Economic Theory (2020) 70:607–632
https://doi.org/10.1007/s00199-019-01231-6

RESEARCH ARTICLE

Costly self-control and limited willpower

Meng-Yu Liang[1] · **Simon Grant**[2] · **Sung-Lin Hsieh**[3]

Received: 24 February 2019 / Accepted: 30 September 2019 / Published online: 22 October 2019

Abstract

In Gul and Pesendorfer (Econometrica 69(6):1403–1435, 2001), a decision-maker, when facing a choice among menus, evaluates each menu in terms of the maximum value of its commitment utility net of self-control costs. This paper extends the model such that this maximum is constrained by the condition that the cost of self-control cannot exceed the decision-maker's stock of willpower w. Four of the five axioms of our characterization are as in their Theorem 3 except that the independence axiom is restricted to a subset of menus. We add one new axiom to regulate willpower as a limited (cognitive) resource in which the available "stock" does not vary across menus. In our characterization, choices within menus that satisfy WARP reveal a constant trade-off between commitment and temptation utilities. However, it is the discontinuity of preferences over menus (along with *violations* of WARP for choices within menus) that *reveals* w (*measured in units of temptation utility*), allowing for a behaviorally meaningful comparative measure of self-control *across* individuals.

Keywords Temptation · Self-control · Willpower · Revealed preference

JEL Classifications D81 · D91 · D11

Liang would like to acknowledge financial support from Taiwan's Ministry of Science and Technology under Grant MOST 105-2410-H-001-012.

✉ Simon Grant
simon.grant@anu.edu.au

Meng-Yu Liang
myliang@econ.sinica.edu.tw

Sung-Lin Hsieh
slhsieh@umich.edu

1 Institute of Economics, Academia Sinica, Taipei, Taiwan

2 Australian National University, Canberra, ACT, Australia

3 Department of Economics, University of Michigan, Ann Arbor, MI, USA

1 Introduction

Self-control, as the psychologists Muraven and Baumeister (2000) put it, "involves overriding or inhibiting competing urges, behaviors or desires." Gailliot and Baumeister (2007) describe it as "the conscious and effortful form of self-regulation." Furthermore, they report physiological research that indicates exercise of such self-control relies on some sort of *limited* energy source. This can be viewed as providing a physiological foundation to the metaphoric concept of an individual's (cognitive) resource of *willpower*. In this paper we propose the following representation for preferences over menus of lotteries that captures this notion of a limited stock of willpower:

$$\begin{aligned} U(A) = &\max_{x \in A} [u(x) + v(x)] - \max_{y \in A} v(y), \\ &\text{s.t. } v(x) \geq \max_{y \in A} v(y) - w, \end{aligned} \tag{1}$$

where A is a (compact) set of lotteries with generic elements denoted by x and y, u and v are von Neumann–Morgenstern utility functions over lotteries that describe the individual's commitment ranking and temptation ranking, respectively, and a nonnegative number w, that measures the individual's *stock of willpower*.

This is the representation characterized by Gul and Pesendorfer (2001) (hereafter, GP) *with the addition of a willpower constraint requiring just the one additional parameter* w. We interpret it as saying that the individual anticipates when she comes to make a choice from a menu she has previously selected, she does so by choosing the alternative from that menu that maximizes the "compromise utility" (that is, the sum of the commitment and temptation utilities), subject to its temptation utility being at least within w of the temptation utility of the most tempting alternative. Denoting this element of A by x^*, the "utility" of the menu which guides her initial choice over menus is then given by the commitment utility $u(x^*)$ less the amount $\max_{y \in A} v(y) - v(x^*)$, which like GP, we interpret as the (utility) cost of self-control. Thus, the stock of willpower w represents the *upper bound* on the self-control cost the individual is able to bear. We refer to the family of preferences over menus that admit a representation of the form given in (1) as *self-control preferences with limited willpower.*

We observe that temptation preferences (with and without self-control) characterized by Theorem 3 of GP may be viewed as the following two polar sub-families of preferences that admit a representation of the form given in (1).[1]

(i) The stock of willpower w is sufficiently large so that it never binds, that is, *unlimited willpower.*

(ii) The stock of willpower w equals zero, that is, *no self-control.* This is readily seen to be equivalent to the Strotz (1955) model of (overwhelming) temptation where the representation may be expressed in the form:

$$U(A) = \max_{x \in A} u(x) \text{ s.t. } v(x) \geq v(y), \text{ for all } y \in A.$$

[1] Throughout we refer to the representation result in Theorem 3 of GP (p1413), rather than Theorem 1 (p1409) which is the one usually cited in the literature.

One of the main contributions of this paper is our representation result, Theorem 1, that provides a characterization of the intermediate case in which w is strictly positive and the willpower constraint is strictly binding for at least some menus. The principal distinction between our extension and the original GP model is that there is now an additional reason a decision-maker might give in to temptation. In GP's model, the decision-maker selects a more tempting alternative only if its compromise utility exceeds that of all other alternatives in the menu. In our model, the decision-maker may also select a more tempting alternative even though there is another alternative that has higher compromise utility because the self-control cost incurred by selecting that alternative with the higher compromise utility exceeds her willpower.

Furthermore, unlike the two polar sub-families of GP's Theorem 3, not only can this intermediate case accommodate violations of the weak axiom of revealed preferences (WARP) by the decision-maker in the "hot" state, it is also possible to classify different choice patterns by the decision-maker as manifestations of distinct types of addictive behavior. We demonstrate this by means of the examples presented in Sect. 2.

While choices within menus that satisfy WARP reveal a constant trade-off between the commitment utility u and the temptation utility v, we wish to emphasize, it is the *discontinuity* of preferences over menus along with *violations* of WARP for choices within menus that reveals the limit w (*measured in units of temptation utility*) of an individual's willpower in exercising costly self-control. Hence, in our characterization, it is possible to obtain a behaviorally meaningful comparative measure of self-control *across* individuals.

We also make a technical contribution by clarifying the role of the mixture space theorem in obtaining the (cardinally unique) utility function U defined over menus.[2] As we shall see in the sequel, a setting with limited willpower naturally leads to violations of the independence axiom. Hence, we cannot directly invoke the mixture space theorem over the space of all menus. Instead, we consider a particular subset of menus for which the preference relation restricted to this subset satisfies independence. Exploiting this property of independence along with the set betweenness axiom, we are able to construct a mixture operator that when paired with this subset of menus makes it a mixture space. Now the mixture space theorem can be applied to obtain the existence of a utility function defined over this subset of menus that is linear with respect to our mixture operator and unique up to positive affine transformations. We also introduce a new axiom to regulate the boundary of self-control across different menus which leads to a menu-invariance willpower capacity represented by the parameter w.

1.1 Related literature

The characterization of self-control preferences has inspired a number of extensions and generalizations. One strand maintains the independence axiom and relaxes set betweenness to allow for agents who are either uncertain about the temptation they

[2] In his characterization of finite additive utility representations for preferences over menus, Kopylov (2009) provides further clarification of the role of the mixture space theorem.

will face in the future and/or may face multiple temptations.[3] Another strand retains set betweenness while relaxing independence. For example, Noor and Takeoka (2010) characterize two models, a general self-control representation and a convex self-control representation. At first glance, our representation might be viewed as the limit of a sequence of convex continuous self-control representations that converge to a self-control cost that has constant marginal cost up to the limit w after which the "marginal cost" explodes to infinity. However, even though we also retain set betweenness and relax independence, we also relax continuity. As we shall see in the sequel, it is this *discontinuity* of the preferences over menus that enables us to identify and calibrate the parameter w, which is the upper bound on the (utility) costs associated with exercising self-control she is able to incur. There is no equivalent feature in either of their models.

As we noted above, our characterization can accommodate violations of WARP. However, unlike other extensions of GP that can accommodate violations of WARP, we do so without requiring either a hot–cold empathy gap or a menu dependent trade-off between commitment and temptation utilities.[4] Hence, from the ex ante point of view, as was the case in the GP model, there is only a single preference relation in the hot state that will be revealed ex post by the choices made within menus. Moreover, we retain GP's one value system by measuring not only the temptation utility in terms of the commitment utility but also the stock of willpower in terms of the temptation utility. This allows us to obtain in our Theorem 3, conditions under which one decision-maker may be said to be able to exert more self-control than another.

To the best of our knowledge, Masatlioglu et al. (2014) was the first paper to explain violations of WARP in the hot state by a willpower constraint. They also confine their ex post choices from menus that are guided by a single preference relation over all possible "prizes" in the hot state. However, as there is no cost for self-control in their model, there exists no trade-off between the commitment and temptation utilities.

1.2 Roadmap

The rest of the paper is organized as follows. In Sect. 2 we provide a graphical depiction of the mechanics of the costly self-control with limited willpower representation to illustrate how we can accommodate behaviors not possible in the GP model and its extant extensions. The framework of GP's preference model over menus is formally presented in Sect. 3. In Sect. 4 we begin with four axioms similar to those that appear in GP, except the domain of independence is suitably restricted. We also present and motivate our additional axiom that ensures the revealed stock of willpower does not vary across menus, leading to the main representation result in which the stock of willpower is strictly positive and the willpower constraint is strictly binding for at least some menus. We also demonstrate how the discontinuity in the preferences over menus along with violations of WARP for the subsequent choices of lotteries

[3] See, for example, Chatterjee and Vijay Krishna (2009), Dekel et al. (2009), Stovall (2010), Kopylov (2009) and Kopylov (2012).

[4] For a discussion about the hot–cold empathy gap in decision-making, see Loewenstein (2000). The possibility of learning leading to perfect foresight of this gap is considered by Bénabou and Tirole (2004) and Ali (2011). For an extension allowing for uncertainty about the subjective state of temptation, see Dekel et al. (2009) and Stovall (2010).

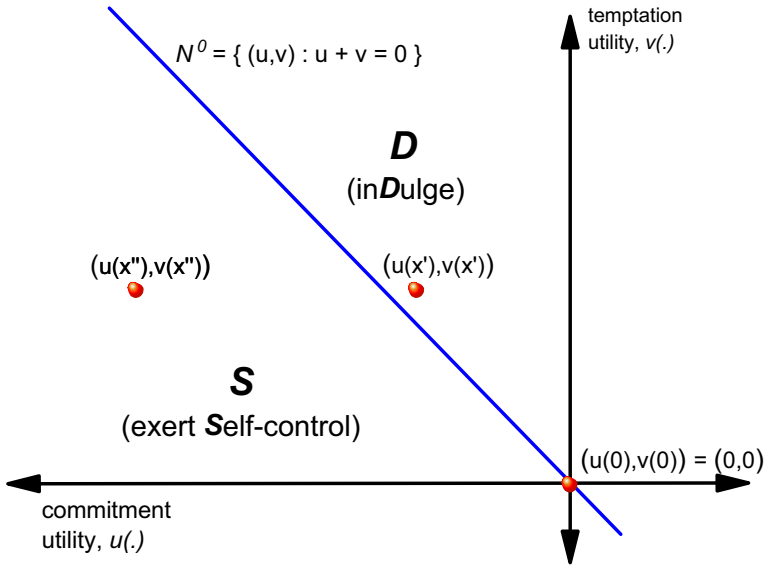

Fig. 1 Normative-temptation utility pairs

within a menu allow us to calibrate the stock of willpower. This in turn allows us in Sect. 5 to define a behaviorally meaningful comparative measure of self-control across individuals and provide a partial characterization of this comparative measure in terms of the representation. We conclude in Sect. 6.

