup \{x\}) \sim (x, \{x\})\}$.

[23] A binary relation $\succcurlyeq$ on X (along with its asymmetric components $\succ$) is *independent* if $x \succ y$ and $\alpha \in (0, 1)$ imply $\alpha x + (1 - \alpha) z \succ \alpha y + (1 - \alpha) z$. It is *continuous* if $\{x \in X : x \succcurlyeq y\}$ and $\{x \in X : y \succcurlyeq x\}$ are closed.

Note that, by Proposition 6, the set $\{x \in \text{conv}(A) : (x, A \cup \{x\}) \sim (x, \{x\})\}$ is nonempty. If it is not a singleton, then any element may be chosen.

The following result shows that the second-order preference $\succsim^Q$ can be characterized by the ideal preference and the reference function.

Proposition 7. *Suppose $\succsim^Q$ is the second-order preference induced by $\succsim$. Then, for each $(x, A), (y, B) \in \mathbb{C}$, the induced ideal preference $\succcurlyeq^{\mathcal{I}}$ and the reference function $r(\cdot)$ satisfy*

$$\tfrac{1}{2}x + \tfrac{1}{2}r(B) \succcurlyeq^{\mathcal{I}} \tfrac{1}{2}r(A) + \tfrac{1}{2}y \textit{ if and only if } (x, A) \succsim^Q (y, B).$$

Proof. By Proposition 6 and the indifference of vacuous choices, the second-order preference $\succsim^Q$ satisfies $(r(A), A \cup \{r(A)\}) \sim^Q (r(B), B \cup \{r(B)\})$ for any $A, B \in \mathbb{M}$. Lemma 1 implies that $(x, A) \succsim^Q (y, B)$ is equivalent to $(x, A \cup \{r(A)\}) \succsim^Q (y, B \cup \{r(B)\})$ since $r(A) \in \text{conv}(A)$ and $r(B) \in \text{conv}(B)$. By Axiom 2, it follows that

$$\tfrac{1}{2}(x, A \cup \{r(A)\}) + \tfrac{1}{2}(r(B), B \cup \{r(B)\}) \succsim^Q \tfrac{1}{2}(y, B \cup \{r(B)\}) + \tfrac{1}{2}(r(A), A \cup \{r(A)\})$$

In other words,

$$\begin{aligned} &\left(\tfrac{1}{2}x + \tfrac{1}{2}r(B),\ \tfrac{1}{2}A \cup \{r(A)\} + \tfrac{1}{2}B \cup \{r(B)\}\right) \\ &\qquad \succsim^Q \left(\tfrac{1}{2}r(A) + \tfrac{1}{2}y,\ \tfrac{1}{2}A \cup \{r(A)\} + \tfrac{1}{2}B \cup \{r(B)\}\right) \end{aligned}$$

which holds if and only if $\frac{1}{2}x + \frac{1}{2}r(B) \succcurlyeq^{\mathcal{I}} \frac{1}{2}r(A) + \frac{1}{2}y$ by Proposition 5. *Q.E.D.*

According to Proposition 7, *preferring x to all else in A* is preferred to *preferring y to all else in B* if and only if the coin toss between x and $r(B)$ is ideally preferred to the coin toss between $r(A)$ and y.[24] This implies that the quality of the choice (x, A) itself increases as the ideal value difference between the chosen option x and the reference of A increases.[25]

[24] An early attempt to formalize preferences over preferences appears in Halldén (1980), who proposed a rule for judging which preference ranking is more important to keep in mind. Proposition 7 captures his rule as a special case.

[25] Individuals' tendency to assess an outcome of a choice in contrast with a reference has been discussed previously. Kőszegi and Rabin (2006)'s reference-dependent preference captures a loss-averse agent's tendency to assess an outcome of a choice in contrast with his expectation about the outcome, which arises from uncertainty. Yet my model remains within expected utility theory, and the reference stems from the evaluator's taste. Therefore, the term "reference" here fundamentally differs from the term "reference point" used in the literature.

As a consequence, the consistent ranking of menus in Proposition 2(b) is entirely attributable to the reference function.

Corollary 1. *$(x, A) \succsim (x, B)$ if and only if $r(B) \succcurlyeq^{\mathcal{I}} r(A)$.*

3. Representation

This section presents the representation theorem for preference over choices. I begin with a few definitions. Given a binary relation $\succcurlyeq_1$ on X, a function $f : X \to \mathbb{R}$ represents $\succcurlyeq_1$ if $f(x) \geq f(y)$ if and only if $x \succcurlyeq_1 y$, and it is affine if $f(\alpha x + (1-\alpha)\, y) = \alpha f(x) + (1-\alpha)\, f(y)$ for all $x, y \in X$ and $\alpha \in [0,1]$. A choice function $c : \mathbb{M} \to X$ is *affine with respect to f* if $f(c(\lambda A + (1-\lambda)\, B)) = f(\lambda c(A) + (1-\lambda)\, c(B))$ for all $A, B \in \mathbb{M}$ and $\lambda \in (0,1)$.[26]

Definition 7. A representation of a *preference over choices* (PC) is a tuple (u, v, r) where u, v are affine functions of lotteries and r is an affine choice function with respect to v such that $\succsim$ is represented by $U_{u,v,r}(x, A) \coloneqq u(x) + v(x) - v(r(A))$.

The following theorem is the main result.

Theorem 1. *$\succsim$ satisfies Axioms 1-3 and Relativity if and only if $\succsim$ has a PC representation (u, v, r). Furthermore, the first-order preference $\succcurlyeq_1^*$ is represented by u, the ideal preference $\succcurlyeq^{\mathcal{I}}$ is represented by v, and the second-order preference $\succsim^Q$ is represented by the function $Q_{v,r} : \mathbb{C} \to \mathbb{R}$ of the form*

$$Q_{v,r}(x, A) \coloneqq v(x) - v(r(A)).$$

Proof. The "if" part is straightforward. For the "only if" part, first note that by Proposition 1, there is a continuous affine function $U : \mathbb{C} \to \mathbb{R}$ representing $\succsim$. Proposition 2 implies that U is of the additively separable form $U(x, A) = h(x) - R(A)$ for some continuous affine functions $h : X \to \mathbb{R}$ of lotteries and $R : \mathbb{M} \to \mathbb{R}$ of sets.[27]

[26] This condition requires affinity only in the values assigned by f; it does not require $c(\lambda A + (1-\lambda)\, B) = \lambda c(A) + (1-\lambda)\, c(B)$.

[27] A real-valued function R of sets is affine if $R(\lambda A + (1-\lambda)\, B) = \lambda R(A) + (1-\lambda)\, R(B)$ for $A, B \in \mathbb{M}$ and $\lambda \in (0,1)$.

To identify the two functions h and R, pick any lottery $x_0 \in X$ and define four real-valued functions $\tilde{h}, u, v$ of lotteries, and $\tilde{R}$ of sets, by

$$u(x) \coloneqq U(x, \{x\}); \qquad \tilde{h}(x) \coloneqq U(x, X) - U(x_0, X);$$
$$v(x) \coloneqq \tilde{h}(x) - u(x); \qquad \tilde{R}(A) \coloneqq \tilde{h}(x) - U(x, A).$$

Note that, by construction, $\tilde{R}$ is a function of sets alone: $\tilde{h}(x) - U(x, A) = \tilde{h}(y) - U(y, A)$ for any $x, y \in A$. Also, all functions $\tilde{h}, u, v$ and $\tilde{R}$ are affine because Axioms 1-3 imply that $\succsim$ restricted to either $\{(x, \{x\}) : x \in X\}$ or $\{(x, X) : x \in X\}$ also admits a continuous affine representation. In particular, $\tilde{h}$ is merely an affine transformation of h, and thus we can safely assume $h = \tilde{h}$ and $R = \tilde{R}$.

