A Representation of Preference over Preferences and Choices

Jun Hyun Ji*

January 30, 2026

Abstract

The quality of a choice is relative to constraints: it is evaluated not only by its outcome but also by what the choice reveals about the decision-maker's preferences. I develop an axiomatic framework for preferences over choices, where an *evaluator* values each observed choice for its outcome and for the revelation that the chosen outcome is preferred to all other feasible alternatives. I define "second-order preference" through an axiom that captures the relative nature of choice quality. 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. This provides a revealed-preference characterization of preferences over preferences, with implications for menu choice, paternalistic policy, and welfare evaluation.

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

*Ph.D. student, 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 am also grateful to Sven Neth and Kevin Zollman for many insightful discussions. I received valuable comments from In-Koo Cho, Antonio Penta, Michael Woodford, and participants at the 2024 Asian School of Economic Theory at NYU Abu Dhabi; Jawwad Noor, Evan Piermont, and all participants at the 2025 BSE Summer Forum (Choice and Decision); David Ahn, David Huffman, Colin Sullivan, Gerelt Tserenjigmid, and my colleague Bruno Kömel. I gratefully acknowledge the financial support by 2023 Tamara Horowitz Memorial Fund.

1. Introduction

The quality of a choice is often judged not only by its outcome but also by what it reveals about the decision-maker's (DM's) preferences. People tend to favor choices that reveal good preferences independently of their consequences, praising those aligned with their ideals, values, or virtues. For example, legal systems impose harsher penalties for intentional crimes rather than for acts committed under duress. Across many institutional contexts—courts, classrooms, insurance claims, politics, and recruiting—evaluators appear to care about what the DM willingly chose to give up. In other words, a choice is evaluated not just by its consequence, but also by all other things that could have been chosen. This motivates the question: how do people assess the quality of a choice separately from its consequence?

To address this question, I develop an axiomatic framework for preference over choices within the standard expected utility (EU) theory. Suppose an *evaluator* (she) observes and ranks choices made by one or more DMs (he), each facing distinct menus of options. I define a choice by the outcome-menu pair (x, A) where $x \in A$ is a lottery over sure-outcomes.[1] The evaluator has a preference $\succsim$ that assesses each (x, A) in two dimensions: the outcome and the revelation that the DM prefers x to all else in A. I propose an axiom that characterizes the general concept of preference for the quality of choices, which I refer to as *the second-order preference*. Combined with the axioms of the EU theory, my axioms yield a unique representation that separately identifies the evaluator's preference over outcomes and second-order preference.

The representation is a function $U_{u,v,r}$ of the form:

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

where u, v are von Neumann-Morgenstern (vNM) utility functions over lotteries, and r is a choice function that selects a mixture $r(A)$ of options on menu A. The function u is a ranking of choices from singleton menus, representing the evaluator's outcome preference. The second-order preference 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 of x to that of y regardless of how they are chosen, whereas

[1] A choice need not be a single lottery chosen from a single menu. In the lottery space, a choice can be a contingency plan over two or more menus: e.g., committing to choose x from menu A and y from menu B, before knowing which of A or B will be the true menu. Alternatively, a choice can be interpreted as empirically observed choices across different menus. (See Ok and Tserenjigmid (2022)'s interpretation of stochastic choices.)

$v(x) > v(y)$ implies that she wants the DM to give up y for x. Similarly, $v(x) - v(r(A)) > v(y) - v(r(B))$ reads as "*preferring x to all else in A* is preferred to *preferring y to all else in B*." 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*. The representation above does not explicitly specify what the reference is for each menu; the reference function obeys a minimal set of rules, but 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. In Section 5, I provide a special case where the reference function has an explicit form.

My identification approach lies in the decomposition of the evaluator's preference into two parts. The first part is the outcome preference characterized by the ranking of choices from singleton menus. The second part is the preference that is unrelated to outcomes. It is identified by fixing a stochastic choice environment in which each choice yields the same expected outcome, but reveals different preferences. Since the evaluator conforms to the EU principles, any strict preference in this environment reflects an outcome-irrelevant (OI) concern. In this sense, the preference restricted to this environment is called the OI-preference.

At this point, the OI-preference is not necessarily related to the quality of choices. I say it is the evaluator's second-order preference only if my key axiom Relativity is satisfied. The axiom captures the intuition that the quality of a choice is relative to constraints, and therefore it cannot be determined by looking at either the chosen outcome or the menu in isolation. In other words, every menu comes with a "good" and "bad" option: the former is the one that the evaluator would like the DM to want, and the latter is what she would not like him to want. Formally, the axiom simply requires 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., $(y, \{y\}) \succsim (y, A)$.

A violation of Relativity implies that the evaluator is interested in external factors beyond choices.[2] In such cases, we can find two menus A, B such that every choice from A is strictly preferred to every choice from B in the OI environment, which implies that the quality of choices is independent of choices. I refer to this phenomenon as *menu-favoritism*, an evaluative bias toward the features of the menu itself, independent of outcomes or preferences. I show that Relativity holds if and only if menu-favoritism is absent. As an illustration, I

[2] While the representation is derived under the EU framework, Relativity is independent of the EU theory: it remains well defined even for evaluators who violate the conventional independence or continuity axioms.

present an example in which a judge, violating Relativity, bases the sentence solely on the defendants' influence within the criminal organization rather than on their actual choices.

The representation also requires a notion of consistency. An inconsistent preference over choices involves second-order preference reversals: e.g., the evaluator wants the DM to prefer x to y, but when a third option becomes available, she wants him to prefer y to x. Ruling out such reversals, my main theorems (Theorems 1-2) state that the consistent preference over choices satisfies the axioms of the EU theory and Relativity if and only if it uniquely admits the representation in (1).

Next, I illustrate how the model can incorporate temptation and self-control as determinants of choice quality. Building on Gul and Pesendorfer (2001), I characterize a special case in which the evaluator assigns higher quality to choices that either exceed her expectation or require self-control. I introduce the axiom called *Choice Betweenness* which yields a representation that identifies the evaluator's expectations about the DM's choices, temptations, and self-control—factors that jointly determine the reference of each menu. This special case illustrates that the quality of a choice involves more than simply choosing a good option over a bad one. The evaluator prefers preferences that 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.

The novelty of this special case is that menus that offer good outcomes constituting easy choices fundamentally differ from menus that offer relatively bad outcomes but hard choices of high quality. This creates a tradeoff between the best outcome versus the best choice, a tension that is absent in the standard menu preference framework. I present two applications of this special case to a menu-choice environment. The first is from the DM's perspective, where a dictator chooses a menu of allocations while anticipating how the evaluator's preference over his choices will affect his future payoff. The second is from the evaluator's perspective, illustrated by a parenting example.

My results have several important implications. First, this is the first revealed-preference characterization of preferences over preferences, built directly on outcome–menu primitives. Existing studies on higher-order preferences have largely taken a normative stance, assuming that preferences are objects of choice and therefore prescriptive (see Pivato, 2025, and references therein). I do not model preference over *preference relations* directly but over stochastic choices that reveal preferences either partially or entirely.[3] Thus, I shift the

[3] In Section 7.1, I focus on the choices that reveal the DM's entire preference relation.

attention to the quality of actual choices rather than internal desires.[4] This approach is descriptive and suggests a positive framework: I show that *if* choices over choices are observable, the underlying second-order preference can be identified. This raises an obvious difficulty in experimentation, as incentivizing choices over choices in laboratory settings may be unnatural. Yet the challenge lies in controlled incentivization, not in observability. Many real-world environments involve decisions that rely on evaluating the choices made by others (e.g., court rulings, dating, employee assessments, parenting, and voting).

Second, my framework also bears on the normative domain. It plays a role similar to that of the classical EU theory, providing a normative benchmark for second-order preferences, and a notion of "attitude toward preferences." I discuss in detail how the conventional implications of the EU theory extend to preference over choices. Moreover, when the evaluator is regarded as the DM himself, my representation can also be interpreted as a welfare measure. When the quality of choices matters in one's welfare, outcome-based paternalistic actions and welfare policies would obviously fail. My model provides a preference-based utility framework for the welfare associated with the quality of one's own choices. As I explain the related literature in detail in Section 2, this could further enrich the study on welfare measures when the choice itself is welfare-relevant (see Bernheim et al., 2024).

Third, my model can guide menu design from either the DM's or the evaluator's perspective. Suppose the DM knows the evaluator's preference. If he can choose his own menu, he may design it to reveal or conceal aspects of his preference that the evaluator likes or dislikes. This situation closely parallels models of temptation and self-control that exploit choices over menus (Gul and Pesendorfer, 2001). In that literature, the evaluator is the DM's future self: he removes tempting options so that his future self can make more preferable choices (e.g., those free from self-control or guilt). I do not model preferences over menus directly. However, since menus are part of the object of interest, preferences over choices contain richer information about the value of menus.

My framework can also be an alternative approach to modeling menu choices, by considering the evaluator's menu designs. Suppose the evaluator either expects certain choices to be made, or observes the DM's past choices from a menu, after which she decides whether to restrict his options. That is, the menu choice is conditioned on expectations or observed choices. For example, most parents want their child to make the "right choice" before they

[4] Throughout this paper, I use the phrase "preferring a preference" to mean "preferring to *behave* as if one holds that preference." If an evaluator instead cares about the DM's internal desires—e.g., wanting him to desire x and not desire y, regardless of how he chooses—then choice behavior may be irrelevant: the DM may choose x while still desiring y to some extent. Such cases fall outside the scope of this paper. For the same reason, I rule out cases in which the evaluator's assessment depends on counterfactual or unobserved preferences that do not manifest in choice (e.g., Bob is troubled not by Amy's acceptance of his proposal, but by his belief that she would have chosen someone else had a superior alternative been available).

enforce it: e.g., willingly giving up television for homework, which reveals the preference they hope to see. Hence, they might restrict the child's menu after repeatedly observing (or expecting) poor choices often enough. Prior studies on menu choices implicitly assume that the future self either mistakenly perceives the menu as exogenous or fails to recognize menu choices as *choices*—an assumption that is unverifiable from menu-choice data (see Bernheim et al., 2024). My approach permits the evaluator and DM to be two distinct individuals, and thus the unverifiable assumption is rendered moot.[5]

Lastly, my model can guide the DM's behavior. Misidentifying preferences over choices can easily lead to misbehavior. Suppose the DM faces menu $\{x, y\}$ and wants to please the evaluator, but only learns that the evaluator prefers $(x, \{x, y\})$ to $(y, \{y, z\})$. In a simple world, the natural inference is that the evaluator prefers the outcome of x to that of y, and thus the DM should choose x over y. However, once he allows for the possibility that the evaluator also cares about his preferences, he cannot infer whether or not she wants him to choose x over y. My model characterizes what additional information the DM must learn in order to correctly identify what the evaluator wants him to do.

