BEHAVIORAL CHARACTERIZATIONS OF NAIVETE FOR TIME-INCONSISTENT PREFERENCES

By

David S. Ahn, Ryota Iijima, Yves Le Yaouanq, and Todd Sarver

April 2015

Revised November 2018

COWLES FOUNDATION DISCUSSION PAPER NO. 2074R

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Behavioral Characterizations of Naivete for Time-Inconsistent Preferences*

David S. Ahn† Ryota Iijima‡ Yves Le Yaouanq§ Todd Sarver¶

First Version: April 28, 2015
Current Draft: November 18, 2018

Abstract

We propose nonparametric definitions of absolute and comparative naivete. These definitions leverage ex-ante choice of menu to identify predictions of future behavior and ex-post (random) choices from menus to identify actual behavior. The main advantage of our definitions is their independence from any assumed functional form for the utility function representing behavior. An individual is sophisticated if she is indifferent ex ante between retaining the option to choose from a menu ex post or committing to her actual distribution of choices from that menu. She is naive if she prefers the flexibility in the menu, reflecting a mistaken belief that she will act more virtuously than she actually will. We propose two definitions of comparative naivete and explore the restrictions implied by our definitions for several prominent models of time inconsistency.

KEYWORDS: Naive, sophisticated, time inconsistent, comparative statics

*Ahn and Sarver acknowledge the financial support of the National Science Foundation through Grants SES-1357719 and SES-1357955, and Le Yaouanq acknowledges financial support from the Corps des Mines and from the Deutsche Forschungsgemeinschaft (through CRC TRR 190). We thank the editor, three anonymous referees, Ned Augenblick, Roland Bénabou, Drew Fudenberg, Christian Gollier, David Laibson, Bart Lipman, Takeshi Murooka, Jawwad Noor, Wolfgang Pesendorfer, Matthew Rabin, Philipp Sadowski, Ran Spiegler, Tomasz Strzalecki, Norio Takeoka, Jean Tirole, and numerous seminar participants for helpful comments and discussions. This paper incorporates results previously circulated under the titles "Comparative Measures of Naivete" by Ahn, Iijima, and Sarver and "Anticipating Preference Reversals" by Le Yaouanq.

†Department of Economics, University of California, Berkeley, 530 Evans Hall #3880, Berkeley, CA 94720-3880. Email: dahn@econ.berkeley.edu.

‡Department of Economics, Yale University, 30 Hillhouse Ave, New Haven, CT 06510. Email: ryota.iijima@yale.edu.

§Seminar für Organisationsökonomik, Ludwig-Maximilians-Universität, Kaulbachstr. 45, München 81825, Germany. Email: yves.leyaouanq@econ.lmu.de.

¶Department of Economics, Duke University, 213 Social Sciences/Box 90097, Durham, NC 27708. Email: todd.sarver@duke.edu.

1 Introduction

Models of dynamic inconsistency play an important role in a wide-ranging set of economic applications, and there is strong and increasing interest in the implications of naivete where individuals mispredict their future behavior.[1] While naivete often yields surprising and significant consequences, so far these effects are usually understood within the context of specific utility representations, where the existence and comparison of naivete are defined and tested through parameters like discount factors or probabilities.

In this paper, we introduce general nonparametric definitions of naivete and sophistication, as well as comparative measures of naivete. We then characterize the implications of these definitions for a broad class of utility specifications. Our behavioral definitions leverage two pieces of choice data. First, we use preference for commitment to measure *anticipated* behavior from an ex-ante perspective before the realization of temptation. Formally, the individual's preferences over different option sets (or menus) capture her demand for commitment and allow an inference of her beliefs regarding her future behavior. Second, we use choices from option sets to measure *actual* behavior from an ex-post perspective under the influence of temptation and after the level of commitment is fixed. Since uncertainty about future behavior seems especially compelling under naivete and is increasingly relevant in applied work, we formally accommodate this uncertainty by modeling ex-post behavior as a random choice rule.

For a simple illustration of our approach, consider first an individual who makes deterministic choices without randomization. Her ex-ante ranking of option sets is given by a preference $\succsim$, and her ex-post choice from any menu is given by a choice function $\mathcal{C}$.[2] When choosing between two options p and q, an individual may prefer p if committing ex ante, $\{p\} \succ \{q\}$, yet choose q if given the option ex post, $\mathcal{C}(\{p,q\}) = q$. This pattern is indicative of time inconsistency and has been documented in numerous contexts, for example, preferences to maintain a healthy diet, decrease spending, or engage timely effort in a difficult task that go unfulfilled ex post. Still, additional information is needed to determine whether the individual is sophisticated or naive about this inconsistency. If we also observe a strict preference to retain the option p ex ante, $\{p,q\} \succ \{q\}$, then we can further infer that she (incorrectly) anticipates that p will be her ex-post choice from the menu $\{p,q\}$ and hence she is naive. Thus, observing preferences to retain the

[1] A recent survey of empirical applications can be found in Section 2.1 of DellaVigna (2009) and a survey of some theoretical applications in contract theory can be found in Kőszegi (2014).

[2] We focus throughout the paper on choice functions rather than correspondences, which presumes the individual uses some tie-breaking procedure to select between equally attractive options. Our primitives for stochastic choice make similar implicit assumptions. Importantly, our results do not depend in any way on how ties are broken. Hence, while our results can easily be extended to deal with choice correspondences (and their stochastic generalizations), it is a strength of the current analysis that knowledge of the complete set of possible options that the individual is willing to choose from a menu is not required.

flexibility of multiple options provides the additional information needed to delineate between sophisticated and naive beliefs. Similarly, in the more general case of stochastic choice, if p is chosen with probability α from the menu $\{p, q\}$ at the ex-post stage, then the relevant ex-ante comparison is between the menu $\{p, q\}$ and commitment to the appropriately weighted mixture $\{\alpha p + (1 - \alpha)q\}$. A strict preference for the former indicates biased beliefs that overestimate the probability of choosing the ex-ante more appealing alternative p to be greater than α.

Our behavioral definitions extend this approach to arbitrary choice sets. To test absolute naivete and sophistication, we compare an individual's predicted value for a menu x of different options against the actual value of her ex-post choice $\mathcal{C}(x)$ from that menu. Ex ante, a sophisticate correctly anticipates her future choice and is indifferent between maintaining the flexibility to choose from x later or committing to her eventual choice $\mathcal{C}(x)$ now, that is, $x \sim \{\mathcal{C}(x)\}$. In contrast, a naif mistakenly anticipates making a more virtuous choice and prefers to maintain the flexibility in x, that is, $x \succsim \{\mathcal{C}(x)\}$. In the case of uncertain temptations and random choice, we maintain this same basic intuition by comparing her preference for the menu versus committing to the lottery over outcomes induced by her distribution of choices.

Using one of the most comprehensive models of time-inconsistent preferences available, the *random Strotz representation*, we show that our behavioral definitions of sophistication and naivete characterize sharp and intuitive parametric restrictions. The random Strotz model is general enough to include the majority of all utility representations for time-inconsistent preferences that appear in the applied literature,[3] including the naive quasi-hyperbolic discounting model of O'Donoghue and Rabin (1999, 2001) and its stochastic extensions, and the parametric restrictions implied by our definitions boil down to the functional-form restrictions that the literature has proposed for these models. Our first contribution is thus in unifying the different parametric notions of naivete that have been explored for various models of time inconsistency by illuminating their common underlying behavioral theme: underdemand for commitment.

Our second and most significant contribution is in developing behavioral definitions of comparative naivete. Comparative measures of naivete rooted in choice behavior have been essentially unexplored in the prior literature. Moreover, even restricting attention to specific utility representations such as naive quasi-hyperbolic discounting, the proper functional-form restrictions that capture increases in naivete are not fully understood or agreed upon. In Section 2, we discuss existing proposals of parametric restrictions for comparing naivete that have been suggested for the quasi-hyperbolic discounting model, and we provide examples that demonstrate why these prior proposals may lead

[3]One important exception is models that incorporate costly self-control. We apply our definitions to the random self-control representation as an extension in Section 5.1, and we explore alternative definitions of naivete for self-control preferences in a companion paper Ahn, Iijima, and Sarver (2016).

to counterintuitive behaviors: An individual ranked as more naive according to one of these statistics may nonetheless make better use of available commitment devices or be less subject to exploitation in market interactions, behavior that strikes one as clearly more sophisticated.

To avoid such counterintuitive possibilities, we take the opposite approach. Our starting point is instead to consider the *behavior* that seems to most reasonably capture increases in naivete. We explore two possible notions of comparative naivete. The first is based on comparing underdemand for commitment. Using deterministic choice to illustrate simply, a commitment to the singleton menu $\{p\}$ is beneficial if $\{p\} \succ \{\mathcal{C}(x)\}$, that is, if p is preferred ex ante to the outcome $\mathcal{C}(x)$ that would actually be chosen from x. However, a naive individual may at the same time exhibit the ranking $x \succ \{p\}$, so she strictly prefers to maintain the flexibility of x due to the mistaken belief that she will ultimately make a more virtuous choice. Thus, a beneficial commitment is declined if $x \succ \{p\} \succ \{\mathcal{C}(x)\}$. Our first definition of comparative naivete is that an individual is more naive than another if she declines more beneficial commitments. Our second definition compares individuals' believed and actual indirect utilities from menus and classifies an individual as more naive if the difference between her ex-ante anticipated utility from x and her utility from the actual choice $\mathcal{C}(x)$ is larger. This comparison manifests behaviorally as a greater willingness to overpay for the menu x, and hence the overvaluation of menus provides another natural metric for naivete. In the case of random choice, both comparative definitions extend by replacing the deterministic choice with the induced lottery over outcomes. We characterize the parametric restrictions corresponding to each of these definitions within the random Strotz model, show how the two definitions are related, and discuss when each might be most appropriate.

Our paper is related to two strands of literature. In the empirical literature on time inconsistency and naivete, our use of ex-ante and ex-post choice behavior has several precedents. DellaVigna and Malmendier (2006) study both the choice of gym membership, which determines the feasible set of attendance/payment pairs, and subsequent attendance levels; Shui and Ausubel (2005) observe consumers' choices of credit card contracts and their subsequent borrowing behavior; Giné, Karlan, and Zinman (2010) offer subjects commitment contracts that incentivize smoking cessation and later test whether or not the subjects smoked; Kaur, Kremer, and Mullainathan (2015) allow subjects to choose wage contracts that constrain their feasible future effort/consumption pairs and then observe actual effort ex post; Augenblick, Niederle, and Sprenger (2015) ask subjects to choose an intertemporal allocation of effort and a probability of being committed to it and then observe whether subjects wish to revise that plan when the first date of task completion arrives. Not only do these papers use similar choice data, but those that test for naivete identify it using behavior that is closely related to our definition.

Our work also relates to papers in decision theory that use behavior at different time

periods to capture sophistication under time inconsistency, as surveyed by Lipman and Pesendorfer (2013). Noor (2011) considers preferences in a recursive domain that includes ex-ante and ex-post choice as projections; he pioneered the approach of using temporal choice as a domain for explicitly testing the sophistication implicitly assumed in most ex-ante axiomatic models of temptation. Kopylov (2012) relaxes Noor's sophistication condition and considers agents who choose flexibility ex ante that is subsequently unused ex post. Kopylov eschews mistaken or naive beliefs, but rather interprets the relaxation of sophistication as reflecting a direct psychic benefit of maintaining positive self-image. Finally, Dekel and Lipman (2012) observe that ex-ante and ex-post choice can be combined to empirically distinguish random Strotz representations from others that involve costly self-control. Much of the technical apparatus from Dekel and Lipman (2012) ends up being useful in studying naivete, as we will explain in the body of the paper.

