# Reference-dependent self-control: Menu effects and behavioral choices

Abhinash Borah and Raghvi Garg*

August 12, 2022

*(Job market paper)*

*(Revise and resubmit at Journal of Economic Behavior and Organization)*

**Abstract**

As is well-known, choices of a decision maker (DM) who attempts self-control in the face of temptation may exhibit menu effects and "non-standard" patterns. Existing models can accommodate some of these patterns but not others; e.g., they can explain self-control undermining menu effects, but not self-control enhancing ones. We introduce a model of self-control with the goal of better understanding and accounting for such effects. The basic idea underlying our model is that the DM experiences a psychological cost if she succumbs to temptation and chooses in a manner that is totally antithetical to her commitment preferences. To mitigate such costs, in any menu, her expression of self-control involves, first, eliminating a subset of alternatives that are worst according to her commitment preferences, with the elimination process being reference-dependent. Then, amongst the remaining alternatives, she chooses the best one according to her temptation preferences. Besides studying menu effects, we characterize the model behaviorally based on a novel axiom called WARP with norms. We also show that the model is well-identified.

**Keywords:** self-control; temptation; normative elimination; reference dependence; menu effects

**JEL codes:** D01, D91

## 1 Introduction

This paper re-visits the theme of temptation and self-control with the goal of studying menu effects that emerge when a decision maker (DM) faced with temptation exercises

*Department of Economics, Ashoka University, Sonepat, Haryana - 131029, India. Emails: abhinash.borah@ashoka.edu.in, raghvi.garg_phd17@ashoka.edu.in.

self-control. To that end, we introduce a decision model that captures the behavior of such a DM. As is well-known, following the seminal paper of Gul and Pesendorfer (2001), several models have been introduced in the literature that highlight this very theme; e.g., Fudenberg and Levine (2006), Noor and Takeoka (2010), Noor and Takeoka (2015), Liang, Grant, and Hsieh (2020), and Masatlioglu, Nakajima, and Ozdenoren (2020), amongst others. Our key motivation in adding to this impressive body of work is the observation that the exercise of self-control may produce a rich pattern of empirically relevant menu effects and non-standard choices. Existing models can accommodate some of these patterns but not others. For instance, evidence shows that there may exist both self-control undermining menu effects as well as self-control enhancing ones. Whereas existing models account for the first type, they cannot accommodate the second. Our goal here through the decision model that we introduce is to add to the understanding of these different patterns of self-control induced menu effects and non-standard choices.

To illustrate our point, take for instance, the following experimental evidence from a study by Sharpe, Staelin, and Huber (2008) that looked at subjects' choices when it comes to the size of soft drinks consumed. Their evidence is indicative of menu effects with a significant portion of subjects switching their choice of what size soft drink to consume when the drink sizes available in a menu were altered by adding a larger or a smaller drink size. Specifically, in their "high-condition," they varied the baseline menu composed of 16, 21 and 32-ounce drink sizes by adding a larger drink size of 44-ounce. In this menu, among the subjects who consumed a 32-ounce drink, 28% consumed a smaller drink size of 21-ounce in the baseline menu. In other words, when a larger-sized drink was added to the menu, a menu effect involving a shift in choice towards greater consumption of soft drink was observed amongst a significant number of DMs. At the same time, another channel through which menu effects were observed was when a smaller-sized alternative was added to the menu. In their "low-condition," the experimenters added a 12-ounce drink to the baseline menu. Now among the subjects who consumed a 16-ounce drink in the low-condition, 23% opted for a larger drink size of 21-ounce in the baseline menu. That is, the menu effect displayed by a significant number of DMs in this case involved decreasing their consumption of soft drink.

We interpret this behavior displaying menu effects as suggestive of DMs faced with a self-control problem, who from a normative perspective would prefer committing to healthier choices like consuming less soft drinks but often struggle to do so in the face of temptation. In the context of such DMs, what this evidence illustrates is the possibility of two qualitatively distinct types of menu effects, one which undermines their ability to exercise self-control and the other which enhances it. *Self-control undermining menu effects* occur when, on a menu being expanded, choice amongst the existing alternatives shifts towards one that is worse according to the DM's commitment preferences. In the example above, this is the case when the menu consisting of {16 oz, 21 oz, 32 oz} drink sizes is expanded to include the normatively worse 44 oz option. In this case, a significant portion of DMs'

choices shifted from the 21 oz drink size to the 32 oz one that is, presumably, worse according to their commitment preferences. In other words, this type of menu effect is associated with a deterioration in the ability to exercise self-control. On the other hand, *self-control enhancing menu effects* emerge when, on a menu being expanded, choice amongst the existing alternatives shifts towards one that is an improvement according to her commitment preferences. In the example above, this is the case when the menu consisting of {16 oz, 21 oz, 32 oz} drink sizes is expanded to include the normatively better 12 oz option. In this case, a significant portion of DMs' choices shifted from the 21 oz drink size to the 16 oz one, demonstrating an improvement in the ability to exercise self-control.[1] Self-control undermining menu effects have been highlighted in the literature and the models mentioned above can accommodate them. However, self-control enhancing menu effects have been less discussed and are not consistent with the aforementioned models. In particular, there is an asymmetry in these models when it comes to accommodating menu effects. Whereas menu effects marked by a deterioration of choices as per commitment preferences are possible when menus are expanded to include normatively worse alternatives, no such effects in the opposite direction of choices improving when menus are expanded to include normatively better alternatives can occur.[2] One motivation for the model of self-control that we introduce here is that it provides a unified framework for accommodating these different types of self-control related menu effects. More generally, we see merit in drawing from the behavioral choice literature to better understand and classify the patterns of non-standard choices that may emerge when a DM exercises self-control and we attempt to do so using our decision model.

The basic idea underlying our model is that, when faced with temptation, the DM experiences a psychological cost if she completely succumbs to it and chooses in a manner that is totally antithetical to her commitment preferences. To mitigate such costs, in any menu, her expression of self-control involves eliminating a subset of alternatives that are worst according to her commitment preferences. Once she has assuaged her concerns about not making a normatively inferior choice by eliminating these alternatives from the menu, she views the remaining alternatives in comparison as normatively acceptable. Accordingly, she feels psychologically unconstrained to choose the best amongst them according to her temptation preferences. Our key innovation in modeling this process of exercising self-control via elimination of normatively inferior choices is in highlighting how attitudes towards what is normatively acceptable and what is not may be reference-dependent. Accordingly, we refer to this choice procedure that the DM employs to arrive at her choices as the *reference-dependent self-control heuristic (RSH).*

[1] Note that in this paper when we talk about menu effects like self-control undermining or enhancing ones, to keep the discussion organized, we will always do so in the context of menus being expanded, similar to how leading examples of such effects like the compromise and attractions effects are presented.

[2] This is an observation that Lipman and Pesendorfer (2013) highlight when discussing the connection that these models establish between exercising self-control and the compromise effect. We re-visit this observation below.

We incorporate reference dependence by building on the hypothesis that the best and the worst alternatives in a menu according to the DM's commitment preferences serve as reference points for what should be eliminated and not chosen in that menu and, accordingly, for what is normatively acceptable. In particular, what this means is that whether the DM considers an alternative to be normatively acceptable or not is not determined absolutely but rather relative to the best and the worst commitment alternatives of the menu in which it appears. The best alternative serves as the benchmark for normatively the most desirable choice in the menu. The further away is any alternative in the menu from it according to commitment preferences, the less normatively appealing it appears in comparison. A similar relative assessment applies the closer is any such alternative to the worst alternative representing the least desirable choice in the menu from a normative perspective. Therefore, when these reference points vary, the DM's normative assessment of the same alternative may change. When one or both of them improve according to commitment preferences, it may strengthen the normative standards against which the alternative is assessed and, accordingly, the alternative may appear normatively less appealing in comparison. The opposite may be true when these reference points deteriorate. Accordingly, an alternative may appear normatively acceptable and not get eliminated in one menu but may in another depending on the reference points. For example, for a DM who wants to make healthy choices by her commitment preferences, if the most healthy alternative in a menu is a 16 oz cola, consuming a 21 oz cola may be perceived as normatively acceptable. However, if the menu is expanded to include a more healthy option of a 12 oz cola, then, in comparison, the 21 oz cola may appear unhealthy, and consuming it may be deemed normatively unacceptable. On the other hand, if the worst alternative in a menu according to her commitment preferences is a 32 oz cola, consuming it may be viewed as normatively unacceptable. However, if the menu were to be expanded and a less healthy option of a 44 oz cola added, the 32 oz cola may no longer appear that bad in comparison and may be considered normatively acceptable. As we will show, this structure of reference dependence implies a natural inter-menu restriction for the elimination process, which plays a crucial role in the analysis of the model. In turn, this opens the door to the different types of menu effects mentioned above.

In the next section, we formally define the RSH. Then, in Section 3, we discuss the main empirical content of the model in terms of its implications for menu effects and non-standard choices. We compare and contrast these implications with those of other leading models of self-control in the literature mentioned above. We show that a key difference between our model and these models when it comes to accounting for menu effects can be understood in terms of a well-known axiom of the behavioral choice literature called weak WARP.[3] Section 4 provides a behavioral characterization of the RSH. We show that

[3]WARP or the weak axiom of revealed preferences provides the behavioral foundation for the standard rational choice model. It basically requires that there should be no choice reversals, i.e., if $x$ is chosen in the presence of $y$, then $y$ should never be chosen in the presence of $x$. Weak WARP is a weakening of WARP that allows choice reversals but rules out re-reversals.

the model can be characterized by a novel axiom, which we call WARP with norms, that appropriately generalizes WARP, where the generalization is precisely to account for the process of eliminating alternatives on normative grounds that the DM engages in. Section 5 outlines the extent to which the parameters of the model can be uniquely identified from choice data. We show that the commitment preferences of the DM can be precisely identified. On the other hand, under plausible assumptions, the temptation preferences can be identified with a fair degree of precision. Proofs of results appear in the Appendix.

## 2 Reference-dependent self-control heuristic

We consider a DM and the choices she makes in different choice problems. Formally, let $X$ be a finite set of alternatives with typical elements denoted by $x$, $y$, $z$ etc. $\mathcal{X}$ denotes the set of non-empty subsets of $X$ with typical elements denoted by $S$, $T$, etc., which we refer to as menus. A choice function $c : \mathcal{X} \to X$ is a mapping that, for any $S \in \mathcal{X}$, specifies the alternative $c(S) \in S$ that the DM chooses in that menu.