I now claim that $R(A) = v(r(A))$ for all A. This is trivially true when A is a singleton. Now, for an arbitrary $A \in \mathbb{M}$, we have $U(y, A) = U(y, A \cup \{r(A)\})$ for any $y \in A$ by Lemma 1. It follows from $U(y, A) = h(y) - R(A)$ that

$$R(A) = R(A \cup \{r(A)\}). \tag{4}$$

Proposition 6 implies that $u(r(A)) = h(r(A)) - R(A \cup \{r(A)\})$. Then, by (4), we have $u(r(A)) = h(r(A)) - R(A)$. Finally, by the definition of v, my claim that $R(A) = v(r(A))$ is true.

It remains to show that $r(\cdot)$ is affine with respect to v. Since R and v are affine functions, we have $v(r(\lambda A + (1 - \lambda) B)) = R(\lambda A + (1 - \lambda) B) = \lambda R(A) + (1 - \lambda) R(B) = \lambda v(r(A)) + (1 - \lambda) v(r(B)) = v(\lambda r(A) + (1 - \lambda) r(B))$. Finally, let $U_{u,v,r} = U$. This proves the existence of the representation of $\succsim$.

It is straightforward to see that u represents $\succcurlyeq_1^*$. The representations of $\succcurlyeq^{\mathcal{I}}$ and $\succsim^Q$ follow directly from Proposition 5 and Proposition 7. *Q.E.D.*

By the theorem, it is straightforward to see that if the evaluator does not care about outcomes at all (i.e., when u is constant), her preference $\succsim$ is the second-order preference itself, which is formally stated as follows.

Corollary 2. *Suppose $\succsim$ admits a PC representation. Then $\succsim = \succsim^Q$ if and only if $(x, \{x\}) \sim (y, \{y\})$ for all $x, y \in X$.*

Analogous to the standard EU theory, the PC representation is unique up to positive affine transformations. In this setting, the only subtlety concerns the reference function. By definition, the function $r(\cdot)$ itself need not be unique because the set $\{x \in \mathrm{conv}(A) : (x, A \cup \{x\}) \sim (x, \{x\})\}$ may contain multiple elements. Nevertheless, any two selections from this set yield equivalent representations as long as the chosen references are ideally indifferent under any PC representation of $\succsim$.

Theorem 2 (Uniqueness)**.** *Suppose $\succsim$ admits the PC representation (u, v, r). Then, (u', v', r') represents $\succsim$ if and only if $u' = \alpha u + \beta_1$ and $v' = \alpha v + \beta_2$ for some $\alpha > 0$ and $\beta_1, \beta_2 \in \mathbb{R}$, and $v(r(A)) = v(r'(A))$ for each $A \in \mathbb{M}$.*

Proof. See Appendix. *Q.E.D.*

3.1. Measure of Attitude Toward Preferences

I define a simple comparative measure of attitude toward preferences in terms of the PC representations. Define for each preference $\succsim$ the most ideal and the least ideal options available on a menu A as $b_{\succsim}(A) \in \{x \in A : x \succcurlyeq^{\mathcal{I}} y \ \forall y \in A\}$ and $w_{\succsim}(A) \in \{x \in A : y \succcurlyeq^{\mathcal{I}} x \ \forall y \in A\}$. For any $\alpha \in [0, 1]$, let

$$\bar{x}_{\succsim}(\alpha) \coloneqq \alpha b_{\succsim}(A) + (1 - \alpha)\, w_{\succsim}(A)$$

denote a mixture between $b_{\succsim}(A)$ and $w_{\succsim}(A)$. I say menu A is *quality-conflicting* for the preference $\succsim$ if its ideal preference $\succcurlyeq^{\mathcal{I}}$ satisfies $x \succ^{\mathcal{I}} y$ for some $x, y \in A$.

Definition 8. The preference $\succsim$ has a *higher degree of evaluative strictness* than $\succsim'$ if for all $\alpha \in [0, 1]$ and every convex $A \in \mathbb{M}$ that is quality-conflicting for both $\succsim, \succsim'$,

$$\Big(\bar{x}_{\succsim'}(\alpha), \{\bar{x}_{\succsim'}(\alpha)\}\Big) \succsim' \Big(\bar{x}_{\succsim'}(\alpha), A\Big) \implies \Big(\bar{x}_{\succsim}(\alpha), \{\bar{x}_{\succsim}(\alpha)\}\Big) \succsim \Big(\bar{x}_{\succsim}(\alpha), A\Big).$$

Intuitively, whenever the less strict evaluator $\succsim'$ prefers to restrict a choice to its singleton rather than let it be chosen from a menu, the stricter evaluator $\succsim$ does so as well. For instance, let the mother's preference be $\succsim$ and the father's be $\succsim'$. Suppose they have been observing their child's daily choices from menu $A = \mathrm{conv}(\{y, z\})$, and the child has chosen y over z half the time: i.e., let $\bar{x}_{\succsim}(0.5) = \bar{x}_{\succsim'}(0.5) = \frac{1}{2}y + \frac{1}{2}z$. (Note that each parent's ideal preference could differ.) If the father (the less strict evaluator) would rather restrict the child's choice to the singleton $\{\frac{1}{2}y + \frac{1}{2}z\}$ than allow the same lottery to be chosen from A, then the mother would also prefer such a restriction.

In Theorem 3 below, I characterize the evaluative strictness in terms of the PC representations. For any function $v : X \to \mathbb{R}$ and reference function r, define the function $\alpha^*(A; v, r)$ by

$$v(r(A)) \coloneqq \alpha^*(A; v, r) \max_{x \in A} v(x) + (1 - \alpha^*(A; v, r)) \min_{x \in A} v(x)$$

where $\alpha^*(A; v, r) \in [0, 1]$ is well-defined whenever v is not constant on A. (Hence, I rule out singleton menus.) The parameter $\alpha^*(A; v, r)$ measures how close the evaluator's reference of A lies to the most ideal rather than the least ideal element of A in terms of v.

Theorem 3. *Suppose two preferences $\succsim, \succsim'$ admit the PC representations (u, v, r) and (u', v', r'), respectively. Then, $\succsim$ has a higher degree of evaluative strictness than $\succsim'$ if and only if $\alpha^*(A; v, r) \geq \alpha^*(A; v', r')$ for all convex $A \in \mathbb{M}$ that are quality-conflicting for both $\succsim, \succsim'$.*

Proof. See Appendix. *Q.E.D.*

When $\succsim$ and $\succsim'$ share the same ideal preference ($v = v'$), the comparison simplifies to $v(r(A)) \geq v(r'(A))$ for all $A \in \mathbb{M}$. Consider two extreme attitudes. If $\alpha^*(A; v, r) = 1$ for all A, then $v(r(A)) = \max_{x \in A} v(x)$: the evaluator judges each choice by how much it falls short of the most ideal outcome. This represents the *strictest* evaluative attitude. At the other extreme, if $\alpha^*(A; v, r) = 0$, then $v(r(A)) = \min_{x \in A} v(x)$: the evaluator judges by how much a choice exceeds the least ideal option, which is the *most lenient* attitude.

4. Special Case: Temptation and Quality of Choice

Theorem 1 accommodates perhaps the broadest range of attitudes within the EU framework, imposing minimal structure on the reference function, but does not explicitly specify how the reference is determined for each menu. In particular, the measure $\alpha^*(\cdot; v, r)$—the evaluator's attitude toward preferences—is not necessarily constant across menus. I adopt Gul and Pesendorfer (2001)'s model of temptation and self-control, and present a special case in which the evaluator prefers preferences that induce *hard choices*—i.e., those made by resisting strong temptations. I impose the following axiom, which resembles the *set betweenness* axiom of Gul and Pesendorfer (2001).[28]

Axiom 5 (Choice Betweenness)**.** *$(x, A) \succsim (x, B)$ implies $(x, A) \succsim (x, A \cup B) \succsim (x, B)$.*

Choice Betweenness (CB) states that, conditional on the same chosen option x, if giving up A is preferred to giving up B, then giving up both A and B is at least as bad as giving up A and at least as good as giving up B.