2 A graphical depiction of the willpower constraint; compulsive behavior (both dissonant and consonant); and comparative self-control

Consider a decision-maker, say Sandra, in a "cold" state at night contemplating purchasing a cup of coffee on her way to work when she will be in a "hot" state craving coffee. Let $\{0, x'\}$ denote a two-element menu where 0 (respectively, x') is the degenerate lottery of purchasing and consuming with probability one a zero (respectively, some strictly positive) quantity of coffee. Without loss of generality, let us normalize the normative and temptation utilities so that $(u(0), v(0)) = (0, 0)$. Figure 1 depicts the normative-temptation utility pairs $(u(x), v(x))$ for other lotteries x that in a two-element menu $\{0, x\}$, are worse in terms of their normative utility as well as being more tempting than 0, that is, $u(x) < 0 < v(x)$.

The level set $N^0 = \{(u, v) : u + v = u(0) + v(0)\}$ divides this set of utility pairs into two regions: D, comprising those pairs whose sum is greater than zero; and, S, comprising those pairs whose sum is less than or equal to zero. Since

$$u(x) + v(x) > 0 = u(0) + v(0) \Longleftrightarrow u(x) > u(0) - (v(x) - v(0)),$$

in GP's model, region D (respectively, region S) corresponds to utility pairs associated with those tempting alternatives for which Sandra is *not* willing (respectively, is willing) to incur the self-control cost required to resist choosing that tempting alternative. That is, if the utility pair $(u(x), v(x))$ associated with the tempting alternative x lies in region D, then Sandra in the hot state will choose x from the menu $\{0, x\}$. We refer to this as "self-indulgent" consumption. Alternatively, if the utility pair associated with x lies in S, then Sandra in the hot state will choose to exercise (costly) self-control by selecting 0 from the menu $\{0, x\}$.

However, according to Elster and Skog (1999), one common criterion for a decision-maker to be considered prone to compulsive consumption (or "addicted"), is for the decision-maker to exhibit an *inability to resist* choosing to consume it in the hot state. A decision-maker prone to compulsive consumption in this sense, may in turn be classified into one of two sub-groups, dissonant or consonant. A dissonant compulsive consumer is one who wishes to resist the tempting alternative while a consonant compulsive consumer is one who does not. Neither type of compulsive consumer, however, can be modeled in GP's framework nor in any of the existing extensions like Noor and Takeoka (2010).

To see how they can be modeled in our framework, suppose lottery x', associated with the utility pair $\left(u\left(x'\right), v\left(x'\right)\right)$ in Fig. 1, is the lottery that assigns probability one to the outcome "paying \$1 to consume one cup of coffee" and lottery x'', associated with the utility pair $\left(u\left(x''\right), v\left(x''\right)\right)$, is the lottery that assigns probability one to the outcome "paying \$10 to consume one cup of coffee." From Fig. 1, we see that the reduction in normative utility arising from the increase in the price of coffee is enough to induce Sandra in the hot state to switch from choosing to consume the cup of coffee to abstaining. Indeed, provided there is always a way to reduce the normative utility sufficiently, say by increasing the monetary cost required to undertake the tempting activity, then any lottery associated with a utility pair in region D can be modified in such a way that leads to the utility pair associated with the modified lottery residing in region S. That is, in the model of GP, Sandra in the hot state chooses self-indulgent consumption because it provides a greater compromise utility not because she is unable to resist temptation.

However, self-control preferences *with limited willpower* allow for consumption choices that can be classified as having been made by a compulsive consumer, both dissonant and consonant. In the context of the consumption of coffee example, Fig. 2 is constructed from Fig. 1 by adding the DM's willpower constraint. This constraint says the DM cannot resist selecting a tempting alternative x from the menu $\{0, x\}$ if the cost of self-control required to select 0 (that is, the temptation utility $v(x)$) exceeds w her stock of willpower. This creates a new region I that lies between regions D and S, in which the horizontal ray $v(x) = w$ going left from the point $(-w, w)$ on the line N^0 becomes part of the boundary of the modified region S.

The interpretations for the regions D and S are the same as above, except notice that for some points in D, like the one associated with the lottery x', even though the inequality

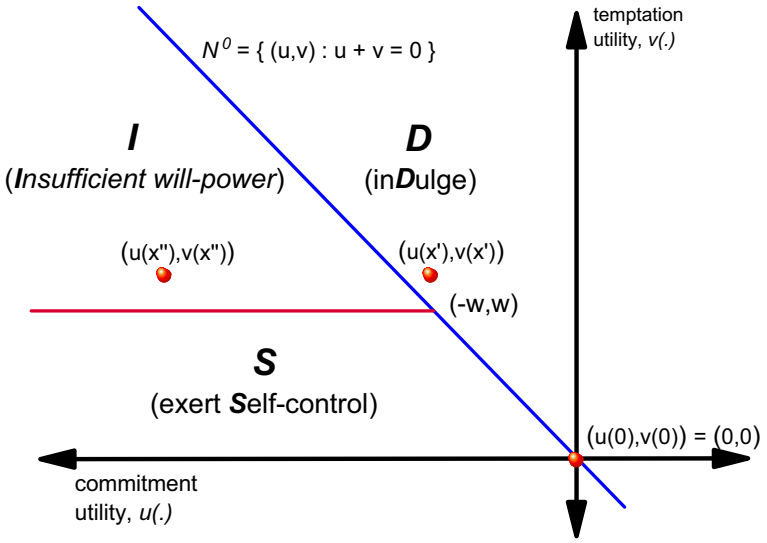

Fig. 2 Normative-temptation utility pairs with willpower constraint

$$u\left(x'\right) > u\left(0\right) - \left(v\left(x'\right) - v\left(0\right)\right),$$

still implies that Sandra *wants* to choose x' from the menu $\{0, x'\}$, in actual point of fact she is *unable* to select 0, since $v\left(0\right) \left(= 0\right) < v\left(x'\right) - w$. That is, the self-control cost required to select 0 in the presence of the tempting alternative x' exceeds her available stock of willpower w. In the sense of Elster and Skog (1999) Sandra is a *consonant* compulsive consumer of coffee.

In contrast to region D, region I may be viewed as those tempting alternatives for which Sandra wants to *but has insufficient willpower to be able to exert self-control*. We call the choice of such a tempting alternative in the hot state as the act of a "wretched man," evoking Saint Paul's sentiment,

> "I do not do the good I want, but the evil I do not want is what I do." (Romans 7:19).

As an example, the utility pair $\left(u\left(x''\right), v\left(x''\right)\right)$ associated with the lottery x'' which was previously in region S in Fig. 1, now resides in region I in Fig. 2. That is, despite the compromise utility of x'' being less than zero, Sandra is unable to choose 0 because $v\left(x''\right)$, the cost of exerting the self-control required for such a choice, exceeds w, her stock of willpower. She will thus pay \$10, albeit reluctantly, for that one cup of irresistibly tempting coffee! In the sense of Elster and Skog (1999) her behavior in choosing x'' from the menu $\{0, x''\}$ is that of a *dissonant* compulsive consumer of coffee.

Our framework also provides a natural way to compare the relative degree of self-control for two self-control preference relations with limited willpower. These may correspond to the two relations of two *different* individuals or the *same* decision-maker at different points in time or in different situations. Fixing a pair of self-control

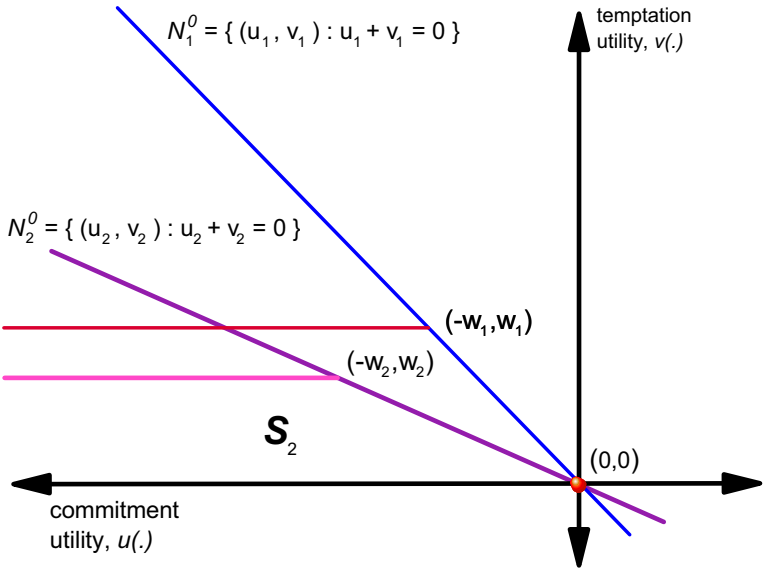

Fig. 3 Normative-temptation utility pairs of decision-maker 2 who has less self-control than decision-maker 1

preference relations with limited willpower, $\succsim_1$ and $\succsim_2$, and an alternative 0, let S_1 (respectively, S_2) denote the region where decision-maker 1 (respectively, decision-maker 2) is willing and able to exercise self-control by selecting 0 over the tempting alternative. In Sect. 5 we provide a behavioral definition for one preference relation $\succsim_1$ to be said to exhibit more self-control than another preference relation $\succsim_2$ and show that it holds whenever S_2 is a subset of S_1. As Fig. 3 illustrates, there are two *distinct* parametric changes that can make S_2 a subset of S_1: either $w_2 \leq w_1$ or the slope of the line $N_2^0 = \{(u_2, v_2) : u_2 + v_2 = 0\}$ is flatter than the slope of the line $N_1^0 = \{(u_1, v_1) : u_1 + v_1 = 0\}$. Intuitively, the first corresponds to a tightening of the willpower constraint, while the second to temptation utility receiving more weight in the compromise utility.

Baumeister et al. (1994) and Roy and Baumeister (2003), among others, have demonstrated that individuals who perform an act requiring self-control in one experiment tend to behave in the follow-up experiment as if they have less self-control. Some psychologists have interpreted this as providing experimental evidence of depletion of willpower as the cause of the reduction in self-control. Viewing $\succsim_2$ as the preferences of the same individual at some later point in time, our Fig. 3 suggests there are at least two distinct channels by which self-control might be diminished via a prior action: *(i)* a reduction in w, that is, a depletion in the stock of available willpower; and *(ii)* a change in the compromise utility that alters the trade-offs the decision-maker is willing to make in resisting a tempting alternative.[5] Hence, our example suggests a need for

[5] It is this second channel that "erodes" the self-control of an individual in Gul and Pesendorfer (2007) for their model of addiction via compulsive consumption.

more nuanced designs of experiments that are able to distinguish between these two channels.

3 Framework and definitions

We consider a two-period decision problem similar to the setting in GP. Let (Z, d) be a compact metric space, where Z is the set of (final) prizes (or consequences). Let $\Delta(Z)$ denote the set of all lotteries defined on Z, with generic elements x, y, a, b, et cetera. That is, $\Delta(Z)$ may be taken to be the set of all probability measures on $\mathscr{B}_Z$, the Borel σ-algebra of subsets of Z. We endow $\Delta(Z)$ with the weak topology generated by some metric d_p. As is standard, for any pair of lotteries x, y in $\Delta(Z)$ and any α in $[0, 1]$, let $\alpha x + (1-\alpha) y$ denote the lottery that assigns to each subset of prizes $B \in \mathscr{B}_Z$, the probability $\alpha x(B) + (1-\alpha) y(B)$.