In the model that we develop, the DM is subject to an intrapersonal conflict between her commitment and temptation preferences, captured by two linear orders, $\succcurlyeq_c \subseteq X \times X$ and $\succcurlyeq_t \subseteq X \times X$, respectively. Faced with this conflict, she is able to exercise self-control and our sequential choice procedure captures this. Specifically, when considering any non-singleton menu of alternatives, in the first stage, the DM eliminates a non-empty, strict subset of these alternatives that are worst according to her commitment preferences, $\succcurlyeq_c$. These are precisely the alternatives that the DM feels she should not choose in that menu so as to avoid experiencing the psychological cost associated with behaving in a way that is completely antithetical to her normative goals. Therefore, eliminating them from her consideration is an expression of self-control. Then, in the second stage, amongst the remaining alternatives, which are deemed normatively acceptable in comparison, she picks the best one according to her temptation preferences, $\succcurlyeq_t$. In the way of notation, for any menu $S$, denote the best and the worst alternatives according to commitment preferences, $\succcurlyeq_c$, by $\bar{z}_S$ and $\underline{z}_S$, respectively.

The key idea underlying the model is that attitude towards what is normatively acceptable and what is not is reference-dependent. As discussed in the Introduction, the best and the worst alternatives according to commitment preferences in any menu serve as reference points determining normative standards in the menu against which the DM's perception about whether an alternative is normatively acceptable or not is formed. A change in these reference points may, therefore, bring about a change in the DM's normative assessment of the same alternative. When one or both of them improve according to commitment preferences, strengthening the normative standards in the corresponding menu, the same alternative may appear normatively less appealing in comparison. The opposite may be

true when the reference points deteriorate. Given this, a natural inter-menu connection suggests itself when it comes to the elimination process. Consider menus $T$ and $S$ with $\bar{z}_T \succcurlyeq_c \bar{z}_S$ and $\underline{z}_T \succcurlyeq_c \underline{z}_S$ and any alternative $x \in S \cap T$. Suppose in the menu $S$, $x$ is not deemed normatively acceptable and gets eliminated. Then, given that the best and the worst alternatives in $T$ according to $\succcurlyeq_c$ are no worse than their respective counterparts in $S$, the normative standards in $T$ are no lower than in $S$. This means that, normatively, $x$ cannot be any more appealing in $T$ than in $S$. Therefore, it seems natural to assume that $x$ would not be deemed normatively acceptable in $T$ and get eliminated from this menu as well. This observation is the key idea in our definition of a normative elimination mapping, which identifies, in any menu, the alternatives that the DM eliminates on normative grounds. In the way of notation, let $\mathcal{X}^+$ denote the set of non-singleton menus in $\mathcal{X}$.

**Definition 2.1.** *$g : \mathcal{X}^+ \to \mathcal{X}$ is a normative elimination mapping w.r.t. $\succcurlyeq_c$ if*

1. *$\forall S \in \mathcal{X}^+$, $\emptyset \neq g(S) \subsetneq S$ and $x \in S \setminus g(S), y \in g(S)$ implies $x \succcurlyeq_c y$*
2. *$\forall S, T \in \mathcal{X}^+$ with $\bar{z}_T \succcurlyeq_c \bar{z}_S$ and $\underline{z}_T \succcurlyeq_c \underline{z}_S$, $g(S) \cap T \subseteq g(T)$.*

That is, for any menu $S \in \mathcal{X}^+$, the set of alternatives eliminated, $g(S)$, is a *non-empty strict subset* of $S$ containing its $k$ worst alternatives according to $\succcurlyeq_c$, where $0 < k < |S|$.[4] Further, the process of elimination respects the inter-menu restriction suggested above. We can now formally define the choice procedure. In the way of notation, note that we adopt the convention that $g(S) = \emptyset$ when $S$ is a singleton.

**Definition 2.2.** *A choice function $c : \mathcal{X} \to X$ is a reference-dependent self-control heuristic (RSH) if there exists an ordered pair of linear orders $(\succcurlyeq_c, \succcurlyeq_t)$ on $X$ and a normative elimination mapping $g : \mathcal{X}^+ \to \mathcal{X}$ w.r.t. $\succcurlyeq_c$, such that for any $S \in \mathcal{X}$, $c(S)$ is the $\succcurlyeq_t$-best alternative in $S \setminus g(S)$, i.e.,*

$$\{c(S)\} = \{x \in S \setminus g(S) : x \succcurlyeq_t y, \forall y \in S \setminus g(S)\}$$

One final comment on notation before concluding this section. Note that in the subsequent analysis, we often abuse notation by suppressing set delimiters; e.g., we write $S \setminus x$ instead of $S \setminus \{x\}$, $c(xy)$ instead of $c(\{x, y\})$, etc.

## 3 Menu effects and non-standard choices

We can now go ahead and highlight the empirical content of our model when it comes to menu effects and non-standard choices, and compare and contrast these with leading

[4] Following standard notation, $|.|$ denotes the cardinality of a set.

models of self-control in the literature. A good starting point for this exercise is to re-visit the two types of menu effects that we discussed in the Introduction. Suppose $S \subseteq T$, $c(T) \in S$ but $c(S) \neq c(T)$ such that a menu effect exists. We say that the menu effect is *self-control undermining* if, on the menu being expanded from $S$ to $T$, choice deteriorates according to commitment preferences, i.e., $c(S) \succcurlyeq_c c(T)$; and *self-control enhancing* if choice improves according to commitment preferences, i.e., $c(T) \succcurlyeq_c c(S)$. Both types of menu effects can occur in our model. Self-control undermining menu effects may occur in our model when a menu is expanded to include an alternative that is worse than any of the ones in the original menu according to the DM's commitment preferences. This may result in a weakening of normative standards in the expanded menu. Accordingly, alternatives which were deemed normatively unacceptable and eliminated in the original menu may no longer get eliminated in the expanded menu and one of them may get chosen. On the other hand, self-control enhancing menu effects may occur when a menu is expanded to include an alternative that is better than any of the ones in the original menu according to the DM's commitment preferences. This may result in a strengthening of normative standards in the expanded menu. Accordingly, the chosen alternative in the original menu, which was deemed normatively acceptable and not eliminated may be no longer deemed so and may get eliminated in the expanded menu. In its place a different alternative from the original menu, which is better by the DM's commitment preferences may get chosen in the expanded menu.

**Example 3.1.** Re-visit the soft drinks example discussed in the Introduction with different drink sizes of 12 oz, 16 oz, 21 oz, 32 oz, and 44 oz. Suppose an RSH-type DM's commitment and temptation preferences are, respectively, decreasing and increasing in drink sizes, i.e., $12 \succcurlyeq_c 16 \succcurlyeq_c 21 \succcurlyeq_c 32 \succcurlyeq_c 44$, and $44 \succcurlyeq_t 32 \succcurlyeq_t 21 \succcurlyeq_t 16 \succcurlyeq_t 12$. When such a DM is faced with the menu $S = \{16, 21, 32\}$, say, $g(S) = \{32\}$ and accordingly, $c(S) = 21$. Now, suppose this menu is expanded to $T = \{16, 21, 32, 44\}$. Since the introduction of the larger, less healthy 44 oz drink potentially weakens normative standards in the menu $T$ compared to the menu $S$, it is possible that the 32 oz drink no longer appears normatively unacceptable in $T$, i.e., $32 \notin g(T) = \{44\}$. In that case, $c(T) = 32$, producing a self-control undermining menu effect. Next, suppose the menu $S = \{16, 21, 32\}$ is expanded to the menu $T' = \{12, 16, 21, 32\}$. Now, the addition of the smaller 12 oz drink may strengthen normative standards in the menu $T'$ compared to $S$, and it is possible that with reference to it, the 21 oz drink appears normatively unacceptable in $T'$, i.e., $21 \in g(T') = \{21, 32\}$. In that case, $c(T') = 16$, thus producing a self-control enhancing menu effect.

One key difference between our model and leading self-control models in the literature when it comes to menu effects is that the latter can only accommodate self-control undermining menu effects, but not self-control enhancing ones. To explain this point as well as for other discussion pertaining to non-standard choices, we first briefly introduce some of these models that have been developed within the framework of the seminal paper by Gul and Pesendorfer (2001) [GP]. Drawing inspiration from an approach pioneered in Kreps

(1979), GP model a two-stage decision problem confronting a DM who faces temptation while choosing from a menu but is also sophisticated to recognize this at an ex ante stage while choosing between menus to face at the second stage. Our model connects to the second stage of the GP decision problem, where like in our set-up, the DM is able to exercise self-control. Specifically, their model conceives two utility functions over the set of alternatives, $u$ and $v$, representing the DM's commitment and temptation preferences, respectively, such that choice in the second stage from a menu is specified by:[5]

$$c(S) = \underset{x \in S}{\operatorname{argmax}} \left\{ u(x) - \big[ \max_{y \in S} v(y) - v(x) \big] \right\} = \underset{x \in S}{\operatorname{argmax}} \big\{ u(x) + v(x) \big\}$$

In other words, the value to the DM of choosing an alternative $x$ from the menu $S$ is given by the commitment utility, $u(x)$, adjusted for the cost of exercising self-control, $\max_{y \in S} v(y) - v(x)$, involved in choosing $x$ from a menu whose most tempting alternative has a temptation level of $\max_{y \in S} v(y)$. Observe that the GP formulation implies, unlike in our set-up, that choices from menus satisfy WARP. This difference between our model and the GP framework has been narrowed by subsequent innovations introduced to this framework, which we discuss next. To keep the set-up of these models comparable with ours, in the subsequent analysis, we will assume that the commitment and temptation utility functions, $u$ and $v$, represent linear orders and choice in any menu is decisive.

In Noor and Takeoka (2010), the cost of exerting self-control is convex implying that the marginal cost of exerting self-control is increasing. Choice in any menu $S$ is determined by:

$$c(S) = \underset{x \in S}{\operatorname{argmax}} \left\{ u(x) - \varphi\big( \max_{y \in S} v(y) - v(x) \big) \right\},$$

where $\varphi$ is a strictly increasing and convex function.[6] Noor and Takeoka (2015) introduce self-control costs which are menu-dependent. Formally,

$$\begin{aligned} c(S) &= \underset{x \in S}{\operatorname{argmax}} \left\{ u(x) - \psi\big( \max_{y \in S} v(y) \big) \big( \max_{y \in S} v(y) - v(x) \big) \right\} \\ &= \underset{x \in S}{\operatorname{argmax}} \left\{ u(x) + \psi\big( \max_{y \in S} v(y) \big) v(x) \right\}, \end{aligned}$$

where $\psi(.) \geq 0$ is increasing. That is, the function $\psi(.)$ scales up or down the self-control cost associated with a menu depending on the most tempting alternative available in it. Both Masatlioglu, Nakajima, and Ozdenoren (2020) and Liang, Grant, and Hsieh (2020) model a DM who has a limited stock of willpower which determines the extent of self-control she can exert in a menu in terms of the alternatives that are psychologically feasible for her to choose. In Masatlioglu, Nakajima, and Ozdenoren (2020), choice from a menu is

[5] In the GP framework, alternatives in a menu are lotteries.