The remainder of the paper is organized as follows. In Section 2, we use the special case of quasi-hyperbolic discounting to illustrate some of the problems with existing proposals and to outline our approach. Then, the following sections contain our formal results. We begin with the special case of deterministic choice in Section 3 to introduce and ground concepts, and then move on to the more general case of stochastic choice in Section 4. Finally, Section 5 discusses several extensions of our analysis.

2 Examples and Motivation

We preview our definitions by focusing attention to the naive quasi-hyperbolic discounting model introduced by O'Donoghue and Rabin (1999, 2001). In this model, the agent would ideally discount future utility by the factor δ. But she is tempted by instantaneous gratification and at the time of choice will discount the future by an additional present-bias factor β, leading to overconsumption in the present and underconsumption in the future relative to her ideal plan. If she is sophisticated, she correctly anticipates this present bias. The innovation of O'Donoghue and Rabin (1999, 2001) is to allow the agent to incorrectly anticipate the magnitude of present bias and instead believe that she will use the present-bias factor $\hat{\beta}$.

As suggested by O'Donoghue and Rabin (1999, 2001), sophistication is intuitively captured by the parametric restriction $\hat{\beta} = \beta$, while naivete is captured by $\hat{\beta} \geq \beta$.[4] We provide foundations for these parametric restrictions. Our proposal for a behavioral criterion for sophistication is that the decision-maker is indifferent between choosing from a menu of available options and committing to the particular option that she will actually choose from that menu, correctly anticipating her future choice. Our criterion for naivete is that she prefers the menu to her eventual selection, incorrectly anticipating making a

[4]Here, we mean (weak) naivete to include the boundary case of sophistication.

more virtuous choice. We show that these criteria are respectively equivalent to $\hat{\beta} = \beta$ and $\hat{\beta} \geq \beta$ for the quasi-hyperbolic model. Moreover, we demonstrate that these criteria also correspond to intuitive restrictions in a broad class of models extending beyond quasi-hyperbolic discounting. Our contribution to understanding absolute naivete is in providing behavioral foundations that apply across a variety of models, thus illuminating a common structure that they share.

While the notion of absolute naivete for a single quasi-hyperbolic agent is relatively unambiguous, how to compare naivete across individuals is more controversial. In the literature, various notions of "more naive" have been proposed for the quasi-hyperbolic model. For example, DellaVigna and Malmendier (2004) and Bousquet (2017) suggest an agent is more naive if the statistic $\hat{\beta} - \beta$ is greater, while Augenblick and Rabin (2015) suggest an agent is more naive if the statistic $\frac{1-\hat{\beta}}{1-\beta}$ is smaller. These proposals appear intuitively plausible at first glance, but the following examples suggest that an agent ranked as more naive according to these statistics may nonetheless engage in behavior that seems patently more sophisticated.

2.1 Problems with Existing Approaches

To illustrate the counterintuitive behaviors associated with existing proposals for comparing naivete, we use two examples. The first is a stylized consumption-savings problem where the agent is given the opportunity to advantageously place assets in an illiquid account as a commitment device for saving. Illiquid savings instruments that preempt instantaneous gratification seem among the most canonical and oft-mentioned examples of policy interventions motivated by insights from behavioral economics, so they seem like a natural first test of how well different rankings perform under real-world applications.

Example 1 (Consumption-savings problem)**.** Consider two risk-neutral individuals facing a consumption-savings problem. The agents are quasi-hyperbolic discounters with a common discount factor $\delta = 1$, a common period 0 (ex-ante) utility function $u_i(c_1, c_2) = c_1 + c_2$, and period 1 utility functions $v_i(c_1, c_2) = c_1 + \beta_i c_2$. Both individuals are strictly naive. At date 0, individual 1 believes that her future β equals $\hat{\beta}_1 = 0.9$, while the true value is $\beta_1 = 0$. Individual 2 believes that her future β equals $\hat{\beta}_2 = 0.98$, while the true value is $\beta_2 = 0.9$. Note that $\hat{\beta}_1 - \beta_1 > \hat{\beta}_2 - \beta_2$ and $\frac{1-\hat{\beta}_1}{1-\beta_1} < \frac{1-\hat{\beta}_2}{1-\beta_2}$, so agent 1 would be considered more naive under the discussed parametric proposals.

In period 0, both individuals are endowed with unit wealth of 1 dollar and have the opportunity to commit to a savings plan which forces them to save all consumption until period 2. This commitment plan has an interest rate of 2%. If they refuse the savings plan, then in period 1 they have the opportunity to save for period 2 and earn 3% interest; in other words, they face the choice set $x = \{(c_1, (1.03)(1 - c_1)) : 0 \leq c_1 \leq 1\}$ at date 1.

Since $\beta_i \times 1.03 < 1$ for $i = 1, 2$, both individuals would decide to consume all their endowment immediately in period 1 if given the opportunity: $\mathcal{C}_i(x) = (1, 0)$ for $i = 1, 2$. Consider now the behavior of the agents at date 0. Since $\hat{\beta}_1 \times 1.03 < 1$, individual 1 correctly anticipates that $\mathcal{C}_1(x) = (1, 0)$. Preferring the consumption plan $(0, 1.02)$ to the consumption plan $(1, 0)$, she therefore commits to the forced savings plan at date 0. In contrast, since $\hat{\beta}_2 \times 1.03 > 1$, individual 2 believes that she will select $(0, 1.03)$ from x, and thus she forgoes the profitable commitment opportunity $(0, 1.02)$. In this decision problem, individual 1 perfectly forecasts her future behavior without commitment, while individual 2 optimistically believes she will save her income when in fact she will not. So although individual 1 would be considered more naive than individual 2 by some parametric criteria in the literature, its seems insensible to call individual 1 more naive than individual 2. ■

Example 1 illuminates that some existing comparisons will lead to scenarios where the supposedly more naive individual accepts the commitment device and leaves herself better off, while the supposedly more sophisticated individual rejects the advantageous commitment and instead ends up consuming a worse alternative. In the next example, adapted from DellaVigna and Malmendier (2004), we show that these proposals also lead to scenarios where the individual deemed as more sophisticated will be subject to more exploitation from a profit-maximizing monopolist. This is because the difference between the anticipated value and the actual value of a contract for the purportedly "more naive" individual is in fact less than than the difference for the "more sophisticated" individual.

Example 2 (Monopoly profit)**.** A monopolist produces a service whose consumption results in delayed benefits, for example, a fitness club offers access to exercise that provides future health benefits. The firm offers a two-part tariff at period 0 that specifies (L, p) where L is a fixed payment like monthly dues for gym membership and p is the price of using the service like a per-visit fee at the gym. The consumer decides at date 0 whether to accept the contract. If she accepts, the consumer decides at date 1 whether to use the service. Both payments L and p are made at date 1, but the benefits of the service are not realized until date 2. Specifically, if the consumer uses the service at date 1, she receives a delayed benefit b at date 2. The cost of providing the service equals c for the firm. We assume that $b > c$, meaning that the surplus from the service is positive. The utility received by the consumer if she does not take up the contract is normalized to 0.

Now consider a quasi-hyperbolic consumer $(\hat{\beta}, \beta, \delta)$, and take $\delta = 1$ for simplicity. Suppose that the consumer considers the contract (L, p). She anticipates using the service in date 1 if and only if $\hat{\beta} b \geq p$. Therefore, at date 0, she accepts the contract if and only if

$$-L + (b - p)\mathbb{1}_{\hat{\beta} b \geq p} \geq 0. \tag{1}$$

The firm knows $(\hat{\beta}, \beta)$ and offers a contract (L, p) that maximizes its expected profit of

$$L + (p - c)\mathbb{1}_{\beta b \geq p}, \tag{2}$$

subject to Equation (1). The inequality in Equation (1) must be binding since otherwise the firm could raise L and increase its profits. This pins down the value of L as a function of p, $L = (b - p)\mathbb{1}_{\hat{\beta} b \geq p}$. Substituting this value into the profit function in Equation (2) yields

$$(b - p)\mathbb{1}_{\hat{\beta} b \geq p} + (p - c)\mathbb{1}_{\beta b \geq p},$$

which simplifies to

$$\underbrace{(b - c)\mathbb{1}_{\beta b \geq p}}_{\text{social surplus}} + \underbrace{(b - p)\mathbb{1}_{\hat{\beta} b \geq p > \beta b}}_{\text{overvaluation}}. \tag{3}$$

The first term in Equation (3) is the social surplus generated from the contract, and the second term captures the consumer's overvaluation of the surplus that she will receive from the contract due to her underestimation of her future impatience. Now consider two quasi-hyperbolic discounters such that $\hat{\beta}_1 = 1, \beta_1 b > c > \beta_2 b$, and $\hat{\beta}_2 > \beta_2$ is sufficiently close to β_2 to imply that individual 2 would be considered less naive than individual 1 based on either of the measures $\hat{\beta} - \beta$ or $(1 - \hat{\beta})/(1 - \beta)$. Since $\beta_1 b > c$, it is easy to see from Equation (3) that the profit-maximizing contract to individual 1 is the same as the optimal contract under sophistication (for instance, $L = b - c$, $p = c$), and individual 1 does not incur any welfare loss due to her naivete. But since $\beta_2 b < c$, the firm can earn more than the total social surplus by offering an exploitative contract to individual 2 that sets a price $p = \beta_2 b + \epsilon$ for some small $\epsilon > 0$. This exploitative contract guarantees that individual 2 naively accepts the contract in anticipation of using the gym, but in actuality does not use the service ex post.[5] This example shows that the firm's ability to exploit the consumer's misprediction is not monotonic in either $\hat{\beta} - \beta$ or $(1 - \hat{\beta})/(1 - \beta)$, but instead depends on the consumer's overvaluation of contracts.[6] ■

2.2 Our Proposed Alternative

These examples demonstrate that existing parametric comparisons of naivete for the quasi-hyperbolic discounting model may lead to situations where the individual forgoes

[5] The use of an exploitative contract by the firm in this example does not rely on the assumption of monopoly power. It is not difficult to show that introducing competition between firms will drive down the fixed fee L, but firms will continue to set a price p that both deters individual 2 from using the service ex post and causes her to have the incorrect ex-ante belief that she will use the service.

[6] Note that in their related analysis, DellaVigna and Malmendier (2004) fix the value of β. While this example shows that monopoly profits are not, in general, monotone in the difference $\hat{\beta} - \beta$, in the special case of fixed β our results will imply that increasing $\hat{\beta}$ leads to an increase in overvaluations and hence monopoly profits.