I develop two extensions. First, to show how my model essentially involves preference over preference relations, I model preference over complete rankings of sure outcomes by fixing a stochastic choice environment that randomizes over all binary menus. I illustrate how the same logic generates a preference over risk attitudes. Second, I enlarge the choice space so that the DM can declare indifference over sets of options. I reproduce the baseline model by characterizing a *no preference for indifference* condition, and discuss its limitations in this extended framework.

The next section reviews related literature. In Section 3, I introduce the model, discuss the implications of my axioms, and characterize the second-order preference. Section 4 presents the representation, uniqueness result, and comparative measures. Sections 5 and 6 present the special case and its applications to menu choices, respectively. The two extensions are in Section 7. Section 8 concludes. Proofs (if omitted) are collected in the Appendix.

2. Related Literature

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.[6] While several

[5] 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").

[6] Regret (Bell, 1982; Loomes and Sugden, 1982) is a classic example.

strands of research explore related ideas, none address the exact question studied here.

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. They axiomatically characterize the choices in the first period when the agent anticipates a mental cost of giving up certain options for another in period 2. The mental cost usually refers to self-control.[7]

Through the lens of my framework, suppose the DM chooses menus "as if" his utility function is M of the form:

$$M(A) \coloneqq \max_{x \in A} U_{u,v,r}(x, A)$$

where $U_{u,v,r}(x, A)$ is the representation in (1). The evaluator can be regarded as the agent's future self and the DM as the agent in period 1 who wants his future self to make better choices. In this sense, the function $U_{u,v,r}$ 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. 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.[8] This suggests that individuals—facing temptation and anticipating guilt or self-control—choose menus as if they have a certain preference over their future choices.

The literature mostly focuses on preferences for commitment—e.g., valuing menus that exclude either tempting options or normative goals. By contrast, my model emphasizes preferences for the quality of choices in general. In the special case I study, I show that the most preferable choice may be available only when the most preferable outcome (e.g., the most tempting *and* important normative goal) is unavailable—a feature that cannot be rationalized by the standard menu preferences in deterministic contexts.[9]

Some later models even take higher-order menus (menus of menus of outcomes, and

[7] Dillenberger and Sadowski (2012) model shame, and Saito (2015) guilt and pride.

[8] As another example, the function M also nests Kopylov (2012)'s representation of a perfectionist's preference who benefits from having normative goals on his menu. The representation is the function M_K of 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 $M = M_K$.

[9] 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 is normative or 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).

so on) as primitives (see Noor, 2011; Noor and Ren, 2023).[10] Such constructions capture higher-order self-control but remain within the DM's perspective. That is, the framework still requires the unverifiable assumption that the future self either perceives the menu as exogenous or fails to recognize the higher-order menu choice as a choice.

The unverifiable assumption was formally characterized by Bernheim et al. (2024)'s study on welfare measures (see also Kőszegi and Rabin, 2008). In their model, a social planner seeks to determine which menu would make the DM happiest based on observing the DM's choices, and they show that this is impossible if the DM's choice itself is welfare-relevant. My model is related to their work in two ways. First, they assume that the DM's welfare is directly associated with *the act of choosing*, which may refer to the quality of his own choices. In this sense, my framework is conceptually consistent with their impossibility result, which I do not attempt to tackle.[11] Instead, in my model, the evaluator can be interpreted as a social planner who has a preference over the DM's outcomes and preferences—e.g., a parent's paternalistic decisions based on children's choices, a policy-maker's decision based on people's choices, etc. Second, Bernheim et al. (2024) assume an arbitrary welfare function of choices for the DM, and do not articulate how exactly choices affect welfare. 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.

There is a vast literature in philosophy and economics on metapreferences and freedom of will, which generally takes a normative stance. 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. These approaches assume that higher-order preferences are objects of choice, and often explore the resulting infinite regress (preferences of order n and beyond). By contrast, my framework takes a descriptive, choice-theoretic stance and studies only second-order preferences, assuming that the evaluator does not possess preferences of higher orders. Therefore, I use the term "second-order preference" differently from the way it is typically used in the literature. Consequently, this paper is the first to formally capture the menu-dependent nature of second-order preferences. Previous studies have largely focused on a fixed choice problem

[10] In those frameworks, the DM reasons recursively—e.g., *"I should do x but I am tempted by y; if I remove x from my menu, I will avoid guilt when doing y; yet if I choose to remove x, I may feel guilty for allowing myself to avoid guilt, and thus I should disable myself from choosing to remove x in the first place,"* and so on.

[11] According to my theorems, if the DM has a preference over his own choices but we only observe his choices over outcomes, then his second-order preference is not identifiable. We may extend this logic and conjecture that if he has a preference over *choices of menus* but we only observe his choices over menus and outcomes, his second-order preference over menu choices is likewise not identifiable, and so on.

(e.g., the heavy smoker of Jeffrey (1974) choosing from {smoke, abstain}), effectively fixing an arbitrary menu. Notice that the representation $U_{u,v,r}$ in (1) is entirely captured by the function $u + v$ of lotteries when the menu is not subject to change. In other words, by observing choices over choices from a fixed menu, second-order preferences cannot be identified.

My model relates to Suzumura and Xu (2009)'s study on preferences for freedom, who likewise take the outcome–menu pair as primitive (see also Sen, 1988; Pattanaik and Xu, 1990, 1998). Their agent, however, values either freedom or outcomes rather than the quality of choices: e.g., the agent characterized as the extreme non-consequentialist prefers (x, A) to (y, B) whenever A contains more alternatives than B. The only direct similarity is that their axiom *Indifference of No-Choice Situations* requires any two choices from singleton menus to be indifferent because they are made without freedom. The second-order preference in my model also satisfies this condition; however, our motivations differ: the indifference occurs because those choices do not reveal preferences. Moreover, these frameworks are defined on deterministic outcomes. I employ the lottery space, which is essential for my result.

I do not assume any utility-relevant context beyond outcomes or menus, hence choices. Dietrich and List (2016)'s model of reason-based choices explains bounded rationality by introducing context that influences how an outcome is valued (see also Shafir et al., 1993).[12] In contrast, my model is unrelated to bounded rationality.[13]

3. Model

3.1. 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 whose elements $A, B \in \mathbb{M}$ are called *menus*.[14] Let $\mathrm{conv}(A)$ denote *the convex hull of* A. Naturally, we have $\mathrm{conv}(A) \in \mathbb{M}$ for all $A \in \mathbb{M}$, which can be interpreted as the menu that offers every randomization of the options in A.

[12] For example, a consumer may prefer apples to bananas, but not "the last apple in a basket" to a banana. Such models address the economist's problem that these contexts are often unobservable.

[13] The evaluator's preferences remain fully rational in the standard sense (e.g., the weak axiom of revealed preference holds when her preference induces an outcome choice rule).

[14] I endow $\mathbb{M}$ with the Hausdorff metric

$$d_H(A, B) := \max\left\{\max_{x\in A}\min_{y\in B} d(x, y),\ \max_{y\in B}\min_{x\in A} d(x, y)\right\}.$$

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 in the set $\{(x, \{x\}) : x \in X\}$ represents an exogenous consumption of a single lottery, and therefore is called a *vacuous choice*. The choice $(x, \{x\})$ indicates that the DM *vacuously chooses* x. 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 consider two ways to 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." Alternatively, α may represent relative frequency: menu 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.[15] Together, these interpretations extend Ok and Tserenjigmid (2022)'s account of stochastic choice—originally defined for a single deterministic menu—to the space of lotteries where both choices and menus are stochastic.

3.2. 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.[16] Likewise, if the evaluator's preference over choices conforms to the EU principles, then she finds no value in *giving up* randomization. Suppose

[15] The empirical interpretation abstracts from scale, whereas outcome evaluation may be sensitive to absolute counts. For example, a DM who chooses x 800 times and y 200 times may be evaluated very differently from one who chooses x 9 times and y once, even though the latter exhibits a higher relative frequency of choosing x. When outcome consequences accumulate with the number of choice opportunities, meaningful comparison across DMs would require normalizing the total number of choices given to each DM.

[16] Axiom 2 is consistent with the assumption that the evaluator remains impartial concerning the timing of uncertainty resolution, as implied by the Independence axiom imposed on menu preferences (see Gul and Pesendorfer, 2001; Dekel et al., 2001, 2007).

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 $(x, \{x, y\}) \sim (x, \text{conv}(\{x, y\}))$ where $\text{conv}(\{x, y\}) = \{\alpha x + (1 - \alpha) y : \alpha \in [0, 1]\}$.

The following result formalizes this implication.

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

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

As a sketch of proof, let $A = \{x, y\}$ for simplicity. The DM, for instance, can condition his choice on good and bad weather, each occurring with equal probability, in which he plans to choose differently from $\{x, y\}$. His menu effectively becomes $\frac{1}{2}\{x, y\} + \frac{1}{2}\{x, y\}$. Yet, if he ultimately chooses x 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. By the same reasoning, the evaluator remains indifferent to any further decomposition of states, as long as the DM chooses x in each one:

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

As the DM considers more finely divided contingencies, his choice approaches $(x, \text{conv}(\{x, y\}))$. By Axiom 3, the indifference is preserved in the limit: $(x, \{x, y\}) \sim (x, \text{conv}(\{x, y\}))$. For any $B \subseteq \text{conv}(\{x, y\})$, notice that $\text{conv}(\{x, y\} \cup B) = \text{conv}(\{x, y\})$. Hence, the result follows similarly.[17]

3.3. 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:

[17] 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.