[6] Noor and Takeoka (2010) also discuss a more general version of this model in which the cost function takes a general form, $\tilde{\varphi}\big(v(x), \max_{y \in S} v(y)\big)$.

determined by:

$$c(S) = \underset{x \in S}{\operatorname{argmax}}\, u(x) \text{ subject to } \max_{y \in S} v(y) - v(x) \leq w,$$

where $w \geq 0$ measures the DM's stock of will power.[7] On the other hand, choices from menus in Liang, Grant, and Hsieh (2020) are like in Gul and Pesendorfer (2001) but with the willpower constraint. Specifically,

$$c(S) = \underset{x \in S}{\operatorname{argmax}}\{u(x) + v(x)\} \text{ subject to } \max_{y \in S} v(y) - v(x) \leq w$$

All of the aforementioned papers contain examples showing that the models therein can accommodate self-control undermining menu effects. However, they cannot accommodate self-control enhancing menu effects.

**Proposition 3.1.** *The self-control models of Noor and Takeoka (2010), Noor and Takeoka (2015), Masatlioglu, Nakajima, and Ozdenoren (2020) and Liang, Grant, and Hsieh (2020) cannot accommodate self-control enhancing menu effects.*

**Proof:** Please refer to Section A.1.1.

The result highlights an asymmetry in these models regarding accommodating menu effects. As the literature points out, one of the contributions of this class of models has been to provide a new perspective on menu effects like the compromise effect. For instance, consider a DM who, by her commitment preferences, wants to eat healthy and reduce her calorie intake but finds desserts tempting. Suppose she has a 240-calorie apple pie and a 500-calorie cheesecake in front of her. The cheesecake is less healthy than the apple pie but presumably more tasty and tempting. Additionally, suppose a 1000-calorie banoffee Nutella waffle, which is even less healthy but tastier, is added to the menu. In that case, the cheesecake may emerge as a natural compromise being in the middle on both dimensions of health and taste. This type of a compromise effect is consistent with these models but with an important caveat. Whereas the compromise effect is symmetric, the effect generated in these models is not (Lipman and Pesendorfer, 2013). For instance, under these models, if, say, a healthy but less tasty 90-calorie fruit yogurt is added to the menu, then no such effect involving the compromise shifting towards the apple pie is possible. This, of course, speaks to the inability of these models to accommodate self-control enhancing menu effects.

This difference between our model and these ones in the literature can be understood in terms of a key axiom of the behavioral choice literature called weak WARP.

[7] Masatlioglu, Nakajima, and Ozdenoren (2020) also discuss a more general version of this model in which the will power stock varies with the chosen alternative.

**Definition 3.1.** *A choice function $c : \mathcal{X} \to X$ satisfies weak WARP if for all $S, T \in \mathcal{X}$ and $x, y \in X$:*

$$\big[\{x, y\} \subseteq S \subseteq T, x = c(xy) = c(T)\big] \Rightarrow y \neq c(S)$$

Weak WARP is a generalization of WARP[8] and allows for several kinds of non-standard behavior while, at the same time, putting meaningful restrictions on such behavior. Essentially, what it says is that as we expand menus, we can have choice reversals but not re-reversals, i.e., choice can flip from, say, $x$ to $y$, but not back again from $y$ to $x$. Several important models in the behavioral choice literature are characterized by weak WARP, e.g., the Rationalization model of Cherepanov, Feddersen, and Sandroni (2013) and the Categorize then Choose (CTC) model of Manzini and Mariotti (2012). As it turns out, one key difference between our model and the other ones of self-control that has a bearing on the question of accommodating different types of menu-effects is that all the others satisfy weak WARP but our model may not.

**Proposition 3.2.** *If a choice function is rationalizable by the self-control models of Noor and Takeoka (2010), Noor and Takeoka (2015), Masatlioglu, Nakajima, and Ozdenoren (2020) and Liang, Grant, and Hsieh (2020), then it satisfies weak WARP. On the other hand, an RSH may not satisfy weak WARP. However, under an RSH, if $S, T \in \mathcal{X}$ such that $\overline{z}_T = \overline{z}_S$, then $\{x, y\} \subseteq S \subseteq T, x = c(xy) = c(T) \implies y \neq c(S)$.*

**Proof:** Please refer to Section A.1.2.

The result clarifies in choice-theoretic terms what's driving the difference between our model and the others when it comes to accommodating the two types of menu effects. As mentioned earlier, weak WARP allows for choice reversals but not re-reversals. So, if we have alternatives $x, y \in X$ with, say, $x \succcurlyeq_c y$, $x \neq y$, then, in the presence of weak WARP, we cannot have $\{x, y\} \subseteq S \subseteq T$ and $\{x, y\} \subseteq S' \subseteq T'$ with $c(S) = x$, $c(T) = y$, and $c(S') = y, c(T') = x$. Since either $c(xy) = x$ or $c(xy) = y$, that is precisely the type of behavior weak WARP rules out. In other words, in these models, unlike ours, we cannot have a situation with choice shifting from $x$ to $y$ under one menu expansion, giving rise to a self-control undermining menu effect; and, at the same time, choice shifting from $y$ to $x$ under another, resulting in a self-control enhancing menu effect. What the result also clarifies is that, under our model, as long as we restrict attention to expansions involving a weakening of normative standards, i.e., $S \subseteq T$, with $\overline{z}_T = \overline{z}_S$ and $\underline{z}_S \succcurlyeq_c \underline{z}_T$, the conclusion of weak WARP holds. Such expansions, of course, permit only self-control undermining menu effects. In other words, the conflict with weak WARP comes about when it comes to accommodating self-control enhancing menu effects.

[8]Recall that WARP imposes the following consistency on a DM's choices: for all $S, T \in \mathcal{X}$ and $x, y \in S \cap T$, $x \neq y$, if $c(S) = x$ then $c(T) \neq y$. That is, if $x$ is chosen in the presence of $y$, then $y$ should never be chosen in the presence of $x$. If choices satisfy WARP, then they can be rationalized by a single strict preference ranking that can be uniquely elicited from these choices.

We now elaborate on other types of non-standard choice behavior that an RSH and the self-control models discussed above can and cannot accommodate. As a way of organizing the discussion, note that when it comes to violations of rational choice theory, i.e., choices that are not in accordance with WARP, three prominent classes of violations, as seen in experiments and field studies, have been highlighted in the literature. These are violations of the conditions of always chosen (AC), no binary cycles (NBC) and never chosen (NC).

**Definition 3.2.** *A choice function $c : \mathcal{X} \to X$ satisfies:*

1. *AC if for all $S \in \mathcal{X}$ and $x \in S$: $[c(xy) = x, \forall y \in S \setminus x] \implies c(S) = x$*
2. *NBC if for all $x_1, \ldots, x_{n+1} \in X$: $[c(x_i x_{i+1}) = x_i, i = 1, \ldots, n] \implies c(x_1 x_{n+1}) = x_1$*
3. *NC if for all $S \in \mathcal{X}$ and $x \in S$: $[x \neq c(xy), \forall y \in S \setminus x] \implies c(S) \neq x$*

Manzini and Mariotti (2007) show that all violations of WARP can be categorized as either violations of AC or violations of NBC (or both). The RSH model, like the other self-control models, can accommodate violations of AC. However, as the following result establishes, it cannot accommodate violations of NBC nor those of NC, thus making falsifiability of the model transparent in the context of well-known patterns of non-standard choice behavior.

**Proposition 3.3.** *An RSH satisfies NBC and NC.*

**Proof:** Please refer to Section A.1.3.

On the other hand, whereas choices in all of the four models mentioned above satisfy NC, they can violate the condition of NBC. This observation makes for an interesting testable distinction between our approach to self-control and that of these models. For instance, in choice problems like the soft drink one discussed above, both our model and these models would tend to predict that between having a small (say, 12 oz) and a medium-sized (say, 21 oz) soft drink, the DM will choose the small one; and between the medium and a large-sized (say, 44 oz) one, she will choose the medium. However, when it comes to a choice between the small and large drink sizes, whereas our model predicts that she will be able to exercise self-control and choose the small one, these models will often predict that she will choose the large one.[9] Beyond the significance that this distinction has from

[9] To illustrate this, consider the convex self-control cost model of Noor and Takeoka (2010), with the three alternatives of small drink ($x$), medium drink ($y$) and large drink ($z$). Say, $u(x) = 13, u(y) = 7, u(z) = 2$ and $v(x) = 1, v(y) = 3, v(z) = 5$. Further, $\varphi(r) = r^2$. Then in the menu $\{x, y\}$,

$$u(x) - \varphi\big(\max_{w \in \{x,y\}} v(w) - v(x)\big) = 13 - (3-1)^2 = 9 > 7 = 7 - (3-3)^2 = u(y) - \varphi\big(\max_{w \in \{x,y\}} v(w) - v(y)\big)$$

Hence, $c(xy) = x$. Similar calculations in the menu $\{y, z\}$ give us that "utilities" from $y$ and $z$ are 3 and 2, respectively; hence $c(yz) = y$. Finally, in the menu $\{x, z\}$, utilities from $x$ and $z$ are $-3$ and 2, respectively; hence, $c(xz) = z$. Together these choices violate NBC. In the limited willpower model of Masatlioglu,

the perspective of the broader taxonomy of non-standard choices, it also has practical ramifications. As is well known, DMs who make cyclical choices may be subject to money pumps. Therefore, an implication of these four models is that the exercise of self-control in the way it is envisaged in them may leave the DM vulnerable to money pumps. On the other hand, in our model, the exercise of self-control does not expose the DM to money pumps made possible by such cyclical patterns of choice.