To interpret this axiom in more detail, I introduce definitions similar to Gul and Pesendorfer (2001)'s notions of temptation and self-control, but interpret them as the evaluator's beliefs.[29]

Definition 9. The preference $\succsim$ "believes"

(a) *menu A tempts the DM from menu B* if $(x, A \cup B) \succ (x, B)$;

(b) *the DM exerts self-control* if $(x, A) \succ (x, A \cup B) \succ (x, B)$.

In case (a), if the evaluator believes that the options in A are relatively more tempting than those in B, then $(x, A \cup B)$ is perceived as more difficult than (x, B). In case (b), if she believes that A tempts the DM away from B but expects him to exert self-control when facing $A \cup B$, she expects him to make a better choice from $A \cup B$ than from A. Thus, (x, A) more closely meets her expectation than $(x, A \cup B)$. If $(x, A \cup B) \sim (x, B)$, she believes that A does not tempt the DM away from B; possibly, B offers an obviously dominant choice that does not require self-control. Therefore, when B is removed, choosing x becomes less obvious: $(x, A) \succ (x, A \cup B) \sim (x, B)$.

[28] If $\alpha^*(\cdot; v, r)$ is constant, then the value of the reference $v(r(A))$ takes the form of the α-maxmin utility function of sets of lotteries presented by Olszewski (2007) in his characterization of attitudes toward ambiguity. He developed a menu choice framework, with Nature ultimately choosing a lottery from the chosen menu. My model relates to this environment only by featuring the evaluator as an entity that does not make the lottery choice. The α-maxmin utility function, also known as the "α-MEU" decision rule, can be traced back to the Hurwicz criterion (Hurwicz, 1951; Arrow and Hurwicz, 1972; Gilboa and Schmeidler, 1989).

[29] In Gul and Pesendorfer (2001)'s setting, temptation and self-control are identified as follows: (i) y is more tempting than x if the agent strictly prefers committing to x (i.e., $\{x\}$ is strictly preferred to $\{x, y\}$); and (ii) the menu $\{x, y\}$ requires self-control if the agent strictly prefers having the tempting option on his menu to exogenously consuming it (i.e., $\{x, y\}$ is strictly preferred to $\{y\}$).

Definition 10. A representation of a *temptation-adjusted preference over choices* (TPC) is a tuple (u, v, τ) of affine functions of lotteries such that $\succsim$ is represented by

$$U_{u,v,\tau}(x, A) \coloneqq u(x) + v(x) - \left[\max_{y \in A} \Big\{ v(y) + \tau(y) \Big\} - \max_{z \in A} \tau(z)\right].$$

The function $Q_{v,\tau}(x, A) \coloneqq U_{u,v,\tau}(x, A) - u(x)$ represents the second-order preference $\succsim^Q$, for which the reference value function $v(r(A)) = \max_{y \in A}\{v(y) + \tau(y)\} - \max_{z \in A} \tau(z)$ takes the form of Gul and Pesendorfer (2001)'s representation of preference for commitment.[30] The interpretation differs slightly. The functions τ and $v + \tau$ capture the evaluator's expectations about the DM's temptation and his preference, respectively. That is, $y^*_A \in \arg\max_{y \in A} v(y) + \tau(y)$ is a choice that she expects the DM to make.

The function $Q_{v,\tau}(x, A)$ can be rewritten as

$$Q_{v,\tau}(x, A) = \underbrace{\Big[v(x) - v(y^*_A)\Big]}_{\text{Expectation gap}} + \underbrace{\Big[\max_{z \in A} \tau(z) - \tau(y^*_A)\Big]}_{\text{Anticipated mental cost}}.$$

The value difference $v(x) - v(y^*_A)$ between the DM's actual choice x and the expected choice y^*_A is called *the expectation gap*; the larger this gap, the more preferable the choice (x, A) becomes. I call the term $\max_{z \in A} \tau(z) - \tau(y^*_A)$ the *anticipated mental cost*, which measures how far the expected choice deviates from the menu's strongest temptation; the greater this deviation, the higher the mental cost the evaluator expects the DM to incur when he faces this menu.[31]

Let $R_{v,\tau}(A) \coloneqq \max_{y \in A}\{v(y) + \tau(y)\} - \max_{z \in A} \tau(z)$. A few brief remarks on this function follow. The TPC representation shows that the quality of (x, A) can be positive (i.e., $Q_{v,\tau}(x, A) > 0$) in two ways. First, the DM can make a choice that exceeds the evaluator's expectation. Second, even when the expectation gap is zero, the quality can still be positive for menus that entail a high anticipated mental cost. This is possible whenever the expected choice y^*_A is neither the most ideal nor the most tempting option.[32] By contrast, if the evaluator does not expect any mental cost

[30] More precisely, the pair (v, r^τ) represents $\succsim^Q$ where the choice rule r^τ is defined by $r^\tau(A) \in \{x \in \mathrm{conv}(A) : v(x) = \max_{y \in A}\{v(y) + \tau(y)\} - \max_{z \in A} \tau(z)\}$.

[31] Note that if the expected choice y^*_A is not unique, the expectation gap and the mental cost are not unique, even though their sum $Q_{v,\tau}(x, A)$ remains uniquely identified.

[32] To see this, note that $y^*_A \notin (\arg\max_{y \in A} v(y)) \cup (\arg\max_{y \in A} \tau(y))$ implies $R_{v,\tau}(A) < v(y^*_A)$, which yields $Q_{v,\tau}(y^*_A, A) > 0$.

(because she believes the DM will pick the most tempting option), then a choice that meets that expectation produces choice quality equal to that of vacuous choices.[33] Moreover, the evaluator has the highest degree of evaluative strictness whenever she expects the DM to choose the most ideal option, which happens to be the most tempting one. Suppose $v = \tau$. Then, it follows that $R_{v,\tau}(A) = \max_{y \in A} v(y)$. Conversely, she has the lowest degree of strictness whenever she expects the DM to choose the most tempting option, which happens to be the least ideal option: e.g., when $v = -\tau$, we have $R_{v,\tau}(A) = \min_{y \in A} v(y)$. Otherwise, she adopts a non-extreme attitude.

The theorem below states that CB yields the TPC representation.

Theorem 4. *Suppose $\succsim$ admits a PC representation. Then $\succsim$ satisfies CB if and only if $\succsim$ has a TPC representation.*

Proof. The "if" part is straightforward. The proof of the "only if" part is an indirect application of Gul and Pesendorfer (2001)'s Theorem 1. First, define a binary relation $\succeq_r$ on $\mathbb{M}$ as $A \succeq_r B$ if and only if $r(A) \succcurlyeq^{\mathcal{I}} r(B)$. Because $\succsim$ admits a PC representation (u, v, r), the function $R^*(A) \coloneqq v(r(A))$ represents $\succeq_r$. Since $r(\cdot)$ is affine with respect to v, $R^* : \mathbb{M} \to \mathbb{R}$ is an affine function of sets. Furthermore, because $R^*(A) = u(x) + v(x) - U(x, A)$ and the functions U, u, and v are continuous, R^* is continuous on $\mathbb{M}$ under the Hausdorff metric.

Next, I show that $A \succeq_r B$ implies $A \succeq_r A \cup B \succeq_r B$, which is the *set betweenness* axiom of Gul and Pesendorfer (2001). By Corollary 1, CB implies that, for any $A, B \in \mathbb{M}$ satisfying $A \cap B \neq \varnothing$, $A \succeq_r B$ implies $A \succeq_r A \cup B \succeq_r B$. It remains to extend this implication to two disjoint menus. Suppose $A \succeq_r B$ for any A, B, and define

$$A' \coloneqq \tfrac{1}{2}A + \tfrac{1}{2}X \quad \text{and} \quad B' \coloneqq \tfrac{1}{2}B + \tfrac{1}{2}X.$$

These menus are not disjoint: for any $x \in A$ and $y \in B$, the lottery $\frac{1}{2}x + \frac{1}{2}y$ belongs to both A' and B'. Since R^* is affine and represents $\succeq_r$, $A \succeq_r B$ implies $R^*(A') = \frac{1}{2}R^*(A) + \frac{1}{2}R^*(X) \geq \frac{1}{2}R^*(B) + \frac{1}{2}R^*(X) = R^*(B')$, and hence $A' \succeq_r B'$. By CB, we have $A' \succeq_r A' \cup B' \succeq_r B'$.