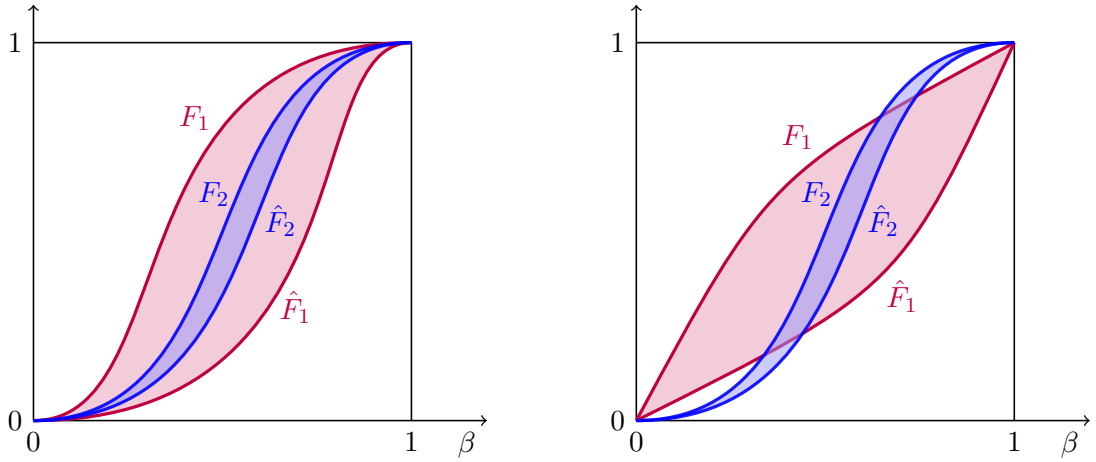
a beneficial commitment or gives up more rents to a monopolist, despite being deemed as more sophisticated. These examples were specifically motivated by and designed to foreshadow our proposed nonparametric comparisons. We propose two approaches to comparing naivete across individuals. In the first approach, we compare underdemand for commitment and say an individual is more naive if she rejects more beneficial commitments than the other. Comparing demand for commitment is economically relevant because commitment devices are often proposed as a policy intervention to manage self-control problems, as in Example 1. In the second approach, we say an individual is more naive if her overvaluation for a menu, measured as the difference between her anticipated indirect utility and her actual indirect utility, is larger than the other's overvaluation for that menu. This approach is economically relevant because a monopolist can extract the overvaluation through a fixed-fee component of a contract, as demonstrated in Example 2. We will show in Section 3 that for the case of deterministic choice these two approaches to comparing naivete converge, so an agent forgoes more advantageous commitments if and only if she overvalues menus more. In particular, for the quasi-hyperbolic discounting model, comparing underdemand for commitment and comparing overvaluation both yield the same parametric restriction: either $\hat{\beta}_1 \geq \hat{\beta}_2 \geq \beta_2 \geq \beta_1$, or individual 2 is sophisticated ($\hat{\beta}_2 = \beta_2$). Note that this is more demanding than the mentioned criteria of comparing differences or ratios of $\hat{\beta}$ and β. That is, our criterion is less finely ordered and leaves some pairs of individuals as unordered that these prior quantitative comparisons would erroneously rank.

2.3 Stochastic Present Bias

Our general results compare the naivete of individuals who exhibit randomness in their ex-post choices due to uncertainty about the nature or degree of time-inconsistency. The relationship between our two suggested approaches to comparing naivete—comparing underdemand for commitment and comparing overvaluation of menus—is more subtle in the general case than suggested by the special deterministic case. When choice is possibly random, the equivalence breaks down and the comparison of underdemand for commitment is a strictly more demanding criterion than the comparison of overvaluations.

Here, we preview these differences by continuing to focus attention on quasi-hyperbolic discounting. Consider a generalization of the $(\hat{\beta}, \beta, \delta)$ model where the level of present bias β is stochastic, governed by the cumulative distribution function (abbreviated as cdf) F. To model naivete, the individual's belief about her future behavior is also stochastic, but governed by the distribution $\hat{F}$.

The standard extension of an order on a deterministic space to the space of beliefs is through stochastic dominance. This turns out to nicely extend the absolute definition of naivete. Recall that a deterministic quasi-hyperbolic discounter is naive if and only



(a) Equations (4) and (5) are both satisfied. **(b)** Equation (5) is satisfied, but not (4).

Figure 1: Comparisons of naivete for stochastic choice.

if $\hat{\beta} \geq \beta$. We will show that a random quasi-hyperbolic discounter whose present bias actually follows distribution F, but is believed to follow distribution $\hat{F}$, satisfies our behavioral definition of naivete if and only if $\hat{F}(\beta) \leq F(\beta)$ for all β. That is, she is naive if and only if her belief $\hat{F}$ first-order stochastically dominates the distribution F.

While prior notions of comparative naivete were not without controversy in the case of deterministic choice (as our prior examples illustrated), comparisons of naivete for stochastic choice have been largely unexplored in the existing literature. Our results help to remedy this gap. We will show that our two proposed criteria for comparing naivete lead to novel parametric restrictions. First, an individual has greater underdemand for commitment than another if and only if either

$$\hat{F}_1(\beta) \leq \hat{F}_2(\beta) \leq F_2(\beta) \leq F_1(\beta) \tag{4}$$

for all β, or the second individual is sophisticated ($\hat{F}_2 = F_2$). When each of these distributions is concentrated on a single value, this first-order stochastic dominance relationship between the distributions specializes to the aforementioned restriction for the deterministic case: $\hat{\beta}_1 \geq \hat{\beta}_2 \geq \beta_2 \geq \beta_1$. Second, the comparison of overvaluations across individuals yields a different, and more permissive, ordering in stochastic environments. We show that an individual has greater overvaluation than another for every menu if and only if

$$F_1(\beta) - \hat{F}_1(\beta) \geq F_2(\beta) - \hat{F}_2(\beta) \tag{5}$$

for all β.

Equation (5) is a strictly weaker restriction than Equation (4), as Figure 1 illustrates. Thus, the approach of ranking naivete by the level of overvaluation is strictly more

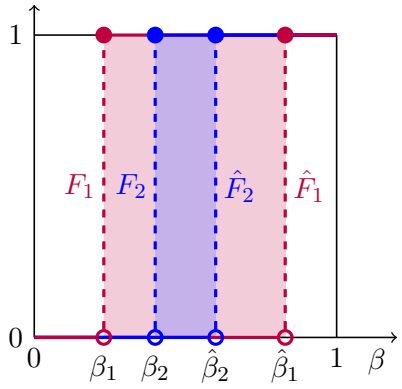


Figure 2: Equation (5) implies Equation (4) for degenerate distributions (deterministic choice), provided $\hat{\beta}_2 > \beta_2$.

general and will order more pairs of individuals than the approach of ranking naivete by underdemand for commitment. However, as noted previously, the two approaches are equivalent in the special case of deterministic choice. When each distribution is a deterministic Dirac measure, the differences $F_i(\beta) - \hat{F}_i(\beta)$ can only take values of 1 or 0, and both of our parametric restrictions become equivalent to $\hat{\beta}_1 \geq \hat{\beta}_2 \geq \beta_2 \geq \beta_1$ (or $\hat{\beta}_2 = \beta_2$). Figure 2 illustrates why these generally different orderings become equivalent without randomness.

The following examples illustrate how our two comparative measures can be used in practice. Example 3 shows that for some applications, the more permissive ordering of naivete captured by Equation (5) is the appropriate comparative, and it is not necessary to resort to the more restrictive ordering of naivete from Equation (4). However, Example 4 then illustrates that for other applications, our weaker ordering can lead to counterintuitive behavior, and the more restrictive ordering is instead appropriate.

Example 3 (Monopoly profit with random choice)**.** Consider the setting of Example 2, the only difference being that the agent is a random quasi-hyperbolic discounter who believes that her future β is distributed according the cdf $\hat{F}$, while the true distribution is given by F. The incentive-compatibility constraint from Equation (1) is modified into

$$-L + \int_{p/b}^{1} (b - p)\, d\hat{F}(\beta) \geq 0, \tag{6}$$

while the expected profit of the firm now equals

$$L + \int_{p/b}^{1} (p - c)\, dF(\beta). \tag{7}$$

Rewriting Equation (7) by substituting the value of L obtained in Equation (6) shows

that the firm's expected profit equals

$$\underbrace{\int_{p/b}^{1}(b-c)\,dF(\beta)}_{\text{social surplus}}+\underbrace{\int_{p/b}^{1}(b-p)\,d(\hat{F}(\beta)-F(\beta))}_{\text{overvaluation}}. \tag{8}$$

When F and $\hat{F}$ have continuous cumulative distribution functions, the monopoly profit can be written as

$$\underbrace{(b-c)\Big(1-F\Big(\frac{p}{b}\Big)\Big)}_{\text{social surplus}}+\underbrace{(b-p)\Big(F\Big(\frac{p}{b}\Big)-\hat{F}\Big(\frac{p}{b}\Big)\Big)}_{\text{overvaluation}}.$$

This final expression makes most clear why overvaluation is increasing in $F(\cdot)-\hat{F}(\cdot)$, although this is generally true even without a continuous cdf. ■

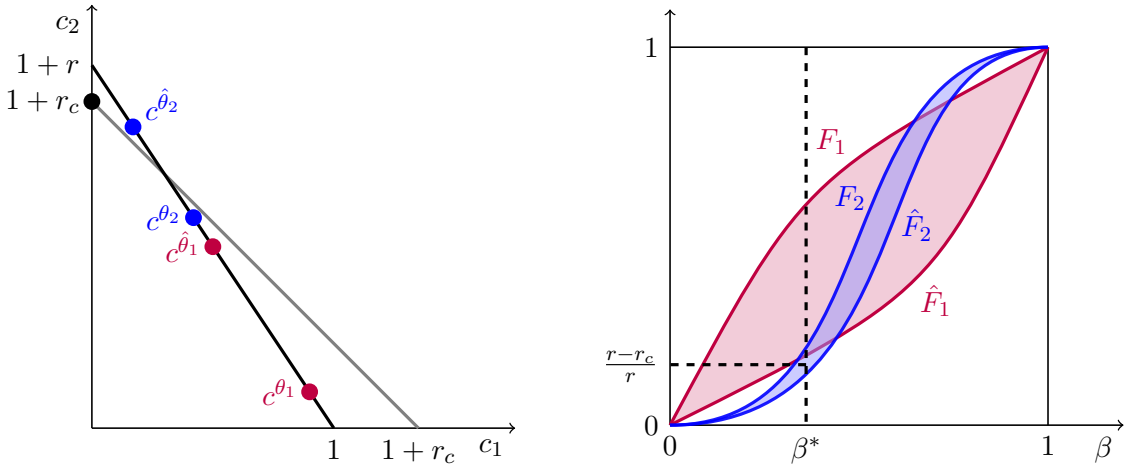
As mentioned, while Equation (5) is equivalent to comparing overvaluation, it is too broad of a criterion to capture underdemand for commitment. That is, one individual might have greater overvaluations for all menus, yet fail to exhibit greater underdemand for commitment. Instead, to compare take-up of commitment devices, the appropriate parametric restriction is the one in Equation (4). We illustrate this point by revisiting the consumption-savings problem from Example 1, but with a simple stochastic element.

Example 4 (Consumption-savings problem with random choice)**.** As in Example 1, suppose $x=\{(c_1,(1+r)(1-c_1)):0\leq c_1\leq 1\}$. The savings commitment plan forces savings and has a return r_c, giving consumption $(1+r_c)$ in period 2. Individuals $i=1,2$ have no present bias ($\beta=1$) with probability θ_i and have the common present bias parameter $\beta=\beta^*$ with the remaining probability $1-\theta_i$.[7] The individuals have naive beliefs that they will instead have no present bias with probability $\hat{\theta}_i\geq\theta_i$ and will have present bias β^* with probability $1-\hat{\theta}_i$. The scope for naivete is thus in overoptimism about the probability of avoiding present bias, and not in the level of that present bias if it actualizes. Assuming $\beta^*(1+r)<1$, individual i will therefore choose $(c_1,c_2)=(0,1+r)$ with probability θ_i and $(c_1,c_2)=(1,0)$ with probability $1-\theta_i$. Denote the vector of expected values of actual consumption from x in each period by

$$c^{\theta_i}=(1-\theta_i,\theta_i(1+r)),$$

and likewise let $c^{\hat{\theta}_i}$ denote the anticipated expected values of consumption from x. Since the utility functions are linear in consumption, the actual ex-ante expected utility for an

[7] That is, $F_i(\beta)=0$ for $\beta\in[0,\beta^*)$, $F_i(\beta)=1-\theta_i$ for $\beta\in[\beta^*,1)$, and $F_i(1)=1$. This simple binary model of stochastic temptation was previously studied by Eliaz and Spiegler (2006) and Chatterjee and Krishna (2009).