Another key difference between our model and these four is in relation to how overwhelmed a DM's ability to exercise self-control can get in the presence of highly tempting alternatives. These four models have the feature that in such scenarios, the DM may be unable to exercise any self-control and may end up choosing the most tempting alternative.[10] This is along the lines of the prediction in Strotz (1955)'s pioneering model, where the DM is unable to exercise any self-control at the time of making choices. As opposed to this, in our model, the DM is always able to exercise some level of self-control even in the presence of highly tempting alternatives and she never ends up choosing the most tempting alternative whenever it happens to be the worst option according to her commitment preferences.

These differences in behavior point toward a key difference in the psychological driver underlying the scope of self-control in the two approaches. In particular, the two approaches point in slightly different directions regarding the psychological cost that is prioritized. In our set-up, the key psychological cost that the DM considers comes from behaving in a manner that is antithetical to her commitment preferences and making choices that are normatively inferior. To mitigate this cost, she manages to muster the willpower to eliminate the normatively worst alternatives according to her commitment preference in any menu. In contrast, the underlying psychology behind decision making in the other four models highlights the point that DMs may struggle to muster such willpower. The cost of exercising self-control that plays a central role in these models is that associated with having to resist temptation and the anxiety it produces. In other words, one way of seeing the difference between the two approaches is by noting that whereas in these four models, the primary driver is the psychological cost associated with exercising self-control, in our model it is the psychological cost associated with *not* exercising any self-control. In reality, when faced with self-control problems, both types of DMs presumably exist. Based on the above discussion, what is worth noting is that the primacy of one of these psychologi-

Nakajima, and Ozdenoren (2020) with, say, a willpower stock of $w = 3$ and the same values of $u$ and $v$, in the menu $\{x, y\}$, since the self-control cost of choosing $x$ is $v(y) - v(x) = 3 - 1 = 2 < 3$, both $x$ and $y$ are feasible to choose and, accordingly, $c(xy) = x$. Similarly both $y$ and $z$ are feasible in the menu $\{y, z\}$ and, accordingly, $c(yz) = y$. But, in the menu $\{x, z\}$, since $v(z) - v(x) = 4 > 3$, $x$ is not feasible. Hence, $c(xz) = z$, resulting in a violation of NBC. The same can be shown for the other two models.

[10] To see this, consider once again the convex self-control model of Noor and Takeoka (2010) and refer back to the soft drinks example from the last footnote with the following change. Let $v(z) = 6$ now with the other values for the $u$ and $v$ functions same as the ones used above. Then, it is straightforward to verify that utilities from choosing $x$, $y$ and $z$ in the menu $\{x, y, z\}$ are $-12$, $-2$ and 2, respectively. Accordingly, $c(xyz) = z$, the most tempting alternative in this menu.

cal drivers over the other produces different implications for the pattern of non-standard choices.

# 4 Behavioral characterization

An RSH can be behaviorally characterized by a single axiom that weakens WARP. The key to the weakening is to draw on the DM's choices to compare menus based on the normative standards they impose. We say that menu $T$ is revealed to impose as high a normative standard as menu $S$ if: $(i)$ $\bar{x} \in S \cup T$ s.t. $c(\bar{x}y) = \bar{x}$, for all $y \in S \cup T \implies \bar{x} \in T$; and $(ii)$ $\underline{x} \in S \cup T$ s.t. $c(\underline{x}y) = y$ for all $y \in S \cup T \implies \underline{x} \in S$. The basis of this elicitation in the context of a DM who avoids choosing the normatively least appealing alternatives in a menu is the following. Note that $\bar{x}$, the "always chosen" alternative, is revealed to be the normatively most appealing alternative in $S \cup T$ as there is no alternative $y$ in this set in whose presence $\bar{x}$ appears normatively unappealing in the menu $\{\bar{x}, y\}$ so as to be not chosen. On the other hand, $\underline{x}$, the "never chosen" alternative, is revealed to be the normatively least appealing alternative in $S \cup T$ as there is no alternative $y$ in this set which, in the presence of $\underline{x}$ in the menu $\{\underline{x}, y\}$, appears normatively unappealing so as to be not chosen. Accordingly, if the reference and context for normative standards in a menu is set by the normatively most appealing and least appealing alternatives in that menu, then $\bar{x} \in T$ and $\underline{x} \in S$ implies that the normatively best and worst alternatives of $T$ can be no worse than their respective counterparts in $S$; hence, $T$ is revealed to impose as high a normative standard as $S$. For any menu $S$, denote the set of all menus that are revealed to impose as high a normative standard as it by $\mathcal{U}(S)$. Note that $S \in \mathcal{U}(S)$.

We can now state the axiom that characterizes an RSH, which we refer to as WARP with norms.

**Axiom 4.1** (*WARP with norms*)**.** $\forall S, T, U \in \mathcal{X}$, *with* $U \in \mathcal{U}(T)$, *and* $x, y \in S \cap T$, $x \in U$, $x \neq y$,

$$\big[x = c(S), y = c(xy), x = c(\{c(U), x\})\big] \implies y \neq c(T)$$

The axiom imposes a similar restriction on the DM's behavior as WARP, once we take account of her concern about only making choices that are normatively acceptable. To elaborate on it, first, note that since $x$ is chosen in $S$, it is clearly normatively acceptable in this menu. Further, since $y = c(xy)$, we know that $y$ is at least as normatively appealing as $x$. Hence, if $x$ is normatively acceptable in $S$ so must be $y$. In other words, in the menu $S$, $x$ is chosen when $y$ is normatively acceptable. Further, since $x = c(\{c(U), x\})$, a similar argument establishes that $x$ is normatively acceptable in $U$.[11] In addition, $U$ is revealed

[11] Note that $x = c(\{c(U), x\})$ subsumes two cases: $(i)$ $x \neq c(U)$, and $(ii)$ $x = c(U)$, in which case $c(\{c(U), x\}) = c(\{x, x\}) = c(\{x\}) = x$.

to impose as high a normative standard as $T$. As such, if $x$ is normatively acceptable in $U$, it must be so in $T$ as well. Therefore, WARP-like consistency demands that $y$ cannot be the chosen alternative in $T$.

**Theorem 4.1.** *A choice function $c$ is an RSH if and only if it satisfies WARP with norms.*

**Proof:** Please refer to Section A.2.

# 5 Identification

We next address the question of how uniquely the parameters underlying an RSH can be identified. The commitment preferences of the DM can be uniquely identified from pairwise choice comparisons.

**Proposition 5.1.** *If $(\succcurlyeq_c, \succcurlyeq_t, g)$ and $(\tilde{\succcurlyeq}_c, \tilde{\succcurlyeq}_t, \tilde{g})$ are both RSH representations of a choice function $c$, then $\succcurlyeq_c = \tilde{\succcurlyeq}_c$.*

The proof is immediate since under an RSH, for any $x, y \in X$, $x \neq y$,

$$x \succcurlyeq_c y \iff y = g(xy) \iff x = c(xy) \iff y = \tilde{g}(xy) \iff x \tilde{\succcurlyeq}_c y$$

On the other hand, the extent of identification of the temptation preferences depends on the degree of intrapersonal conflict between the DM's commitment and temptation selves, as we show below. First note, given that the commitment preferences are uniquely identified, if the intrapersonal conflict involving these preferences is total with them going in exactly the opposite direction, then temptation preferences are also exactly identified. To understand the extent of identification when the intrapersonal conflict is not total, first, note that if $c$ is an RSH and for some menu $S$ and $x, y \in S$, $x \neq y$, $c(S) = x$ and $c(xy) = y$, then it clearly manifests an intrapersonal conflict involving $x$ and $y$. Specifically, $y \succcurlyeq_c x$ and $x \succcurlyeq_t y$ under any RSH representation $(\succcurlyeq_c, \succcurlyeq_t, g)$ of $c$. For any such $x, y$ with $c(xy) = y$ and $c(S) = x$, for some $S$ with $y \in S$, define, based on pairwise choice comparisons, the following set of alternatives that form a "preference interval" of commitment preferences, sandwiched between $x$ and $y$:[12]

$$X_{xy} = \{z \in X : c(yz) = y, c(xz) = z\}$$

The following result is the key to understanding the identification of temptation preferences under an RSH.

[12]Note that $x, y \in X_{xy}$. To see that $X_{xy}$ is a preference interval of commitment preferences, note that if $\succcurlyeq_c$ is the commitment preference relation under such an RSH $c$, then $X_{xy} = \{z \in X : y \succcurlyeq_c z \succcurlyeq_c x\}$.

**Proposition 5.2.** *Let $(\succcurlyeq_c, \succcurlyeq_t, g)$ and $(\tilde{\succcurlyeq}_c, \tilde{\succcurlyeq}_t, \tilde{g})$ be both RSH representations of a choice function $c$. If $x, y \in S \in \mathcal{X}$, $x \neq y$, such that $c(S) = x$ and $c(xy) = y$, then $\succcurlyeq_t = \tilde{\succcurlyeq}_t$, when restricted to the set $X_{xy} = \{z \in X : c(yz) = y, c(xz) = z\}$.*

**Proof:** Please refer to Section A.3.1

In other words, temptation preferences restricted to any such set $X_{xy}$, where $x$ and $y$ have an intrapersonal conflict, is uniquely identified. Accordingly, the greater are intrapersonal conflicts involving pairs of alternatives, especially, alternatives that are "distant" according to commitment preferences, the greater is the extent of identification of the DM's temptation preferences. In particular, the following is an immediate corollary of the above result, to present which we introduce the following notation. Let $\underline{z}$ and $\overline{z}$ be, respectively, the never chosen and always chosen alternatives of $X$, i.e., $c(x\underline{z}) = x$, for all $x \in X$ and $c(x\overline{z}) = \overline{z}$, for all $x \in X$. As noted above, any choice function that is an RSH satisfies the condition of NBC, which guarantees the existence of these alternatives for an RSH. Under such a choice function these alternatives are, respectively, the worst and the best alternative in $X$ according to the DM's commitment preferences.