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 expected outcome, but reveals different preferences. Hence, 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 outcomes. However, the two choices reveal 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 $\succsim^*$ on $\mathbb{C}$ is called *the OI-preference over choices* if

$$(x, A) \succsim^* (y, B) \text{ whenever } \left(\tfrac{1}{2}x + \tfrac{1}{2}y, \tfrac{1}{2}A + \tfrac{1}{2}\{y\}\right) \succsim \left(\tfrac{1}{2}y + \tfrac{1}{2}x, \tfrac{1}{2}B + \tfrac{1}{2}\{x\}\right).$$

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 instruct 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, $\succsim^*$ captures how $\succsim$ behaves when outcomes are irrelevant because the expected 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 $\succsim^*$ reflects the evaluator's outcome-irrelevant concern.

By construction, the OI-preference $\succsim^*$ is what remains after removing the outcome preference from $\succsim$. The following result shows that as long as $\succsim$ satisfies the standard vNM

axioms, $\succsim^*$ treats all vacuous choices as indifferent (the proof is straightforward and omitted for brevity).

Proposition 2 (Indifference of Vacuous Choices)**.** *If $\succsim$ satisfies Axioms 1, 2, 3 then the induced OI-preference $\succsim^*$ satisfies $(x, \{x\}) \sim^* (y, \{y\})$ 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 reveal anything about the DM's preferences.

3.4. Second-order Preference and Menu-favoritism

I now establish the definition of second-order preference. At this point, the OI-preference is not necessarily related to the quality of choices; by construction, it can reflect preferences for anything unrelated to outcomes. The following axiom serves as the condition under which the OI-preference can be regarded as a second-order preference. That is, it 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 $x, y \in A$.*

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" and "bad" option: the former is the one that the evaluator would like the DM to want, and the latter is what she would not like him to want. 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 $\succsim^*$ is the evaluator's second-order preference if Relativity holds.

Definition 3. The OI-preference $\succsim^*$ induced by $\succsim$ is called *the second-order preference over choices* if $\succsim$ satisfies Axioms 1, 2, 3, and Relativity.

If $\succsim^*$ is a second-order preference, then $(x, A) \succsim^* (y, B)$ denotes that *"preferring x to all else in A"* is preferred to *"preferring y to all else in B."* Naturally, $\succsim^*$ also satisfies Axioms 1, 2, 3, and Relativity.[18]

[18] 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 the conventional independence or continuity axioms (Axioms 2-3).

Violations of Relativity are a clear sign that the evaluator is interested in external factors beyond choices. Suppose the axiom does not hold for some non-singleton menu B.[19] This means either (i) that every possible choice from B is strictly preferred to the corresponding vacuous choice or (ii) that every vacuous choice of an option in B is strictly preferred to the corresponding choice from B: i.e., either (i) $(x, B) \succ (x, \{x\})$ for all $x \in B$ or (ii) $(x, \{x\}) \succ (x, B)$ for all $x \in B$. As an 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.

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 $(x, A) \succ^* (y, B)$ for all $x \in A$ and $y \in B$ where $\succsim^*$ is the OI-preference induced by $\succsim$.

Suppose $(x, A) \succ^* (y, B)$ 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 3. *Suppose $\succsim$ satisfies Axioms 1, 2, and 3. Then, $\succsim$ violates Relativity if and only if $\succsim$ exhibits menu-favoritism.*

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

The "if" part is trivial. For the "only if" part, the idea of the proof can be illustrated as follows. Consider a judge (she) deciding which of two suspects (DM_1 and DM_2) deserves a

[19] Relativity holds trivially for singleton menus as long as $\succsim$ is complete.

[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.

harsher sentence based on the stochastic choice data presented by the prosecutor—suppose she imposes the harsher sentence on the suspect whose choice she finds less preferable. Let $B = \text{conv}(\{c, s\})$, where c and s denote committing a crime and staying out of it, respectively. Suppose the judge's preference satisfies $(c, \{c\}) \succ (c, \{c, s\})$ and $(s, \{s\}) \succ (s, \{c, s\})$, thereby violating Relativity. Using this fact alone, I show that the judge exhibits menu-favoritism. Suppose the prosecutor reports that each DM_i $(i = 1, 2)$ faced a stochastic menu M_i given by

$$M_1 = \tfrac{1}{2}\big[\lambda\{c\} + (1-\lambda)B\big] + \tfrac{1}{2}\{y_2\} \quad \text{and} \quad M_2 = \tfrac{1}{2}B + \tfrac{1}{2}\big[\lambda\{c\} + (1-\lambda)\{y_1\}\big]$$

where $y_i \in B$ is the option that DM_i chose willingly. To the judge, the outcomes are irrelevant: the two menus yield the same expected outcome given any $y_1, y_2 \in B$.

The interpretation is as follows. With probability $\frac{\lambda}{2}$, DM_1 is forced to engage in the crime; with probability $\frac{1-\lambda}{2}$, he can decide whether to commit the crime or not. Otherwise, he is dragged into whatever DM_2 does (y_2) with probability $\frac{1}{2}$. DM_2, by contrast, can decide for himself with probability $\frac{1}{2}$; with probability $\frac{\lambda}{2}$, he too is forced into the crime; and with probability $\frac{1-\lambda}{2}$, DM_1's choice (y_1) determines his outcome. The parameter λ thus measures DM_2's *relative influence*. When λ is large, DM_2 effectively becomes the "boss" of the operation: his own decision carries greater weight in shaping the other's outcome.

Notice that when comparing the DMs' choices, the probability assigned to $\{c\}$ (namely, $\frac{\lambda}{2}$ in both M_1 and M_2) is irrelevant by the standard notion of Independence. Then, it follows that DM_1's choice from M_1 is strictly preferred to DM_2's choice from M_2 if and only if

$$(1-\hat{\lambda})(y_1, B) + \hat{\lambda}(y_2, \{y_2\}) \succ \hat{\lambda}(y_2, B) + (1-\hat{\lambda})(y_1, \{y_1\}) \tag{2}$$

where $\hat{\lambda} = \frac{1}{2-\lambda}$. If λ is large enough, then for DM_1, the singleton menu $\{y_2\}$ is assigned a large probability relative to B, whereas for DM_2, the singleton $\{y_1\}$ is assigned a small probability relative to B. Since the judge strictly prefers a vacuous choice to choices made willingly, the standard vNM axioms imply that there must be a large enough λ such that (2) holds for all $y_1, y_2 \in B$. Hence, we have $(x, M_1) \succ (y, M_2)$ for all $x \in M_1$ and $y \in M_2$. By definition of the OI-preference $\succsim^*$, we have $(x, \lambda\{c\} + (1-\lambda)B) \succ^* (y, B)$ for all $x \in \lambda\{c\} + (1-\lambda)B$ and $y \in B$. In other words, the judge would assign a harsher sentence to DM_2 solely because of his greater influence, without needing to know the suspects' choices. This completes the intuitive argument underlying the "only if" part of the proof.

3.5. Ideal Preference and Consistency

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 choice function called the reference function. The motivation for the first component is as follows. Since the evaluator wants the DM to hold certain preferences in each decision environment, some notion of an "ideal DM"—whose choice is always of the highest quality given any menu—must be in her mind. Since $(x, \{x, y\}) \succsim^* (y, \{x, y\})$ 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." Recall the evaluator's first-order preference $\succcurlyeq_1$: $x \succcurlyeq_1 y$ means the evaluator prefers the outcome of x to that of y regardless of how they are chosen. In contrast, $x \succcurlyeq^{\mathcal{I}} y$ means she wants the DM to give up y for x, which relies on menus.

Definition 5. Let $\succsim^*$ 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 $(x, \{x, y\}) \succsim^* (y, \{x, y\})$.

Yet the above definition specifies the ideal DM's behavior only for doubleton menus, and thus is an incomplete characterization. To demonstrate the limitation of this incompleteness, consider the following rankings: $(x, \{x, y\}) \succ^* (y, \{x, y\})$ and $(y, \{x, y, z\}) \succ^* (x, \{x, y, z\})$. 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. I rule out these reversals, and impose the following notion of consistency.

Axiom 5 (Consistency)**.** *For all* $(x, A), (y, A), (x, B), (y, B) \in \mathbb{C}$, $(x, A) \succsim (y, A)$ *if and only if* $(x, B) \succsim (y, B)$.

Consistency 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."[21] Note that this does not determine the ranking *across* menus: i.e., $(x, A) \succsim (y, A)$ does not guarantee $(x, A) \succsim (y, B)$. This fundamentally differs from individuals who care only about outcomes because they would prefer (x, A) to (y, B) whenever x yields a better outcome than y.

Consistency implies that if the evaluator's preference over choices satisfies the standard vNM axioms, then her ideal DM is an expected utility maximizer.