(a) Expected values of consumption for perceived and actual stochastic choice.

(b) Individual 2 mistakenly forgoes the savings commitment device.

Figure 3: Stochastic naivete and the uptake of savings commitment devices.

individual is therefore[8]

$$(1-\theta_i)+\theta_i(1+r)=1+\theta_i r.$$

Given her naive beliefs, individual i instead anticipates utility $1+\hat{\theta}_i r$ from the menu x. Suppose

$$\frac{r_c}{r}>\theta_1,\theta_2 \qquad \text{and} \qquad \hat{\theta}_2>\frac{r_c}{r}>\hat{\theta}_1. \tag{9}$$

The first inequality implies that $r_c>\theta_i r$ for $i=1,2$, so both individuals would benefit from the savings commitment device. The second inequality implies that individual 2 will forgo the commitment device, while individual 1 will make use of the device. Thus, 2 exhibits greater underdemand for commitment than 1 in this decision problem. However, it is easy to see that the parameter restrictions in Equation (9) can be satisfied even when individual 2 has lower overvaluations than 1, that is, when $\hat{\theta}_2-\theta_2\leq\hat{\theta}_1-\theta_1$. This example demonstrates that individual 2 can have lower overvaluations than individual 1, yet still exhibit greater underdemand for commitment.[9] Figure 3a illustrates anticipated and actual expected values of consumption that satisfy these conditions together with the ex-ante indifference curve through the commitment consumption plan $(0,1+r_c)$. ■

Note that the basic argument in Example 4 applies to any perceived and actual cumulative distributions functions for present bias factors β that violate Equation (4). To illustrate, suppose F_i and $\hat{F}_i$ are continuous for $i=1,2$ and let $\beta^*=1/(1+r)$. Then, when confronted with the consumption-savings problem from this example, individual i

[8]Risk neutrality is assumed in this example for expositional simplicity, but is in no way central to the main qualitative conclusions. Our main results neither require nor assume risk neutrality.

[9]In contrast, Equation (4), which corresponds to $\hat{\theta}_1\geq\hat{\theta}_2\geq\theta_2\geq\theta_1$ in this example, is incompatible with Equation (9) and hence requires that individual 1 have greater underdemand for commitment.

will choose to save with probability $\theta_i = 1 - F_i(\beta^*)$, but naively believes that she will save with probability $\hat{\theta}_i = 1 - \hat{F}_i(\beta^*)$. Thus, by Equation (9), if $(r - r_c)/r < F_1(\beta^*), F_2(\beta^*)$ then the savings commitment device is beneficial to both individuals, and if $\hat{F}_2(\beta^*) < (r - r_c)/r < \hat{F}_1(\beta^*)$ then individual 2 passes up this beneficial commitment device and only individual 1 utilizes it. Figure 3b illustrates how this is possible even in the case where individual 1 has greater overvaluations than individual 2.[10]

The main distinction between the two previous examples is that in Example 3, the monopolist can design a custom contract for each individual in the population using knowledge of their distributions F and $\hat{F}$. This assumption is reasonable if the population is roughly homogeneous or if the firm has detailed consumer data on potential customers. In this case, the exploitation of an individual due to naivete is increasing in her overvaluations (equivalently, the parametric comparison in Equation (5)). In contrast, Example 4 considers the welfare impact of a single fixed policy instrument offered across an entire population of individuals who might be heterogeneous in their distributions F and $\hat{F}$. In this case, the welfare loss due to naivete increases with greater underdemand for commitment (equivalently, Equation (4)), but not necessarily in the level of overvaluations. Thus, which of our two proposed comparisons of naivete is most suitable depends on the specifics of the information structure and whether contracts (or commitment devices) are designed at the individual or population level.

3 Deterministic Choice

We begin our analysis by examining naivete in the context of deterministic choice and beliefs. The next section will study the more general environment of random choice. We feel that random choice is an important consideration given that naivete regards possibly incorrect beliefs about future behavior. For now, this section focuses on the special case of deterministic choice to establish intuition and avoid some of the additional technical details required to formalize random dynamic inconsistency. While some insights are general and extend to the random case, others are interestingly limited to deterministic choice and have subtle variations when considering random choice.

3.1 Primitives

We study a two-stage model with an agent who initially decides on a menu of several options and subsequently selects a particular option from that menu.

[10] As might be evident, another possibility when Equation (4) fails to hold is that $\hat{F}_1(\beta^*), \hat{F}_2(\beta^*) < (r - r_c)/r$ and $F_1(\beta^*) < (r - r_c)/r < F_2(\beta^*)$. In this case, both individuals pass up the savings commitment device, but this decision is only a mistake for individual 2.

Let C be a compact and metrizable space of outcomes. Let $\Delta(C)$ denote the set of lotteries (countably additive Borel probability measures) over C, with typical elements $p, q, \ldots \in \Delta(C)$. When it causes no confusion, we slightly abuse notation and write c in place of the degenerate lottery $\delta_c \in \Delta(C)$ supported on c. Let $\mathcal{K}(\Delta(C))$ denote the family of nonempty compact subsets of $\Delta(C)$ with typical elements $x, y, \ldots \in \mathcal{K}(\Delta(C))$. These sets are interpreted as menus or budget sets. The menu determines the level of flexibility versus commitment, with larger menus providing more flexibility and smaller menus providing more commitment. An *expected-utility function* is a continuous function $u : \Delta(C) \to \mathbb{R}$ such that $u(\alpha p + (1-\alpha)q) = \alpha u(p) + (1-\alpha)u(q)$ for all lotteries p, q. A function is *nontrivial* if it is not constant. We write $u \approx v$ when u and v are expected-utility functions and u is a positive affine transformation of v. For a fixed expected-utility function u and menu x, let $B_u(x) \equiv \operatorname{argmax}_{p \in x} u(p)$.

We consider a pair of behavioral primitives. The first primitive is a preference relation $\succsim$ on $\mathcal{K}(\Delta(C))$, with indifference $\sim$ and strict preference $\succ$ defined as usual. This primitive provides insight into the agent's projection regarding her future behavior. The behavior encoded in $\succsim$ is taken before the direct experience of temptation but while (possibly incorrectly) anticipating its future occurrence. This is an economically important primitive, because it also captures demand for commitment, which is an important consideration when analyzing commitment instruments without mandatory take-up. The second primitive is a (deterministic) choice function $\mathcal{C} : \mathcal{K}(\Delta(C)) \to \Delta(C)$.[11] This is the standard model for economic choice, and it records the individual's actual choices from menus while experiencing temptation.

3.2 Absolute Naivete

We now introduce our nonparametric definition for absolute naivete. In a nutshell, an individual is naive if she overvalues a menu relative to the actual choice that she would ultimately make from that menu.

Definition 1. *An individual is* sophisticated *if* $x \sim \{\mathcal{C}(x)\}$ *for all menus* x*. An individual is* naive *if* $x \succsim \{\mathcal{C}(x)\}$ *for all menus* x*. An individual is* strictly naive *if she is naive and not sophisticated.*

A sophisticated individual correctly anticipates choosing $\mathcal{C}(x)$ from x. A naive individual erroneously overvalues the option to retain the other alternatives in x, thinking that her final choice will be more virtuous than $\mathcal{C}(x)$. Many decisions that open or

[11] Note that all of our definitions and theorems can be modified to accommodate choice correspondences instead of choice functions. We choose to work with choice functions in order to make our primitives as undemanding as possible in terms of richness of the required choice data.

restrict future options can be modeled as menus and can therefore be related to our definitions. For example, purchasing an unlimited gym membership can be modeled as the option set that includes any number of monthly visits, each paired with the fixed cost of the membership. Similarly, many financial decisions, like opening a line of credit or placing savings in an illiquid retirement account, can be viewed as adding or removing options from future decisions. In these examples, we argue that some consumers may strictly prefer x to $\mathcal{C}(x)$, indicating a lack of sophistication in the form of excess optimism about their future choices. The examples in Section 2 provide concrete illustrations of such violations of sophistication. Of course, the opposite violation with overdemand for commitment, where $\{\mathcal{C}(x)\} \succ x$, is also potentially interesting and certainly indicates a violation of sophistication (in this case, in the form of excess pessimism). Many of our results have straightforward analogous statements with appropriate changes in signs for this opposite case. However, this direction receives less attention and seems less empirically relevant, so we focus our analysis on traditional naivete in the form of underdemand for commitment throughout the paper.[12]

Our definition of sophistication is similar to the Independence of Redundant Alternatives axiom that Gul and Pesendorfer (2005) use to study deterministic choice in a finite-outcome setting, but our definition of naivete has not been considered in the literature.[13] After presenting the main result of this section, we will discuss the connection to other related papers, most notably Noor (2011), and we will also touch on some of the assumptions implicit in our definition.

The ubiquitous Strotz model of dynamic inconsistency offers a general application for these concepts. The sophisticated Strotz model is specified by two preferences. The first is her ex-ante commitment preference over future consumption, as represented by the utility function u. The second is her temptation preference that governs her actual consumption choices at the ex-post stage, as represented by the utility function v. Naivete requires divergence between believed and actual consumption. Specification of a naive Strotz individual therefore requires a third preference to capture her possibly erroneous beliefs about her future behavior, as represented by the utility function $\hat{v}$.[14]

Definition 2. *A* Strotz representation *of* $(\succsim, \mathcal{C})$ *is a triple* $(u, v, \hat{v})$ *of nontrivial expected-utility functions such that the function* $U : \mathcal{K}(\Delta(C)) \to \mathbb{R}$ *defined by*

$$U(x) = \max_{p \in B_{\hat{v}}(x)} u(p)$$

[12] What is difficult and we leave open is what happens when violations of sophistication are not uniformly in the same direction, so $\{\mathcal{C}(x)\} \succ x$ for some menus x and $\{\mathcal{C}(y)\} \prec y$ for other menus y. We suspect that not much can be said in that case, at least not using the techniques in this paper.

[13] Grant, Kajii, and Polak (2000) and Siniscalchi (2011) employ similar ideas to formalize sophistication in different settings of belief updating.

[14] Recall that a utility function is nontrivial if it is not constant, and $B_v(x) = \operatorname{argmax}_{q \in x} v(q)$.

is a utility representation of $\succsim$ *and*

$$\mathcal{C}(x) \in B_u(B_v(x)).$$

While she anticipates maximizing $\hat{v}$, a naive Strotzian agent's ex-post behavior $\mathcal{C}$ actually maximizes v. Note that both the domain of choice and the representation itself are quite general. For example, C could be a set of infinite-horizon consumption streams, and hence quasi-hyperbolic discounting (β-δ preferences) is a special case of the Strotz representation (see Section 3.4).

The following result demonstrates that our behavioral definition of naivete characterizes sharp parametric restrictions on $\hat{v}$ and v. A naive individual believes that her future behavior will be more virtuous than it actually is. For the parameters of the Strotz model, this means that the anticipated temptation utility $\hat{v}$ is more aligned with the commitment utility u than the actual utility v that will govern future consumption. The alignment has a specific structure: $\hat{v}$ is a linear combination of u and v, that is, $\hat{v} \approx \alpha u + (1-\alpha)v$.[15] The belief $\hat{v}$ puts additional unjustified weight on the normative utility u, but aggregates u with v in a linear manner. Other cases where the believed temptation differs from the actual temptation in a less structured way are also allowed in our primitives and correspond to a form of misprediction, but these cases fail to satisfy our proposed behavioral test of naivete. For example, our definition excludes an individual who actually will be tempted to indulge in sweet treats but believes she will be tempted to indulge in salty treats. This structure also relies crucially on the linear structure of the domain of lotteries and the assumed expected-utility functions.