**Corollary 5.1.** *Let $c$ be an RSH. Further, let $Y = \{y \in X : y = c(\underline{z}y\overline{z})\}$ and $\underline{y}$ be the never chosen alternative of $Y$, i.e., the worst alternative in $Y$ according to the DM's commitment preferences. Then, the DM's temptation preferences are identified uniquely, when restricted to the set $X_{\underline{y}\overline{z}}$. In particular, if $\underline{y}$ happens to be the never chosen alternative in the set $X \setminus \underline{z}$, i.e., the second worst alternative in $X$ according to commitment preferences, then temptation preferences are identified uniquely up to the set $X \setminus \underline{z}$.*

In other words, one way to analyze the question of identification of temptation preferences is to look at which alternatives from $X$ are chosen in three alternative menus consisting of itself, and the worst and the best alternatives in $X$ according to commitment preferences, $\underline{z}$ and $\overline{z}$, respectively. If $\underline{y}$ is the worst of such alternatives according to commitment preferences, then temptation preferences are exactly identified at least up to the set $X_{\underline{y}\overline{z}}$ consisting of $\underline{y}$ and all alternatives better than it according to commitment preferences. In particular, if it so happens to be the case that $\underline{y}$ is the second worst alternative according to commitment preferences, then temptation preferences are uniquely identified when restricted to the set $X \setminus \underline{z}$. That is, these preferences are identified almost uniquely with only the position of the worst alternative under commitment preferences being indeterminate under it.

Finally, note that beyond the exact identification of temptation preferences restricted to such preference intervals, further invariant inferences about these preferences can be made by building chains of such inferences. Let the binary relation $Q \subseteq X \times X$ denote the revealed temptation preference relation. Specifically, for any $x, y \in X$, $xQy$ if under

any RSH representation $(\succcurlyeq_c, \succcurlyeq_t, g)$ of $c$, we have $x \succcurlyeq_t y$; i.e., $Q$ captures the extent of identification of temptation preferences.

**Proposition 5.3.** *Let $c$ be an RSH. Then $xQy$ iff there exists sets $X_{\underline{z}_i\overline{z}_i} \subseteq X$, $i = 1, \ldots, k$, and $x = x_0, x_1, \ldots, x_{k-1}, x_k = y$, such that $x_{i-1}, x_i \in X_{\underline{z}_i\overline{z}_i}$ and $x_{i-1} \succcurlyeq_t x_i$, $i = 1, \ldots, k$, under any RSH representation $(\succcurlyeq_c, \succcurlyeq_t, g)$ of $c$.*

**Proof:** Please refer to Section A.3.2

As far as the identification of the normative elimination mapping is concerned, this too may not be identified exactly. But, for any menu $S$, we can provide bounds on the set $g(S)$, that for a rich enough set of outcomes can be quite tight. To that end, define for any $S \in \mathcal{X}^+$, the set:

$$D(c(S)) = \{x \in S : c(\{c(S), x\}) = c(S), x \neq c(S)\}$$

Since pairwise choice comparisons allow us to uniquely identify the DM's commitment preferences, the set $D(c(S))$ contains those alternatives in the menu $S$ that are worse according to these preferences than the chosen alternative from it, $c(S)$. Further, since $c(S)$ is not eliminated in $S$ and the ones that are must be worse than it according to these preferences, we can conclude that, under any RSH representation of $c$, $g(S) \subseteq D(c(S))$. Next, for any menu $S$, recall the collection $\mathcal{U}(S)$ of menus that are revealed to impose as high a normative standard as $S$. It is straightforward to verify that, under an RSH, the $\succcurlyeq_c$-worst and $\succcurlyeq_c$-best alternatives in any menu $T \in \mathcal{U}(S)$ are, respectively, no worse than the $\succcurlyeq_c$-worst and $\succcurlyeq_c$-best alternatives in $S$. In other words, for any such menu $S$, the intra-menu restriction that a normative elimination mapping imposes applies to the menus in $\mathcal{U}(S)$. Accordingly, we have $g(S) \cap T \subseteq g(T)$ for any such $T \in \mathcal{U}(S)$. Further, $g(T) \subseteq D(c(T))$. Hence, if $x \in g(S) \cap T$, then $x \in D(c(T))$. Now, define for any $x \in S$, $\mathcal{U}(S; x) = \{T \in \mathcal{U}(S) : x \in T\}$. Putting all of this together, therefore, allows us to conclude that in any RSH representation of $c$, if $x \in g(S)$ then $x \in \bigcap_{T \in \mathcal{U}(S;x)} D(c(T))$. Further, for any $S \in \mathcal{X}$, let:

$$Z(S) = \{z \in S : c(xz) = x, \forall x \in S \setminus z\}$$

It should be obvious that under any RSH representation of $c$, $Z(S)$ is a singleton and contained in $g(S)$. As such, the following result follows:

**Proposition 5.4.** *Let $c$ be an RSH. Then for any RSH representation $(\succcurlyeq_c, \succcurlyeq_t, g)$ of $c$ and any $S \in \mathcal{X}^+$, $Z(S) \subseteq g(S) \subseteq \{x \in S : x \in \bigcap_{T \in \mathcal{U}(S;x)} D(c(T))\}$.*

# A Appendix

## A.1 Proofs for Section 3: Menu effects and non-standard choices

### A.1.1 Proof of Proposition 3.1

Let $c$ be a choice function rationalized by any one of the four models, with $u$ and $v$ representing $\succcurlyeq_c$ and $\succcurlyeq_t$, respectively. Suppose towards a contradiction that there is a self-control enhancing menu effect under it, i.e., for some $\{x, y\} \subseteq S \subseteq T$, $x := c(T) \succcurlyeq_c c(S) =: y$, $x \neq y$; i.e., $u(x) > u(y)$. We know that if $\hat{z}_T := \text{argmax}_{z \in T}\, v(z) = \text{argmax}_{z \in S}\, v(z) =: \hat{z}_S$, then under any of these models we have no menu effects. So assume, $\hat{z}_T \succcurlyeq_t \hat{z}_S$, $\hat{z}_T \neq \hat{z}_S$; i.e., $v(\hat{z}_T) > v(\hat{z}_S)$.

$(i)$ Let $c$ be rationalized by Noor and Takeoka (2010). Then:

$$c(T) = x \implies u(x) - \varphi(v(\hat{z}_T) - v(x)) > u(y) - \varphi(v(\hat{z}_T) - v(y)) \tag{1}$$
$$c(S) = y \implies u(y) - \varphi(v(\hat{z}_S) - v(y)) > u(x) - \varphi(v(\hat{z}_S) - v(x)) \tag{2}$$

Equations (1) and (2) imply:

$$\begin{aligned} u(x) - u(y) &> \varphi(v(\hat{z}_T) - v(x)) - \varphi(v(\hat{z}_T) - v(y)) \\ u(x) - u(y) &< \varphi(v(\hat{z}_S) - v(x)) - \varphi(v(\hat{z}_S) - v(y)) \end{aligned}$$

Further, since $u(x) > u(y)$ and $\varphi$ is strictly increasing, we have $\varphi(v(\hat{z}_S) - v(x)) > \varphi(v(\hat{z}_S) - v(y)) \implies v(\hat{z}_S) - v(x) > v(\hat{z}_S) - v(y) \implies v(y) > v(x)$. Putting everything together, we have:

$$\begin{aligned} &\varphi(v(\hat{z}_T) - v(x)) - \varphi(v(\hat{z}_T) - v(y)) < \varphi(v(\hat{z}_S) - v(x)) - \varphi(v(\hat{z}_S) - v(y)) \\ \implies\ &\varphi(v(\hat{z}_T) - v(\hat{z}_S) + v(\hat{z}_S) - v(x)) - \varphi(v(\hat{z}_T) - v(\hat{z}_S) + v(\hat{z}_S) - v(y)) \\ &\qquad < \varphi(v(\hat{z}_S) - v(x)) - \varphi(v(\hat{z}_S) - v(y)) \end{aligned}$$

Letting $k = v(\hat{z}_T) - v(\hat{z}_S) > 0$, we then have:

$$\varphi(k + v(\hat{z}_S) - v(x)) - \varphi(k + v(\hat{z}_S) - v(y)) < \varphi(v(\hat{z}_S) - v(x)) - \varphi(v(\hat{z}_S) - v(y))$$

But for a strictly increasing and convex $\varphi$, with $\varphi(0) = 0$, and for any $k > 0$, $a > b \geq 0$, we must have $\varphi(k+a) - \varphi(k+b) \geq \varphi(a) - \varphi(b)$. Letting, $a = v(\hat{z}_S) - v(x)$ and $b = v(\hat{z}_S) - v(y)$, we arrive at a contradiction!

$(ii)$ Let $c$ be rationalized by Noor and Takeoka (2015). Then for any increasing $\psi \geq 0$,

$$c(T) = x \implies u(x) + \psi(v(\hat{z}_T))v(x) > u(y) + \psi(v(\hat{z}_T))v(y) \tag{3}$$

and

$$c(S) = y \implies u(y) + \psi(v(\hat{z}_S))v(y) > u(x) + \psi(v(\hat{z}_S))v(x) \tag{4}$$

Equations (3) and (4) imply:

$$u(x) - u(y) > \psi(v(\hat{z}_T))(v(y) - v(x))$$
$$u(x) - u(y) < \psi(v(\hat{z}_S))(v(y) - v(x))$$

Further, since $u(x) > u(y)$ and $\psi \geq 0$, we have $v(y) - v(x) > 0$. Putting everything together, we have:

$$u(x) - u(y) > \psi(v(\hat{z}_T))(v(y) - v(x)) \geq \psi(v(\hat{z}_S))(v(y) - v(x)) > u(x) - u(y)!$$

(*iii*) Let $c$ be rationalized by Masatlioglu, Nakajima, and Ozdenoren (2020). Since $u(x) > u(y)$ and $c(S) = y$, this implies $v(\hat{z}_S) - v(x) > w$, i.e., $x$ is not feasible in $S$. Since $v(\hat{z}_T) > v(\hat{z}_S)$, this implies $v(\hat{z}_T) - v(x) > w$. That is, $x$ is not feasible in $T$ and, hence, $c(T) \neq x$!

(*iv*) Let $c$ be rationalized by Liang, Grant, and Hsieh (2020). Since $c(T) = x$, we have $v(\hat{z}_T) - v(x) \leq w$. Accordingly, $v(\hat{z}_T) > v(\hat{z}_S) \implies v(\hat{z}_S) - v(x) \leq w$, i.e., $x$ is feasible in $S$. Since $c(S) = y$, this implies $u(y) + v(y) > u(x) + v(x)$. Accordingly, $0 < u(x) - u(y) < v(y) - v(x)$, which implies $v(y) > v(x)$. Then $v(\hat{z}_T) - v(y) \leq w$, i.e., $y$ is feasible in $T$. This implies $c(T) = y$!