Proposition 4. *Suppose* $\succsim$ *satisfies Axioms 1, 2, 3 and Consistency. Let* $\succsim^*$ *and* $\succcurlyeq^{\mathcal{I}}$ *be the induced OI-preference and the ideal preference of* $\succsim$*, respectively. Then,*

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

(a) $\succsim^*$ *satisfies Consistency,*

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

(c) $\succcurlyeq^{\mathcal{I}}$ *is complete, transitive, continuous and independent.*[22]

Proof. Part (a) is straightforward. For (b), note that (a) implies

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

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

Furthermore, Consistency imposes a consistent ranking of menus conditional on outcomes, which is formally stated as follows:

Proposition 5. *Suppose* $\succsim$ *satisfies Axioms 1, 2, 3 and Consistency. Then, for all* $(x, A), (y, A), (x, B), (y, B) \in \mathbb{C}$*,* $(x, A) \succsim (x, B)$ *if and only if* $(y, A) \succsim (y, B)$*.*

Proof. For the sake of contradiction, suppose there are A, B such that $(x, A) \succsim (x, B)$ and $(y, B) \succ (y, A)$ for some $x, y \in A \cap B$. There are two cases to consider: (i) $(x, A) \succsim (y, A)$ and (ii) $(y, A) \succ (x, A)$. By Consistency, (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 Axioms 1, 2, 3. 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. The two choices are identical since $\frac{1}{2}x + \frac{1}{2}y = \frac{1}{2}y + \frac{1}{2}x \in X$ and $\frac{1}{2}A + \frac{1}{2}B = \frac{1}{2}B + \frac{1}{2}A \in \mathbb{M}$. A similar contradiction is reached in case (ii).[23] *Q.E.D.*

Roughly speaking, Proposition 5 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. Technically, it also implies that the evaluator's utility function must be additively separable in its arguments—the outcomes and the menus.

[22] A binary relation $\succcurlyeq$ on X—along with its asymmetric and symmetric components $\succ$ and $\sim$)—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.

[23] Independence (Axiom 2) also implies that the evaluator is indifferent between the DM's choice of a compound lottery and a simple lottery, a condition known as *the reduction of compound lotteries axiom* (Samuelson, 1952). Notice that menu $\frac{1}{2}A + \frac{1}{2}A$ contains two compound lotteries $\frac{1}{2}x + \frac{1}{2}y$ and $\frac{1}{2}y + \frac{1}{2}x$, which may be two different contingency plans from the DM's perspective. However, the evaluator would not distinguish them since they both yield the same expected outcome.

3.6. Reference Function

I now characterize the second component of the second-order preference: the reference function. The ideal preference $\succcurlyeq^{\mathcal{I}}$ alone does not determine how the second-order preference ranks choices. The reference function resolves this: it is a choice function that selects a reference against which each chosen option is evaluated.[24]

To characterize the reference function, I use the following result.

Proposition 6. *Suppose $\succsim$ satisfies Axioms 1, 2, 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, 2, 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 Proposition 1 that $(x, \lambda A + (1 - \lambda) A) \sim (x, A \cup \{x\})$. Then 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. The choice function $r : \mathbb{M} \to X$ is *the reference function* induced by $\succsim$ if $r(A) \in \{x \in \text{conv}(A) : (x, A \cup \{x\}) \sim (x, \{x\})\}$.

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

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

[24] 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 captured a loss-averse agent's tendencies 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.

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

$$\tfrac{1}{2}x + \tfrac{1}{2}r(B) \succcurlyeq^{\mathcal{I}} \tfrac{1}{2}r(A) + \tfrac{1}{2}y \iff (x, A) \succsim^* (y, B).$$

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

$$\begin{aligned}\Big(\tfrac{1}{2}x + \tfrac{1}{2}r(B), &\tfrac{1}{2}A \cup \{r(A)\} + \tfrac{1}{2}B \cup \{r(B)\}\Big) \\ &\succsim^* \big(\tfrac{1}{2}r(A) + \tfrac{1}{2}y, \tfrac{1}{2}A \cup \{r(A)\} + \tfrac{1}{2}B \cup \{r(B)\}\big)\end{aligned}$$

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 4. *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. 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]

As a consequence, the consistent ranking of menus implied by Consistency in Proposition 5 is entirely attributed to the reference function.

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

4. Representation

I use the standard definitions of preference representations and *affine* functions. Given a preference relation $\succcurlyeq_1$ on X, I say the function v represents $\succcurlyeq_1$ when $v(x) \geq v(y)$ if and only if $x \succcurlyeq_1 y$. The function v is *affine* if $v(\alpha x + (1-\alpha)\, y) = \alpha v(x) + (1-\alpha)\, v(y)$ for all $x, y \in X$ and $\alpha \in [0, 1]$. Analogously, 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 $\lambda \in [0, 1]$. I define a notion of an affine choice function, as follows:

[25] An early attempt to formalize preferences over preferences appears in Halldén (1980), who proposed a rule for judging which preference ranking is more valuable not to forget. Proposition 7 captures his rule as a special case.

Definition 7. A choice function $c : \mathbb{M} \to X$ is *affine with respect to* a function $v : X \to \mathbb{R}$ if $v(c(\lambda A + (1-\lambda) B)) = v(\lambda c(A) + (1-\lambda) c(B))$ for all $\lambda \in (0,1)$.

I now introduce the representation of the preference over choices.[26]

Definition 8. The 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, 2, 3, Consistency 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 $\succsim^*$ is represented by the function $V_{v,r} : \mathbb{C} \to \mathbb{R}$ of the form*

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

Proof. The "if" part is straightforward. For the "only if" part, first note that by the result of Herstein and Milnor (1953), Axioms 1-3 are equivalent to the existence of a continuous affine function $U : \mathbb{C} \to \mathbb{R}$ representing $\succsim$. Consistency and Proposition 5 imply 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]

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 $R = \tilde{R}$.

[26] Note that a choice function c that is affine with respect to v does not necessarily satisfy $c(\lambda A + (1-\lambda) B) = \lambda c(A) + (1-\lambda) c(B)$ for all $\lambda \in (0,1)$, which is a stronger condition.

[27] The 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 $\lambda \in (0,1)$.

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

$$R(A) = R(A \cup \{r(A)\}). \tag{3}$$

Proposition 6 implies that $u(r(A)) = h(r(A)) - R(A \cup \{r(A)\})$. Then, by (3), 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 $\succsim^*$ follow directly from Proposition 4 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 $\succsim = \succsim^*$ 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 $\{x \in \text{conv}(A) : (x, A \cup \{x\}) \sim (x, \{x\})\}$ 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.*

4.1. 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 the preference $\succsim$ is *regular* if for all non-singleton A, $(x, A \cup \{x\}) \not\sim (x, \{x\})$ for some $x \in \text{conv}(A)$. Equivalently, if (u, v, r) represents a regular $\succsim$, then v is not constant on every non-singleton A.

Definition 9. The preference $\succsim$ has a *higher degree of evaluative strictness* than $\succsim'$ if they are both regular, and for all convex $A \in \mathbb{M}$ and $\alpha \in [0,1]$,

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

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 $A = \text{conv}(\{y, z\})$, 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 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. The parameter $\alpha^*(A; v, r)$ measures how close the evaluator's reference of A lies to the best rather than the worst element of A in terms of v.

Theorem 3. *Suppose the two regular 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 $A \in \mathbb{M}$.*

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

Hence, evaluative strictness corresponds to placing greater weight on the most ideal outcome within each menu. When $\succsim$ and $\succsim'$ share the same ideal preference ($v = v'$), the comparison simplifies to

$$v(r(A)) \geq v(r'(A)) \quad \forall A \in \mathbb{M}.$$

Consider two extreme attitudes. If $\alpha^*(A; v, r) = 1$ for all A, then $r(A) = \max_{x \in A} v(x)$: 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 $r(A) = \min_{x \in A} v(x)$: the evaluator judges by how much a choice exceeds the least ideal option. This represents the *most lenient* attitude.

5. Special Case: Temptation and Quality of Choice

Theorem 1 does not specify how the reference function is determined. In particular, the measure $\alpha^*(\cdot; v, r)$—the evaluator's attitude toward preferences—is not necessarily constant across menus.[28] 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).

Axiom 6 (Choice Betweenness (CB))**.** $(x, A) \succsim (x, B)$ *implies* $(x, A) \succsim (x, A \cup B) \succsim (x, B)$.

The axiom 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. Intuitively, one can infer that B contains options that are relatively more ideal than those in A.

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

Definition 10. 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 $(x, A) \succ (x, A \cup B) \succ (x, B)$.

[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's criterion (Hurwicz, 1951; Arrow and Hurwicz, 1972; Gilboa and Schmeidler, 1989).

[29] In Gul and Pesendorfer (2001)'s setting, the 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 $\{y\}$).

In case (a), the evaluator perceives the choice of x from $A \cup B$ as more difficult than from B: when A is available in addition to B, she expects the DM to either choose poorly or choose ideally by exerting costly self-control. Thus, A contains options that tempt the DM away from the ideal options in B. If $(x, A \cup B) \sim (x, B)$, she believes that adding A does not meaningfully change the DM's choice situation. In case (b), she expects self-control and thus the DM to make a better choice when facing $A \cup B$ than when facing A alone. Holding x fixed, she therefore prefers giving up A to giving up $A \cup B$. When $(x, A) \succ (x, A \cup B) \sim (x, B)$, the evaluator does not expect self-control, believing that the DM would succumb to temptation in A whether facing A alone or $A \cup B$. She therefore regards the two choice situations as essentially equivalent.

Definition 11. The 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 $V_{v,\tau}(x, A) \coloneqq U_{u,v,\tau}(x, A) - u(x)$ represents the second-order preference $\succsim^*$ where 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 slightly differs. The functions τ and $v + \tau$ capture the evaluator's expectations about the DM's temptation and the DM's preference, respectively. Let $y_A^* \in \arg\max_{y \in A} v(y) + \tau(y)$. Then y_A^* is the evaluator's expectation about the DM's choice.

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

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

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 $\max_{z \in A} \tau(z) - \tau(y_A^*)$ *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 to incur when the DM faces this menu.

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 arise (i.e., $V_{v,\tau}(x, A) > 0$)

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

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 could still arise from a difficult menu—one that incurs high anticipated mental cost. This is possible whenever the expected choice y_A^* is neither the most ideal nor the most tempting option.[31] By contrast, if the evaluator expects zero mental cost (because she believes the DM will pick the most tempting option), then a choice that meets that expectation produces the quality commensurate to vacuous choices.[32]

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 a direct application of Gul and Pesendorfer (2001)'s Theorem 1. Notice that by Corollary 1, CB holds if and only if $r(A) \succcurlyeq^{\mathcal{I}} r(B)$ implies $r(A) \succcurlyeq^{\mathcal{I}} r(A \cup B) \succcurlyeq^{\mathcal{I}} r(B)$. Define a binary relation $\succeq_r$ on $\mathbb{M}$ as $A \succeq_r B$ if and only if $r(A) \succcurlyeq^{\mathcal{I}} r(B)$. Then, the following lemma holds (I leave the proof in the appendix).