Let $\mathscr{A}$ denote the set of menus which we take to be the set of all compact subsets of $\Delta(Z)$ with generic elements A, B, C. We endow $\mathscr{A}$ with the (Hausdorff) topology generated by the metric

$$d_h(A, B) = \max\left\{\max_{x\in A}\min_{y\in B} d_p(x, y), \max_{x\in B}\min_{y\in A} d_p(x, y)\right\}.$$

For any pair of menus A, B in $\mathscr{A}$ and any α in $[0, 1]$, let $\alpha A + (1-\alpha) B$ denote the menu in $\mathscr{A}$ given by $\{\alpha x + (1-\alpha) y : x \in A, y \in B\}$.

The preferences $\succsim$ of the decision-maker (hereafter, DM) are defined on $\mathscr{A}$. As is standard, $\succ$ (respectively, $\sim$) denotes the asymmetric (respectively, symmetric) parts of $\succsim$. We consider the restriction of $\succsim$ to singleton lotteries as the DM's *commitment preferences* defined over the set of lotteries $\Delta(Z)$. That is, the DM is deemed to weakly prefer lottery x to lottery y (in terms of her commitment preferences) if $\{x\} \succsim \{y\}$.

The following is a general class of preferences over menus that captures a notion of the DM having a limited stock of willpower to resist temptation when exercising costly self-control.

Definition 3.1 A preference relation over menus $\succsim$ is a member of the family of *self-control preferences with limited willpower*, if there exists linear utility functions $u, v : \Delta(Z) \to \mathbb{R}$ and a stock of willpower $w \geq 0$, such that $\succsim$ can be represented by the functional given in (1).

As we noted in the introduction, this is the representation characterized by GP with the addition of a willpower constraint. In order to formalize the main differences between our model and that of GP, it is convenient to introduce the following concepts and attendant notation regarding lotteries for which the individual can and cannot exert self-control with respect to her commitment preferences.

For the lottery $x \in \Delta(Z)$, we take the set of tempting alternatives for which the DM fails to exert self-control as ones for which the DM is unable or unwilling to exert self-control in a two-element menu comprising x and that alternative. That is, an alternative y is deemed *a tempting alternative to x for which the DM fails to exert self-control*, if, despite strictly preferring, according to her commitment preferences,

lottery x to lottery y, she is indifferent between $\{y\}$ and the menu $\{x, y\}$. We interpret the latter indifference as reflecting her (rational) anticipation that if in the hot state she faces a choice from the menu $\{x, y\}$, she will be unable or unwilling to exert (sufficient) self-control to choose x. Formally,

$$T(x) := \{y \in \Delta(Z) : \{x\} \succ \{x, y\} \sim \{y\}\}.$$

Correspondingly, we define the set of *tempting alternatives to x for which the DM can exert costly self-control* as being,

$$S(x) := \{y \in \Delta(Z) : \{x\} \succ \{x, y\} \succ \{y\}\}.$$

Now for any lottery y in $T(x)$, there are two reasons for why the DM might anticipate she would not choose x from the menu $\{x, y\}$: (i) the alternative y is at least as good a "compromise" alternative as x, or (ii) despite x being the better compromise candidate, the individual has insufficient willpower to resist choosing y. In the latter case, by considering another alternative $(1-\alpha)x + \alpha y$ that is formed by taking a convex combination of x and y sufficiently close to x, that is, α is sufficiently close to 0, the DM will be able (and willing) to exert self-control when making a choice from the menu $\{x, (1-\alpha)x + \alpha y\}$. Hence, we divide $T(x)$ into two sets, $I(x)$ and $D(x)$, where

$$I(x) := \{y \in T(x) : (1-\alpha)x + \alpha y \in S(x), \text{ for some } \alpha \in (0, 1)\}, \text{ and}$$
$$D(x) := \{y \in T(x) : (1-\alpha)x + \alpha y \notin S(x), \text{ for all } \alpha \in (0, 1)\}.$$

We view $I(x)$ as containing those alternatives in $T(x)$ for which the DM *would like to but has insufficient willpower to* exert the self-control necessary to resist the tempting alternative, while $D(x)$ contains those alternatives for which the DM prefers to indulge herself. If $I(x) = \varnothing$ for all $x \in \Delta(Z)$, then the DM's preferences should reduce to one of the two polar sub-families of GP's Theorem 3.

As we shall see in the next section, a setting with limited willpower naturally leads to violations of the independence axiom. So instead, we consider a particular subset of menus, $\mathscr{B}(\succsim)$, for which the preference relation restricted to this subset satisfies independence. We consider the collection of all singleton menus as well as those two-element menus in which either there is self-control or any anticipated failure of self-control does not arise as a result of insufficient willpower. Formally, set

$$\mathscr{B}(\succsim) := \{\{x\} : x \in \Delta(Z)\} \cup \{\{x, y\} : \{x\} \succ \{y\} \text{ and } y \notin I(x)\}.$$

In addition, like GP, we cannot directly invoke the mixture space theorem, either. Hence, it will be convenient to consider the collection of all singleton menus as well as those two-element menus in which there is a tempting alternative for which the DM can exert costly self-control. Formally, set

$$\mathscr{M}(\succsim) := \{\{x\} : x \in \Delta(Z)\} \cup \{\{x, y\} : \ y \in S(x)\}.$$

In our construction, we will exploit this restricted version of independence axiom along with the set betweenness axiom to construct a mixture operator that when paired with $\mathscr{M}\left(\succsim\right)$ makes it a mixture space. Now the mixture space theorem can be applied to obtain the existence of a utility function defined over $\mathscr{M}\left(\succsim\right)$ that is linear with respect to our mixture operator and unique up to positive affine transformations.

4 The representation

We begin by imposing similar axioms to those in Theorem 3 of GP except that, for reasons we shall explain below, the independence axiom is restricted to $\mathscr{B}\left(\succsim\right)$.

Axiom 1 (Ordering) $\succsim$ *is a complete and transitive binary relation.*

Axiom 2a (Upper Semi-Continuity) *The sets* $\left\{B \in \mathscr{A} : B \succsim A\right\}$ *are closed.*

Axiom 2b (Lower von Neumann–Morgenstern Continuity) $A \succ B \succ C$ *implies* $\alpha A + (1-\alpha) C \succ B$ *for some* $\alpha \in (0,1)$.

Axiom 2c (Lower Singleton Continuity) *The sets* $\{x : \{y\} \succsim \{x\}\}$ *are all closed.*

Axiom 3 (Restricted Independence) *For any* $A, B, C \in \mathscr{B}\left(\succsim\right)$, $A \succ B$ *and* $\alpha \in (0,1)$ *implies* $\alpha A + (1-\alpha) C \succ \alpha B + (1-\alpha) C$.

Axiom 4 (Set Betweenness) $A \succsim B$ *implies* $A \succsim A \cup B \succsim B$.

To see why we cannot impose a more standard but stronger notion of continuity, note that when a sequence of menus passes through the willpower constraint, a representation of the form in expression (1) will in general entail a corresponding discontinuity in the preferences over menus. For example, consider u and v over lotteries x, $\{x_k\}_{k=1}^{\infty}$ and y, where $\lim_{k\to\infty} x_k = x$, $u(x) + v(x) > u(y) + v(y)$, $v(y) - v(x_k) > w$ for all k, and $v(y) - v(x) = w$. Hence, we see that lower semi-continuity is violated, since

$$\lim_{k\to\infty} U(\{x_k, y\}) = \lim_{k\to\infty} u(y) = u(y) < u(x) - w = U(\{x, y\}).$$

Therefore, we adopt continuity axiom used by GP for their theorem 3. We retain upper semi-continuity (Axiom 2a) but relax lower semi-continuity (Axiom 2b) except for the subset of singleton menus (Axiom 2c).

Unlike GP, however, we do not require independence to hold on the entire relation. Instead we restrict its application to $\mathscr{B}\left(\succsim\right)$. For example, suppose $A \succ B$ because all the "better" options in B are considered by the DM not available for her to choose from that menu owing to insufficient willpower. Conceivably, there might exist some other menu C and a (sufficiently small) weight α such that for the set $\alpha B + (1-\alpha) C$, the mixtures it contains with those better options in B might not exceed the DM's willpower. This in turn might result in $\alpha B + (1-\alpha) C \succsim \alpha A + (1-\alpha) C$, that is, a violation of independence.

Finally, to obtain a representation that allows for menus in which a failure of self-control may be due to insufficient willpower, we propose one new axiom that allows us to interpret willpower as a limited resource in which the available "stock" does not vary across menus.

For any subset $A \subset \Delta(Z)$, let $\overline{A}$ denote its closure.

Axiom 5 (Menu-Invariance Willpower Capacity) *For any lotteries* $a, b, x, y \in \Delta(Z)$, *if* $b \in \overline{I(a)} \cap S(a)$ *and* $y \in S(x)$, *then* $\{\frac{1}{2}a + \frac{1}{2}x, \frac{1}{2}a + \frac{1}{2}y\} \succsim \{\frac{1}{2}x + \frac{1}{2}a, \frac{1}{2}x + \frac{1}{2}b\}$.

To appreciate what this axiom entails, notice that since $b \in S(a)$ (respectively, $y \in S(x)$), means that for the menu $\{a, b\}$ (respectively, $\{x, y\}$), the individual has sufficient willpower to exert costly self-control to select in the hot state a from $\{a, b\}$ (respectively, x from $\{x, y\}$). However, since $b \in \overline{I(a)}$, exerting such self-control exhausts her entire stock of willpower. Hence, selecting a instead of b must incur at least as much self-control cost as choosing x instead of *any lottery* $y' \in S(x)$. Hence, the self-control cost in resisting temptation in the menu $\{a, b\}$ can be no less than it is in resisting temptation in the menu $\{x, y\}$. Now by Lemma A.4 it follows that $\frac{1}{2}a + \frac{1}{2}y \in S\left(\frac{1}{2}a + \frac{1}{2}x\right)$ and $\frac{1}{2}x + \frac{1}{2}b \in S\left(\frac{1}{2}x + \frac{1}{2}a\right)$. This means that for both menus $\{\frac{1}{2}a + \frac{1}{2}x, \frac{1}{2}a + \frac{1}{2}y\}$ and $\{\frac{1}{2}x + \frac{1}{2}a, \frac{1}{2}x + \frac{1}{2}b\}$, the individual would choose the lottery $\frac{1}{2}a + \frac{1}{2}x$. However, the axiom requires that the cost of exercising such self-control should be no more in the case of choosing $\frac{1}{2}a + \frac{1}{2}x$ from the menu $\{\frac{1}{2}a + \frac{1}{2}x, \frac{1}{2}a + \frac{1}{2}y\}$ than it is in the case of choosing $\frac{1}{2}a + \frac{1}{2}x$ from the menu $\{\frac{1}{2}x + \frac{1}{2}a, \frac{1}{2}x + \frac{1}{2}b\}$.

In conjunction with the other axioms, it provides us with a characterization of the family of preferences over menus that admit a self-control with limited willpower representation.

Theorem 1 *Suppose* $I(a) \neq \varnothing$ *for some* $a \in \Delta(Z)$. *Then the preference relation* $\succsim$ *satisfies Axioms 1–5 if and only if it admits a representation of the form in expression (1), where neither u nor v is constant and v is not an affine transformation of u except for when* $v(\cdot) = -\alpha u(\cdot) + \beta$ *for some* $\alpha \in (0, 1)$ *and* $\beta \in \mathbb{R}$.
Furthermore, the triple (u, v, w) *in the representation is unique in the sense that if* (u', v', w') *represent the same preferences as* (u, v, w) *then* $u' = \alpha u + \beta$, $v' = \alpha v + \beta'$ *and* $w' = \alpha w$ *for some* $\alpha > 0$ *and* $\beta,\ \beta' \in \mathbb{R}$.

The full proof appears in two appendices below but here we provide a sketch of the main ideas and arguments.