### A.1.2 Proof of Proposition 3.2

Let $c$ be a choice function rationalized by any one of the four models. Let $\{x, y\} \subseteq S \subseteq T$, $c(xy) = c(T) = x$. To show these models satisfy weak WARP, we need to show $c(S) \neq y$. Suppose not. Since we have shown in Proposition 3.1 that none of these models can accommodate self-control enhancing menu effects, any menu effects under them must be self-control undermining ones. Then, $c(xy) = x, c(S) = y \implies u(x) > u(y)$ and $c(S) = y, c(T) = x \implies u(y) > u(x)$!

The following example shows that an RSH may violate weak WARP. Consider $X = \{x, y, z, w\}$ and the choice function $c$ specified in the table.

| | $xy$ | $xz$ | $xw$ | $yz$ | $yw$ | $zw$ | $xyz$ | $xyw$ | $xzw$ | $yzw$ | $xyzw$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $c(.)$ | $x$ | $x$ | $x$ | $y$ | $y$ | $z$ | $y$ | $y$ | $x$ | $z$ | $y$ |
| $g(.)$ | $y$ | $z$ | $w$ | $z$ | $w$ | $w$ | $z$ | $w$ | $zw$ | $w$ | $zw$ |

Clearly $c$ violates weak WARP as $c(yz) = c(xyzw) = y$ and $c(yzw) = z$. It is also straightforward to verify that with preferences $(\succcurlyeq_c, \succcurlyeq_t)$ given by $x \succcurlyeq_c y \succcurlyeq_c z \succcurlyeq_c w$ and

$z \succcurlyeq_t y \succcurlyeq_t x \succcurlyeq_t w$, and normative elimination mapping $g : \mathcal{X}^+ \to \mathcal{X}$ specified in the table, $c$ is an RSH.

Further, consider an RSH $c$; and $S, T \in \mathcal{X}$ with $\overline{z}_T = \overline{z}_S$, $\{x, y\} \subseteq S \subseteq T, c(xy) = c(T) = x$. Suppose towards a contradiction, $y = c(S)$. Since $c(S) = y$, we know $y \notin g(S)$. Further, $c(xy) = x$ implies $x \succcurlyeq_c y$. Hence, $x \notin g(S)$ and we have $y \succcurlyeq_t x$. Now, $\overline{z}_S = \overline{z}_T$ and, accordingly, $\overline{z}_S \succcurlyeq_c \overline{z}_T$. Further, since $S \subseteq T$, $\underline{z}_S \succcurlyeq_c \underline{z}_T$. Therefore, we have $g(T) \cap S \subseteq g(S)$. Since $x, y \notin g(S)$, it follows that $x, y \notin g(T)$. But then $y \succcurlyeq_t x \implies c(T) \neq x$!

### A.1.3 Proof of Proposition 3.3

We postpone the proof till next section. The proof follows from two conclusions that we establish there: (i) an RSH is characterized by the condition WARP with norms that we define in Section 4 (Theorem 4.1); and (ii) a choice function that satisfies WARP with norms also satisfies NBC and NC (Lemma A.2).

## A.2 Proofs for Section 4: Behavioral characterization

### A.2.1 Preliminaries

We first prove a few preliminary lemmas that we draw on in the proof of characterization. First, we define for any choice function $c$, a binary relation $P_c$ on $X$ as follows: For any $x, y \in X$, $x \neq y$, $xP_cy$ if $\exists S \in \mathcal{X}$ with $x, y \in S$ and $T \in \mathcal{U}(S)$, $y \in T$ s.t. $x = c(S)$ and $y = c(\{c(T), y\})$.[13] It is straightforward to establish that if $c$ satisfies WARP with norms, then $P_c$ is asymmetric.

**Lemma A.1.** *If a choice function $c$ satisfies WARP with norms, then $P_c$ is asymmetric.*

*Proof.* Suppose $xP_cy$ and $yP_cx$, for some $x, y \in X$, $x \neq y$. That is, there exists $S', T', S'', T''$ with $T' \in \mathcal{U}(S')$, $T'' \in \mathcal{U}(S'')$, $x, y \in S' \cap S''$, $x \in T''$, $y \in T'$ s.t. $x = c(S')$ and $y = c(\{c(T'), y\})$, and $y = c(S'')$ and $x = c(\{c(T''), x\})$. Now, either, $x = c(xy)$ or $y = c(xy)$, say, the former. Then, $x, y \in S' \cap S''$, $y = c(S'')$, $x = c(xy)$, $y = c(\{c(T'), y\})$, $y \in T' \in \mathcal{U}(S')$ but $x = c(S')$ implies a violation of WARP with norms. □

Recall the two choice consistency conditions referred to as no binary cycles (NBC) and never chosen (NC) defined earlier.

[13] As pointed out earlier in the main text, $y = c(\{c(T), y\})$ subsumes two cases: $(i)$ $y \neq c(T)$, and $(ii)$ $y = c(T)$, in which case $c(\{c(T), y\}) = c(\{y, y\}) = c(y) = y$.

**Definition A.1.** *A choice function $c : \mathcal{X} \to X$ satisfies:*

1. *NBC if for all $x_1, \ldots, x_{n+1} \in X$: $[c(x_i x_{i+1}) = x_i, i = 1, \ldots, n] \implies c(x_1 x_{n+1}) = x_1$*
2. *NC if for all $S \in \mathcal{X}$ and $x \in S$: $[x \neq c(xy), \forall y \in S \setminus x] \implies c(S) \neq x$*

We next establish that these two conditions follow from WARP with norms.

**Lemma A.2.** *If a choice function satisfies WARP with norms, then it satisfies NBC and NC.*

*Proof.* First, we show that if a choice function $c$ satisfies WARP with norms, then it satisfies NBC. We do so by proving the contrapositive, i.e., $c$ violating NBC implies WARP with norms is violated. The proof is by induction on the number of alternatives involved in the NBC violation, denote this number by $k$. First, consider the case of $k = 3$. Let $c(x_1x_2) = x_1$, $c(x_2x_3) = x_2$, $c(x_1x_3) = x_3$ and suppose w.l.o.g. that $c(x_1x_2x_3) = x_1$. Since $\{x_1, x_2, x_3\} \in \mathcal{U}(\{x_1, x_2, x_3\})$, $c(x_1x_2x_3) = x_1$ and $c(x_1x_3) = x_3$ implies $x_1 P_c x_3$. At the same time, since there is no always chosen or never chosen alternative in $\{x_1, x_2, x_3\}$, it trivially follows that $\{x_1, x_2, x_3\} \in \mathcal{U}(\{x_1, x_3\})$. Accordingly, $c(x_1x_3) = x_3$ and $c(x_1x_2x_3) = x_1$ implies $x_3 P_c x_1$. Hence, $P_c$ is not asymmetric and by Lemma A.1 it follows that $c$ violates WARP with norms. Now suppose the result has been proven for all $k \leq n-1$. That is any violation of NBC with $k \leq n-1$ alternatives implies a violation of WARP with norms. We wish to prove the same for $k = n$. To that end, let $c(x_1x_2) = x_1$, $c(x_2x_3) = x_2$, $\ldots$, $c(x_{n-1}x_n) = x_{n-1}$ and $c(x_1x_n) = x_n$. Now, either $(a)$ $c(x_1x_3) = x_3$ or $(b)$ $c(x_1x_3) = x_1$. If $(a)$, then $c(x_1x_2) = x_1$, $c(x_2x_3) = x_2$, $c(x_1x_3) = x_3$. This is a violation of NBC with 3 alternatives and the conclusion that WARP with norms is violated follows from the case of $k = 3$. If $(b)$, then $c(x_1x_3) = x_1$, $c(x_3x_4) = x_3$, $\ldots$, $c(x_{n-1}x_n) = x_{n-1}$ and $c(x_1x_n) = x_n$. This is a violation of NBC with $n-1$ alternatives and the conclusion that WARP with norms is violated follows from the case of $k = n-1$.

Next, we show that if $c$ satisfies WARP with norms then $c$ satisfies NC; or equivalently, if $c$ violates NC then it violates WARP with norms. So assume that for some menu $S$ and $x \in S$, $x \neq c(xy)$ for all $y \in S \setminus x$ and $c(S) = x$. Consider the menu $\{x, \hat{y}\}$ with $c(x\hat{y}) = \hat{y}$, for some $\hat{y} \in S \setminus x$. It is straightforward to see that $S \in \mathcal{U}(\{x, \hat{y}\})$ as $x$ is the never chosen alternative of $S \cup \{x, \hat{y}\} = S$.[14] Accordingly, $c(x\hat{y}) = \hat{y}$ and $c(S) = x$ implies that $\hat{y} P_c x$. At the same time, $S \in \mathcal{U}(S)$; and $c(S) = x$, $c(x\hat{y}) = \hat{y}$ implies $x P_c \hat{y}$. That is, $P_c$ is not asymmetric and by Lemma A.1 it follows that $c$ violates WARP with norms. □

Note that one consequence of a choice function $c$ satisfying NBC is that in any menu, an

[14] Of course, if $S \cup \{x, \hat{y}\} = S$ has an always chosen alternative, it must belong to $S$.

always chosen and never chosen alternative is guaranteed to exist, i.e., for any $S$, there exists $\bar{x}, \underline{x} \in S$ s.t. $c(\bar{x}y) = \bar{x}$, for all $y \in S$, and $c(\underline{x}y) = y$ for all $y \in S$.

Finally, we establish that $c$ satisfying WARP with norms is equivalent to the binary relation $P_c$ being acyclic.

**Lemma A.3.** *A choice function $c$ satisfies WARP with norms if and only if $P_c$ is acyclic.*

*Proof.* It is straightforward to see that $P_c$ acyclic implies WARP with norms is satisfied. To see this, suppose $x = c(S)$ and $y = c(xy)$, for some $S$ and $x, y \in S$. This implies $xP_cy$, since $S \in \mathcal{U}(S)$. Now, if there exists $T$ with $x, y \in T$, and $U \in \mathcal{U}(T)$ with $x \in U$ s.t. $x = c(\{c(U), x\})$, then clearly $y \neq c(T)$; otherwise, we have $yP_cx$, violating the acyclicity of $P_c$.