Lemma 1. *$\succeq_r$ is complete, transitive, continuous and independent.*

By the result of Herstein and Milnor (1953), Lemma 1 holds if and only if $\succeq_r$ has a unique continuous affine representation $R^* : \mathbb{M} \to \mathbb{R}$. By construction, the ranking $\succeq_r$ of singletons follows $\succcurlyeq^{\mathcal{I}}$ and thus, $R^*(\{x\}) = v(x)$ for all $x \in X$ where v is the representation of $\succcurlyeq^{\mathcal{I}}$. Since $r(A) \succcurlyeq^{\mathcal{I}} r(B)$ is equivalent to $A \succeq_r B$ which is represented by $R^*(A) \geq R^*(B)$, it follows that $R^*(A) = v(r(A))$ for all $A \in \mathbb{M}$.

Clearly, CB holds if and only if $A \succeq_r B$ implies $A \succeq_r A \cup B \succeq_r B$, which is the *set betweenness* axiom of Gul and Pesendorfer (2001). By 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.*

[31] 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 $V_{v,\tau}(y_A^*, A) > 0$.

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

5.1. 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 $(x, \{x, y, z\}) \succ^* (y, \{x, y, z\}) \succ^* (z, \{x, y, z\})$ and $(x, \{x\}) \succ (y, \{y\}) \succ (z, \{z\})$ so that x is the best outcome among the three alternatives. Consider the following ranking:

$$(y, \{y, z\}) \succ (x, \{x, y, z\}) \succ (y, \{x, y, z\}) \succ (z, \{x, y, z\}). \tag{4}$$

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 (4) implies that $u(y) + v(y) > u(z) + v(z)$, and $r(\cdot)$ satisfies

$$u(x) - u(y) < \underbrace{[v(y) - v(r(\{y, z\}))]}_{V_{v,r}(y,\{y,z\})} - \underbrace{[v(x) - v(r(\{x, y, z\}))]}_{V_{v,r}(x,\{x,y,z\})} \tag{5}$$

which means that the outcome value difference between x and y is smaller than the quality difference between choosing y over z and choosing x over y and z.[33]

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 (4) cannot be rationalized.[34] However, the TPC representation (u, v, τ) can rationalize it. Suppose $\tau(x) > \tau(z) > \tau(y)$. Then, we have $v(r(\{x, y, z\})) = v(x)$, implying that $(x, \{x, y, z\})$ is equivalent to a vacuous choice that yields only consumption value. In contrast, $v(r(\{y, z\}))$ is either $v(y) - v(z)$ or $\tau(z) - \tau(y)$, which captures the strictly positive value of preferring y to z. Intuitively, the value $v(r(\{x, y, z\}))$ moves closer to $v(z)$, the more z tempts the DM relative to y; by comparison, choosing x over y and z is an obvious decision since x is both tempting and ideal. Therefore, if the outcome difference between x and y is sufficiently small, then the value of preferring y to z can outweigh the value of consuming x.

In the menu preference framework, the ranking (4) implies that the menu $\{y, z\}$ is preferred to $\{x, y, z\}$ even though x is preferred to y and z. The 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. The standard models that assume deterministic contexts can only rationalize (4) by identifying x as temptation that is normatively

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

[34] 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, (5) implies $u(x) - u(y) < (1 - \alpha^*)(v(y) - v(x))$ which contradicts the assumption that x maximizes both u and v.

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 had 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.

6. Applications

This section provides two examples illustrating how the TPC representation can be used to model menu choices, and therefore does not contain any general analytic result. In the first part of this section, 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.

6.1. 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 $t \in [-10, 10]$, so 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 that Dictator 2 is unaware of: in period 0, Dictator 1 is allowed to privately choose his menu A from an exogenously given set $\mathcal{A} \subseteq \mathbb{R}^2$ 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, while Dictator 2 may not know Dictator 1's.

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 $t \in [-10, 10]$, after which, 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 let $x_p^w = (6+w, 6+w)$ denote a Pareto optimal (PO) allocation 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 $t \in [-10, 10]$ because t does not affect her payoff.

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

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

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 $t = V_{\delta,k}(x,A)$ where $V_{\delta,k}(x,A)$ measures the quality of the choice, evaluated by Dictator 2. That is, the more favorably she evaluates this choice, the more she wishes to reciprocate in period 2.[35] Then in the subgame perfect Nash equilibrium, Dictator 1's payoff depends only on his own choice: $m_1(x, V_{\delta,k}(x,A)) = U_{\delta,k}(x,A)$.

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 $\tau + v = \delta y_1 + k y_2$ reflects her expectation about Dictator 1's actual behavior. Then Dictator 1's payoff $U_{\delta,k}$ over the three allocations can be summarized as follows:

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

where $\frac{k}{\delta}$ captures how much Dictator 2 values Dictator 1's altruistic preference, relative to

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

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* $\succeq_M$*, as:* $A \succeq_M B$ *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 $V_{\delta,k}(x_p^w, B^w) = k(6 + w) - (\delta + k)(6 + w) + \delta(6 + w) = 0$. That is, Dictator 2 does not reward Dictator 1 for choosing the PO allocation. Intuitively, the choice (x_p^w, B^w) does not reveal anything about Dictator 1's preference that Dictator 2 cares about. 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 foregone monetary gain from the PO allocation, then Dictator 1 strictly prefers B^w to A_0 for all $w > 0$. This follows from $m_1(x_s, 0) \geq U_{\delta,k}(x_s, A_0)$, which holds because $V_{\delta,k}(x_s, A_0)$ is at most zero: x_s is the least ideal option in A_0, and thus Dictator 2 always evaluates this choice unfavorably. Consequently, for any $w > 0$, the payoff from the PO allocation exceeds $m_1(x_s, 0) = 6$.

Part (b) of Proposition 8, on the other hand, 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, if $U_{\delta,k}(x_s, A_0) \leq U_{\delta,k}(x_f, A_0)$, which is equivalent to $u(x_f) + v(x_f) \geq u(x_s) + v(x_s)$ and 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 $\min\{\delta, k\} = V_{\delta,k}(x_f, A_0)$, Proposition 8(b) follows. This reflects that 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 Dictator 2's evaluative strictness is minimal: she expects no self–control from Dictator 1 and thus views the choice of the fair allocation 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)'s menu preference representation. They consider a similar framework (without period 2), and model Dictator 1's menu preference in period 0. They capture what they refer 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 their model, menu A_0 is never preferred to B^w. Through the lens of his framework, this means that the value of pride cannot outweigh the value of committing to the best outcome.[36]

In experimental literature, the idea that a person can derive values from both outcomes and opportunities is not new. Opportunities are vital information when fathoming others' selfishness (List, 2007; Bardsley, 2008) or 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 arbitrary belief-based payoff structure and do not articulate how different menus can reveal different quality of preferences or intentions.

6.2. 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

$$(h, \{h, t\}) \succsim (h, \{h\}) \succ (t, \{t\}) \succsim (t, \{h, t\}).$$

[36] 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 $U_S(A) = W(5) + \alpha W(6)$. By choosing x_f over x_s, the dictator gains the sense of pride measured by $\alpha W(6)$. 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 $A \subseteq A^*$,

$$U_\delta(x, A) := \delta v(x) + \underbrace{v(x) - v(r(A))}_{:=V_\delta(x,A)} \tag{6}$$

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 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 now can 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

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

And 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. For 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 one should be ideally preferred to another. 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

$$(\tilde{\alpha}'h + (1 - \tilde{\alpha}')\, t', A') \sim (h, \{h\}) \quad \text{and} \quad \tilde{\alpha} > \tilde{\alpha}'. \tag{7}$$

According to Theorem 4, (7) 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 stronger self-control than giving up t.

7. Extensions

In this section, I develop two separate extensions. First, to show how my model essentially involves preference over preference relations, I focus on the second-order preference $\succsim^*$ restricted to the choices that reveal the DM's entire preference relation on the set of sure outcomes. I illustrate how the same logic generates a preference over risk attitudes. Second, I enlarge the choice space so that the DM can declare indifference over sets of options. I reproduce the baseline model by characterizing a *no preference for indifference* condition, and discuss its limitations in this extended framework.

7.1. Preference over Rankings

Given the finite set Z of sure outcomes, let $n = \binom{|Z|}{2}$ denote the number of two-element subsets of Z. Let $\mathbb{D}(Z) = \{D_1, \ldots, D_n\}$ be the set of these two-element menus. Define the stochastic menu

$$D^* = \sum_{D_i \in \mathbb{D}(Z)} \tfrac{1}{n} D_i$$

which implies that the DM's menu is D_i with probability $\frac{1}{n}$ for each $i = 1, \ldots, n$. Each choice made from D^* specifies a complete contingency plan for what would be chosen in every possible binary choice situation. The set of these choices is

$$\mathbb{C}^{\mathbb{D}} = \left\{ \left(\tfrac{1}{n} x_1 + \cdots + \tfrac{1}{n} x_n, D^* \right) : x_i \in D_i \text{ for each } i \in \{1, \ldots, n\} \right\}.$$

Let $\mathbb{B}(Z)$ be the set of all complete binary relations on Z. The choice rule induced by $P \in \mathbb{B}(Z)$ is $\mathcal{C}(D, P) \coloneqq \{x \in D : (x, y) \in P \text{ for all } y \in D\}$ for each $D \in \mathbb{D}(Z)$ which is

well-defined since P is complete and D is a doubleton. Given a second-order preference $\succsim^*$, I define another choice rule $\mathcal{C}^*$, which I refer to as *the evaluator's preferred choice rule*, by

$$\mathcal{C}^*(D, P, \succsim^*) := \{x \in \mathcal{C}(D, P) : (x, D) \succsim^* (y, D) \text{ for all } y \in \mathcal{C}(D, P)\}$$

for each $D \in \mathbb{D}(Z)$ and $P \in \mathbb{B}(Z)$. The rule $\mathcal{C}^*$ essentially breaks the indifference induced by P. For example, if $\mathcal{C}(\{x, y\}, P) = \{x, y\}$ and $(x, \{x, y\}) \succ^* (y, \{x, y\})$, then we have $\mathcal{C}^*(\{x, y\}, P, \succsim^*) = \{x\}$. The interpretation is that if the DM's preference is P, then he breaks his indifference by choosing the option that aligns with the evaluator's second-order preference.