Note that $A' \cup B' = \frac{1}{2}(A \cup B) + \frac{1}{2}X$.[34] Therefore, $A' \succeq_r A' \cup B' \succeq_r B'$ implies

$$\tfrac{1}{2}R^*(A) + \tfrac{1}{2}R^*(X) \geq \tfrac{1}{2}R^*(A \cup B) + \tfrac{1}{2}R^*(X) \geq \tfrac{1}{2}R^*(B) + \tfrac{1}{2}R^*(X),$$

[33] Formally, $y_A^* \in \arg\max_{z \in A} \tau(z)$ implies $R_{v,\tau}(A) = v(y_A^*)$.

[34] Observe that $\frac{1}{2}(A \cup B) + \frac{1}{2}X = \{\frac{1}{2}x + \frac{1}{2}z : x \in A \cup B,\ z \in X\} = \{\frac{1}{2}x + \frac{1}{2}z : x \in A,\ z \in X\} \cup \{\frac{1}{2}y + \frac{1}{2}z : y \in B,\ z \in X\} = (\frac{1}{2}A + \frac{1}{2}X) \cup (\frac{1}{2}B + \frac{1}{2}X) = A' \cup B'$.

which yields $R^*(A) \geq R^*(A \cup B) \geq R^*(B)$. Hence $A \succeq_r B$ implies $A \succeq_r A \cup B \succeq_r B$ for arbitrary $A, B \in \mathbb{M}$. Thus, $\succeq_r$ satisfies Gul and Pesendorfer (2001)'s axiom. By their Theorem 1, there exists a continuous affine function τ of lotteries such that $R^*(A) = \max_{y \in A} v(y) + \tau(y) - \max_{z \in A} \tau(z)$, which completes the proof. *Q.E.D.*

4.1. The Best Choice vs. The Best Outcome

The novelty of the TPC representation is that the most preferable choice cannot be made when the menu offers the most preferable outcome—the one that yields both the highest outcome value and the most ideal value. For instance, suppose $\succsim$ satisfies $(x, \{x, y, z\}) \succ^Q (y, \{x, y, z\}) \succ^Q (z, \{x, y, z\})$ and $(x, \{x\}) \succ (y, \{y\}) \succ (z, \{z\})$ so that x is the best outcome among the three alternatives. Consider the following ranking:

$$(y, \{y, z\}) \succ (x, \{x, y, z\}) \succ (y, \{x, y, z\}) \succ (z, \{x, y, z\}). \tag{5}$$

Clearly, the best choice is $(y, \{y, z\})$, which is made when x is unavailable. Given a PC representation (u, v, r), the best outcome x satisfies $u(x) \geq \max\{u(y), u(z)\}$ and $v(x) \geq \max\{v(y), v(z)\}$. The ranking (5) implies that $u(y) + v(y) > u(z) + v(z)$, and that $r(\cdot)$ satisfies

$$u(x) - u(y) < \underbrace{[v(y) - v(r(\{y, z\}))]}_{Q_{v,r}(y,\{y,z\})} - \underbrace{[v(x) - v(r(\{x, y, z\}))]}_{Q_{v,r}(x,\{x,y,z\})}. \tag{6}$$

This means that the outcome value difference between x and y is smaller than the quality difference between $(y, \{y, z\})$ and $(x, \{x, y, z\})$.[35]

If the evaluator's degree of evaluative strictness is constant—i.e., $\alpha^*(A; v, r) = \alpha^*$ for some $\alpha^* \in [0, 1]$ for all A—then (5) cannot be rationalized.[36] However, the TPC representation (u, v, τ) can rationalize it. Suppose $\tau(x) > \tau(z) > \tau(y)$. Then, we have $Q_{v,r}(x, \{x, y, z\}) = 0$. In contrast, $Q_{v,r}(y, \{y, z\})$ is either $v(y) - v(z)$ or $\tau(z) - \tau(y)$, both of which are strictly positive. If this is larger than $u(x) - u(y)$, then (5) occurs. Intuitively, The more z tempts the DM relative to y, the closer the reference value $v(r(\{y, z\}))$ moves to $v(z)$; by comparison, $(x, \{x, y, z\})$ is an obvious decision since

[35] (6) is derived directly from the inequality $U_{u,v,r}(y, \{y, z\}) > U_{u,v,r}(x, \{x, y, z\})$.

[36] Suppose there is $\alpha^* \in [0, 1]$ such that $v(r(A)) = \alpha^* \max_{y \in A} v(y) + (1 - \alpha^*) \min_{y \in A} v(y)$ for all A. Then, (6) implies $u(x) - u(y) < (1 - \alpha^*)(v(y) - v(x))$ which contradicts the assumption that x maximizes both u and v.

x is both tempting and ideal.

In the menu preference framework, the ranking (5) implies that the menu $\{y, z\}$ is preferred to $\{x, y, z\}$ even though x is preferred to y and z. The standard models of temptation using menu preferences in the literature cannot rationalize the tendency to remove an option from a menu that the DM would otherwise have chosen.[37]

5. Applications

This section provides two examples illustrating how the TPC representation can be used to model menu choices, rather than providing a general analysis. In the first part, I illustrate how the DM's awareness of the evaluator's preference over his choices can guide his menu choices. I consider a simple setting in which the evaluator's preference affects the DM's ultimate payoff, and thus the DM strategically chooses his menu to either reveal or conceal aspects of his preference that the evaluator likes or dislikes. In the second part, I illustrate the evaluator's design of the DM's menus based on the DM's choices, using a parenting example.

5.1. The DM's Menu Choice: A Dictator Example

Consider two dictators, each choosing an allocation of wealth. In period 1, Dictator 1 (the DM) publicly chooses an allocation $x = (x_1, x_2) \in \mathbb{R}^2$ from menu A of allocations between himself (who gets x_1) and Dictator 2 (the evaluator who gets x_2). In period 2, Dictator 2 chooses an allocation $(t, 0)$ where $t \in \mathbb{R}$: she can either give to or take from Dictator 1, with no effect on her own payoff. Following the dictator game context adopted by Dillenberger and Sadowski (2012) and Saito (2015), I assume there is an ex ante stage of which Dictator 2 is unaware: in period 0, Dictator 1 is allowed to privately choose his menu A from an exogenously given set $\mathcal{A}$ that contains menus of allocations (see Table 1). Hence, Dictator 2 perceives Dictator 1's menu A as exogenous. I also assume that Dictator 1 is fully aware of Dictator 2's preference.

[37] The standard models that assume deterministic contexts can only rationalize (5) by identifying x as a temptation that is normatively inferior—e.g., against the agent's long-term goal (see Gul and Pesendorfer, 2001). However, in my example, x is both the most tempting and ideal option. Furthermore, the tendency to remove a normatively superior option from a menu has been rationalized by the agent's motivation to avoid a sense of guilt that stems from not choosing it (see Kopylov, 2012). This can only make sense here if either y or z would be chosen over x, thereby validating the anticipated guilt. Yet, clearly, x is the best outcome to choose.

Period 0:	Dictator 1 privately chooses a menu $A \in \mathcal{A}$ of allocations.
Period 1:	Dictator 1 publicly chooses an allocation $x = (x_1, x_2) \in A$.
Period 2:	Dictator 2 chooses $t \in \mathbb{R}$, and the allocation $(x_1 + t, x_2)$ is distributed.