GP established that given the preference relation admits a functional representation, imposing set betweenness means that each finite menu can be shown to be indifferent to an appropriately selected *two-element subset* of that menu. Moreover, upper semi-continuity and set betweenness together are enough to allow us to extend the representation defined over this subset of menus to arbitrary compact sets. Hence, this function has a unique extension to the entire domain of menus.

Instead of considering all menus, we only impose independence and the willpower capacity axioms to those menus with at most two elements. In "Appendix A," we reproduce GP's theorem 3 using axioms 1–4 for the case when $I(a) = \phi$ for all $a \in \Delta(Z)$. Notice that in this case $\mathscr{B}\left(\succsim\right)$ is the collection of all singleton and two-element menus. Hence, we are back to GP's framework. However, for the independence axiom

to generate a constant trade-off between commitment and temptation utilities, we remove those two-element menus that the DM does not exert costly self-control in her ex post choices. That is, we use independence along with the set betweenness axiom to construct a mixture operator to pair with the domain $\mathscr{M}\left(\succsim\right)$. Hence, the mixture space theorem can be applied to obtain the existence of a utility function U defined over $\mathscr{M}\left(\succsim\right)$ that is linear with respect to our mixture operator and unique up to positive affine transformations. For any x, y that satisfies $\{x\} \succ \{x, y\} \succ \{y\}$, we have $\{x, y\} \in \mathscr{M}\left(\succsim\right)$. If we normalize $v(x) = 0$ then we can set $v(y) = U(\{x\}) - U(\{x, y\})$. Furthermore from the linearity of $U(\cdot)$ it follows that for any lottery z such that $v(z) = \alpha v(y)$, we have $v(z) = \alpha(U(\{x\}) - U(\{x, y\})$. Therefore, the degree of temptation that affects the DM's ex post choices can be measured by her ex ante utility representation U. We follow GP's construction to extend v for all lotteries in $\Delta(Z)$ and show that this construction of v is menu independent.

In "Appendix B," we consider the case when $I(a) \neq \varnothing$ for some $a \in \Delta(Z)$. Hence, $\mathscr{B}\left(\succsim\right)$ does not contain all two-element menus and for those two-element menus for which the willpower constraint is binding, we impose axiom 5 to regulate the boundary of the self-control region $\overline{I(a)} \cap S(a)$ to be invariant to the choice of a. Since $I(a) \neq \varnothing$, we can select a lottery $b \in I(a)$. From the definition of $I(a)$, it follows that there exists some $\overline{\alpha} \in (0, 1)$ such that the function $V(\alpha) := U(\{\alpha b + (1 - \alpha)a, a\})$ has a discontinuity at $\alpha = \overline{\alpha}$. From our construction, we have $w = v(\overline{\alpha}b + (1 - \overline{\alpha})a) - v(a) = \alpha_1 v(y)$ for some $\alpha_1 \in R_+$. Hence, we obtain $w = \alpha_1(U(\{x\}) - U(\{x, y\})$ which is also measured by her ex ante utility representation U. Thus, once U is fixed, we only have one degree of freedom for v, that is, the initial selection $\{x, y\}$ from $\mathscr{M}\left(\succsim\right)$ for the construction of v from which we began and normalized $v(x) = 0$.

4.1 Regular preferences

We say a costly self-control with limited willpower preference relation is "regular," if for some alternative a the willpower constraint strictly binds for at least one other tempting alternative (that is, $I(a) \neq \varnothing$) and for at least another tempting alternative, although the DM is able to exert self-control to resist it, she rather prefers to indulge herself (that is, $D(a) \neq \varnothing$). Hence, if a triple (u, v, w) represents a regular costly self-control with limited willpower preference then v is not an affine transformation of u.

Definition 4.1 A costly self-control with limited willpower preference relation $\succsim$ is deemed *regular* if there exists some $a \in \Delta(Z)$, such that $I(a) \neq \varnothing$ and $D(a) \neq \varnothing$.

Unlike the two polar sub-families of GP's Theorem 3, this intermediate case can accommodate violations of the weak axiom of revealed preferences (WARP) for the choice function that describes the choice the DM makes from a menu of lotteries.[6]

[6] For the case where the stock of willpower is sufficiently large so that it never binds (respectively, the case where the stock of willpower is zero) the choice function describing the choice the DM makes from a menu is the one generated by the utility function $u(x) + v(x)$ (respectively, $v(x)$ with $u(x)$ only used to break ties).

For instance, consider example 2 from Dekel et al. (2009). It concerns choices among menus that contain selections from the three food items: broccoli (b), (high-fat) ice cream (i) and (low-fat) yogurt (y). We suggest that in a setting where the decision-maker can resist temptation at a cost of exerting self-control, albeit only up to the limit of her willpower, it might be natural for her to express the following ranking of menus:

$$\{b\} \succ \{b, y\} \succ \{y\} \succ \{y, i\} \succ \{i\} \sim \{b, i\}.$$

The two strict preferences $\{b\} \succ \{b, y\}$ and $\{b, y\} \succ \{y\}$ reflect the willingness and the ability of the DM to resist choosing the more tempting option y from the menu $\{b, y\}$. Similarly, the two strict preferences $\{y\} \succ \{y, i\}$ and $\{y, i\} \succ \{i\}$ reflect her willingness and ability to resist choosing the more tempting option i from the menu $\{y, i\}$. The indifference $\{i\} \sim \{b, i\}$ arises because of the inability of the DM to resist choosing the tempting option i from the menu $\{b, i\}$ since the self-control cost required to select option b exceeds her stock of willpower. Thus, the violation of WARP arises from:

$$\begin{aligned} &u(b) + v(b) > u(y) + v(y) > u(i) + v(i) \\ &\quad \& \max\{v(y) - v(b),\ v(i) - v(y)\} \leq w < v(i) - v(b). \end{aligned} \tag{2}$$

Furthermore, from (2) we can also conclude that $i \in I(b)$. Letting δ_z denote the (degenerate) lottery that assigns probability 1 to the single prize $z \in Z$ obtaining, from the definition of $I(b)$ it follows that there exists some $\overline{\alpha} \in (0, 1)$ such that $(1-\alpha)\delta_b + \alpha\delta_i \in S(b)$ for all $\alpha < \overline{\alpha}$, and $(1-\alpha)\delta_b + \alpha\delta_i \in I(b)$, for all $\alpha > \bar{\alpha}$. That is,

$$\begin{aligned} &v(b) \geq v((1-\alpha)\delta_b + \alpha\delta_i) - w, \text{ for all } \alpha < \overline{\alpha}; \text{ and,} \\ &v(b) < v((1-\alpha)\delta_b + \alpha\delta_i) - w, \text{ for all } \alpha > \overline{\alpha}. \end{aligned}$$

Hence, exploiting the linearity of $v(\cdot)$ for convex mixtures and rearranging we obtain $w = \overline{\alpha}(v(i) - v(b))$.

The following theorem summarizes this insight that it is the discontinuity of the preferences over menus that enables us to calibrate the stock of willpower parameter w.

Theorem 2 *Suppose the preference relation $\succsim$ admits a costly self-control with limited willpower representation $U(\cdot)$ characterized by the triple (u, v, w). For any three lotteries $x_1, x_2, x_3 \in \Delta(Z)$, if*

$$\{x_1\} \succ \{x_1, x_2\} \succ \{x_2\} \succ \{x_2, x_3\} \succ \{x_3\} \sim \{x_1,\ x_3\}$$

then $x_3 \in I(x_1)$ and $v(x_3) - v(x_1) > w$. Moreover, there exists some $\overline{\alpha} \in (0, 1)$ such that $\overline{\alpha}(v(x_3) - v(x_1)) = w$, and the function $V(\alpha) := U(\{(\alpha x_1 + (1-\alpha)x_3, x_1\})$ has a discontinuity at $\overline{\alpha}$.

5 A comparative measure of self-control

In this section, we define a comparative measure for a preference for self-control. The following two definitions are taken from GP.

Definition 5.1 The preference $\succsim$ has self-control at A if there exists B, C such that $A = B \cup C$ and $B \succ A \succ C$. The preference $\succsim$ has self-control if $\succsim$ has self-control at some $A \in \mathscr{A}$.

Definition 5.2 The preference $\succsim_1$ has more self-control than $\succsim_2$ if, for all $A \in \mathscr{A}$, $\succsim_2$ has self-control at A implies $\succsim_1$ has self-control at A.

Theorem 9 of GP considers situations such that $\{a\} \succ_2 \{a, b\} \succ_2 \{b\}$ implies $\{a\} \succ_1 \{a, b\} \succ_1 \{b\}$ or $\{b\} \succ_1 \{a, b\} \succ_1 \{a\}$. Hence, they obtain the condition that $u_2 + v_2$ and v_2 are convex combinations of $u_1 + v_1$ and v_1 as the characterization for $\succsim_1$ having more self-control than $\succsim_2$.

To make the comparison between $\succsim_1$ and $\succsim_2$ tractable in our framework, we provide the following partial characterization in which we restrict attention to pairs of decision-makers that have the same ranking and intensity for tempting goods thereby allowing us to measure their temptation utilities with the same units and, by so doing, permitting us to make a meaningful comparison of their respective willpower limits.

Theorem 3 *Let $\succsim_1$, $\succsim_2$ be two regular costly self-control with limited willpower preferences. Let (u_1, v_1, w_1) be a representation of $\succsim_1$. Then, $\succsim_1$ has more self-control than $\succsim_2$ if there exists u_2, v_2 and w_2 such that (u_2, v_2, w_2) represents $\succsim_2$ and satisfies the following:*

$$\begin{pmatrix} u_2 + v_2 \\ v_2 \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ 0 & 1 \end{pmatrix} \begin{pmatrix} u_1 + v_1 \\ v_1 \end{pmatrix},$$

for some $\alpha > 0$, $\beta \geq 0$ and $w_2 \leq w_1$.

Notice that $u_2 = \alpha u_1 + (\alpha + \beta - 1)\, v_1$. Therefore, if $\alpha + \beta$ is equal to 1, then $u_2 = \alpha u_1$ making the line I_2 flatter than line I_1 in Fig. 3 simply as a consequence of the change in the relative scales of differences in commitment utility versus differences in temptation utility. However, if $\alpha + \beta$ is not equal to 1, then the commitment utility of the representation of preferences $\succsim_2$ will be affected by temptation utility and as a result the lines $u_2 = 0$ and $v_2 = 0$ need not be perpendicular to each other. However, as long as the line I_2 is flatter than the line I_1 and $w_2 \leq w_1$, we have $S_2 \subset S_1$. Hence, using the above theorem we are still able to compare their willpower limits under our representation.

6 Conclusion

In our representation, given a strictly positive willpower parameter $w > 0$, the discontinuity in preferences over menus should not be regarded as arising from overwhelming temptation. Rather it can be viewed as a manifestation of insufficient willpower on

the part of an individual to resist a tempting alternative. An implication of this may be seen in situations where an individual faces in her daily routine the temptation not to perform (or at least, postpone) an unpleasant (but important) task. If the limitation of her willpower is of concern, then measures to increase the commitment utility of the task she needs to perform will be to no avail. In this regard, such improvements are akin to "pushing on a string." Rather, she should endeavor to increase the temptation utility of the task in order to relax her willpower constraint. For example, adding an immediate reward (such as watching a movie) for performing a difficult and unpleasant task (such as injecting yourself with a needle) is more effective than trying to persuade yourself about the long-run (medical) benefits, as Dan Ariely (2011) attests, citing his own personal experience in strictly following an eighteen month drug regimen to treat a disease he contracted from a contaminated blood transfusion. Indeed, for some rewards, such as "sugar-coating" medicine, the bundling of the rewarding sugar with the unpleasant medicine might make the individual better off through the relaxation of her willpower constraint, even though the compromise utility of the bundle might be strictly less than the compromise utility of the medicine alone.