I define the evaluator's preference $\succsim^{\mathbb{B}}$ over the complete binary relations on Z, as follows:

Definition 12. The preference relation $\succsim^{\mathbb{B}}$ on $\mathbb{B}(Z)$ is called *the preference over the rankings* of Z induced by the second-order preference $\succsim^*$, if for all $P, Q \in \mathbb{B}(Z)$, $x_i \in \mathcal{C}^*(D_i, P, \succsim^*)$ and $x_i' \in \mathcal{C}^*(D_i, Q, \succsim^*)$ for each $i \in \{1, ..., n\}$,

$$P \succsim^{\mathbb{B}} Q \iff \left(\tfrac{1}{n}x_1 + \cdots + \tfrac{1}{n}x_n, D^*\right) \succsim^* \left(\tfrac{1}{n}x_1' + \cdots + \tfrac{1}{n}x_n', D^*\right).$$

The definition relies on the premise that the evaluator prefers the preference P to Q whenever she prefers the complete contingency plan induced by P to the one induced by Q. One may object that this construction relies exclusively on binary choice situations and thus does not fully describe a DM's behavior. Indeed, a choice from D^* does not specify how the DM would choose from menus with three or more elements. Nonetheless, observing the DM's contingent choices across all binary menus provides strong information about the underlying preference relation. Ranking these contingency plans therefore induces a meaningful ranking over preference relations and the evaluator's second-order preference.

The preference $\succsim^{\mathbb{B}}$ admits the following representation.

Theorem 5. *Suppose the second-order preference $\succsim^*$ is represented by the pair (v, r) as in Theorem 1. Then, the preference $\succsim^{\mathbb{B}}$ over the rankings of Z induced by $\succsim^*$ is represented by $V^* : \mathbb{B}(Z) \to \mathbb{R}$ of the form:*

$$V^*(P) := \sum_{D \in \mathbb{D}(Z)} \left(\max_{x \in \mathcal{C}(D,P)} v(x) \right).$$

Proof. Let the function $V(x, A) = v(x) - v(r(A))$ represent $\succsim^*$. Then it follows that $\mathcal{C}^*(D, P, \succsim^*) = \arg\max_{x \in \mathcal{C}(D,P)} v(x)$ for any $P \in \mathbb{B}(Z)$ and $D \in \mathbb{D}(Z)$. Suppose $P \succsim^{\mathbb{B}} Q$. For each $i = 1, \ldots, n$, choose any $x_i \in \arg\max_{x \in \mathcal{C}(D_i,P)} v(x)$ and $x_i' \in \arg\max_{x \in \mathcal{C}(D_i,Q)} v(x)$. Then, by the definition of $\succsim^{\mathbb{B}}$ and Proposition 4, $P \succsim^{\mathbb{B}} Q$ is equivalent to $\sum_{i=1}^n \frac{1}{n}x_i \succcurlyeq^{\mathcal{I}}$

$\sum_{i=1}^n \frac{1}{n} x_i'$ where $\succcurlyeq^{\mathcal{I}}$ is the ideal preference of $\succsim^*$. Since v is affine and represents $\succcurlyeq^{\mathcal{I}}$, it follows that $\sum_{i=1}^n v(x_i) \geq \sum_{i=1}^n v(x_i')$, which is the desired result. *Q.E.D.*

For example, let $Z = \{z_1, z_2, z_3\}$ and $D^* = \frac{1}{3}\{z_1, z_2\} + \frac{1}{3}\{z_2, z_3\} + \frac{1}{3}\{z_1, z_3\}$. Suppose there are two decision-makers who strictly prefer z_1 to z_3. They differ on z_2: DM_1 strictly prefers z_3 to z_2, whereas DM_2 strictly prefers z_2 to z_1. Then DM_1 chooses the plan $\frac{1}{3}z_1 + \frac{1}{3}z_3 + \frac{1}{3}z_1$ from D^*, and DM_2 chooses $\frac{1}{3}z_2 + \frac{1}{3}z_2 + \frac{1}{3}z_1$. Let $P, Q \in \mathbb{B}(Z)$ denote the preferences of DM_1 and DM_2, respectively. Suppose the evaluator's ideal preference representation v satisfies $v(z_1) = 1$, $v(z_2) = \bar{v} \in (0, 1)$, and $v(z_3) = 0$. Then Theorem 5 implies that $V^*(P) = 2$ and $V^*(Q) = 1 + 2\bar{v}$, and thus $P \succsim^{\mathbb{B}} Q$ whenever $\bar{v} \leq \frac{1}{2}$.

Theorem 5 has a broader methodological implication. It shows that if one fixes a stochastic choice environment that elicits certain preferences of interest, then the evaluator's second-order preference induces a well-defined preference over the resulting class of preferences. The example below adapts this idea by changing the elicitation environment D^* to characterize preferences over risk preferences.

Example 1 (Preference over Risk Preferences)**.** *Fix outcomes z_1, z_2, z_3, and for each $p \in [0, 1]$, consider the binary menu $D(p) := \{z_2, pz_1 + (1-p)\, z_3\}$. Suppose an expected-utility maximizing DM_i ($i = 1, 2$) has a utility function u_i satisfying $u_i(z_1) = 1 \geq u_i(z_2) \geq u_i(z_3) = 0$ for each $i = 1, 2$. Let each DM's menu be $D(p)$ where p is uniformly distributed on $[0, 1]$. Then, DM_i's contingency plan for this menu is a mapping $p \mapsto x_i(p) \in D(p)$ and his risk attitude is characterized by a cutoff $\bar{u}_i := u_i(z_2)$: he chooses z_2 for $p \leq \bar{u}_i$ and chooses the lottery $pz_1 + (1-p)\, z_3$ for $p > \bar{u}_i$.*

Suppose the evaluator's ideal DM's cutoff is $\bar{v}$. That is, her ideal preference representation v satisfies $v(z_1) = 1$, $v(z_3) = 0$, and $v(z_2) = \bar{v} \in (0, 1)$. By affinity, $v\,(pz_1 + (1-p)z_3) = p$. Therefore, a DM_i with cutoff $\bar{u}_i$ yields the value $\bar{v}$ whenever $p \leq \bar{u}_i$ and p whenever $p > \bar{u}_i$. It follows that her preference over the DMs' risk attitudes can be characterized by the function V_{risk} of the form:

$$V_{\mathrm{risk}}(\bar{u}) = \int_0^{\bar{u}} \bar{v}\, dp + \int_{\bar{u}}^1 p\, dp = \bar{u}\,\bar{v} + \frac{1 - \bar{u}^2}{2}.$$

Then, it is straightforward to conclude that $V_{\mathrm{risk}}(\bar{u}_1) \geq V_{\mathrm{risk}}(\bar{u}_2)$ whenever $|\bar{v} - \bar{u}_1| \leq |\bar{v} - \bar{u}_2|$. That is, the evaluator prefers the risk preference associated with $\bar{u}_1$ to that associated with $\bar{u}_2$ if and only if $\bar{u}_1$ is closer to $\bar{v}$ than $\bar{u}_2$.

7.2. Preference over Indifference

As the second extension, I allow the evaluator to observe the DM's indifference directly by enlarging the choice space. Given a menu A, the DM is now allowed to choose any subset

$\mathbf{a} \subseteq A$, which is interpreted as declaring indifference among all elements of $\mathbf{a}$. The final outcome choice is then delegated to the evaluator, who selects $x \in \mathbf{a}$. For example, when facing menu $\{x, y\}$, the DM can choose to declare:

$$\{x\} : \textit{"I choose } x\textit{,"} \quad \{y\} : \textit{"I choose } y\textit{,"} \quad \{x, y\} : \textit{"I'm indifferent between } x \textit{ and } y\textit{."}$$

Formally, let $\mathbb{M}(A)$ denote the set of all nonempty compact subsets of A. An *extended choice* is a pair $(\mathbf{a}, A)$ where $\mathbf{a} \in \mathbb{M}(A)$. The evaluator's extended preference $\succsim$ is now defined on the set of extended choices $\overline{\mathbb{C}} = \{(\mathbf{a}, A) : \mathbf{a} \in \mathbb{M}(A) \text{ and } A \in \mathbb{M}\}$.

This extended framework is important for three reasons. First, it applies to environments in which indifference is directly observable and the act of declaring indifference matters beyond the eventual outcome. The baseline model, by contrast, implies that if two DMs choose the same outcome from the same menu, then the quality of their choices is identical, regardless of whether the outcome was strictly preferred or chosen out of indifference. Second, declaring indifference is not the same as randomizing (e.g., flipping a coin). Although both actions can induce the same distribution over outcomes, they might reflect different attitudes. Declaring indifference involves a willingness to delegate the final choice, which may itself be evaluated positively or negatively. Whereas choosing a coin flip may instead reflect beliefs about future states or moods. A person who is not addicted to alcohol may declare indifference between beer and coffee, whereas an alcoholic might plan to drink beer when he is in a bad mood. Third, in the space of lotteries, indifference may come from the DM's imprecise beliefs, lack of decisiveness, or open-mindedness. For instance, the DM may declare indifference over any randomization between x and y as long as the probability of x lies between 0.4 and 0.6, suggesting imprecise beliefs about future circumstances. In contrast, a truly indifferent person would be indifferent among any randomization. When the evaluator cares about such features, extended choices can reveal more about the DM's underlying preferences.