Definition 3. *Let* $u, v, \hat{v}$ *be expected-utility functions. Then* $\hat{v}$ *is* more u-aligned *than* v*, written as* $\hat{v} \gg_u v$*, if either* $\hat{v} \approx \alpha u + (1-\alpha)v$ *for some* $\alpha \in [0,1]$ *or* $v \approx -u$*.*

Any strict convex combination of u and v is more u-aligned than v. One case that is tedious is when $v = -u$ because u and $-u$ have, up to positive affine transformations and excluding trivial preferences, no convex combinations except u and $-u$ themselves. We therefore adopt as convention that any expected-utility function is more u-aligned than $-u$, since $-u$ is maximally divergent from u.[16]

Theorem 1. *Suppose* $(\succsim, \mathcal{C})$ *has a Strotz representation* $(u, v, \hat{v})$*. Then the individual is naive if and only if* $\hat{v} \gg_u v$ *(and is sophisticated if and only if* $\hat{v} \approx v$*).*

[15] Recall that $\hat{v} \approx \alpha u + (1-\alpha)v$ means that $\hat{v}$ is a positive affine transformation of $\alpha u + (1-\alpha)v$ and hence the two functions share the same set of maximizers.

[16] Incorporating this special exception for this boundary case into the definition of the order $\gg_u$ also tidies the conclusions of the following characterization theorems, that would otherwise have to read "$\hat{v} \gg_u v$ or $v \approx -u$."

Theorem 1 is a special case of one of the main results of Section 4 (see Theorem 3), where we turn to the more general case of random choice and uncertain beliefs.

We close this section with a discussion of several important assumptions that are implicit in our definitions of sophistication and naivete, which will also help to further clarify how this paper connects with the related literature. First, our definitions assume there is not a nontrivial preference for flexibility, that is, there is no uncertainty about what constitutes virtuous behavior. Suppose an agent faces no temptation but is unsure what her future tastes will be. Then she may prefer to keep the flexibility of the menu rather than be forced to choose ex ante, in order to maintain the option value of waiting to see what taste realizes. While the possibility of uncertain normative preferences is substantively important, we suppress that consideration here and focus attention exclusively on misprediction of temptation.[17] Throughout this paper, we implicitly assume that the normative preference or taste has already been realized or is known to the decision maker. Of course, in many policy applications, balancing the benefits of flexibility and of commitment is important, as in deciding penalties for early withdrawals from retirement accounts. However, even a parametric model of naive choice with both kinds of uncertainty is still outstanding. Section 5.2 discusses some of these issues in more depth.

Second, in our definition, inferring sophistication from $x \sim \{\mathcal{C}(x)\}$ assumes consequentialism, that is, the individual is indifferent between committing to her (correctly) anticipated choice $\mathcal{C}(x)$ from x at the ex-ante stage and selecting the menu x with the belief that she will choose $\mathcal{C}(x)$ ex post. Put differently, adding or removing unchosen options has no effect on the evaluation of a menu. In contrast, an individual who exerts costly willpower to avoid choosing tempting options as in Gul and Pesendorfer (2001) does not evaluate a menu only by its choice consequences. In this case, she may strictly prefer to remove these unchosen temptations.[18] In Section 5.1, we show that if individuals can exert costly self-control then our behavioral test of naivete can lead to false negatives, but not false positives: Satisfying our definition of naivete in the presence of costly self-control implies *a fortiori* that the individual is naive; however, violating our definition of naivete does not preclude the possibility that an individual with Gul and Pesendorfer (2001) preferences is in fact naive.

Since our definitions do not tightly characterize sophistication and naivete in the case of the self-control preferences of Gul and Pesendorfer (2001), a natural question is whether a tight characterization is in fact possible. The answer is affirmative, but only in the case of deterministic choice. As an axiom en route to characterizing a recursive

[17] Ahn and Sarver (2013) characterize sophistication for the case of uncertain normative preferences and demand for flexibility, rather than the demand for commitment that arises with temptation, and propose the failure to anticipate future preference realizations as a form of unforeseen contingencies.

[18] Alternatively, an agent that derives self-satisfaction from exercising willpower may strictly prefer to include tempting options that she will not consume.

model of temptation, Noor (2011) proposes a definition of sophistication for the self-control model.[19] Noor's definition requires that whenever $\{p\} \succ \{q\}$ (that is, p is the preferred ex-ante commitment), p is the unique ex-post choice from the menu containing both lotteries if and only if $\{p, q\} \succ \{q\}$. While not intended for the consequentialist models we have in mind in this paper, there is nonetheless a close connection between his sophistication condition and the one proposed in this paper when applied to the Strotz model. A companion paper, Ahn, Iijima, and Sarver (2016), modifies the definition of sophistication from Noor (2011) to provide a tight behavioral characterization of naivete for both deterministic self-control preferences and deterministic Strotz preferences. In that paper, it is shown that Noor's definition of sophistication is equivalent to the one in our Definition 1 for Strotz preferences, but the two definitions diverge for self-control preferences. Moreover, that paper extends Noor's work by proposing a definition of naivete and characterizing a recursive model of self-control that allows for naivete as well as sophistication. However, Ahn, Iijima, and Sarver (2016) also show that a tight characterization of naivete is impossible when self-control and random choice are simultaneously permitted; this impossibility result is closely related to the lack of unique identification of parameters in a random self-control representation. In contrast, Definition 1 in this paper extends easily to the case of random choice and characterizes sophistication and naivete within the random Strotz model, as we will establish in Section 4.

3.3 Comparative Naivete

In this section, we introduce two definitions for comparing naivete across individuals. The first naturally extends our proposed test for absolute naivete by comparing the sets of forgone opportunities for beneficial commitment. The second directly measures the difference between anticipated and actual indirect utilities for menus. We show that, in the deterministic case, both definitions turn out to be equivalent.

Recall that a naive agent satisfies $x \succsim \{\mathcal{C}(x)\}$, that is, there is a potential gap between her value for the menu x and the value of her eventual choice $\mathcal{C}(x)$. To compare the degree of naivete across agents, our first definition measures the size of this gap using underdemand for commitment.

Definition 4. *Individual 1 is* more naive *than individual 2 if, for all menus x and lotteries p,*

$$x \succ_2 \{p\} \succ_2 \{\mathcal{C}_2(x)\} \implies x \succ_1 \{p\} \succ_1 \{\mathcal{C}_1(x)\}.$$

Any singleton $\{p\}$ that is ex-ante ranked strictly between a menu x and its resulting choice $\{\mathcal{C}(x)\}$ presents a welfare-enhancing commitment that will be unfortunately declined: p is preferred ex ante to $\mathcal{C}(x)$ yet the individual chooses to maintain the flexibility

[19] A variation of this axiom is also used by Kopylov (2012).

of x due to the naive belief that she will make a more virtuous choice. Thus, the beneficial opportunity to commit to consuming p instead of $\mathcal{C}(x)$ will be naively rejected. Definition 4 classifies an individual as more naive than another if she has greater underdemand for commitment, that is, if she forgoes more beneficial commitments. Singleton menus like $\{p\}$ are especially useful in comparing naivete because there is no ambiguity about the eventual choice from such menus.[20]

Our second proposal for comparing naivete is based on the utility difference between a menu x and the actual choice $\mathcal{C}(x)$. In many applications of time inconsistency and naivete to industrial organization and contract theory, the firm's ability to extract excess surplus is tied to the extent to which the individual overestimates the utility that she will receive from a set of options or from an action-dependent contract.[21] For example, if a monopolist can charge a fixed fee, then it can extract the difference between the anticipated and actual indirect utility for a contract, above the standard extraction of the social surplus. This motivates the following definition.

Definition 5. *Suppose* $(\succsim, \mathcal{C})$ *has a Strotz representation* $(u, v, \hat{v})$. *The* coefficient of overvaluation *of a menu* x *is defined by:*

$$OV(x) = \underbrace{\max_{p \in B_{\hat{v}}(x)} u(p)}_{\text{believed indirect utility}} - \underbrace{\max_{p \in B_v(x)} u(p)}_{\text{actual indirect utility}}.$$

Our second definition of comparative naivete requires that $OV_1(x) \geq OV_2(x)$ for all menus x, reflecting the idea that individual 1 makes a larger mistake when she contemplates her future behavior than individual 2. Early in this paper, Example 2 studied a monopolist designing contracts for naive agents. There, the coefficient of overvaluation appeared in the monopolist's profit in Equation (3), where it was interpreted as the extra profit, above and over the standard social surplus, extracted by the monopolist because of the agent's mistaken beliefs. Thus, another way to interpret this comparison is that a monopolist can extract more excess surplus from individual 1 (above social surplus) than from individual 2. Note that even without the assumption of quasilinear preferences

[20] By definition, an agent is less naive if she takes better advantage of full commitments, that is, of singleton commitments. However, this does not imply that a less naive individual will make better use of partial commitment devices. That is, suppose $x \subset y$ and $\mathcal{C}(x) \succ \mathcal{C}(y)$, but x is not a singleton. Then it is entirely possible for the more naive agent to take up the beneficial partial commitment $(x \succ y)$, while the less naive agent forgoes that partial commitment $(y \succ x)$. Several models in the literature demonstrate that naive individuals may mistakenly pay (in the form of money, effort, or foregone options) for partial commitment devices that are too weak to actually be effective, e.g., Heidhues and Kőszegi (2009). In these situations, becoming less naive in the sense of Definition 4 may make an individual worse off: She becomes sophisticated enough to recognize the potential benefits of commitment, but not sophisticated enough to recognize that a partial commitment will leave temptations that she will be unable to resist.

[21] Some applications are reviewed in Spiegler (2011) and Kőszegi (2014, Section 6).

that was used in Example 2, this measure $OV(x)$ of overvaluation has cardinal meaning because we consider lotteries and expected-utility preferences.

Underdemand for commitment provides one criterion for comparing naivete. Overvaluations provide another criterion. The following theorem shows that both approaches are equivalent in the deterministic setting. It also shows that these comparisons of naivete for two individuals are equivalent to a direct restriction on the parameters of the Strotz representations for those individuals.

Theorem 2. *Suppose* $(\succsim_1, \mathcal{C}_1)$ *and* $(\succsim_2, \mathcal{C}_2)$ *are naive and have Strotz representations* $(u, v_1, \hat{v}_1)$ *and* $(u, v_2, \hat{v}_2)$*. Then the following are equivalent:*

1. *Individual 1 is more naive than individual 2.*
2. $OV_1(x) \geq OV_2(x)$ *for all menus* x*.*
3. $\hat{v}_1 \gg_u \hat{v}_2 \gg_u v_2 \gg_u v_1$*, or* $\hat{v}_2 \approx v_2$ *(individual 2 is sophisticated).*

Theorem 2 is a corollary of our more general results for stochastic choice in Section 4 (see Theorems 4 and 5).

When comparing a pair of naive individuals, individual 1 always has greater underdemand for commitment whenever individual 2 is sophisticated. This is because individual 2 never forgoes a beneficial commitment, thus rendering the required implication in Definition 4 vacuously true. That is, all sophisticated individuals are less naive than all naive individuals. In all other cases, where both individuals are strictly naive, the last condition in Theorem 2 imposes sharp and intuitive restrictions on the believed and actual temptations of both agents in the Strotz model: While they share common normative preferences over singleton commitments, individual 1 is more optimistic about her future behavior than individual 2, as reflected in the requirement $\hat{v}_1 \gg_u \hat{v}_2$. However, individual 1's actual ex-post choices are even less virtuous than individual 2's choices, as reflected in $v_2 \gg_u v_1$. A more naive individual is more optimistic about the virtuousness of her future behavior while actually exercising less virtue. In our view, since naivete concerns the difference between believed and actual behavior, $\hat{v}$ and v, both parameters should be implicated in comparing naivete.[22] Geometrically, Figure 1 illustrates that if individual 2 is strictly naive then comparative naivete implies that both individuals' anticipated temptations $\hat{v}_i$ and actual temptations v_i are convex combinations of the shared commitment utility u and the more naive individual's actual temptation v_1, progressively located on the arc connecting u and v_1.