Table 1: Timeline of the dictator game

Let m_i be Dictator i's terminal payoff function, where $i \in \{1, 2\}$. Then Table 1 implies $m_1(x, t) = x_1 + t$ and $m_2(x, t) = x_2$. Consider three allocations: a fair allocation $x_f = (5, 5)$, a selfish allocation $x_s = (6, 4)$, and a Pareto-optimal (PO) allocation $x_p^w = (6 + w, 6 + w)$ with an increment $w > 0$. Suppose $\mathcal{A} = \{A_0, B^w\}$ where $A_0 = \{x_f, x_s\}$ and $B^w = \{x_f, x_s, x_p^w\}$. Under standard selfish preferences, Dictator 1 would choose B^w and x_p^w for any $w > 0$, and Dictator 2 would choose any $t \in \mathbb{R}$ because t does not affect her payoff.

Alternatively, suppose Dictator 1's payoff function is as follows:

$$U_{\delta,k}(x, A) \coloneqq x_1 + \underbrace{kx_2 - \left[\max_{(y_1,y_2)\in A} \left\{ \delta y_1 + ky_2 \right\} - \max_{(z_1,z_2)\in A} \delta z_1 \right]}_{\coloneqq Q_{\delta,k}(x,A)}$$

where $\delta, k \geq 0$. The interpretation is that Dictator 1 knows that Dictator 2 has a preference over *Dictator 1's* choice in period 1, which affects the choice of t. More specifically, in response to Dictator 1's choice of (x, A), assume Dictator 2 chooses $t = Q_{\delta,k}(x, A)$, where $Q_{\delta,k}(x, A)$ measures the quality of the choice, evaluated by Dictator 2.[38] That is, the more favorably she evaluates this choice, the more she wishes to reciprocate in period 2.[39] Then in the subgame perfect Nash equilibrium, Dictator 1's payoff depends only on his own choice: $m_1(x, Q_{\delta,k}(x, A)) = U_{\delta,k}(x, A)$.

Notice that the function $U_{\delta,k}$ is a TPC representation (u, v, τ) with a slightly modified interpretation: Dictator 1's outcome preference is $u(x) = x_1$; Dictator 2's "ideal dictator's preference" is $v(x) = kx_2$, reflecting altruism with a constant weight k; Dictator 2's expectation about Dictator 1's temptation ranking is $\tau(x) = \delta u(x)$ with a constant weight δ; and the ranking $(\tau+v)(y) = \delta y_1+ky_2$ reflects her expectation

[38] Hence, Dictator 2 is a non-strategic player who responds only to the quality of Dictator 1's choice by choosing $t = Q_{\delta,k}(x, A)$ without forming any beliefs (e.g., about Dictator 1's true character).

[39] For instance, let Dictator 2's utility function be $U_2(x, t; A) \coloneqq m_2(x, t) + Q_{\delta,k}(x, A)t - \frac{1}{2}t^2$ where the term $Q_{\delta,k}(x, A)t - \frac{1}{2}t^2$ reflects her utility from giving t to Dictator 1 in response to his choice (x, A), which exhibits diminishing marginal utility. Then the maximizer is $t^* = Q_{\delta,k}(x, A)$. The assumption that the choice t is exactly equal to $Q_{\delta,k}(x, A)$ is therefore specific to this example.

about Dictator 1's actual behavior. Then Dictator 1's payoff $U_{\delta,k}$ over the three allocations can be summarized as follows:

	Outcome (u)	Ideal (v)	Temptation (τ)
Allocation $x_f = (5,5)$	5	$5k$	5δ
Allocation $x_s = (6,4)$	6	$4k$	6δ
Allocation $x_p^w = (6+w, 6+w)$	$6+w$	$(6+w)k$	$(6+w)\delta$

where $\frac{k}{\delta}$ captures how much Dictator 2 values Dictator 1's altruistic preference, relative to how much she believes Dictator 1 is tempted by wealth.

The following results hold.

Proposition 8. *Define Dictator 1's menu preference in period 0, denoted $\succeq_M$, by $A \succeq_M B$ if and only if $\max_{x\in A} U_{\delta,k}(x,A) \geq \max_{y\in B} U_{\delta,k}(y,B)$. Then,*

(a) $U_{\delta,k}(x_s,A_0) > U_{\delta,k}(x_f,A_0)$ implies $B^w \succ_M A_0$ for all $w > 0$;

(b) $U_{\delta,k}(x_s,A_0) \leq U_{\delta,k}(x_f,A_0)$ implies $A_0 \succ_M B^w$ whenever $\min\{\delta,k\} > 1+w$.

First, notice that Dictator 1's payoff from (x_p^w, B^w) is $U_{\delta,k}(x_p^w, B^w) = m_1(x_p^w, 0) = 6+w$ since $Q_{\delta,k}(x_p^w, B^w) = k(6+w) - (\delta+k)(6+w) + \delta(6+w) = 0$. That is, Dictator 2 does not reward Dictator 1 for choosing the PO allocation. Intuitively, she believes that any rational person would obviously choose x_p^w from B^w, and thus the choice (x_p^w, B^w) itself does not impress her. Indeed, giving up Pareto-inferior allocations entails no sacrifice: he gives up neither being altruistic nor being selfish.

Part (a) of Proposition 8 states that if the reward for the fair allocation is insufficient to offset the forgone monetary gain from the PO allocation, then Dictator 1 strictly prefers B^w to A_0 for all $w > 0$. This follows from $m_1(x_s,0) = 6 \geq U_{\delta,k}(x_s,A_0)$, which holds because $Q_{\delta,k}(x_s,A_0)$ is at most zero: x_s is the least ideal option in A_0, and thus Dictator 2 always evaluates this choice unfavorably. On the other hand, for any $w > 0$, the payoff from the PO allocation exceeds 6.

Part (b) of Proposition 8 shows that if "being nice" yields a sufficiently high reward, Dictator 1 may prefer A_0 to B^w even though A_0 excludes the PO allocation. In particular, $U_{\delta,k}(x_s,A_0) \leq U_{\delta,k}(x_f,A_0)$ is equivalent to $u(x_f)+v(x_f) \geq u(x_s)+v(x_s)$, which holds whenever $k \geq 1$. Then Dictator 1 chooses A_0 over B^w whenever the reward from (x_f, A_0) exceeds the monetary sacrifice from excluding x_p^w. Since

$$U_{\delta,k}(x_f,A_0) = 5 + 5k - \max\{5(\delta+k), 6\delta+4k\} + 6\delta = 5 + \min\{k,\delta\},$$

where $\min\{\delta, k\} = Q_{\delta,k}(x_f, A_0)$, Proposition 8(b) follows. This illustrates how the value of the most preferable choice can outweigh that of the most preferable outcome.

Turning to Dictator 2's evaluation, notice that her reference value function is $R(A) = \max_{y \in A}\{\delta y_1 + k y_2\} - \max_{z \in A} \delta z_1$. Then the reference value of menu A_0 is $R(A_0) = \max\{5k - \delta, 4k\}$. Hence, if $\delta \geq k$, then $R(A_0) = \min_{x \in A_0} v(x) = 4k$, meaning that Dictator 2's evaluative strictness is minimal: she expects no self-control from Dictator 1 and thus views (x_f, A_0) as maximally praiseworthy, which is reflected in how much she rewards it in period 2.

The model most closely related to this application is Saito (2015), which considers a similar framework but without period 2, and characterizes Dictator 1's menu preference in period 0. He captures what he refers to as *impure altruism*, which is exhibited when an "intrinsically selfish" dictator behaves altruistically in order to feel pride and to avoid the shame of acting selfishly. However, even in his model, menu A_0 can never be preferred to B^w.[40] Some experimental studies recognize that opportunities are vital information when assessing others' selfishness (List, 2007; Bardsley, 2008) or the intentions behind their actions (Falk et al., 2003, 2008). The theory of reciprocity by Falk and Fischbacher (2006) has a richer game-theoretic framework that could capture the dynamic interactions discussed in this example. However, they exploit an arbitrary belief-based payoff structure and do not articulate how different menus can lead to differences in the quality of intentions.

5.2. The Evaluator's Menu Choice: A Parenting Example

Let $A^* = \text{conv}(\{h, t\})$, where h denotes doing homework and t watching television. Consider parents (the evaluator) who not only want their child to do homework rather than watch television, but also want him to willingly make the right choice (i.e., they want their child to strictly prefer h to t). Formally, their preference $\succsim$ over the child's choices satisfies $(h, \{h, t\}) \succsim (h, \{h\}) \succ (t, \{t\}) \succsim (t, \{h, t\})$.