As a simple extension, I focus on reproducing the original PC representation within this extended framework. I show that doing so requires the assumption that the evaluator has *no preference for indifference* (NPI). I characterize this NPI and then discuss its limitations for the resulting representation.

I first modify the axioms as follows:

Axiom 1*. *$\succsim$ on $\overline{\mathbb{C}}$ is complete and transitive.*

Axiom 2*. *For all $\lambda \in (0, 1)$, $(\mathbf{a}, A) \succ (\mathbf{b}, B)$ implies $\lambda(\mathbf{a}, A) + (1 - \lambda)(\mathbf{c}, C) \succ \lambda(\mathbf{b}, B) + (1 - \lambda)(\mathbf{c}, C)$.*

Axiom 3*. $\{(\mathbf{a}, A) : (\mathbf{a}, A) \succsim (\mathbf{b}, B)\}$ *and* $\{(\mathbf{a}, A) : (\mathbf{b}, B) \succsim (\mathbf{a}, A)\}$ *are closed.*

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

Axiom 5* (Consistency*)**.** *For all* $(\mathbf{a}, A), (\mathbf{b}, A), (\mathbf{a}, B), (\mathbf{b}, B) \in \overline{\mathbb{C}}$, $(\mathbf{a}, A) \succsim (\mathbf{b}, A)$ *if and only if* $(\mathbf{a}, B) \succsim (\mathbf{b}, B)$.

Relativity* extends Relativity to the extended domain in which the DM may choose a subset $\mathbf{a} \subseteq A$. The axiom states that, for every menu A, there exist two extended choices $\mathbf{a}, \mathbf{b} \in \mathbb{M}(A)$ that play the roles of a "good" and a "bad" option, as in the baseline model. First, $(\mathbf{a}, A) \succsim (\{x\}, \{x\})$ for all $x \in \mathbf{a}$ means that choosing $\mathbf{a}$ from A is at least as good as the exogenous consumption of any element of $\mathbf{a}$. Second, $(\{y\}, \{y\}) \succsim (\mathbf{b}, A)$ for all $y \in \mathbf{b}$ means that $(\mathbf{b}, A)$ is at least as bad as any corresponding enforced outcome. The remaining axioms extend their baseline counterparts in a straightforward way.

The NPI condition is characterized by the following axiom.

Axiom 6* (No Preference for Indifference (NPI))**.** $(\mathbf{a}, A) \succsim (\mathbf{b}, A)$ *implies* $(\mathbf{a}, A) \sim (\mathbf{a} \cup \mathbf{b}, A)$.

NPI states that, for a fixed menu A, if declaring indifference over $\mathbf{a}$ is preferred to declaring indifference over $\mathbf{b}$, then declaring indifference over $\mathbf{a} \cup \mathbf{b}$ is indifferent to declaring indifference over $\mathbf{a}$. This axiom effectively captures the idea that the evaluator does not care about whether the chosen option is strictly preferred or selected due to indifference.

To describe the reference function in the extended framework, I need the following definitions. I say $\overline{c} : \mathbb{M} \to \mathbb{M}$ is a *(stochastic) choice correspondence* if $\overline{c}(A) \in \text{conv}(\mathbb{M}(A))$ for all $A \in \mathbb{M}$.[37] And $\overline{c}$ is affine with respect to a function $f : \mathbb{M} \to \mathbb{R}$ if $f(\overline{c}(\lambda A + (1 - \lambda) B)) = \lambda f(\overline{c}(A)) + (1 - \lambda) f(\overline{c}(B))$ for $\lambda \in [0, 1]$. Let $H(A) \coloneqq \{x \in A : (\{x\}, A) \succsim (\{x'\}, A)\ \forall x' \in A\}$ which collects every option $x \in A$ such that choosing the single option x from A is preferred to choosing any other single option x' from A. Define $\mathcal{R}(A) \coloneqq \{\mathbf{a} \in \text{conv}(\mathbb{M}(A)) : (\mathbf{a}, A \cup \mathbf{a}) \sim (\{x\}, \{x\})\ \forall x \in H(\mathbf{a})\}$. The extended version of the reference function in Theorem 1 is as follows:

Definition 13. The choice correspondence $\overline{r} : \mathbb{M} \to \mathbb{M}$ is *the extended reference correspondence* induced by $\succsim$ if $\overline{r}(A) \in \mathcal{R}(A)$.

The reference in the extended framework is a mixture over extended choices. For example, the reference of menu $\{x, y\}$ may be a randomization between choosing x and declaring

[37] Note that $\text{conv}(\mathbb{M}(A))$ is a set of menus—i.e., $\text{conv}(\mathbb{M}(A)) \subseteq \mathbb{M}$. For example, if $A = \{x, y\}$, then $\mathbb{M}(A) = \{\{x\}, \{y\}, \{x, y\}\}$ and $\text{conv}(\mathbb{M}(A)) = \{\lambda_1\{x\} + \lambda_2\{y\} + \lambda_3\{x, y\} : \lambda_1 + \lambda_2 + \lambda_3 = 1 \text{ and } \lambda_1, \lambda_2, \lambda_3 \geq 0\}$.

indifference. Technically, $\overline{r}(\{x,y\})$ is a set of lotteries: e.g., $\overline{r}(\{x,y\}) = \lambda\{x\}+(1-\lambda)\{x,y\} = \{x, \lambda x + (1-\lambda)\, y\}$ for some $\lambda \in [0,1]$.

I now present the PC representation that exhibits NPI.

Definition 14. The representation of a *NPI preference over choices* (NPI-PC) is a tuple $(u, v, \overline{r})$ where u, v are affine functions of lotteries, and $\overline{r}$ is an affine extended reference correspondence with respect to v such that $\succsim$ is represented by

$$\overline{U}_{u,v,\overline{r}}(\mathbf{a}, A) \coloneqq \max_{x\in\mathbf{a}} u(x) + v(x) - \max_{y\in\mathcal{H}(\overline{r}(A))} v(y)$$

where $\mathcal{H}(A) = \arg\max_{x\in A} u(x) + v(x)$.

Theorem 6. *$\succsim$ satisfies Axioms 1*, 2*, 3*, Consistency*, Relativity*, and NPI if and only if $\succsim$ has a NPI-PC representation.*

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

The function $\overline{U}_{u,v,\overline{r}}$ inherits the essential structure of $U_{u,v,r}$ in Theorem 1, but now incorporates the NPI condition. When the DM declares indifference over $\mathbf{a} \subseteq A$, the evaluator chooses the final option $x^* \in \arg\max_{x\in\mathbf{a}} u(x)+v(x)$ which captures her compromise between the outcome utility and ideal ranking. The last term in $\overline{U}_{u,v,\overline{r}}$ serves as the reference value. The reference correspondence $\overline{r}(A)$ specifies the set of outcomes that are relevant for forming a single reference. The operator $\mathcal{H}$ retains only the options in $\overline{r}(A)$ that maximize $u+v$, and thus determines the ultimate reference $r^* \in \arg\max_{y\in\mathcal{H}(\overline{r}(A))} v(y)$. If the DM is not allowed to declare indifference, the chosen set $\mathbf{a}$ is a singleton, and therefore $\overline{U}_{u,v,\overline{r}}$ effectively becomes the baseline PC representation $U_{u,v,r}$. The only difference is that the original reference function $r(\cdot)$ does not take indifference into account.

To discuss the implications of NPI on the second-order preference, suppose the function u is constant, in which case, the preference $\succsim$ is the second-order preference itself.

Corollary 3. *Suppose $\succsim$ admits a NPI-PC representation $(u, v, \overline{r})$ and $(\{x\},\{x\}) \sim (\{y\},\{y\})$ for all $x, y \in X$. Then there is a function $\overline{v}$ of sets such that $\succsim$ is represented by*

$$\overline{V}(\mathbf{a}, A) \coloneqq \overline{v}(\mathbf{a}) - \overline{v}(\overline{r}(A)) \quad \textit{where} \quad \overline{v}(\mathbf{a}) \coloneqq \max_{x\in\mathbf{a}} v(x).$$

Corollary 3 shows that the ideal preference in the extended framework is technically a binary relation on sets, represented by the function $\overline{v}$ of sets. Then it follows from NPI that this function takes the form $\overline{v}(\mathbf{a}) = \max_{x\in\mathbf{a}} \overline{v}(\{x\})$.

Without NPI, the ideal ranking $\overline{v}$ can take various functional forms.[38] For example, $\overline{v}$ may assign higher values to larger sets, interpreting a declaration of indifference as desirable flexibility or open-mindedness. Alternatively, $\overline{v}$ may assign higher values to smaller sets, reflecting a preference for resolute attitudes, decisiveness or commitment. More generally, standard results on preferences over sets can be used to characterize a wide range of possible ideal preferences over indifference, whereas the NPI condition collapses all such distinctions by forcing $\overline{v}(\mathbf{a})$ to depend only on the best element in $\mathbf{a}$.[39]

8. Conclusion

I show that second-order preference—a concept long confined to philosophical debate—can be identified through choice. This framework bridges philosophical insights and economic theory within the standard EU paradigm, relying exclusively on observable decisions rather than assuming arbitrary belief-dependent utilities.

Appendix

A. Proof of Proposition 1

I first show that $(x, A) \sim (x, \text{conv}(A))$ holds by using the following result.

Lemma 2 (Shapley-Folkman Theorem)**.** *For any $A \in \mathbb{M}$ and $n \in \mathbb{N}$, define $A_n \coloneqq \sum_{s=1}^{n} \lambda_s A$ where $\lambda_s = \frac{1}{n}$ for $s = 1, \ldots, n$. Then $A_n \to \text{conv}(A)$ in the Hausdorff metric, as $n \to \infty$.*

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

Since $\mathbb{C}$ is a mixture space, by the result of Herstein and Milnor (1953), Axioms 1-3 imply that $(x, A) \sim \lambda(x, A) + (1 - \lambda)(x, A)$ for all $(x, A) \in \mathbb{C}$ and $\lambda \in (0, 1)$. Define A_n as in Lemma 2. By iteration, we have $(x, A) \sim (x, A_n)$ for all n. Then, the result follows by Lemma 2 and Axiom 3. Next, pick any $B \subseteq \text{conv}(A)$. To show that $(x, A) \sim (x, A \cup B)$, notice that the above result implies $(x, A \cup B) \sim (x, \text{conv}(A \cup B))$. Since $\text{conv}(A \cup B) = \text{conv}(A)$, the result follows by transitivity.