[22]See also Lemma 3 in the appendix, which shows that with shared commitment preferences, individual 1 is more naive than individual 2 if and only if she is less temptation averse and rejects more commitments ex ante (that is, $\{p\} \succ_2 x$ whenever $\{p\} \succ_1 x$) and she makes less virtuous choices from every menu ex post (that is, $\{p\} \succ_1 \{\mathcal{C}(x)\}$ whenever $\{p\} \succ_2 \{\mathcal{C}(x)\}$).

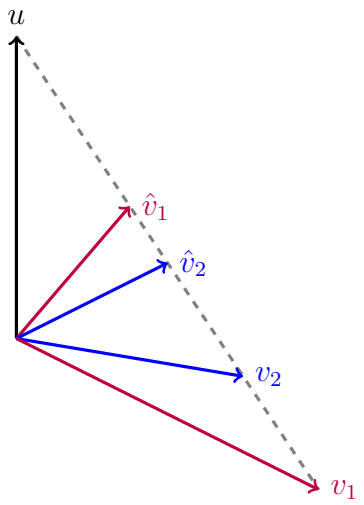


Figure 4: Alignment of believed and actual utilities implied by comparative naivete in the deterministic Strotz representation, in the case where individual 2 is strictly naive (Theorem 2).

As suggested in Section 2 and formally proven in Section 4, with random choices, while having greater underdemand for commitment will imply having higher overvaluations, the converse fails and one can have higher overvaluation without having greater underdemand for commitment. That is, in general, comparing naivete through underdemand for commitment is a more strenuous criterion that orders fewer individuals than comparing overvaluations, and the equivalence stated in Theorem 2 fails to generalize to random choice.

3.4 Application to Quasi-Hyperbolic Discounting

To illustrate the usefulness of our definitions, in this subsection we consider their implications for the ubiquitous quasi-hyperbolic discounting model. These applications extend the insights observed in the examples in Section 2.

Let $C = [a, b]^{\mathbb{N}}$ be a set of infinite-horizon consumption streams, with elements $c = (c_1, c_2, \dots) \in C$.[23] A lottery $p \in \Delta(C)$ resolves immediately and yields a consumption stream. We focus on the simple case with one-shot resolution of uncertainty for expositional parsimony, but all of the following results generalize to richer settings that incorporate temporal lotteries or true dynamic choice.[24] In these more general dynamic environments, simple atemporal lotteries over consumption streams provide sufficient

[23] The product topology on C is compact and metrizable.

[24] Kreps and Porteus (1978) provide the first complete analysis of dynamic choice with uncertainty that resolves gradually through time (i.e., temporal lotteries). The models of temptation in Gul and Pesendorfer (2004) and Noor (2011) use an infinite-horizon version of such a setting.

choice observations to apply the following comparative statics. Note that our treatment here is not fully dynamic, because the entire stream of consumption is settled immediately. This allows us to finesse the agent's assessments of her future beliefs, her future beliefs about even further future beliefs, and so on. Our point here is that sophistication and naivete can be distinguished without appeal to these higher-order epistemic assessments.

Suppose the commitment preference is represented by an expected-utility function whose values $u(c) = u(\delta_c)$ over deterministic streams (that is, whose Bernoulli utility indices) comply with exponential discounting,

$$u(c) = \sum_{t=1}^{\infty} \delta^{t-1} w(c_t), \tag{10}$$

for some instantaneous utility function $w : [a, b] \to \mathbb{R}$. The quasi-hyperbolic discounting model captures present bias with an additional discount factor applied to all future periods: If the present-bias factor is β, then ex-post (period 1) choice from a menu of consumption streams x will maximize

$$v_\beta(c) = w(c_1) + \beta \sum_{t=2}^{\infty} \delta^{t-1} w(c_t). \tag{11}$$

In the deterministic quasi-hyperbolic discounting model, the individual's ex-ante (period 0) behavior may reflect an incorrect belief that her future present-bias parameter is $\hat{\beta}$, while her ex-post behavior actually uses the present-bias parameter β. It is immediate that this choice procedure corresponds to a special case of the deterministic Strotz representation.

Definition 6. *A* quasi-hyperbolic (QH) representation *of* $(\succsim, \mathcal{C})$ *is a tuple* $(w, \beta, \hat{\beta}, \delta)$ *of a continuous and nontrivial function* $w : [a, b] \to \mathbb{R}$ *and scalars* $\beta, \hat{\beta} \in [0, 1]$ *and* $\delta \in (0, 1)$*, such that* $(u, v_\beta, v_{\hat{\beta}})$ *defined as in Equations* (10) *and* (11) *for these parameters is a Strotz representation for* $(\succsim, \mathcal{C})$*.*

Corollaries 1 and 2 illustrate the implications of our absolute and comparative definitions in the QH representation. These results follow immediately from Theorems 1 and 2, respectively, together with the observation that $v_{\hat{\beta}} \gg_u v_\beta$ if and only if $\hat{\beta} \geq \beta$.

Corollary 1. *Suppose* $(\succsim, \mathcal{C})$ *has a QH representation* $(w, \beta, \hat{\beta}, \delta)$*. Then the individual is naive if and only if* $\hat{\beta} \geq \beta$ *(and is sophisticated if and only if* $\hat{\beta} = \beta$*).*

Corollary 2. *Suppose* $(\succsim_1, \mathcal{C}_1)$ *and* $(\succsim_2, \mathcal{C}_2)$ *are naive and have QH representations* $(w, \beta_1, \hat{\beta}_1, \delta)$ *and* $(w, \beta_2, \hat{\beta}_2, \delta)$*. Then the following are equivalent:*

1. *Individual 1 is more naive than individual 2.*

2. $OV_1(x) \geq OV_2(x)$ *for all menus* x*.*

3. $\hat{\beta}_1 \geq \hat{\beta}_2 \geq \beta_2 \geq \beta_1$*, or* $\hat{\beta}_2 = \beta_2$ *(individual 2 is sophisticated).*

The parametric restriction for comparative naivete in Corollary 2 includes the special cases where individuals share the same ex-post behavior and differ only in their beliefs (i.e., $\beta_1 = \beta_2$ and $\hat{\beta}_1 \geq \hat{\beta}_2$), and where they share the same ex-ante beliefs can differ only in their actual behavior (i.e., $\hat{\beta}_1 = \hat{\beta}_1$ and $\beta_2 \geq \beta_1$). However, these particular cases of comparative naivete are sometimes too restrictive, in that they permit the comparison fewer individuals. Our characterization of comparative naivete permits a more complete ordering of individuals by allowing differences in both ex-ante beliefs and ex-post behavior, yet still rules out the counterintuitive predictions that could arise from other, more permissive parametric comparisons. In particular, Corollary 2 implies that comparing differences or ratios of parameters is not sensible for the quasi-hyperbolic discounting model, as already suggested in Section 2.

4 General Results

In many environments, temptation is more realistically modeled as a random phenomenon. For example, someone might be motivated to work out at the gym on some days but lack enough willpower on other days. Uncertainty about future behavior is arguably even more compelling when considering naivete about temptation: Even if her actual future behavior is deterministic, a naive agent who cannot precisely predict her behavior might more naturally be modeled as having uncertainty about her future temptation, rather than making a resolute but incorrect prediction.

As is standard in ubiquitous applications, random choice data should be interpreted as an idealization of repeated observations of choices from menus. We stress that the case of random choice is a pure generalization of deterministic choice, since deterministic choice is the special case where the distribution of choices is concentrated on a single object. That is, random choice only increases the range of observable environments relative to the deterministic case, and environments where only a single choice is observed still fall under the purview of our model. That all said, we choose to study the general case because the literature suggests compelling reasons to accommodate randomness, and random temptation has been a part of many recent applications of time inconsistency and naivete, ranging from optimal contracting (Eliaz and Spiegler (2006), Spiegler (2011)) to credit markets (Heidhues and Kőszegi (2010)) to the design of commitment devices (Duflo, Kremer, and Robinson (2011)).

In measuring naivete, as is the case for any model of mistaken beliefs, one important and subtle consideration that arises when making repeated observations of choices from menus is the potential for learning; the agent may learn about her tendency to be tempted from her past choices and therefore become more sophisticated over time (e.g., Ali (2011)). For example, if an agent initially exhibits a preference $x \succ \{p\}$ but finds herself repeatedly and frequently choosing an alternative $q \in x$ that is ex-ante dominated by p, she may revise her beliefs and consequently update her ranking of menus to $\{p\} \succ x$ over the course of an experiment. However, while we must observe repeated choices *from* menus to elicit the entire distribution of random ex-post choice, our analysis of ex-ante choice only concerns the initial beliefs held by the agent at the beginning of the experiment. As such, we only need a single set of observations of her choices *between* menus to elicit her initial deterministic menu preference. That is, for our purposes, it suffices to observe just the initial ranking of x and $\{p\}$ to determine the beliefs held prior to any learning. Any effects of learning on menu preference after the initial ranking can be ignored as outside the purview of our model.

4.1 Primitives

We again consider a pair of behavioral primitives. As before, the first primitive is a preference relation $\succsim$ on $\mathcal{K}(\Delta(C))$, which captures the agent's demand for commitment and hence her ex-ante beliefs about future temptations. However, the second primitive is now a random choice rule $\lambda : \mathcal{K}(\Delta(C)) \to \Delta(\Delta(C))$ such that $\lambda^x(x) = 1$, where $\Delta(\Delta(C))$ denotes the space of lotteries over $\Delta(C)$. The behavior encoded in λ is taken while experiencing temptation ex post. For each $x \in \mathcal{K}(\Delta(C))$, λ^x is a probability measure over lotteries, with $\lambda^x(\{p\})$ denoting the probability of choosing the lottery $p \in x$ out of this menu. More generally, $\lambda^x(y)$ denotes the probability of choosing a lottery in the set $y \subset x$ when the choice set is the menu x.

A random choice rule λ is *deterministic* if λ^x is degenerate for all menus x, that is, $\lambda^x = \delta_p$ for some $p \in x$. Identifying the Dirac measure δ_p with p itself, in this case we can express λ as a standard choice function $\mathcal{C} : \mathcal{K}(\Delta(C)) \to \Delta(C)$ by taking $\mathcal{C}(x) = p$ for $\delta_p = \lambda^x$. Thus the behavioral primitives in this section are a strict generalization of the deterministic environment considered in Section 3.