Appendix A: Preliminary results for representation

We begin our derivation by noting it follows from Rader (1963), the ordering axiom and upper semi-continuity ensure the preferences over compact menus admit a utility representation.

Lemma A.1 *If Axioms 1 and 2a hold, then there exists a function* $U : \mathscr{A} \to \mathbb{R}$ *that represents* $\succsim$.

Note that unlike GP's Lemma 1 (p1421), we do not adopt the independence axiom to establish the existence of a linear representation for the above lemma. Instead, we use our Axiom 3 in conjunction with Axiom 4, set betweenness to obtain a linear representation over $\mathscr{M}\left(\succsim\right)$ in Lemma A.5. As pointed out by GP (p 1413), upper semi-continuity and set betweenness together are enough to extend the representation from finite menus to arbitrary compact sets. (Proof appears in GP, lemma 8, p. 1430.) Hence, we only need to establish our results for finite menus. The next lemma is identical to GP's lemma 2, which enables us to identify the utility of any finite set with an appropriately chosen two-element subset.

Lemma A.2 (GP, Lemma 2, p. 1422). *Let* U *be a function that represents some* $\succsim$ *satisfying Axiom 4. If* $A \in \mathscr{A}$ *is a finite set, then*

$$U(A) = \max_{x \in A} \min_{y \in A} U(\{x, y\}) = \min_{y \in A} \max_{x \in A} U(\{x, y\}).$$

Moreover, there is an x^*, y^* *such that* (x^*, y^*) *solves the maxmin problem and* (y^*, x^*) *solves the minmax problem.*

We use Lemma A.2 to prove a result analogous to Lemma 3 in GP (p1422). But unlike GP we establish this result without assuming that the function representing $\succsim$ is linear. Instead the proof invokes Axiom 3 (restricted independence) directly.

Lemma A.3 *Let U be a function that represents some $\succsim$ that satisfy Axioms 3 and 4 and $A = \alpha\{x, y\} + (1-\alpha)\{a, b\}$.*

$$\{x, y\} \succ \{y\} \text{ and } \{a, b\} \succ \{b\} \text{ implies } U(A) = \min_{y' \in A} U\left(\left\{\alpha x + (1-\alpha)a, y'\right\}\right),$$

and

$$\begin{aligned}&\{x\} \succ \{x, y\},\ \{a\} \succ \{a, b\},\ y \notin I(x) \text{ and } b \notin I(a)\\ &\text{implies } U(A) = \max_{x' \in A} U\left(\left\{x', \alpha y + (1-\alpha)b\right\}\right).\end{aligned}$$

Proof By Lemma A.2, there exists (x^*, y^*) such that $A \sim \{x^*, y^*\}$ and (x^*, y^*) solves the maxmin problem. First, we show that $\{x, y\} \succ \{y\}$ and $\{a, b\} \succ \{b\}$ implies $x^* = \alpha x + (1-\alpha)a$. By Axiom 3, we have

$$\begin{aligned}A &\succ \alpha\{y\} + (1-\alpha)\{a, b\},\\ A &\succ \alpha\{x, y\} + (1-\alpha)\{b\}.\end{aligned}$$

Suppose $x^* = \alpha x + (1-\alpha)b$. Then, since $A \sim \{x^*, y^*\}$ and it solves the maxmin problem, we have

$$A \succ \alpha\{x, y\} + (1-\alpha)\{b\} = \{\alpha x + (1-\alpha)b, \alpha y + (1-\alpha)b\} \succsim A,$$

which yields a contradiction. Similarly, if $x^* = \alpha y + (1-\alpha)a$, then

$$A \succ \alpha\{y\} + (1-\alpha)\{a, b\} = \{\alpha y + (1-\alpha)b, \alpha y + (1-\alpha)a\} \succsim A.$$

If $x^* = \alpha y + (1-\alpha)b$, then

$$A \succ \alpha\{y\} + (1-\alpha)\{a, b\} \succ \{\alpha y + (1-\alpha)b\} = \{\alpha y + (1-\alpha)b, \alpha y + (1-\alpha)b\} \succsim A.$$

Hence, $x^* = \alpha x + (1-\alpha)a$. Suppose that we have $\{x\} \succ \{x, y\}$ and $\{a\} \succ \{a, b\}$ with $y \notin I(x)$ and $b \notin I(a)$. Then we can apply Axiom 3 and obtain

$$\begin{aligned}\alpha\{x\} + (1-\alpha)\{a, b\} &\succ A,\\ \alpha\{x, y\} + (1-\alpha)\{a\} &\succ A.\end{aligned}$$

Then since $A \sim \{y^*, x^*\}$ and it solves the minmax problem, we can use a similar argument as above to show $y^* = \alpha y + (1-\alpha)b$. □

Lemma A.3 enables us to define a mixture operation for the space $\mathscr{M}(\succsim)$, which we recall is comprised of the set of singleton menus and two-element menus in which there is a tempting alternative for which the DM can exert costly self-control. Since any A in $\mathscr{M}(\succsim)$ has at most two elements, it follows that for any pair of menus A and B in $\mathscr{M}(\succsim)$ and any α in $(0, 1)$, the menu $\alpha A + (1-\alpha)B$ has either one, two or

four elements. So consider the following (set-)mixture operator which we denote by $h_\alpha(\cdot,\cdot)$. If $A=\{a,b\}$ and $B=\{x,y\}$ with $\{a\}\succ\{a,b\}\succ\{b\}$ and $\{x\}\succ\{x,y\}\succ\{y\}$, then the $(\alpha,1-\alpha)$-(set-)mixing of A and B consists of taking the $(\alpha,1-\alpha)$-convex combination of the two better alternatives from each set and the $(\alpha,1-\alpha)$-convex combination of the two worse alternatives from each set. Thus, the resulting "mixture" set still contains only two elements. For all other possible configurations, the standard operation leads to at most two elements anyway, so no modification is required in these cases. More formally, we have for any A and B in $\mathscr{M}\left(\succsim\right)$ and any α in $(0,1)$,

$$h_\alpha(A,B):=\begin{cases}\{\alpha a+(1-\alpha)x,\alpha b+(1-\alpha)y\} & \text{if } \begin{array}{l}A=\{a,b\},b\in S(a),\\ B=\{x,y\},y\in S(x),\end{array}\\ \alpha A+(1-\alpha)B & \text{otherwise.}\end{cases}$$

Lemma A.4 *If a preference ordering $\succsim$ satisfies Axioms 3 and 4, then $\left(\mathscr{M}\left(\succsim\right),\{h_\alpha\}_{\alpha\in[0,1]}\right)$ is a mixture space as defined in* Kreps (1988, p. 52).[7]

Proof First, we will show that $h_\alpha(A,B)\in\mathscr{M}\left(\succsim\right)$ for any $A,B\in\mathscr{M}\left(\succsim\right)$. From Lemma A.3, it is known that h_α is either a singleton set or a two-element set. If h_α is a singleton set, then obviously it is in $\mathscr{M}\left(\succsim\right)$. If h_α has two elements, then it only takes one of the two possible forms, either $h_\alpha(\{a,b\},\{x\})$, or $h_\alpha(\{a,b\},\{x,y\})$ with $b\in S(a)$ and $y\in S(x)$. By Axiom 3, we have

$$\begin{aligned}\{\alpha a+(1-\alpha)x\}=\alpha\{a\}+(1-\alpha)\{x\}&\succ\{\alpha a+(1-\alpha)x,\alpha b+(1-\alpha)x)\}\\&\succ\alpha\{b\}+(1-\alpha)\{x\}=\{\alpha b+(1-\alpha)x\}\end{aligned}$$

Hence, $h_\alpha(\{a,b\},\{x\})\in\mathscr{M}\left(\succsim\right)$. By Axiom 3, we also have

$$\begin{aligned}\{\alpha a+(1-\alpha)x\}&\succ\alpha\{a,b\}+(1-\alpha)\{x\}\succ\alpha\{a,b\}+(1-\alpha)\{x,y\}\\&\succ\alpha\{a,b\}+(1-\alpha)\{y\}\succ\alpha\{b\}+(1-\alpha)\{y\}=\{\alpha b+(1-\alpha)y\}\end{aligned}$$

Hence, $h_\alpha(\{a,b\},\{x,y\})\in\mathscr{M}\left(\succsim\right)$ as well.

Next we will show that $h_\alpha(\left(h_\beta(A,B),B\right)=h_{\alpha\beta}(A,B)$ for any $A,B\in\mathscr{M}\left(\succsim\right)$. We only deal with the case when $A=\{x,y\}$ and $B=\{a,b\}$ with $y\in S(x)$ and $b\in S(a)$ because for the rest cases, the argument is similar but easier.

$$\begin{aligned}&h_\alpha\left(h_\beta(\{x,y\},\{a,b\}),\{a,b\}\right)\\&\quad=h_\alpha(\{\beta x+(1-\beta)a,\ \ \beta y+(1-\beta)b\},\{a,b\})\\&\quad=\{\alpha(\beta x+(1-\beta)a)+(1-\alpha)a,\alpha(\beta y+(1-\beta)b)+(1-\alpha)b\}\\&\quad=h_{\alpha\beta}(\{x,y\},\{a,b\}).\end{aligned}$$

□

Since it follows from Lemma A.3 that for any A,B in $\mathscr{M}\left(\succsim\right)$, $h_\alpha(A,B)\sim\alpha A+(1-\alpha)B$, as a consequence of Lemma A.4 we can apply the mixture space theorem Kreps (1988, Theorem 5.11, p. 54) to obtain the following representation of $\succsim$ restricted to $\mathscr{M}\left(\succsim\right)$.

[7] In particular, we have for any $\alpha,\beta\in(0,1)$, and any $A,B\in\mathscr{M}\left(\succsim\right)$, $h_\alpha\left(h_\beta(A,B),B\right)=h_{\alpha\beta}(A,B)$.

Lemma A.5 *A preference relation satisfies Axioms 1–4 if and only if there exists a linear function* $U : \mathscr{M}\left(\succsim\right) \to \mathbb{R}$, *such that for any* A, B *in* $\mathscr{M}\left(\succsim\right)$, $U(A) \geq U(B) \Leftrightarrow A \succsim B$. *Moreover,* U *in the representation is unique up to a positive affine transformation and its restriction to singleton sets is continuous.*

Now to extend the representation obtained in Lemma A.5, notice first it follows from Axiom 4 (set betweenness) that for *any* two-element menu either the menu is indifferent to a singleton menu that consists of just one element from that menu or that menu lies in preference terms strictly between the two singleton menus formed from its two elements. That is, $\{a\} \sim \{a, b\}$ or $\{a, b\} \sim \{b\}$ or $\{a\} \succ \{a, b\} \succ \{b\}$. For the third case, since $\{a, b\}$ is in $\mathscr{M}\left(\succsim\right)$, $U(\{a, b\})$ is already defined. For the other two cases, we can simply set $U(\{a, b\})$ either to $U(\{a\})$ or to $U(\{b\})$. This provides the unique extension of the function $U(\cdot)$ from Lemma A.5 to extend the representation of $\succsim$ to all two-element sets.