[40] In Saito (2015)'s model, if the parameter for the cost of shame is removed and the maximal parameter for the sense of pride is imposed, then the dictator's menu choice is represented by the function U_S of the form: $U_S(A) := \max_{x \in A} \alpha W(x_1) + W(x_2) + \alpha\left[\max_{y \in A} W(y_1) - W(x_1)\right]$ for some $\alpha > 0$ where $W : \mathbb{R} \to \mathbb{R}$ is a ranking of wealth outcomes. Here, the ex post choice of allocation is the most altruistic allocation—i.e., $\arg\max_{x \in A} W(x_2)$. The sense of pride is captured by the term $\max_{y \in A} W(y_1) - W(x_1)$ which is the dictator's maximal wealth outcome available on the menu A minus his chosen wealth outcome $W(x_1)$. When the menu is $A_0 = \{x_f, x_s\}$, we have $U_S(A_0) = W(5) + \alpha W(6)$. When the menu is B^w, we have $U_S(B^w) = W(6 + w) + \alpha W(6 + w)$. Assuming that W is an increasing function, we have $U_S(B^w) \geq U_S(A_0)$.

Consider the following representation U_δ of $\succsim$: for $A \subseteq A^*$,

$$U_\delta(x, A) := \delta v(x) + \underbrace{v(x) - v(r(A))}_{:=Q_\delta(x,A)} \tag{7}$$

where $v(h) > v(t)$, and $\delta > 0$ is the relative weight on the child's outcome. Since v is a vNM utility function, we can assume $v(h) = 1$ and $v(t) = 0$, without loss of generality.

The parents' degree of evaluative strictness, captured by their reference $r(A^*)$, can be identified as follows. Suppose they observed that the child chose homework a fraction $\alpha \in [0, 1]$ of the time and ended up choosing television $1-\alpha$ of the time. That is, his choice was $(\alpha h + (1 - \alpha)\, t, A^*)$. For each $\alpha \in [0, 1]$, suppose they can decide whether to keep allowing the child to choose from A^* or to enforce homework time with probability α and television time with probability $1 - \alpha$. Formally, they are ranking the following two choices: $(\alpha h + (1 - \alpha)\, t, A^*)$ and $(\alpha h + (1 - \alpha)\, t, \{\alpha h + (1 - \alpha)\, t\})$ for each α. Then, by Relativity, there is an $\alpha^* \in [0, 1]$ such that

$$(\alpha^* h + (1 - \alpha^*)\, t, A^*) \sim (\alpha^* h + (1 - \alpha^*)\, t, \{\alpha^* h + (1 - \alpha^*)\, t\}),$$

which implies $r(A^*) = \alpha^* h + (1 - \alpha^*)\, t$.

Next, the parents' relative weight δ on the child's outcome can be identified by considering two cases. First, when $0 \leq \alpha^* < 1$, suppose the parents can now enforce homework time with certainty. At what threshold $\tilde{\alpha} \in [0, 1]$ would the parents switch from enforcing homework (by banning television) to allowing him to keep deciding freely? Formally, they are ranking the two choices: $(\alpha h + (1 - \alpha)\, t, A^*)$ and $(h, \{h\})$ for each α. Then there is $\tilde{\alpha} \in [\alpha^*, 1)$ such that $(\tilde{\alpha} h + (1 - \tilde{\alpha})\, t, A^*) \sim (h, \{h\})$. It is easy to show that $U_\delta(\tilde{\alpha} h + (1 - \tilde{\alpha})\, t, A^*) = U_\delta(h, \{h\})$ implies

$$\delta = \frac{\tilde{\alpha} - \alpha^*}{1 - \tilde{\alpha}}.$$

The threshold is $\tilde{\alpha} = \frac{\alpha^* + \delta}{1 + \delta}$, which increases in both δ and α^*. This implies that the parents become more reluctant to allow the child to choose freely either because the child's outcome is more important than inducing the behavior they prefer, or because they have a higher degree of evaluative strictness toward his preferences. In the other case, when $\alpha^* = 1$, we can similarly identify δ by finding $\tilde{\alpha}$ such that

$(\tilde{\alpha}h + (1-\tilde{\alpha})\,t, A^*) \sim (t, \{t\})$, in which case $\delta = \frac{1-\tilde{\alpha}}{\tilde{\alpha}}$.

I now introduce the child's temptation: let t' denote the child's all-time favorite television show. Suppose the parents do not differentiate between the outcomes of t and t', nor do they think that either should be ideally preferred to the other. That is, $v(t) = v(t')$. In this case, if the parents have a constant degree of evaluative strictness, the same threshold $\tilde{\alpha}$ should apply to the child facing menu $A' = \text{conv}(\{h, t'\})$. Suppose the threshold $\tilde{\alpha}'$ for A' is smaller than $\tilde{\alpha}$. That is, $\tilde{\alpha}'$ satisfies

$$(\tilde{\alpha}'h + (1-\tilde{\alpha}')\,t', A') \sim (h, \{h\}) \quad \text{and} \quad \tilde{\alpha} > \tilde{\alpha}'. \tag{8}$$

According to Theorem 4, (8) is a clear sign that the parents believe t' tempts their child more than t. Intuitively, because they recognize that the child faced a stronger temptation, they believe that his decision to give up t' required more self-control than the decision to give up t did.

6. Discussion

6.1. Elicitation Issue

There is an obvious elicitation issue. The second-order preference cannot be identified from ordinary choice data; identification requires sufficiently rich comparisons among elements of $\mathbb{C}$. Although choices over choices may be unnatural to incentivize in laboratory settings, they are not inherently unobservable: many real-world decisions involve evaluating the choices made by others, including court rulings, dating, employee assessments, parenting, and voting. These settings, however, may involve beliefs, strategic interactions, equilibrium effects, and dynamic considerations that limit the model's direct application. One possible direction is to consider multiple DMs facing different menus and allow the evaluator to choose which DM to reward based on their choices. Even in this setting, constructing an incentive-compatible design with sufficient support over choice objects remains a practical challenge.[41] The contribution of the present analysis is instead to characterize the structure of second-order preference and what can be identified once sufficiently rich comparisons among

[41] The paper also does not claim to identify an individual's second-order preference over her own choices, even with sufficient data on choices over choices—a challenge analogous to that discussed by Bernheim et al. (2024).

choices are available.

6.2. Extension

I discuss two extensions. First, one could construct a stochastic choice environment that records what the DM would choose from every binary menu. Because the resulting contingency plan matches "behaving according to" an entire preference ranking over all alternatives, evaluating such plans would connect the present choice-based framework more directly to preferences over preference relations. More generally, tailoring the choice environment to reveal a particular behavioral trait could generate preferences over risk attitudes or other classes of preferences. Second, the choice domain could be enlarged to allow the DM to declare indifference among a set of options rather than select a single outcome. This distinction matters because the same outcome may be selected under either strict preference or indifference, while the declaration of indifference may itself be evaluated as open-mindedness or instead as indecisiveness or a lack of commitment. Extending the choice space would allow the model to distinguish among evaluators according to how they value these features.[42]

7. Conclusion

This paper develops a model in which choices are evaluated not only by their outcomes but also by how well they align with ideal preferences. Within the standard EU paradigm, I define preference over preferences—a concept long confined to philosophical debate—through a revealed-preference argument, and show that it can either be identified through choices or interpreted as a measure of one's welfare associated with choices.

[42] Some preliminary results for both extensions are available upon request. For the first extension, I provide a representation of preference over complete binary relations and illustrate how the same logic generates a preference over risk attitudes. For the second, I reproduce the baseline model in the extended choice space by characterizing a *no preference for indifference* condition and discuss its limitations in the extended framework.