4.2 Absolute Naivete

The conceptual apparatus just introduced for the deterministic case extends to random choice. For any (compound) lottery $\lambda^x \in \Delta(\Delta(C))$, its *average choice* $m(\lambda^x)$ is the expectation of the identity function under λ^x or, formally, $m(\lambda^x) = \int p \, d\lambda^x \in \Delta(C)$. That is, $m(\lambda^x)$ reduces the compound lottery λ^x into a single lottery in $\Delta(C)$. For example,

suppose lottery p gives outcome c with probability 1, and q gives c with probability $2/3$ and c' with probability $1/3$, so $p = \delta_c$ and $q = (2/3)\delta_c + (1/3)\delta_{c'}$. Suppose also that p is chosen from the menu $\{p, q\}$ with probability $\lambda^{\{p,q\}}(\{p\}) = 1/2$. Then the unconditional probability of outcome c given this menu is $5/6$: $m(\lambda^{\{p,q\}}) = (5/6)\delta_c + (1/6)\delta_{c'}$. This reduction from a distribution over multiple lotteries to a single lottery does not assume any attitude towards risk, such as risk neutrality, over deterministic outcomes in C.[25]

Definition 7. *An individual is* sophisticated *if* $x \sim \{m(\lambda^x)\}$ *for all menus* x. *An individual is* naive *if* $x \succsim \{m(\lambda^x)\}$ *for all menus* x. *An individual is* strictly naive *if she is naive and not sophisticated.*

A sophisticate is indifferent between selecting the choice set x for tomorrow and committing to the actual distribution of outcomes $m(\lambda^x)$ that would result from her choices from that menu. A naif incorrectly anticipates making more virtuous choices and hence expects a more attractive distribution of outcomes from x than will occur in actuality. As noted above, deterministic second-stage choice formalized as a choice function $\mathcal{C} : \mathcal{K}(\Delta(C)) \to \Delta(C)$ is a special case of the random choice framework. The corresponding random choice rule λ satisfies $\lambda^x(\{p\}) = 1$ if and only if $\mathcal{C}(x) = p$, and hence $m(\lambda^x) = \mathcal{C}(x)$. In this case, the conditions for sophistication and naivete in Definition 7 reduce to our prior definitions, $x \sim \{\mathcal{C}(x)\}$ and $x \succsim \{C(x)\}$, respectively.

Our definitions lend themselves to simple tests of violations of sophistication and naivete even when choices are random. Consider a binary menu $\{p, q\}$ where $\{p\} \succ \{q\}$, and let $\alpha = \lambda^{\{p,q\}}(\{p\})$. Then, $m(\lambda^{\{p,q\}}) = \alpha p + (1-\alpha)q$ and thus sophistication (naivete) implies

$$\{p, q\} \sim (\succsim) \{\alpha p + (1 - \alpha)q\}.$$

In other words, a sophisticate is indifferent between the option set $\{p, q\}$ and a mixture of these lotteries that matches her ex-post choice frequencies, whereas a naif prefers keeping her options open. One possible experimental design that implements our approach would be to elicit the ranking of $\{p, q\}$ and $\{\hat{\alpha}p + (1 - \hat{\alpha})q\}$ for various values of $\hat{\alpha}$ and then compare these rankings to the actual choice frequencies α of a group of subjects.[26]

We now apply our general definitions to the random Strotz model, which generalizes the classic Strotz model to allow uncertainty about future temptations. Dekel and Lipman (2012) provide a thorough analysis of the random Strotz model. Just as a single temptation is parametrized using a single expected-utility function in the deterministic Strotz model considered in Section 3, a random temptation is analogously parametrized

[25] Our analysis does implicitly assume indifference to compounding. However, indifference to compounding can be relaxed by considering appropriate certainty equivalents for compound lotteries rather than assuming indifference between λ^x and $m(\lambda^x)$.

[26] This experimental design is implemented in Le Yaouanq (2015) to measure individual-level naivete about memory lapses.

using a probability measure over expected-utility functions. When defining the random Strotz representation, it will be mathematically convenient to associate expected-utility functions with their corresponding Bernoulli utility indices. Formally, let $\mathcal{V}$ denote the set of all continuous functions $v : C \to \mathbb{R}$, and endow $\mathcal{V}$ with the supremum norm and corresponding Borel σ-algebra. Thus, $\mathcal{V}$ is the set of all continuous Bernoulli utility indices on the consumption space C, and each element $v \in \mathcal{V}$ can also be identified (with slight abuse of notation) with its corresponding expected-utility function on $\Delta(C)$ by letting $v(p) \equiv \int_C v(c)\, dp$. Note that $\mathcal{V}$ is a vector space.

Definition 8. *A probability measure μ on $\mathcal{V}$ has* finite-dimensional support *if there exists a finite set of expected-utility functions $\{v_1, \ldots, v_n\} \subset \mathcal{V}$ such that* $\text{supp}(\mu) \subset \text{span}(\{v_1, \ldots, v_n\})$.

We will restrict attention to random Strotz representations with finite-dimensional support. This is arguably a mild restriction, as we are unaware of any application of the random Strotz model that does not have finite-dimensional support. For example, any deterministic Strotz representation (see Section 3) or any random quasi-hyperbolic discounting representation (see Section 4.4) has finite-dimensional support. In addition, if the consumption space C is finite, then any probability measure μ on $\mathcal{V}$ trivially has finite-dimensional support.

Without loss of generality, we also restrict attention to probability measures on $\mathcal{V}$ that are *nontrivial*, in the sense of assigning probability zero to constant functions.[27]

Definition 9. *A* random Strotz representation *of $(\succsim, \lambda)$ is a triple $(u, \mu, \hat{\mu})$ of a nontrivial expected-utility function u and nontrivial probability measures μ and $\hat{\mu}$ over $\mathcal{V}$ with finite-dimensional support such that the function $U : \mathcal{K}(\Delta(C)) \to \mathbb{R}$ defined by*

$$U(x) = \int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\hat{\mu}(v)$$

is a utility representation of $\succsim$ and, for all menus x and all measurable $y \subset x$,

$$\lambda^x(y) = \mu(p_x^{-1}(y))$$

for some measurable selection function $p_x : \mathcal{V} \to x$ with $p_x(v) \in B_u(B_v(x))$ for all $v \in \mathcal{V}$.

The interpretation of the representation of the ex-ante preference $\succsim$ is straightforward. To understand the representation of the ex-post random choice rule λ, note that after the realization of a temptation utility $v \in \mathcal{V}$, the individual's choice of lottery is an element

[27] The restriction to nontrivial measures in the definition of the random Strotz representation is without loss of generality since any weight assigned to constant functions can be moved to the commitment utility u without altering the ex-ante preference or ex-post random choice rule.

of the set $B_u(B_v(x))$ of lexicographic maximizers of v then u. There may be multiple elements in this set for a fixed v, and the individual's tie-breaking procedure among these is modeled using a selection function p_x from the correspondence $v \mapsto B_u(B_v(x))$ mapping temptations to possible choices.[28] Given this mapping from temptation utilities to choices, the distribution of temptation utilities then determines the stochastic choice of the individual. The probability of choosing an element of the subset $y \subset x$ is equal to the probability under μ of an ex-post expected-utility function v for which the optimal choice is in y, that is, $\lambda^x(y) = \mu(\{v \in \mathcal{V} : p_x(v) \in y\})$.

The functional characterization of naivete for the random Strotz representation is the stochastic generalization of the definition for deterministic Strotz. In the deterministic case, naivete implies the believed $\hat{v}$ is more u-aligned than v: $\hat{v} \gg_u v$. In the random case, the believed distribution over all possible temptations stochastically dominates the actual distribution of temptations, where stochastic dominance is with respect to the $\gg_u$ order. As is standard, a stochastically dominant measure puts more weight on the upper contour sets of the basic ordering $\gg_u$ over the state space. The following definitions adapt the technology developed by Dekel and Lipman (2012).

Definition 10. *Let u be an expected-utility function. A measurable set $\mathcal{U} \subset \mathcal{V}$ is a* u-upper set *if, for any $v \in \mathcal{U}$ and $v' \in \mathcal{V}$, if $v' \gg_u v$ then $v' \in \mathcal{U}$.*

We let $\gg_u$ denote both the basic ordering over expected-utility functions and the induced stochastic order over measures on expected-utility functions.

Definition 11. *Let u be an expected-utility function, and let $\mu, \hat{\mu}$ be probability measures over $\mathcal{V}$. Then $\hat{\mu}$ is* more u-aligned *than μ, written as $\hat{\mu} \gg_u \mu$, if $\hat{\mu}(\mathcal{U}) \geq \mu(\mathcal{U})$ for all u-upper sets $\mathcal{U}$.*

Note that $\hat{v} \gg_u v$ (in the determinate sense) is equivalent to $\delta_{\hat{v}} \gg_u \delta_v$ (in the stochastic sense). We write $\hat{\mu} \approx \mu$ whenever both $\hat{\mu} \gg_u \mu$ and $\mu \gg_u \hat{\mu}$, that is, when $\hat{\mu}(\mathcal{U}) = \mu(\mathcal{U})$ for all u-upper sets $\mathcal{U}$. In this case, it can be shown that the measures induce identical distributions over ex-post expected-utility *preferences* and can differ only by affine transformations of the utility functions in their supports.[29] They are therefore identical in every respect that is relevant for both ex-ante and ex-post choice.

Generalizing our earlier result, absolute naivete is equivalent to $\hat{\mu}$ dominating μ in the stochastic order generated by $\gg_u$.

[28] Since there may be a multiplicity of selection functions, there may in turn be multiple maximizing choice probabilities over x for a fixed probability measure μ over $\mathcal{V}$. That is, just as there can be a multiple choice *functions* induced by a choice *correspondence*, there can be multiple random choice rules that maximize the same random Strotz representation. However, this multiplicity is not important for our results since observing any maximizing random choice rule provides sufficient information for our comparatives.

[29] The formal statement and proof of this claim can be found in Dekel and Lipman (2012); in particular, see their Theorem 3 and its proof.

Theorem 3. *Suppose* $(\succsim, \lambda)$ *has a random Strotz representation* $(u, \mu, \hat{\mu})$. *Then the individual is naive if and only if* $\hat{\mu} \gg_u \mu$ *(and is sophisticated if and only if* $\hat{\mu} \approx \mu$*).*

The proof of this result makes use of a characterization by Dekel and Lipman (2012) of comparative temptation aversion for ex-ante preferences with random Strotz representations. They say that $\succsim_2$ is *more temptation averse* than $\succsim_1$ if, for all menus x and lotteries p,[30]

$$\{p\} \succ_1 x \implies \{p\} \succ_2 x.$$

Dekel and Lipman (2012) show that if $\succsim_i$ has a random Strotz representation (u, μ_i) for $i = 1, 2$, then $\succsim_2$ is more temptation averse than $\succsim_1$ if and only if $\mu_1 \gg_u \mu_2$. To prove Theorem 3, we apply this comparative to the measures $\hat{\mu}$ and μ in our two-period random Strotz representation for a single individual. In particular, we show that naivete is equivalent to the condition

$$\int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\hat{\mu}(v) = U(x) \geq u(m(\lambda^x)) = \int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\mu(v), \quad \forall x.$$

This condition implies that the hypothetical ex-ante preference $\succsim^*$ generated by the representation with correct beliefs (u, μ) is more temptation averse than the actual ex-ante preference $\succsim$ with representation $(u, \hat{\mu})$, and hence $\hat{\mu} \gg_u \mu$.

4.3 Comparative Naivete

In this section, we adapt the comparatives introduced in the deterministic case to the stochastic setting. Similar to our strategy for extending the absolute definition of naivete, our basic approach to extending deterministic comparisons of naivete to the random case is based on replacing the deterministic choice with the average choice.

4.3.1 Comparing Underdemand for Commitment

Definition 12 generalizes Definition 4 by detecting failures to choose beneficial commitment opportunities.

Definition 12. *Individual 1 is* more naive *than individual 2 if, for all menus* x *and lotteries* p*,*

$$x \succ_2 \{p\} \succ_2 \{m(\lambda_2^x)\} \implies x \succ_1 \{p\} \succ_1 \{m(\lambda_1^x)\}.$$

[30] This formal definition appears with different interpretations in Ahn (2008) and Sarver (2008). It is also similar in spirit to the behavioral comparisons of ambiguity aversion in Epstein (1999) and Ghirardato and Marinacci (2002), who compare arbitrary acts to unambiguous acts in the same manner that we compare arbitrary menus to singleton menus.