It remains to extend the representation to all menus. Our first step in this task is to define, as do GP, the linear (commitment utility) function $u : \Delta(Z) \to \mathbb{R}$, by setting $u(x) := U(\{x\})$. Next, for any two lotteries a, b and any $\gamma \in (0, 1)$, such that $\{a, b\} \in \mathscr{M}\left(\succsim\right)$ and $\{a, (1-\gamma)b + \gamma x\} \in \mathscr{M}\left(\succsim\right)$ for all $x \in \Delta(Z)$, we define the (temptation utility) function $v : \Delta(Z) \to \mathbb{R}$, as follows:

$$v(x; a, b, \gamma) := \frac{U(\{a, b\}) - U(\{a, (1-\gamma)b + \gamma x\})}{\gamma}.$$

We begin by noting a result that is analogous to GP's Lemma 4 (p1423) holds in our setting even though the domain of the $U(\cdot)$ in its statement is $\mathscr{M}\left(\succsim\right)$ rather than the unrestricted domain $\mathscr{A}$. GP's proof is still valid in our setting since all two-element sets used in their proof are in $\mathscr{M}\left(\succsim\right)$.

Lemma A.6 *Let* U *be a linear function that represents the restriction of some* $\succsim$ *to* $\mathscr{M}\left(\succsim\right)$. *Suppose that* $(1-\gamma)b + \gamma x \in S(a)$ *for all* $x \in \Delta(Z)$. *Then:*

- (i) $\forall\, x$ *such that* $x \in S(a)$, $v(x; a, b, \gamma) = U(\{a, b\}) - U(\{a, x\})$.
- (ii) $v(a; a, b, \gamma) = U(\{a, b\}) - U(\{a\})$.
- (iii) $v\left(\alpha x + (1-\alpha)x'; a, b, \gamma\right) = \alpha v(x; a, b, \gamma) + (1-\alpha)\, v\left(x'; a, b, \gamma\right)$.
- (iv) $v\left(x; a, b, \gamma'\right) = v(x; a, b, \gamma)$, *for all* $\gamma' \in (0, \gamma)$.
- (v) *Suppose that* $(1-\gamma)b' + \gamma x \in S\left(a'\right)$, *for all* $x \in \Delta(Z)$. *Then* $v(x; a, b, \gamma) = v\left(x; a', b', \gamma\right) + v\left(b'; a, b, \gamma\right)$.

Although U is linear on $\mathscr{M}\left(\succsim\right)$, we have not established that it is linear on $\mathscr{B}\left(\succsim\right)$. However, using an argument similar to the proof of Lemma 5.6 in Kreps (1988), we obtain the following weaker version of linearity.

Lemma A.7 *Let* U *be a function that restricted to* $\mathscr{M}\left(\succsim\right)$ *is linear and represents some* $\succsim$ *satisfying Axioms 1–4. If* $\{x, y\} \in \mathscr{B}\left(\succsim\right)$, *then for any* $A \in \mathscr{M}\left(\succsim\right)$, *and any* $\alpha \in (0, 1)$,

$$U(\alpha\{x, y\} + (1-\alpha)A) = \alpha U(\{x, y\}) + (1-\alpha)U(A).$$

Proof If $\{x, y\}$ satisfies $\{x\} \succ \{x, y\} \succ \{y\}$, the linearity is already proven in lemma A.4. We have to deal with $\{x, y\}$ with $\{x\} \sim \{x, y\} \succ \{y\}$ or $\{x\} \succ \{x, y\} \sim \{y\}$. If $\{x, y\}$ satisfies $\{x\} \succ \{x, y\} \sim \{y\}$, we claim that $\alpha\{x, y\} + (1-\alpha) A \sim \alpha\{y\} + (1-\alpha) A$ for all $A \in \mathscr{M}\left(\succsim\right)$. By Axiom 3, $\{x\} \succ \{y\}$ implies $\alpha\{x\} + (1-\alpha) A \succ \alpha\{y\} + (1-\alpha) A$, and Axiom 4 further implies $\alpha\{x\} + (1-\alpha) A \succsim \alpha\{x, y\} + (1-\alpha) A \succsim \alpha\{y\} + (1-\alpha) A$. In this case, we only have to show that $\alpha\{x, y\} + (1-\alpha) A \succ \alpha\{y\} + (1-\alpha) A$ will lead to a contradiction. Let us take $A = \{a, b\}$ with $\{a\} \succ \{a, b\} \succ \{b\}$. Suppose that $\alpha\{x, y\} + (1-\alpha) A \succ \alpha\{y\} + (1-\alpha) A$. In this case, we have $\{x\} \succ \{x, y\} \sim \{y\}$. Since $\{x\} \succ \{y\}$, letting $\alpha, \beta \in (0, 1)$ and applying Axiom 3, we obtain

$$\beta\{x\} + (1-\beta)\{y\} \succ \beta\{y\} + (1-\beta)\{y\} = \{y\} \sim \{x, y\},$$

and

$$\alpha\{x'\} + (1-\alpha)\{a, b\} \succ \alpha\{x, y\} + (1-\alpha)\{a, b\},$$

where $\{x'\} = \beta\{x\} + (1-\beta)\{y\}$.
Since $\alpha\{x'\} + (1-\alpha)\{a, b\} \succ \alpha\{x, y\} + (1-\alpha)\{a, b\} \succ \alpha\{y\} + (1-\alpha)\{a, b\}$, von Neumann–Morgenstern continuity implies there exists some $\gamma \in (0, 1)$ such that

$$\begin{aligned} &\alpha\{x\} + (1-\alpha)\{a, b\} \\ &\quad \succ \gamma\left(\alpha\{x'\} + (1-\alpha)\{a, b\}\right) + (1-\gamma)\left(\alpha\{x'\} + (1-\alpha)\{a, b\}\right). \end{aligned}$$

We only deal with the case when $A = \{a, b\}$ with $\{a\} \succ \{a, b\} \succ \{b\}$ because when A is a singleton set, the argument is similar but easier. Since $\alpha\{x'\} + (1-\alpha)\{a, b\}$ and $\alpha\{y\} + (1-\alpha)\{a, b\}$ are both in $\mathscr{M}\left(\succsim\right)$, we use Lemma A.3 to obtain the first "$\sim$" below

$$\begin{aligned} \alpha\{x, y\} + (1-\alpha)\{a, b\} &\succ \gamma\left(\alpha\{x'\} + (1-\alpha)\{a, b\}\right) + (1-\gamma)\left(\alpha\{y\} + (1-\alpha)\{a, b\}\right) \\ &\sim h_\gamma\left(\alpha\{x'\} + (1-\alpha)\{a, b\}, \alpha\{y\} + (1-\alpha)\{a, b\}\right) \\ &\sim \alpha\left(\gamma\{x'\} + (1-\gamma)\{y\}\right) + (1-\alpha)\{a, b\} \\ &\succ \alpha\{x, y\} + (1-\alpha)\{a, b\}, \end{aligned}$$

which yields a contradiction. The last "$\succ$" uses the fact that $\{x'\} \succ \{y\}$, $\gamma\{x'\} + (1-\alpha)\{y\} \succ \{y\} \sim \{x, y\}$ and Axiom 4. Since $\alpha\{x, y\} + (1-\alpha) A \sim \alpha\{y\} + (1-\alpha) A$, we have $U(\alpha\{x, y\} + (1-\alpha) A) = U(\alpha\{y\} + (1-\alpha) A)$. Using Lemma A.5, we have $U(\alpha\{y\} + (1-\alpha) A) = \alpha U(\{y\}) + (1-\alpha) U(A) = \alpha U(\{x, y\}) + (1-\alpha) U(A)$ for all $A \in \mathscr{M}\left(\succsim\right)$. For $\{x, y\}$ with $\{x\} \sim \{x, y\} \succ \{y\}$, Axioms 3 and 4 imply $\alpha\{x\} + (1-\alpha)\{a, b\} \succsim \alpha\{x, y\} + (1-\alpha)\{a, b\}$ in the previous discussion. By using the fact that $\{x\} \succ \{y\}$, and applying a similar argument, we can rule out the possibility that $\alpha\{x\} + (1-\alpha)\{a, b\} \succ \alpha\{x, y\} + (1-\alpha)\{a, b\}$ and further obtain $U(\alpha\{x, y\} + (1-\alpha) A) = U(\alpha\{x\} + (1-\alpha) A) = \alpha U(\{x\}) + (1-\alpha) U(A) = \alpha U(\{x, y\}) + (1-\alpha) U(A)$ for all $A \in \mathscr{M}\left(\succsim\right)$ □

Next we adapt Lemma 5 of GP(p1424) to our framework and establish, given a certain condition holds, there exists a costly self-control representation over any two-element menu in which the willpower constraint never binds. This certain condition requires that for any lottery a for which $S(a) \neq \varnothing$, we can find a lottery $b \in S(a)$ and a $\gamma > 0$, such that $(1-\gamma)b + \gamma x \in S(a)$ for all $x \in \Delta(Z)$.

Lemma A.8 *Let U be a function over menus. Suppose its restriction to $\mathscr{M}\left(\succsim\right)$ is linear and it represents a preference relation $\succsim$ satisfying Axioms 1–4. Consider a lottery $a \in \Delta(Z)$ for which $I(a) = \varnothing$. Suppose there exist lottery $b \in \Delta(Z)$ and $\gamma \in (0,1)$, such that $(1-\gamma)b + \gamma x \in S(a)$, for all $x \in \Delta(Z)$. Then for any lottery $y \in \Delta(Z)$, such that $U(\{y\}) \leq U(\{a\})$:*

$$U(\{a, y\}) = \max_{x \in \{a,y\}} \{u(x) + v(x; a, b, \gamma)\} - \max_{x' \in \{a,y\}} v\left(x'; a, b, \gamma\right).$$

Proof For the case where $U(\{a\}) > U(\{a, y\}) > U(\{y\})$, by GP(p 1424) we know $v(y; a, b, \gamma) \geq v(a; a, b, \gamma)$ and $u(a) + v(a; a, b, \gamma) - v(y; a, b, \gamma) > u(y) + v(y; a, b, \gamma) - v(y; a, b, \gamma)$. Let $A = (1-\gamma)\{a, b\} + \gamma\{a, y\}$. Since $\{a, b\} \in \mathscr{M}\left(\succsim\right)$, by Lemma A.7 we have $U(A) = (1-\gamma)U(\{a, b\}) + \gamma U(\{a, y\})$ for $\{a, y\} \in \mathscr{B}\left(\succsim\right)$. For the case where $U(\{a\}) = U(\{a, y\}) > U(\{y\})$, the first part of Lemma A.3 establishes that $U(A) = \min_{x' \in A} U\left(\{a, x'\}\right)$. Hence, we have $v(a; a, b, \gamma) \geq v(y; a, b, \gamma)$ by the same argument in GP. For the case where $U(\{a\}) > U(\{a, y\}) = U(\{y\})$ and $y \in D(a)$, we will show $v(y; a, b, \gamma) \geq v(a; a, b, \gamma) + u(a) - u(y)$. From GP this is equivalent to show that

$$U(\{a, (1-\gamma)b + \gamma y\}) \leq (1-\gamma)U(\{a, b\}) + \gamma U(\{a, y\}) = U(A)$$

The above inequality holds because of the second part of Lemma A.3 $U(A) = \max_{w \in A}(\{w, (1-\gamma)b + \gamma y\})$. □

With these preliminary results in hand, we are in a position to characterize the family of preferences over menus for which either the willpower constraint is never binding or temptation is overwhelming, that is, the stock of willpower is zero. As we noted in the previous section, this is GP's Theorem 3 (p1413).

Lemma A.9 *Suppose $I(a) = \phi$ for all $a \in \Delta(Z)$. A preference relation $\succsim$ satisfies Axioms 1–4 if and only if the preference relation $\succsim$ admits a representation of the form in expression (1), where w is either sufficiently large so the constraint is never binding or $w = 0$.*

Appendix B: Proof of Theorem 1

For the case that there exists a lottery a for which $I(a) \neq \varnothing$, we establish two lemmas before we prove Theorem 1. The first lemma extends the representation result from Lemma A.8 to include two-element menus in which the willpower constraint binds.