Appendix

A. Proof of Lemma 1

For each $n \in \mathbb{N}$, define $A_n := \sum_{s=1}^{n} \lambda_s A$ where $\lambda_s = \frac{1}{n}$ for $s = 1, \ldots, n$. Fix $\alpha \in (0, 1/|Z|]$ where $|Z|$ is the number of alternatives in Z, and choose any $A \in \mathbb{M}$. I first prove the following claims.

Claim 1. *$\alpha A + (1-\alpha)\text{conv}(A) = \text{conv}(A)$ and $\alpha A_n + (1-\alpha)\text{conv}(A) = \text{conv}(A)$.*

Proof. For the first equality, the inclusion $\alpha A + (1-\alpha)\text{conv}(A) \subseteq \text{conv}(A)$ follows from the convexity of $\text{conv}(A)$. For the reverse inclusion, take any $z \in \text{conv}(A)$. Since X is the lottery simplex over the finite set Z, Carathéodory's theorem (Rockafellar, 1970) implies that there exist $x_1, \ldots, x_m \in A$ and $\lambda_1, \ldots, \lambda_m \geq 0$, with $m \leq |Z|$ and $\sum_{i=1}^{m} \lambda_i = 1$, such that $z = \sum_{i=1}^{m} \lambda_i x_i$. Since $m \leq |Z|$, there is some j such that $\lambda_j \geq \frac{1}{|Z|}$. Since $\alpha \leq 1/|Z|$, we have $\lambda_j \geq \alpha$, which gives $z = \alpha x_j + (\lambda_j - \alpha)x_j + \sum_{i \neq j} \lambda_i x_i$ and thus $z = \alpha x_j + (1-\alpha)\left[\frac{\lambda_j - \alpha}{1-\alpha} x_j + \sum_{i \neq j} \frac{\lambda_i}{1-\alpha} x_i\right]$ where the expression in brackets belongs to $\text{conv}(A)$, which implies $z \in \alpha A + (1-\alpha)\text{conv}(A)$. Thus, $\alpha A + (1-\alpha)\text{conv}(A) = \text{conv}(A)$. For the second equality, clearly, $A \subseteq A_n \subseteq \text{conv}(A)$. Therefore, $\alpha A + (1-\alpha)\text{conv}(A) \subseteq \alpha A_n + (1-\alpha)\text{conv}(A) \subseteq \alpha\text{conv}(A) + (1-\alpha)\text{conv}(A)$. Since $\alpha A + (1-\alpha)\text{conv}(A) = \text{conv}(A)$ and, by convexity, $\alpha\text{conv}(A) + (1-\alpha)\text{conv}(A) = \text{conv}(A)$, it follows that $\alpha A_n + (1-\alpha)\text{conv}(A) = \text{conv}(A)$. *Q.E.D.*

Claim 2. *$(x, A) \sim (x, A_n)$ for all n.*

Proof. Suppose, toward a contradiction, that $(x, A) \succ (x, A_n)$. By Axiom 2, $\alpha(x, A) + (1-\alpha)(x, \text{conv}(A)) \succ \alpha(x, A_n) + (1-\alpha)(x, \text{conv}(A))$. By Claim 1, both sides are equal to $(x, \text{conv}(A))$, which is a contradiction. The same argument rules out $(x, A_n) \succ (x, A)$. Hence, by Axiom 1, $(x, A) \sim (x, A_n)$. *Q.E.D.*

Combining Claim 2 and the following result with Axiom 3, we have $(x, A) \sim (x, \text{conv}(A))$.

Lemma 2 (Shapley-Folkman Theorem)**.** *For any $A \in \mathbb{M}$ and $n \in \mathbb{N}$, $A_n \to \text{conv}(A)$ in the Hausdorff metric as $n \to \infty$.*

Proof. See Emerson and Greenleaf (1969); Starr (1969). *Q.E.D.*

Finally, pick any $B \subseteq \text{conv}(A)$. To show that $(x, A) \sim (x, A \cup B)$, notice that the above results imply $(x, A \cup B) \sim (x, \text{conv}(A \cup B))$. Since $\text{conv}(A \cup B) = \text{conv}(A)$, the result follows by transitivity.

B. Proof: $\mathbb{C}^c \coloneqq \{(x, A) \in \mathbb{C} : A \text{ is convex}\}$ is a mixture space

First, the operation is closed on $\mathbb{C}^c$: if A and B are nonempty compact convex sets, then $\alpha A + (1-\alpha)B$ is also nonempty, compact, and convex, and $\alpha x + (1-\alpha)y \in \alpha A + (1-\alpha)B$. *Sure mix* and *commutativity* are immediate. It remains to verify one-sided distributivity. For any $\alpha, \beta \in [0, 1]$, the outcome coordinates of $\alpha[\beta(x, A) + (1-\beta)(y, B)] + (1-\alpha)(y, B)$ and $\alpha\beta(x, A) + (1-\alpha\beta)(y, B)$ are both equal to $\alpha\beta x + (1-\alpha\beta)y$. For the menu coordinates, take any $x \in A$ and $y_1, y_2 \in B$. Then $\alpha[\beta x + (1-\beta)y_1] + (1-\alpha)y_2 = \alpha\beta x + (1-\alpha\beta)y$, where $y \coloneqq \frac{\alpha(1-\beta)}{1-\alpha\beta}y_1 + \frac{1-\alpha}{1-\alpha\beta}y_2$ whenever $\alpha\beta < 1$. The two weights defining y are nonnegative and sum to one, so $y \in B$ by convexity. Hence $\alpha[\beta A + (1-\beta)B] + (1-\alpha)B \subseteq \alpha\beta A + (1-\alpha\beta)B$. Conversely, for any $x \in A$ and $y \in B$, setting $y_1 = y_2 = y$ gives $\alpha[\beta x + (1-\beta)y] + (1-\alpha)y = \alpha\beta x + (1-\alpha\beta)y$, so the reverse inclusion also holds. The case $\alpha\beta = 1$ is immediate. Therefore, $\alpha[\beta A + (1-\beta)B] + (1-\alpha)B = \alpha\beta A + (1-\alpha\beta)B$.

C. Proof of Theorem 2

The "if" part is straightforward. For the "only if" part, suppose both (u, v, r) and (u', v', r') are PC representations of $\succsim$. By standard EU theory, it is easy to see that the functions u' and v' are affine transformations of u and v, and thus represent $\succcurlyeq_1^*$ and $\succcurlyeq^{\mathcal{I}}$, respectively. Lastly, to show that $r(A) \sim^{\mathcal{I}} r'(A)$ for all $A \in \mathbb{M}$, note that Lemma 1 and Corollary 1 imply $r(A) \sim^{\mathcal{I}} r(\text{conv}(A))$ and $r'(A) \sim^{\mathcal{I}} r'(\text{conv}(A))$ for all A. Hence, I assume A is already convex so that $r(A), r'(A) \in A$. Then, Proposition 6 implies $(r(A), \{r(A)\}) \sim (r(A), A)$ and $(r'(A), \{r'(A)\}) \sim (r'(A), A)$. If $u(r(A)) = u(r'(A))$, then by transitivity, $(r(A), A) \sim (r'(A), A)$. It follows that $\frac{1}{2}(r(A), A) + \frac{1}{2}(r'(A), \{r'(A)\}) \sim \frac{1}{2}(r'(A), A) + \frac{1}{2}(r(A), \{r(A)\})$ which is equivalent to

$$\left(\tfrac{1}{2}r(A) + \tfrac{1}{2}r'(A), \tfrac{1}{2}A + \tfrac{1}{2}\{r'(A)\}\right) \sim \left(\tfrac{1}{2}r'(A) + \tfrac{1}{2}r(A), \tfrac{1}{2}A + \tfrac{1}{2}\{r(A)\}\right) \tag{9}$$

and thus, $r(A) \sim^{\mathcal{I}} r'(A)$. If $u(r(A)) \neq u(r'(A))$, then we can assume $u(r(A)) > u(r'(A))$ without loss of generality, which means

$$(r(A), \{r(A)\}) \sim (r(A), A) \succ (r'(A), A) \sim (r'(A), \{r'(A)\}).$$

In this case, (9) still holds, and thus $r(A) \sim^{\mathcal{I}} r'(A)$.