We next show that WARP with norms implies that $P_c$ is acyclic. To establish this, we show by induction that no $P_c$ cycle of length $k \geq 2$, i.e., one involving $k$ distinct alternatives, exists. It is straightforward to verify this for $k = 2$, since a $P_c$ two-cycle implies that $P_c$ is not asymmetric and by Lemma A.1, this implies a violation of WARP with norms. Hence, no $P_c$ two-cycle exists. For the inductive step, suppose no $P_c$ cycle involving $k \leq n$ alternatives exists. We need to establish that no such cycle with $n+1$ alternatives exists. Suppose not, say, $x_iP_cx_{i+1}$, for $i = 1, \ldots, n$ and $x_{n+1}P_cx_1$. That is, for all $i = 1, \ldots, n$, $j = i+1$, and $i = n+1, j = 1$, $\exists S_i$ with $x_i, x_j \in S_i$, and $T_i \in \mathcal{U}(S_i)$ with $x_j \in T_i$, s.t., $x_i = c(S_i)$ and $x_j = c(\{c(T_i), x_j\})$. Further, let $\underline{z} \in S_1 \cup \cdots \cup S_{n+1}$ be s.t. $c(z\underline{z}) = z, \forall z \in S_1 \cup \cdots \cup S_{n+1}$. Since, from Lemma A.2, we know that $c$ satisfies NBC when it satisfies WARP with norms, such a $\underline{z}$ exists. Let $S = \{x_1, \ldots, x_{n+1}, \underline{z}\}$. From Lemma A.2, we also know that $c$ satisfies $NC$. This implies $c(S) \neq \underline{z}$. Suppose w.l.o.g. that $c(S) = x_n$. Then note that $c(x_{n-1}x_n) \neq x_{n-1}$ as this would imply $x_nP_cx_{n-1}$, since $S \in \mathcal{U}(S)$. Along with $x_{n-1}P_cx_n$ we arrive at a $P_c$ two-cycle, which is ruled out. More generally, $x_m \neq c(x_mx_n)$, for any $m = 1, \ldots, n-1$. This is so because otherwise we have $x_nP_cx_m$ and this results in a $P_c$ cycle of $n - m + 1 \leq n$ alternatives, which is ruled out. This implies $x_n = c(x_mx_n)$, for $m = 1, \ldots, n-1$. Now in the menu $\{x_n, x_{n+1}\}$ we have two possibilities: either (a) $c(x_nx_{n+1}) = x_n$ or (b) $c(x_nx_{n+1}) = x_{n+1}$. If (a), then $c(x_nx) = x_n$ for all $x \in S$. Accordingly, $S_{n-1} \in \mathcal{U}(S)$, since the always chosen alternative of $S \cup S_{n-1}$ is either $x_n$ or some alternative not in $S$, i.e., in either case it belongs to $S_{n-1}$. On the other hand the never chosen alternative of $S \cup S_{n-1}$ is $\underline{z} \in S$. Then since $c(S_{n-1}) = x_{n-1}$, we have $x_nP_cx_{n-1}$, which along with $x_{n-1}P_cx_n$ contradicts the fact that there are no $P_c$ two-cycles. If (b), then by virtue of $c$ satisfying NBC, $c(x_{n+1}x) = x_{n+1}$ for all $x \in S$. Accordingly, arguing along similar lines as above, it can be verified that $S_{n+1} \in \mathcal{U}(S)$. Further, $T_{n+1} \in \mathcal{U}(S_{n+1})$. But, this means that $T_{n+1} \in \mathcal{U}(S)$. Given that $c(\{c(T_{n+1}, x_1\}) = x_1$, we therefore have that $x_nP_cx_1$, which results in a $P_c$ cycle of length $n$, contradicting the fact that no such cycle exits. Therefore, we can conclude that $P_c$ has no cycle involving $n+1$ alternatives. □

### A.2.2 Proof of Theorem 4.1

**Sufficiency**: Let $c : \mathcal{X} \to X$ satisfy WARP with norms, which is equivalent to the binary relation $P_c$ being acyclic. We show below that we can identify linear orders $\succcurlyeq_c$ and $\succcurlyeq_t$ on $X$ and a normative elimination mapping $g : \mathcal{X}^+ \to \mathcal{X}$ w.r.t. $\succcurlyeq_c$ such that the ordered pair $(\succcurlyeq_c, \succcurlyeq_t)$ and the mapping $g$ represent $c$ as an RSH.

Define $\succcurlyeq_c \subseteq X \times X$ as follows: for any $x, y \in X$, $x \succcurlyeq_c y$ if $x = c(xy)$. We establish that $\succcurlyeq_c$ is a linear order, i.e., $\succcurlyeq_c$ is:
Complete : $c(xy) \neq \emptyset$, for all $x, y \in X$, Thus, either $x \succcurlyeq_c y$ or $y \succcurlyeq_c x$.
Antisymmetric : Suppose, towards a contradiction, $x \succcurlyeq_c y$ and $y \succcurlyeq_c x$ for some $x \neq y$. Then by definition, $x = c(xy)$ and $y = c(xy)$!
Transitive : Let $x \succcurlyeq_c y$ and $y \succcurlyeq_c z$. If $x = y$ or $y = z$, there is nothing to prove. So assume these alternatives are distinct. From the definition of $\succcurlyeq_c$, we have $x = c(xy)$ and $y = c(yz)$. Since $c$ satisfies WARP with norms, it follows from Proposition A.2 that $c$ satisfies NBC. Accordingly, $x = c(xz)$, i.e., $x \succcurlyeq_c z$.

Next, to define the linear order $\succcurlyeq_t \subseteq X \times X$, start with the binary relation $P_c$. Let $P_c^*$ be the transitive closure of $P_c$. Since $P_c$ is acyclic, it follows that $P_c^*$ is a partial order. By Szpilrajn's theorem, we know that this partial order can be extended to a linear order. We define $\succcurlyeq_t$ as this linear order.

To define the normative elimination mapping $g : \mathcal{X}^+ \to \mathcal{X}$, first, define for any $S \in \mathcal{X}^+$ and $x \in S$, the following subset of $\mathcal{U}(S)$:

$$\mathcal{U}(S; x) = \{T \in \mathcal{U}(S) : x \in T\}$$

Further, for any $T \in \mathcal{X}^+$, let:

$$D(c(T)) = \{y \in T : c(\{c(T), y\}) = c(T), y \neq c(T)\}$$

Now, let

$$g(S) = \left\{ x \in S : x \in \bigcap_{T \in \mathcal{U}(S;x)} D(c(T)) \right\}$$

To establish that $g$ is a well defined normative elimination mapping w.r.t. the linear order $\succcurlyeq_c$ defined above, first, note that for any $S \in \mathcal{X}^+$, $c(S) \notin D(c(S))$ and, accordingly, $c(S) \notin g(S)$. Hence $g(S) \subsetneq S$. Next, note that since $c$ satisfies NBC, for any such menu $S$, there exists $\underline{x}_S$ such that $c(x\underline{x}_S) = x$ for all $x \in S$. Moreover, we know that $c$ satisfies NC as well. Hence, $c(S) \neq \underline{x}_S$. That is, $\underline{x}_S \in D(c(S)) \neq \emptyset$. Next consider $T \in \mathcal{U}(S; \underline{x}_S)$. This implies by definition of $\mathcal{U}$ that $c(x\underline{x}_S) = x, \forall x \in T$; and by NC that $\underline{x}_S \neq c(T)$. Hence, $\underline{x}_S \in D(c(T))$ and, accordingly, $\underline{x}_S \in \cap_{T \in \mathcal{U}(S;\underline{x}_S)} D(c(T))$. That is, $\underline{x}_S \in g(S) \neq \emptyset$.

Next we establish that for any $y, z \in S$, if $z \in g(S)$ and $z \succcurlyeq_c y$, i.e., $c(yz) = z$, then $y \in g(S)$. Suppose otherwise. This means that there exists $T \in \mathcal{U}(S; y)$ s.t. $y = c(\{c(T), y\})$. By NBC, $z = c(\{c(T), z\})$. Therefore, $z \notin T$, for otherwise $T \in \mathcal{U}(S; z)$ and $z \notin D(c(T))$, implying $z \notin g(S)$. Now consider the menu $T \cup z$. Note that $T \cup z \in \mathcal{U}(S; z)$, as $T \in \mathcal{U}(S)$. Further, $c(T), z \neq c(T \cup z)$, or else $z \notin D(c(T \cup z))$, implying $z \notin g(S)$. Finally, if $c(T \cup z) = w$ for some $w \neq z, c(T)$, then since $T \cup z \in \mathcal{U}(T)$ and $T \in \mathcal{U}(T \cup z)$,[15] we have $c(T) P_c w$ and $w P_c c(T)$! Accordingly, $x \in S \setminus g(S)$, $y \in g(S) \implies c(xy) = x$, or $x \succcurlyeq_c y$.

Finally, consider any $S, T \in \mathcal{X}^+$ s.t. $\bar{z}_T \succcurlyeq_c \bar{z}_S$ and $\underline{z}_T \succcurlyeq_c \underline{z}_S$.[16] It is straightforward to verify that his means that $T \in \mathcal{U}(S)$. Consider any $x \in g(S) \cap T$. Note that $\mathcal{U}(T; x) \subseteq \mathcal{U}(S; x)$. This means that $x \in \bigcap_{\hat{T} \in \mathcal{U}(S;x)} D(c(\hat{T})) \implies x \in \bigcap_{\hat{T} \in \mathcal{U}(T;x)} D(c(\hat{T})) \implies x \in g(T)$, i.e., $g(S) \cap T \subseteq g(T)$. Hence, we have established that $g$ is a well defined normative elimination mapping w.r.t. the linear order $\succcurlyeq_c$.

To show: $(\succcurlyeq_c, \succcurlyeq_t, g)$ is an RSH representation of $c$.

Pick any menu $S \in \mathcal{X}$ and let $x = c(S)$. First, note that since $x \notin D(c(S))$, therefore, $x \notin g(S)$. Now consider $y \in S$, $y \neq x$, such that $y \notin g(S)$. That is, there exists $T \in \mathcal{U}(S; y)$ s.t., $y \notin D(c(T))$. In other words, $y = c(\{c(T), y\})$. Hence, we have $x P_c y$. Finally, since $P_c \subseteq \succcurlyeq_t$, it follows that $x \succcurlyeq_t y$. Therefore, $c(S) = \{x \in S \setminus g(S) : x \succcurlyeq_t y, \forall y \in S \setminus g(S)\}$.