Lemma B.10 *Let U be a function over menus. Suppose its restriction to $\mathscr{M}\left(\succsim\right)$ is linear and it represents a preference relation $\succsim$ satisfying Axioms 1–5. Consider a lottery a for which $I(a) \neq \varnothing$. Suppose there exist lottery $b \in \Delta(Z)$ and $\gamma \in (0,1)$, such that $(1-\gamma)b + \gamma x \in S(a)$, for all $x \in \Delta(Z)$. Then for any lottery $y \in \Delta(Z)$, such that $U(\{y\}) \leq U(\{a\})$:*

$$U(\{a,y\}) = \max_{x\in\{a,y\}} \{u(x) + v(x;a,b,\gamma)\} - \max_{x'\in\{a,y\}} v\left(x';a,b,\gamma\right),$$
$$\text{s.t.} \max_{x'\in\{a,y\}} v\left(x';a,b,\gamma\right) - v(x;a,b,\gamma) \leq w(a)$$

where $w(a) = \max_{x'\in S(a)} v\left(x';a,b,\gamma\right) - v(a;a,b,\gamma)$.

Proof For the case where $y \in S(a)$, we have $v(y;a,b,\gamma) \leq \max_{z\in S(a)} v(z;a,b,\gamma)$. Hence, by Lemma A.8, we have the desired result for all $\{a,y\} \in \mathscr{B}\left(\succsim\right)$. The remaining part of the proof is to show that if $y \in I(a)$, then we must have $v(y;a,b,\gamma) - v(a;a,b,\gamma) \geq w(a)$, which is equivalent to show that $v(y;a,b,\gamma) > \max_{z\in S(a)} v(z;a,b,\gamma)$. Since $y \in I(a)$, $\exists\bar{\gamma} \in (0,1)$ such that $\gamma' y + \left(1-\gamma'\right)a \in S(a)$ if $\gamma' \leq \bar{\gamma}$ and $\gamma' y + \left(1-\gamma'\right)a \in I(a)$ if $\gamma' > \bar{\gamma}$. Let $\bar{y} = \bar{\gamma}y + (1-\bar{\gamma})a$. Since $\bar{y} \in S(a)$, by Lemma A.6 (iii), we have $\bar{\gamma}v(y;a,b,\gamma) + (1-\bar{\gamma})v(a;a,b,\gamma) = v(\bar{y};a,b,\gamma) \geq v(a;a,b,\gamma)$. Hence, $v(y;a,b,\gamma) \geq v(a;a,b,\gamma)$. Moreover, $u(a) + v(a;a,b,\gamma) > u(\bar{y}) + v(\bar{y};a,b,\gamma) = \bar{\gamma}(u(y) + v(y;a,b,\gamma)) + (1-\bar{\gamma})(u(a) + v(a;a,b,\gamma))$. Hence, $u(a) + v(a;a,b,\gamma) > u(y) + v(y;a,b,\gamma)$. We claim that $\max_{z\in S(a)} v(z;a,b,\gamma) = v(\bar{y};a,b,\gamma)$. For any $b' \in S(a)$, from Axiom 5, we have $\frac{1}{2}\{a\} + \frac{1}{2}\{a,b'\} \succsim \frac{1}{2}\{a\} + \frac{1}{2}\{a,\bar{y}\}$. Since $\{a,b'\}$ and $\{a,\bar{y}\}$ are in $\mathscr{M}\left(\succsim\right)$, we have $\frac{1}{2}u(a) + \frac{1}{2}U\left(\{a,b'\}\right) \geq \frac{1}{2}u(a) + \frac{1}{2}U(\{a,\bar{y}\})$, $U\left(\{a,b'\}\right) = u(a) + v(a;a,b,\gamma) - v\left(b';a,b,\gamma\right)$ and $U(\{a,\bar{y}\}) = u(a) + v(a;a,b,\gamma) - v(\bar{y};a,b,\gamma)$. Hence, $v\left(b';a,b,\gamma\right) \leq v(\bar{y};a,b,\gamma)$. □

In order to be able to apply the above lemmas, we need to establish that for any lottery a in which $S(a) \neq \varnothing$, we can find a lottery $b \in S(a)$ and a $\gamma > 0$, such that $(1-\gamma)b + \gamma x \in S(a)$ for all $x \in \Delta(Z)$. This is the import of GP's claim 1 (p1426) but their proof does not work when there exists a lottery a for which $I(a) \neq \varnothing$. So we prove the following directly.

Lemma B.11 *If there exists some pair $x, y' \in \Delta(Z)$ such that $y' \in S(x)$, then there is a $\gamma > 0$ such that $(1-\gamma)y + \gamma a \in S(x)$ for all $a \in \Delta(Z)$ and $y = \frac{x+y'}{2}$.*

Proof From Lemma A.4 and $\{x,y\} = \frac{1}{2}\{x\} + \frac{1}{2}\{x,y'\}$, we know $y \in S(x)$. Suppose there is a $\gamma_z > 0$ such that $(1-\gamma_z)y + \gamma_z[z] \in S(x)$ for all $z \in Z$. Since Z is finite, letting $\gamma = \min_{z\in Z}\{\gamma_z\}$ and applying Lemma A.4, we can obtain the desired result. Hence, for any $z \in Z$, let $y_i := (1-\gamma_i)y + \gamma_i[z]$, and $x_i := (1-\gamma_i)x + \gamma_i[z]$ and $\gamma_i \to 0$. We will show that $y_i \in S(x)$ for sufficiently large i. Suppose to the contrary that we can find a subsequence $y_{i'}$ from y_i, such that $y_{i'} \notin S(x)$. We show all the possible alternatives will lead to a contradiction.

Case 1. Suppose we have $\{x, y_{i'}\} \sim \{x\}$ for all i'. By Axiom 2a, we have $\{x,y\} \succsim \{x\}$, which contradicting $y \in S(x)$.

Case 2. Suppose we have $y_{i'} \in I(x)$ for all i'. By the definition of $I(x)$, there exists $\alpha \in (0,1)$ such that $(1-\alpha)x + \alpha y_{i'} \in S(x)$. Let $a = \frac{1-\alpha}{2-\alpha} y' + (1 - \frac{1-\alpha}{2-\alpha})((1-\alpha)x + \alpha y_{i'})$. By Lemma A.4, we have $a \in S(x)$. However, we can rewrite $a = \left(1 - \frac{\alpha\gamma_i}{2-\alpha}\right) y + \frac{\alpha\gamma_i}{2-\alpha}[z]$. Hence, for all $\gamma_j \leq \frac{\alpha\gamma_i}{2-\alpha}$, we have $y_j \in S(x)$, which yields a contradiction.

Case 3. Suppose we have $y_{i'} \in D(x)$ and $y \in T(x_{i'})$ for all i'. By Axiom 4, we have

$$U(\{x, y_{i'}, x_{i'}, y\}) \leq \max\{U(\{x, y_{i'}\}), U(\{x_{i'}, y)\} = \max\{U(\{y_{i'}\}), U(\{y\})\}$$

and

$$U(\{x, y_{i'}, x_{i'}, y\}) \geq \min\{U(\{x, y\}), U(\{x_{i'}, y_{i'}\})\}.$$

Note that $\{x_{i'}, y_{i'}\} = (1-\gamma_{i'})\{x, y\} + \gamma_{i'}\{[z]\}$. By Lemma A.7, we have $U(\{x_{i'}, y_{i'}\}) = (1-\gamma_{i'})U(\{x, y\}) + \gamma_{i'}U(\{[z]\})$. Since $U(\{x, y\}) > U(\{y\})$, we obtain a contradiction.

Case 4. Suppose we have $y_{i'} \in D(x)$ and $y \in S(x_{i'})$ for all i'. We apply Lemma A.7 to obtain

$$\frac{1}{2}U(\{x, y_{i'}\}) + \frac{1}{2}U(\{x_{i'}, y\}) = U\left(\frac{1}{2}\{x, y_{i'}\} + \frac{1}{2}\{x_{i'}, y\}\right);$$

therefore, the same argument in GP (Claim 1, p. 1426) follows. □

Now we are ready to prove our main theorem.

Proof of Theorem 1. We show that the axioms imply that the relation $\succsim$ admits a representation of the form given by the function U as defined in Theorem 1. Since there exists a in $\Delta(Z)$ such that $I(a) \neq \varnothing$, we have $S(a) \neq \varnothing$. By Lemma B.11, there exists $b \in S(a)$, $\gamma \in (0,1)$ satisfy $(1-\gamma)b + \gamma a' \in S(a)$ for all $a' \in \Delta(Z)$. By Lemma B.10, we can let $u(a') := U(\{a'\})$, $v(a') := v(a'; a, b, \gamma)$ for all $a' \in \Delta(Z)$ and $w = v(\bar{a}; a, b, \gamma) - v(a; a, b, \gamma)$, where $\bar{a} \in \overline{I(a)} \cap S(a)$. By the first part of the proof of Lemma A.8, we know $u(a) + v(a) > u(b) + v(b)$ and $v(b) > v(a)$. Hence, neither u nor v is constant and $v(\cdot) \neq -\alpha u(\cdot) + \beta$ for some $\alpha \in (-\infty, 0] \cup [1, \infty)$ and $\beta \in R$.

Now consider a set $A = \{x, y'\}$, where x and y' are in the relative interior of $\Delta(Z)$. Assume without loss of generality, that $u(x) \geq u(y')$. Since x is in the interior of $\Delta(Z)$ and $b \in S(a)$, we can select $\alpha \in (0,1)$ and $x' \in \Delta(Z)$ such that $\{x, y\} = \alpha\{x'\} + (1-\alpha)\{a, b\} \in \mathscr{M}(\succsim)$. Hence, by Lemma B.11, there exists $\gamma' \in (0,1)$ such that $(1-\gamma')y + \gamma' a' \in S(x)$ for all $a' \in \Delta(Z)$. If $I(x) = \varnothing$, then $\{x, y'\} \in \mathscr{B}(\succsim)$. Hence, we can apply Lemma A.8 and obtain

$$U(\{x, y'\}) = \max_{x'' \in \{x, y'\}} \{u(x'') + v(x''; x, y, \gamma')\} - \max_{x'' \in \{x, y'\}} v(x''; x, y, \gamma').$$

Note that the willpower constraint is relevant only when $y' \in S(x)$. Hence, we need to show that $v\left(y';x,y,\gamma'\right)-v\left(x;x,y,\gamma'\right)\leq w$ if $y'\in S(x)$. Since $y'\in S(x)$, by Axiom 5, we have $\frac{1}{2}\{a\}+\frac{1}{2}\{x,y'\}\succsim\frac{1}{2}\{x\}+\frac{1}{2}\{a,a^*\}$. By Lemma A.7, we then have $\frac{1}{2}u(a)+\frac{1}{2}U\left(\{x,y'\}\right)\geq\frac{1}{2}u(x)+\frac{1}{2}U(\{a,\bar{a}\})$, where $U\left(\{x,y'\}\right)=u(x)+v\left(x;x,y,\gamma'\right)-v\left(y';x,y,\gamma'\right)$ and $U(\{a,\bar{a}\})=u(a)-w$. Hence, we have $v\left(y';x,y,\gamma'\right)-v\left(x;x,y,\gamma'\right)\leq w$. Let $\gamma^*=\min\{\gamma,\gamma'\}$. By Lemma A.6 (iv), $v(\cdot;a,b,\gamma^*)=v(\cdot;a,b,\gamma)$ and $v(\cdot;x,y,\gamma^*)=v\left(\cdot;x,y,\gamma'\right)$. By Lemma A.6 (v), for an appropriate constant k, $v(\cdot;a,b,\gamma^*)=v(\cdot;x,y,\gamma^*)+k$ and hence it follows that

$$U\left(\{x,y'\}\right)=\max_{x''\in\{x,y'\}}\{u\left(x''\right)+v\left(x''\right)\}-\max_{y''\in\{x,y'\}}v\left(y''\right)$$
$$\text{s.t.}\max_{y''\in\{x,y'\}}v\left(y''\right)-v\left(x''\right)\leq w$$