D. Proof of Theorem 3

By Proposition 6, there exist $\alpha_*, \alpha'_* \in [0,1]$ such that

$$\Big(\bar{x}_{\succsim'}(\alpha'_*), \{\bar{x}_{\succsim'}(\alpha'_*)\}\Big) \sim' \Big(\bar{x}_{\succsim'}(\alpha'_*), A\Big); \quad \Big(\bar{x}_{\succsim}(\alpha_*), \{\bar{x}_{\succsim}(\alpha_*)\}\Big) \sim \Big(\bar{x}_{\succsim}(\alpha_*), A\Big).$$

Notice that, in terms of the representation, $(\bar{x}_{\succsim}(\alpha), \{\bar{x}_{\succsim}(\alpha)\}) \succsim (\bar{x}_{\succsim}(\alpha), A)$ is equivalent to $v(r(A)) \geq \alpha \max_{x\in A} v(x) + (1-\alpha) \min_{x \in A} v(x)$. Clearly, $\alpha'_* = \alpha^*(A; v', r')$ and $\alpha_* = \alpha^*(A; v, r)$. It follows that $\succsim$ has a higher degree of evaluative strictness than $\succsim'$ if and only if, for all $\alpha \in [0,1]$, $\alpha^*(A; v', r') \geq \alpha$ implies $\alpha^*(A; v, r) \geq \alpha$.

References

Arrow, K. J. and Hurwicz, L. (1972). An optimality criterion for decision-making under ignorance. *Uncertainty and expectations in economics : essays in honour of G.L.S. Shackle*.

Bardsley, N. (2008). Dictator game giving: Altruism or artefact? *Experimental Economics*, 11 (2): 122–133.

Bell, D. E. (1982). Regret in Decision Making under Uncertainty. *Operations Research*, 30 (5): 961–981.

Bernheim, B. D., Kim, K., and Taubinsky, D. (2024). Welfare and the Act of Choosing. *National Bureau of Economic Research Working Paper Series*, No. 32200.

Dekel, E. and Lipman, B. L. (2012). Costly Self-Control and Random Self-Indulgence. *Econometrica*, 80 (3): 1271–1302.

Dekel, E., Lipman, B. L., and Rustichini, A. (2001). Representing Preferences with a Unique Subjective State Space. *Econometrica*, 69 (4): 891–934.

Dekel, E., Lipman, B. L., and Rustichini, A. (2009). Temptation-Driven Preferences. *The Review of Economic Studies*, 76 (3): 937–971.

Dekel, E., Lipman, B. L., Rustichini, A., and Sarver, T. (2007). Representing preferences with a unique subjective state space: A corrigendum. *Econometrica*, 75 (2): 591–600.

Dietrich, F. and List, C. (2016). Reason-Based Choice And Context-Dependence: An Explanatory Framework. In *Economics and Philosophy*, volume 32.

Dillenberger, D. and Sadowski, P. (2012). Ashamed to be selfish. *Theoretical Economics*, 7 (1): 99–124.

Emerson, W. R. and Greenleaf, F. P. (1969). Asymptotic Behavior of Products $C^p = C + \cdots + C$ in Locally Compact Abelian Groups. *Transactions of the American Mathematical Society*, 145: 171–204.

Falk, A., Fehr, E., and Fischbacher, U. (2003). On the Nature of Fair Behavior. *Economic Inquiry*, 41 (1): 20–26.

Falk, A., Fehr, E., and Fischbacher, U. (2008). Testing theories of fairness—Intentions matter. *Games and Economic Behavior*, 62 (1): 287–303.

Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54 (2): 293–315.

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, 68 (1): 5–20.

Gilboa, I. and Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18 (2): 141–153.

Gregor, M., editor (1998). *Groundwork of the Metaphysics of Morals.* Cambridge Texts in the History of Philosophy. Cambridge University Press, Cambridge.

Gul, F. and Pesendorfer, W. (2001). Temptation and Self-Control. *Econometrica*, 69 (6): 1403–1435.

Halldén, S. (1980). *The foundations of decision logic.* (Library of Theoria, 14.) Lund: CWK Gleerup.

Hayashi, T. (2024). Meta-preference, endogenous preference formation and dynamic choice. Technical report.

Herstein, I. N. and Milnor, J. (1953). An Axiomatic Approach to Measurable Utility. *Econometrica*, 21 (2): 291.

Hurwicz, L. (1951). Some specification problems and applications to econometric methods. *Econometrica*, 19: 343–344.

Iyengar, S. S. and Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, 79 (6).

Jeffrey, R. C. (1974). Preference Among Preferences. *The Journal of Philosophy*, 71 (13): 377.

Kant, I. (1785). Groundwork of the metaphysics of morals. In Gregor, M., editor, *Groundwork of the Metaphysics of Morals*, Cambridge Texts in the History of Philosophy, pages 1–66. Cambridge University Press, Cambridge.

Kopylov, I. (2012). Perfectionism and Choice. *Econometrica*, 80 (5): 1819–1843.

Kőszegi, B. and Rabin, M. (2006). A Model of Reference-Dependent Preferences. *The Quarterly Journal of Economics*, 121 (4): 1133–1165.

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

Kreps, D. M. (1979). A Representation Theorem for "Preference for Flexibility". *Econometrica*, 47 (3): 565.

List, J. A. (2007). On the Interpretation of Giving in Dictator Games. *Journal of Political Economy*, 115 (3): 482–493.

Loomes, G. and Sugden, R. (1982). Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty. *The Economic Journal*, 92 (368): 805–824.

Nehring, K. (2006). Self-Control through Second-Order Preferences. Technical report.

Noor, J. (2011). Temptation and Revealed Preference. *Econometrica*, 79 (2): 601–644.

Noor, J. and Ren, L. (2023). Temptation and guilt. *Games and Economic Behavior*, 140: 272–295.

Noor, J. and Takeoka, N. (2015). Menu-dependent self-control. *Journal of Mathematical Economics*, 61: 1–20.

Ok, E. A. and Tserenjigmid, G. (2022). Indifference, indecisiveness, experimentation, and stochastic choice. *Theoretical Economics*, 17 (2).

Olszewski, W. (2007). Preferences Over Sets of Lotteries. *The Review of Economic Studies*, 74 (2): 567–595.

Pattanaik, P. K. and Xu, Y. (1990). On Ranking Opportunity Sets in Terms of Freedom of Choice. *Recherches économiques de Louvain*, 56 (3-4).

Pattanaik, P. K. and Xu, Y. (1998). On preference and freedom. *Theory and Decision*, 44 (2).

Pivato, M. (2024). Universal recursive preference structures. Technical report.

Pivato, M. (2025). Autonomy and Metapreferences. *SSRN Working Paper No. 5410003*.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

Saito, K. (2015). Impure altruism and impure selfishness. *Journal of Economic Theory*, 158: 336–370.

Samuelson, P. A. (1952). Probability, Utility, and the Independence Axiom. *Econometrica*, 20 (4): 670.

Sarver, T. (2008). Anticipating Regret: Why Fewer Options May Be Better. *Econometrica*, 76 (2): 263–305.

Sen, A. (1988). Freedom of choice: Concept and content. *European Economic Review*, 32 (2): 269–294.

Shafir, E., Simonson, I., and Tversky, A. (1993). Reason-based choice. *Cognition*, 49 (1): 11–36.

Starr, R. M. (1969). Quasi-Equilibria in Markets with Non-Convex Preferences. *Econometrica*, 37 (1): 25–38.

Stovall, J. E. (2010). Multiple Temptations. *Econometrica*, 78 (1): 349–376.

Stovall, J. E. (2018). Temptation with uncertain normative preference. *Theoretical Economics*, 13 (1): 145–174.

Suzumura, K. and Xu, Y. (2009). Consequentialism and Non-Consequentialism. In *The Handbook of Rational and Social Choice*. Oxford University Press.