**Necessity**: Let $c : \mathcal{X} \to X$ be an RSH with parameters $(\succcurlyeq_c, \succcurlyeq_t, g)$. Consider $S \in \mathcal{X}$ and $x, y \in S$ s.t. $x = c(S)$ and $y = c(xy)$. Since $y = c(xy)$, we know $y \succcurlyeq_c x$. Further $x = c(S)$ implies $x \notin g(S)$ and accordingly, $y \notin g(S)$. This gives us $x \succcurlyeq_t y$. Now consider $T, U \in \mathcal{X}$ with $U \in \mathcal{U}(T)$, $x, y \in T$ and $x \in U$ s.t. $x = c(\{c(U), x\})$. This implies $x \notin g(U)$. Since $U \in \mathcal{U}(T)$, i.e., $\bar{z}_U \succcurlyeq \bar{z}_T$ and $\underline{z}_U \succcurlyeq \underline{z}_T$, we have $g(T) \cap U \subseteq g(U)$. Hence, $x \notin g(U)$ implies $x \notin g(T)$ and, accordingly, $y \notin g(T)$. Therefore, $y \neq c(T)$, since $x \succcurlyeq_t y$, establishing that $c$ satisfies WARP with norms.

## A.3 Proofs for Section 5: Identification

### A.3.1 Proof of Proposition 5.2

In the proof, we draw on the conclusion of Proposition 5.1 that commitment preferences are uniquely identified and, hence, $\succcurlyeq_c = \tilde{\succcurlyeq}_c$. In particular, for any $x', y' \in X$, $x' \succcurlyeq_c y'$ iff

[15] It is straightforward to verify that $T \cup z \in \mathcal{U}(T)$ and $T \in \mathcal{U}(T \cup z)$ as the always chosen and never chosen alternatives of $T \cup z$ and $T$ are the same. This follows since $z$ is neither the always chosen nor the never chosen alternative of of $T \cup z$. It is not the always chosen alternative of $T \cup z$ as the always chosen alternative of $T \cup S$ is in $T$ and $z \notin T$. It is not the never chosen alternative of $T \cup z$ as $z = c(\{c(T), z\})$.

[16] Recall that for any menu $R \in \mathcal{X}$, $\bar{z}_R$ and $\underline{z}_R$ denote the $\succcurlyeq_c$-best and $\succcurlyeq_c$-worst alternatives of $R$, respectively.

$x' = c(x'y')$ iff $x' \tilde{\succcurlyeq}_c y'$.

First, consider $x, y \in X_{xy}$. Clearly, $c(S) = x, y \in S$ and $c(xy) = y$ implies that $y \succcurlyeq_c x$, $y \in S \setminus g(s)$ and, accordingly, $x \succcurlyeq_t y$, under any RSH representation of $c$. Next, consider any $a, b \in X_{xy}$, $a \neq b$, and assume w.l.o.g. that $c(ab) = a$, so that $y \succcurlyeq_c a \succcurlyeq_c b \succcurlyeq_c x$ and $y \tilde{\succcurlyeq}_c a \tilde{\succcurlyeq}_c b \tilde{\succcurlyeq}_c x$.[17] If there exists $S'$ with $a, b \in S'$ and $c(S') = b$, then it follows that $b \succcurlyeq_t a$ under any RSH representation of $c$. On the other hand, if no such $S'$ exists, consider the menu $S'' = \{a, b, \underline{z}\}$, where $\underline{z}$ is the worst alternative in $X$ according to $\succcurlyeq_c$ $= \tilde{\succcurlyeq}_c$. Clearly, $c(S'') \neq \underline{z}$. Further, $c(S'') \neq b$ since no such $S''$ exists. Accordingly, $c(S'') = a$. Now, let $T = \{x, y, b, \underline{z}\}$. It cannot be the case that $c(T) = y$. To see this note that $\overline{z}_S \succcurlyeq_c \overline{z}_T \equiv y$, and $\underline{z}_S \succcurlyeq_c \underline{z}_T \equiv \underline{z}$. Accordingly, by the property of the normative elimination mapping, since $c(S) = x \in S \setminus g(S)$, we have that $x \in T \setminus g(T)$ under any RSH representation. But, if $c(T) = y$, this would imply that $y \succcurlyeq_t x$ under any RSH representation! Hence, $c(T) \neq y$. Clearly, $c(T) \neq \underline{z}$. Accordingly, $c(T) = x$ or $c(T) = b$. Since, $b \succcurlyeq_c x$, in either case, $b \in T \setminus g(T)$ under any RSH representation. Finally, note that $y \equiv \overline{z}_T \succcurlyeq_c \overline{z}_{S''} = a$ and $\underline{z} \equiv \underline{z}_T = \underline{z}_{S''}$. Hence, by the property of the normative elimination mapping, $b \in T \setminus g(T)$ implies $b \in S'' \setminus g(S'')$ and, accordingly, $a \succcurlyeq_t b$ under any RSH representation of $c$. Therefore, temptation preferences restricted to the set $X_{xy}$ are uniquely identified.

### A.3.2 Proof of Proposition 5.3

*If direction*: First, consider $x = x_0, x_1, \ldots, x_{k-1}, x_k = y$, such that $x_{i-1}, x_i \in X_{\underline{z}_i \overline{z}_i}$ and $x_{i-1} \succcurlyeq_t x_i$, $i = 1, \ldots, k$, under any RSH representation $(\succcurlyeq_c, \succcurlyeq_t, g)$ of $c$. Since any such $\succcurlyeq_t$ is transitive, we have that $x \succcurlyeq_t y$ under any RSH representation. Hence, $xQy$.

*Only if direction*: Next, consider any $x, y$ with $xQy$. Recall the binary relation $P_c$ defined above. Let $P_c^*$ be the transitive closure of $P_c$. The proof proceeds in two steps.

*First*, we establish that $xQy \implies xP_c^*y$. Suppose $\neg[xP_c^*y]$. Then, the following two cases are possible: Either $yP_c^*x$ or $\neg[yP_c^*x]$. Consider the first case. Since $P_c^*$ is the transitive closure of $P_c$, $yP_c^*x$ implies that there exists a sequence $(y_i)_{i=1}^k$ in $X$ such that $yP_cy_1$, $y_1P_cy_2$, $\ldots$, $y_kP_cx$. Further, for any $\succcurlyeq_t$ that is part of an RSH representation, we have $P_c \subseteq \succcurlyeq_t$ and $\succcurlyeq_t$ is transitive. Therefore, it follows that $y \succcurlyeq_t x$ under any RSH representation. In the second case, where $\neg[yP_c^*x]$, there exists no sequence $(y_i)_{i=1}^k$ in $X$ such that $yP_cy_1$, $y_1P_cy_2$, $\ldots$, $y_kP_cx$. In this case it is possible to extend $P_c^*$ to a linear order $\succcurlyeq_t$ under which $y \succcurlyeq_t x$. The proof of Theorem 4.1 establishes that any such linear order can be part of an RSH representation. Therefore, in either case, $\neg[xQy]$. Hence, $xQy$ implies that there exists $x = x_0, x_1, \ldots, x_{k-1}, x_k = y$, such that $x_{i-1}P_cx_i$, $i = 1, \ldots, k$.

[17] The case of $a = y$ and $b = x$ has already been covered above. The proof below subsumes the cases when either $a = y$ or $b = x$.

*Second*, we establish that if $x_{i-1}P_c x_i$, then there exists $X_{\underline{z}_i\overline{z}_i} \subseteq X$ with $x_{i-1}, x_i \in X_{\underline{z}_i\overline{z}_i}$. Since $x_{i-1}P_c x_i$, there exists $S$ with $x_{i-1}, x_i \in S$, $x_{i-1} = c(S)$ and $x_i = c(x_i c(T))$, for some $T \in \mathcal{U}(S)$ with $x_i \in T$. If $c(x_{i-1}x_i) = x_i$, then we can take $T = S$. On the other hand, if $c(x_{i-1}x_i) = x_{i-1}$, then some $T \neq S$ exists. Let $\overline{z}_T$ be the always chosen alternative of $T$. Then, since $T \in \mathcal{U}(S)$, in either case it follows that $x_{i-1}, x_i \in X_{c(T)\overline{z}_T}$. That is, our desired conclusion follows with $\underline{z}_i = c(T)$ and $\overline{z}_i = \overline{z}_T$. Since for any $\succcurlyeq_t$ that is part of an RSH representation, we have $x_{i-1} \succcurlyeq_t x_i$, we can therefore conclude that $xQy$ implies there exists sets $X_{\underline{z}_i\overline{z}_i} \subseteq X$, $i = 1, \ldots, k$, and $x = x_0, x_1, \ldots, x_{k-1}, x_k = y$, such that $x_{i-1}, x_i \in X_{\underline{z}_i\overline{z}_i}$ and $x_{i-1} \succcurlyeq_t x_i$, $i = 1, \ldots, k$, under any RSH representation $(\succcurlyeq_c, \succcurlyeq_t, g)$ of $c$.

## References


Cherepanov, Vadim, Timothy Feddersen, and Alvaro Sandroni. 2013. "Rationalization." *Theoretical Economics* 8 (3):775–800.

Fudenberg, Drew and David K Levine. 2006. "A dual-self model of impulse control." *American Economic Review* 96 (5):1449–1476.

Gul, Faruk and Wolfgang Pesendorfer. 2001. "Temptation and self-control." *Econometrica* 69 (6):1403–1435.

Kreps, David M. 1979. "A representation theorem for "Preference for Flexibility"." *Econometrica* 47 (3):565–577.

Liang, Meng-Yu, Simon Grant, and Sung-Lin Hsieh. 2020. "Costly self-control and limited willpower." *Economic Theory* 70 (3):607–632.

Lipman, Barton L and Wolfgang Pesendorfer. 2013. "Temptation." In *Advances in economics and econometrics: Tenth World Congress*, vol. 1. Citeseer, 243–288.

Manzini, Paola and Marco Mariotti. 2007. "Sequentially rationalizable choice." *American Economic Review* 97 (5):1824–1839.

———. 2012. "Categorize then choose: Boundedly rational choice and welfare." *Journal of the European Economic Association* 10 (5):1141–1165.

Masatlioglu, Yusufcan, Daisuke Nakajima, and Emre Ozdenoren. 2020. "Willpower and compromise effect." *Theoretical Economics* 15 (1):279–317.

Noor, Jawwad and Norio Takeoka. 2010. "Uphill self-control." *Theoretical Economics* 5 (2):127–158.

———. 2015. "Menu-dependent self-control." *Journal of Mathematical Economics* 61:1–20.

Sharpe, Kathryn M, Richard Staelin, and Joel Huber. 2008. "Using extremeness aversion to fight obesity: Policy implications of context dependent demand." *Journal of Consumer Research* 35 (3):406–422.

Strotz, Robert Henry. 1955. "Myopia and inconsistency in dynamic utility maximization." *Review of Economic Studies* 23 (3):165–180.