[38] Without NPI, the representation in Theorem 6 takes an even more general form, but the reference correspondence cannot be uniquely identified as in Theorem 2. An axiom such as NPI, which imposes conditions on the evaluator's ideal preference, is needed in order to uniquely identify the reference correspondence.

[39] The function $\overline{v}$ could take the form of Gul and Pesendorfer (2001)'s preference for commitment, which would capture an intermediate preference for indifference: e.g., $\overline{v}(\{x\}) \geq \overline{v}(\{x, y\}) \geq \overline{v}(\{y\})$.

B. Proof of Proposition 3

The "if" part is trivial. For the "only if" part, suppose Relativity fails. I assume without loss of generality, that there is a convex non-singleton menu B such that $(y, \{y\}) \succ (y, B)$ for all $y \in B$. 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. A choice from A_λ is a pair (x_{y_1}, A_λ) where $x_{y_1} := \lambda b + (1-\lambda)\, y_1$ for some $y_1 \in B$. Then, it follows from the definition of $\succsim^*$ that $(x_{y_1}, A_\lambda) \succ^* (y_2, B)$ 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 standard vNM axioms, is also equivalent to

$$\begin{aligned} &\tfrac{1}{2}\Big[\lambda(b,\{b\}) + (1-\lambda)\,(y_1, B)\Big] + \tfrac{1}{2}(y_2, \{y_2\}) \\ &\qquad \succ \tfrac{1}{2}(y_2, B) + \tfrac{1}{2}\Big[\lambda(b,\{b\}) + (1-\lambda)\,(y_1, \{y_1\})\Big]. \end{aligned} \tag{8}$$

Notice that $(b, \{b\})$, with probability weight $\frac{\lambda}{2}$ on both sides, is an irrelevant alternative, and thus (8) is equivalent to

$$\left[\frac{1-\lambda}{2-\lambda}\right](y_1, B) + \left[\frac{1}{2-\lambda}\right](y_2, \{y_2\}) \succ \left[\frac{1}{2-\lambda}\right](y_2, B) + \left[\frac{1-\lambda}{2-\lambda}\right](y_1, \{y_1\}). \tag{9}$$

The vNM axioms imply that if $\lambda > 0$ is large enough, the strict preference in (9) 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 $\succsim^*$, we have menu-favoritism: $(x, A_\lambda) \succ^* (y, B)$ for all $x \in A_\lambda$ and $y \in B$.

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 the PC representations of $\succsim$. Due to the 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 Proposition 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

$$\big(\tfrac{1}{2}r(A) + \tfrac{1}{2}r'(A), \tfrac{1}{2}A + \tfrac{1}{2}\{r'(A)\}\big) \sim \big(\tfrac{1}{2}r'(A) + \tfrac{1}{2}r(A), \tfrac{1}{2}A + \tfrac{1}{2}\{r(A)\}\big) \tag{10}$$

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, (10) 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

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

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 $\alpha_* = \alpha^*(A; v, r)$. Then 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 $\alpha^*(A; v, r) \geq \alpha$.

E. Proof of Lemma 1

Since $\succcurlyeq^{\mathcal{I}}$ is complete and transitive, so is $\succeq_r$. For continuity, since $\mathbb{M}$ is a topological space, it is sufficient to show that $A \succ_r C \succ_r B$ implies there are $\alpha, \beta \in (0,1)$ such that $\alpha A + (1-\alpha) B \succ_r C \succ_r \beta A + (1-\beta) B$. Since $\succcurlyeq^{\mathcal{I}}$ is continuous, there are $\alpha, \beta \in (0,1)$ such that $\alpha r(A) + (1-\alpha) r(B) \succ^{\mathcal{I}} r(C) \succ^{\mathcal{I}} \beta r(A) + (1-\beta) r(B)$. Since $r(\cdot)$ is affine with respect to $\succcurlyeq^{\mathcal{I}}$, we have $r(\alpha A + (1-\alpha) B) \succ^{\mathcal{I}} r(C) \succ^{\mathcal{I}} r(\beta A + (1-\beta) B)$ which is equivalent to our desired result. For Independence, suppose $A \succ_r B$ or equivalently, $r(A) \succ^{\mathcal{I}} r(B)$. Since $\succcurlyeq^{\mathcal{I}}$ is *independent*, $\lambda \in (0,1)$ implies $\lambda r(A) + (1-\lambda) r(C) \succ^{\mathcal{I}} \lambda r(B) + (1-\lambda) r(C)$. By the affinity of $r(\cdot)$, it implies $r(\lambda A + (1-\lambda) C) \succ^{\mathcal{I}} r(\lambda B + (1-\lambda) C)$. By definition, we have $\lambda A + (1-\lambda) C \succ_r \lambda B + (1-\lambda) C$.

F. Proof of Theorem 6

The "if" part is omitted. The "only if" part proceeds similarly as in the proof of Theorem 1, except for a slight modification of the identifying functions. I first prove the following lemma, which is the extended version of Proposition 6.

Lemma 3. *$\mathcal{R}(A) \neq \emptyset$ for all $A \in \mathbb{M}$ if $\succsim$ satisfies Axioms 1*, 2*, 3* and Relativity*.*

Proof. By Relativity*, each $\mathbb{M}(A)$ contains $\mathbf{s}, \mathbf{w} \in \mathbb{M}(A)$ such that $(\mathbf{s}, A) \succsim (\{z\}, \{z\})$ for all $z \in \mathbf{s}$ and $(\{y\}, \{y\}) \succsim (\mathbf{w}, A)$ for all $y \in \mathbf{w}$. Choose $z \in H(\mathbf{s})$ and $y \in H(\mathbf{w})$. Then, Axioms 1*, 2*, 3* ensure the existence of a $\lambda \in [0,1]$ such that $(\lambda \mathbf{s} + (1-\lambda) \mathbf{w}, \lambda A + (1-\lambda) A) \sim (\lambda z + (1-\lambda) y, \{\lambda z + (1-\lambda) y\})$. Let $\mathbf{a} = \lambda \mathbf{s} + (1-\lambda) \mathbf{w}$ and $x = \lambda z + (1-\lambda) y$.

Then, by Axioms 1*, 2*, 3*, we have $x \in H(\mathbf{a})$. Since the irrelevance of forgone randomization continues to hold in this extended framework (the proof is omitted), we have $(\mathbf{a}, A \cup \mathbf{a}) \sim (\{x\}, \{x\})$. *Q.E.D.*

The following lemma shows that NPI yields the following result.

Lemma 4. *Let $U(\mathbf{a}, A) = h(\mathbf{a}) - R(A)$ be the affine representation of $\succsim$. Then $h(\mathbf{a}) = \max_{x \in \mathbf{a}} h(\{x\})$ for all $\mathbf{a} \in \mathbb{M}$.*

Proof. By NPI, $h(\mathbf{a}) \geq h(\mathbf{b})$ implies $h(\mathbf{a}) = h(\mathbf{a} \cup \mathbf{b})$. Choose any $A \in \mathbb{M}$ and let $b_A \in \arg\max_{x \in A} h(\{x\})$. Then, it is easy to show that NPI implies $h(\{b_A\}) = h(\{b_A, x\})$ for all $x \in A$, and by iteration, we can conclude $h(\{b_A\}) \geq h(A)$. Since $A \cup \{b_A\} = A$, we have $h(\{b_A\}) = h(A)$. *Q.E.D.*

Fix an arbitrary $\mathbf{a}_0 \in \mathbb{M}$ and define

$$u(x) \coloneqq U(\{x\}, \{x\}); \qquad h(\mathbf{a}) \coloneqq \max_{x \in \mathbf{a}} U(\{x\}, X) - U(\mathbf{a}_0, X);$$

$$\overline{v}(\mathbf{a}) \coloneqq h(\mathbf{a}) - u(x') \text{ where } x' \in H(\mathbf{a}); \qquad R(A) \coloneqq h(\mathbf{a}) - U(\mathbf{a}, A).$$

$R(A) = \overline{v}(\overline{r}(A))$ holds trivially if A is a singleton. Note that by the irrelevance of forgone randomization, for any $A \in \mathbb{M}$, we have $U(\mathbf{b}, A) = U(\mathbf{b}, A \cup \overline{r}(A))$ for all $\mathbf{b} \in \mathbb{M}(A)$, which implies $R(A) = R(A \cup \overline{r}(A))$. By Lemma 3 and the definition of $\overline{r}(\cdot)$, we also have $U(\{x^*\}, \{x^*\}) = U(\overline{r}(A), A \cup \overline{r}(A))$ for all $x^* \in H(\overline{r}(A))$, which implies $u(x^*) = h(\overline{r}(A)) - R(A)$. Then it follows from the definition of $\overline{v}$ that $R(A) = \overline{v}(\overline{r}(A))$. By Lemma 4, $x^* \in H(\overline{r}(A))$ implies $x^* \in \arg\max_{y \in \overline{r}(A)} h(\{y\})$ and $\max_{y \in \overline{r}(A)} h(\{y\}) = u(x^*) + \overline{v}(\overline{r}(A)) = h(\{x^*\})$. It follows that $\overline{v}(\overline{r}(A)) = \overline{v}(\{x^*\})$. Define $v(y) \coloneqq \overline{v}(\{y\})$ for all $y \in X$. Then $U(\mathbf{a}, A) = \max_{x \in \mathbf{a}} u(x) + v(x) - v(\overline{r}(A)) = \max_{x \in \mathbf{a}} u(x) + v(x) - \max_{y \in H(\overline{r}(A))} v(y)$. All subsequent steps follow as in the original proof and are thus omitted.

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.

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.

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

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.

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*.

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.

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.