If $I(x)\neq\varnothing$, we can apply Lemma B.10,

$$U\left(\{x,y'\}\right)=\max_{x''\in\{x,y'\}}\{u\left(x''\right)+v\left(x'';x,y,\gamma'\right)\}-\max_{y''\in\{x,y'\}}v\left(y'';x,y,\gamma'\right),$$
$$\text{s.t.}\max_{y''\in\{x,y'\}}v(y'';x,y,\gamma')-v(x'';x,y,\gamma')\leq w(x),$$

where $w(x)=\max_{y''\in S(x)}v\left(y'';x,y,\gamma'\right)-v(x;x,y,\gamma')$. Take $\bar{x}\in\overline{I(x)}\cap S(x)$. By Axiom 5, we have $\frac{1}{2}\{a\}+\frac{1}{2}\{x,\bar{x}\}\sim\frac{1}{2}\{x\}+\frac{1}{2}\{a,\bar{a}\}$. By Lemma A.7, we then have $\frac{1}{2}u(a)+\frac{1}{2}U(\{x,\bar{x}\})=\frac{1}{2}u(x)+\frac{1}{2}U(\{a,\bar{a}\})$, where $U(\{x,\bar{x}\})=u(x)+v\left(x;x,y,\gamma'\right)-v\left(\bar{x};x,y,\gamma'\right)$ and $U(\{a,\bar{a}\})=u(a)-w$. Hence, we have $w(x)=v\left(\bar{x};x,y,\gamma'\right)-v\left(x;x,y,\gamma'\right)=w$. Follow the same argument as above, we have

$$U\left(\{x,y'\}\right)=\max_{x''\in\{x,y'\}}\{u\left(x''\right)+v\left(x''\right)\}-\max_{y''\in\{x,y'\}}v\left(y''\right)$$
$$\text{s.t.}\max_{y''\in\{x,y'\}}v\left(y''\right)-v\left(x''\right)\leq w.$$

Now consider an arbitrary finite set A. We know that

$$\begin{aligned}U(A)&=\max_{x\in A}\min_{y\in A}U(\{x,y\})\\&=\max_{x\in A}\min_{y\in A}\left\{\begin{array}{c}\max_{x'\in\{x,y\}}\{u\left(x'\right)+v\left(x'\right)\}-\max_{y'\in\{x,y\}}v\left(y'\right)\\\text{s.t.}\max_{y'\in\{x,y\}}v(y')-v(x')\leq w\end{array}\right\}\\&=\max_{x\in A}\min_{y\in A}\left\{\begin{array}{c}\max_{x'\in\{x,y\}}\{u\left(x'\right)+v\left(x'\right)\}\\\text{s.t.}\max_{y'\in\{x,y\}}v(y')-v(x')\leq w\end{array}\right\}+\min_{y\in A}\{-v(y)\}\end{aligned}$$

Let $y^*\in\arg\max_{y\in A}v(y)$. If, $x\in A$ such that $v(y^*)-v(x)>w$, then x does not solve the constraint maxminmax problem because for the pair $\{x,y^*\}$ we would choose y^* instead of x. Hence, x will not survive after the second requirement, i.e., $\min_{y\in A}$ when we take $y=y^*$. Now consider any $x\in A$ such that $v(y^*)-v(x)\leq w$,

then for any pair $\{x, x'\}$ where $x' \in A$, we have $v\left(x'\right) - v(\ x^*) \leq w$. Hence, if $v\left(y^*\right) - v\left(x'\right) \leq w$, then we choose x over x' only when $u\left(x\right) + v\left(x\right) \geq u\left(x'\right) + v\left(x'\right)$. Hence, we have

$$U\left(A\right) = \max_{x \in A}\{u\left(x\right) + v\left(x\right)\} - \max_{y \in A}\{v\left(y\right)\}$$
$$\text{s.t.} \max_{y \in A} v\left(y\right) - v\left(x\right) \leq w$$

To show that the axioms are necessary, for any $A \in \mathscr{A}$ let us denote the most tempting lottery $y_A \in \arg\max_{y \in A} v\left(y\right)$, the admissible set $D_A = \{x \in A : v\left(x\right) \geq w - v(y_A)\}$ and the best admissible compromise lottery $x_A \in \arg\max_{x \in D_A}\{u\left(x\right) + v\left(x\right)\}$. Note that u and v are both continuous functions and D_A is a compact set. Hence, Axiom 2a holds. For Axiom 2 to hold, note that $\alpha y_A + (1-\alpha)\, y_C \in \arg\max_{y \in \alpha A + (1-\alpha)C} v\left(y\right)$ for all $\alpha \in [0, 1]$ and $A, C \in \mathscr{A}$. Hence, we have $\alpha x_A + (1-\alpha)\, x_C \in D_{\alpha A + (1-\alpha)C}$. By selecting $\alpha > \frac{U(B) - U(C)}{U(A) - U(C)}$, we can verify that preferences represented by U satisfy Axiom 2b. For Axiom 3 to hold, if $A_1, A_2 \in \mathscr{B}\left(\succsim\right)$, we claim that $\alpha x_{A_1} + (1-\alpha)\, x_{A_2} \in \arg\max_{x \in D_{\alpha A_1 + (1-\alpha)A_2}}\{u\left(x\right) + v\left(x\right)\}$. Hence, $U\left(\alpha A_1 + (1-\alpha)\, A_2\right) = \alpha U\left(A_1\right) + (1-\alpha)\, U\left(A_2\right)$, which yields the result. Suppose to the contrary that the above claim does not hold. Since $\alpha x_{A_1} + (1-\alpha)\, x_{A_2} \in D_{\alpha A_1 + (1-\alpha)A_2}$, we must have some $x' \in A_i$ such that $u(x') + v\left(x'\right) > u\left(x_i\right) + v\left(x_i\right)$ for some $i \in \{1, 2\}$ and $x' \notin D_{A_i}$. Since A_i has at most two elements and $x' \neq x_{A_i}$, we must have $x_A = y_A$, $u\left(x'\right) > u\left(x_i\right)$ and $U(A_i) = u\left(x_A\right)$. Hence, $x' \in T\left(x_A\right)$. Since $w > 0$, we can find an $\alpha \in (0, 1)$ such that $\alpha x' + (1-\alpha)\, x_{A_i} \in D_{A_i}$, which implies $\alpha x' + (1-\alpha)\, x_{A_i} \in S\left(x_A\right)$. This contradicts $A_i \in \mathscr{B}\left(\succsim\right)$. To prove Axiom 4, if $v\left(y_A\right) \leq v\left(y_A\right)$, then we have $y_{A \cup B} = y_B$. Hence,$U\left(A \cup B\right) \leq u\left(x_A\right) + v\left(x_A\right) - v\left(y_A\right)$ and $x_B \in D_{A \cup B}$, which implies $A \succsim A \cup B \succsim B$. If $v\left(y_A\right) > v\left(y_B\right)$, then $y_{A \cup B} = y_A$. Hence, we have $U(A) = U(A \cup B)$. To show that Axiom 5 holds, since $b \in S\left(a\right)$ and $y \in S\left(x\right)$, we have $u\left(a\right) > U(\{a, b\}) > u\left(b\right)$ and $u\left(x\right) > U\left(\{x, y\}\right) > u\left(y\right)$. Hence, $0 < v\left(b\right) - v\left(a\right) \leq w$ and $0 < v\left(y\right) - v\left(x\right) \leq w$. Moreover, $b \in \overline{I\left(a\right)}$ implies $v\left(b\right) - v\left(a\right) = w$. Otherwise, we would have an open ball centered at b, denoted as $B\left(b\right)$, such that all $b' \in B\left(b\right)$ we have $v\left(b'\right) - v\left(a\right) < w$ and $b' \in S\left(a\right)$, which contradicts $b \in \overline{I\left(a\right)}$. Hence, by straightforward computation we can conclude that $U(\{\frac{1}{2}a + \frac{1}{2}x, \frac{1}{2}a + \frac{1}{2}y\}) \geq U\left(\{\frac{1}{2}x + \frac{1}{2}a, \frac{1}{2}x + \frac{1}{2}b\}\right)$. □

References

Ali, S.N.: Learning self-control. Q. J. Econ. **126**, 857–893 (2011)

Ariely, D.: The Upside of Irrationality: The Unexpected Benefits of Defying Logic at Work and at Home. Harper Perennial, New York (2011)

Baumeister, R.F., Heatherton, T., Tice, D.: Losing Control: How and Why People Fail at Self-Regulation. Academic Press, Cambridge (1994)

Bénabou, R., Tirole, J.: Willpower and personal rules. J. Polit. Econ. **112**(4), 848–886 (2004)

Chatterjee, K., Vijay Krishna, R.: A "Dual Self" representation for stochastic temptation. Am. Econ. J.: Microecon. **1**, 148–167 (2009)

Dekel, E., Lipman, B.L., Rustichini, A.: Temptation driven preferences. Rev. Econ. Stud. **76**(3), 937–971 (2009)

Elster, J., Skog, O.-J. (eds.): Getting Hooked: Rationality and Addiction. Cambridge University Press, Cambridge (1999)
Gailliot, M.T., Baumeister, R.F.: The physiology of willpower: linking blood glucose to self-control. Personal. Soc. Psychol. Rev. **11**(4), 303–327 (2007)
Gul, F., Pesendorfer, W.: Temptation and self-control. Econometrica **69**(6), 1403–1435 (2001)
Gul, F., Pesendorfer, W.: Harmful addiction. Rev. Econ. Stud. **74**, 147–172 (2007)
Kopylov, I.: Finite additive utility representations for preferences over menus. J. Econ. Theory **144**, 354–374 (2009)
Kopylov, I.: Perfectionism and choice. Econometrica **80**(5), 1819–1843 (2012)
Kreps, D.: Notes on the Theory of Choice. Westview Press, Boulder (1988)
Loewenstein, G.: Emotions in economic theory and economic behavior. Am. Econ. Rev. **90**(2), 426–432 (2000)
Masatlioglu, Y., Nakajima, D., Ozdenoren, E.: Revealed willpower. Working Paper, pp. 1–20 (2014)
Muraven, M., Baumeister, R.F.: Self-regulation and depletion of limited resources: does self-control resemble a muscle? Psychol. Bull. **126**(2), 247–259 (2000)
Noor, J., Takeoka, N.: Uphill self-control. Theor. Econ. **5**(2), 127–158 (2010)
Rader, T.: The existence of a utility function to represent preferences. Rev. Econ. Stud. **30**(3), 229–232 (1963)
Roy, F., Baumeister, K.V.: Willpower, Choice and Self-Control. In: Loewenstein, G., Read, D., Baumeister, R.F. (eds.) Time and Decision. Russell Sage Foundation, New York (2003)
Stovall, J.E.: Multiple temptations. Econometrica **78**(1), 349–376 (2010)
Strotz, R.: Myopia and inconsistency in dynamic utility maximization. Rev. Econ. Stud. **23**(3), 165–180 (1955)