The parametric restrictions implied by Definition 12 in the random Strotz model generalize the result obtained in Theorem 2: Unless individual 2 is sophisticated, individual 1 is more naive if she is both more optimistic and less virtuous.

Theorem 4. *Suppose* $(\succsim_1, \lambda_1)$ *and* $(\succsim_2, \lambda_2)$ *are naive and have random Strotz representations* $(u, \mu_1, \hat{\mu}_1)$ *and* $(u, \mu_2, \hat{\mu}_2)$*. Then individual 1 is more naive than individual 2 if and only if*

$$\hat{\mu}_1 \gg_u \hat{\mu}_2 \gg_u \mu_2 \gg_u \mu_1,$$

or $\hat{\mu}_2 \approx \mu_2$ *(individual 2 is sophisticated).*

4.3.2 Comparing Overvaluation

We now turn to generalizing the quantification of naivete with overvaluation introduced in the deterministic case. The following generalization of Definition 5 introduces the coefficient of overvaluation in the random setting.

Definition 13. *Suppose* $(\succsim, \lambda)$ *has a random Strotz representation* $(u, \mu, \hat{\mu})$*. The* coefficient of overvaluation *of a menu* x *is defined by:*

$$OV(x) = \underbrace{\int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\hat{\mu}(v)}_{\text{believed indirect utility}} - \underbrace{\int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\mu(v)}_{\text{actual indirect utility}}.$$

A natural conjecture given Theorem 2 in the deterministic case is that comparing naivete through underdemand for commitment is equivalent to comparing overvaluations. This is generally false outside the deterministic case: Our behavioral comparative is sufficient but not necessary. With random temptation, an individual can have larger overvaluations than another but fail to have greater underdemand for commitment. The most direct way to see this is to observe that if individual 2 is strictly naive and is less naive than individual 1, then for every menu x,[31]

$$\begin{aligned} \underbrace{\int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\hat{\mu}_1(v)}_{\text{1's believed indirect utility}} &\geq \underbrace{\int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\hat{\mu}_2(v)}_{\text{2's believed indirect utility}} \\ &\geq \underbrace{\int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\mu_2(v)}_{\text{2's actual indirect utility}} \geq \underbrace{\int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\mu_1(v)}_{\text{1's actual indirect utility}}. \end{aligned} \tag{12}$$

[31] By Theorem 4, if individual 2 is strictly naive and is less naive than individual 1, then $\hat{\mu}_1 \gg_u \hat{\mu}_2 \gg_u \mu_2 \gg_u \mu_1$. By the Dekel and Lipman (2012) characterization of temptation aversion (see Theorem 7), this condition is equivalent to Equation (12).

The inequalities in Equation (12) exclude some cases where individual 1 is more susceptible to exploitation than individual 2. For example, consider a binary choice set $\{p, q\}$ where p is ex-ante desirable, but q is tempting. Suppose that individual 1 anticipates that she chooses p from $\{p, q\}$ with probability 90%, while she actually always selects q. Suppose that individual 2 anticipates that she chooses p with probability 91%, but her actual probability equals 89%. Individual 1 is not more naive than individual 2 according to Definition 12 because she is not more optimistic ex ante ($\hat{\mu}_1 \not\gg_u \hat{\mu}_2$), but the overvaluation of the menu $\{p, q\}$ is nonetheless higher for 1 than for 2: $OV_1(\{p, q\}) = 0.9(u(p) - u(q)) > OV_2(\{p, q\}) = 0.02(u(p) - u(q))$.

This example shows that the behavioral comparison of naivete based on greater underdemand for commitment is no longer equivalent to the functional comparison of higher overvaluations once we move beyond the deterministic setting of Section 3. Thus, we will instead develop an alternative behavioral foundation for comparing overvaluations that is valid in both the deterministic and the random choice settings.

Our general model incorporates all relevant dimensions of consumption into the space C. But for the sake of developing intuition for how to calibrate overvaluation from choice data, consider a special quasilinear environment where ex-ante choices are over pairs of a menu $x \in \mathcal{K}(\Delta(C))$ and a money transfer $t \in \mathbb{R}$, and ex-ante utility takes the form $V(x, t) = U(x) + t$. By definition, the overvaluation of the menu x must satisfy

$$(x, 0) \sim (m(\lambda^x), OV(x)).$$

The required monetary premium for x relative to $m(\lambda^x)$ immediately quantifies overvaluation for quasilinear preferences. Moreover, $OV_1(x) \geq OV_2(x)$ is equivalent to the behavioral comparative that individual 1 is willing to overpay more for any menu x than individual 2:

$$(x, 0) \succsim_2 (m(\lambda_2^x), t) \implies (x, 0) \succsim_1 (m(\lambda_1^x), t).$$

Since our general model does not assume quasilinearity, we must take a different approach to calibrating overvaluation. As a side benefit of assuming expected utility, we can use linearity in probabilities in a similar manner to the role of the money numeraire in the previous discussion. That is, we can use additional odds of winning a good prize to quantify the value of x relative to $m(\lambda^x)$. The following definition takes this approach to converting overvaluation into a behavioral measure.

Definition 14. *Fix any lotteries p, q such that $\{q\} \succ \{p\}$. The* probability premium *of a menu x is defined by:*

$$P(x; p, q) = \sup \big\{\alpha \in [0, 1] : (1 - \alpha)x + \alpha\{p\} \succsim (1 - \alpha)\{m(\lambda^x)\} + \alpha\{q\}\big\}.$$

The probability premium indicates how much a menu x can be mixed with an inferior alternative with the individual still preferring it to $m(\lambda^x)$ mixed with a superior alternative. To see its implications, suppose that $\succsim$ admits an affine utility representation. Then, note that $P(x;p,q)=0$ if and only if $x \sim \{m(\lambda^x)\}$. In particular, the individual is sophisticated if and only if $P(x;p,q)=0$ for all x. If instead $x \succ \{m(\lambda^x)\}$, then mixing these menus with the lotteries p and q, respectively, where $\{q\} \succ \{p\}$, could reverse this preference. The weighting α needed for an individual to choose the commitment lottery $(1-\alpha)\{m(\lambda^x)\}+\alpha\{q\}$ over $(1-\alpha)x+\alpha\{p\}$ thus provides a quantitative measure of the difference in the values assigned by the individual to x and $m(\lambda^x)$.

While more permissive than comparing underdemand for commitment, comparing overvaluations still yields several useful equivalent characterizations for random Strotz preferences. First, having greater overvaluations is equivalent to having greater probability premiums. Second, there is an additional interesting parametric restriction that exposes itself in the random case. In particular, if an individual has larger overvaluations than another, then the difference between her anticipated and actual distributions over the realization of temptation will first-order stochastically dominate that difference for the other individual.

Theorem 5. *Suppose $(\succsim_1, \lambda_1)$ and $(\succsim_2, \lambda_2)$ are naive and have random Strotz representations $(u, \mu_1, \hat{\mu}_1)$ and $(u, \mu_2, \hat{\mu}_2)$. Fixing any lotteries p, q with $\{q\} \succ_i \{p\}$, the following are equivalent:*

1. *$P_1(x;p,q) \geq P_2(x;p,q)$ for all menus x.*
2. *$OV_1(x) \geq OV_2(x)$ for all menus x.*
3. *$\hat{\mu}_1(\mathcal{U}) - \mu_1(\mathcal{U}) \geq \hat{\mu}_2(\mathcal{U}) - \mu_2(\mathcal{U})$ for all u-upper sets $\mathcal{U}$; equivalently, $\hat{\mu}_1 - \mu_1 \gg_u \hat{\mu}_2 - \mu_2$.*

The parametric restriction on distributions in the last condition of Theorem 5 is strictly weaker than the ordering of distributions that characterized underdemand for commitment in Theorem 4. We therefore have as a corollary that comparing naivete through underdemand for commitment is always a more selective criterion than comparing overvaluations. Recall, however, that in the deterministic case, the two restrictions become equivalent (see Theorem 2) because then all the distributions are degenerate and their differences can only take the values 0 or 1.

Corollary 3. *Suppose $(\succsim_1, \lambda_1)$ and $(\succsim_2, \lambda_2)$ are naive and have random Strotz representations $(u, \mu_1, \hat{\mu}_1)$ and $(u, \mu_2, \hat{\mu}_2)$. If individual 1 is more naive than individual 2, then $OV_1(x) \geq OV_2(x)$ for all menus x.*

4.4 Application to Random Quasi-Hyperbolic Discounting

The standard quasi-hyperbolic discounting model assumes completely confident beliefs about future behavior, an assumption that seems less palatable under naivete when these beliefs are incorrect. We explore a generalization of the QH representation that allows for naive and uncertain beliefs about β. Several applications in different areas employ naive uncertainty about future present bias. Heidhues and Kőszegi (2010, Section 4) employ random quasi-hyperbolic discounting to explain the structure of credit markets and its consequent welfare implications. In their study of fertilizer adoption decisions by Kenyan farmers, Duflo, Kremer, and Robinson (2011) estimate a specification of random quasi-hyperbolic discounting where naivete is parameterized by a mistakenly believed positive chance of virtuous exponential discounting. Admitting uncertainty about intertemporal substitution also often usefully serves as a reduced-form proxy for a shock in the economy, like wage uncertainty, or for heterogeneity across agents in an aggregate economy, like the distribution of wealth. Similarly, random present-bias can provide a parsimonious channel for capturing uncertainty about external factors that affect present-bias.

In this subsection we establish the implications of our general results for the special case of the random quasi-hyperbolic discounting representation. Definition 15 generalizes Definition 6 by allowing the agent to be uncertain about the future value of her discount factor β.

Definition 15. *A* random quasi-hyperbolic (RQH) representation *of* $(\succsim, \lambda)$ *is a quadruple* $(w, F, \hat{F}, \delta)$ *of a continuous and nontrivial function* $w : [a, b] \to \mathbb{R}$*, a scalar* $\delta \in (0, 1)$*, and cumulative distribution functions* F *and* $\hat{F}$ *on* $[0, 1]$ *such that when* u *and* v_β *are defined as in Equations* (10) *and* (11)*, the function* $U : \mathcal{K}(\Delta(C)) \to \mathbb{R}$ *defined by*

$$U(x) = \int_0^1 \max_{p \in B_{v_\beta}(x)} u(p)\, d\hat{F}(\beta)$$

is a utility representation of $\succsim$ *and, for all menus* x *and all measurable* $y \subset x$*,*

$$\lambda^x(y) = F(p_x^{-1}(y))$$

for some measurable selection function $p_x : [0, 1] \to x$ *with* $p_x(\beta) \in B_u(B_{v_\beta}(x))$ *for all* $\beta \in [0, 1]$.[32]

The following corollaries show how our definitions of absolute and comparative naivete translate into stochastic generalizations of Corollaries 1 and 2 for the RQH representation.

[32] We are abusing notation slightly and using F to also denote the probability measure on $[0, 1]$ that has F as its distribution function. That is, for any measurable set $A \subset [0, 1]$, we write $F(A)$ to denote $\int_A dF(\beta)$. Hence $\lambda^x(y) = \int_0^1 \mathbf{1}_{[p_x(\beta) \in y]}\, dF(\beta)$.

A naive individual underestimates the degree of her present bias, which is reflected in her belief $\hat{F}$ putting more likelihood on larger values of β than the actual distribution F that governs her ex-post choices. Let $\geq_{FOSD}$ denote the usual first-order stochastic dominance order, with $\hat{F} \geq_{FOSD} F$ if $\hat{F}(\beta) \leq F(\beta)$ for all $\beta \in [0,1]$.

Corollary 4. *Suppose $(\succsim, \lambda)$ has a RQH representation $(w, F, \hat{F}, \delta)$. Then the individual is naive if and only if $\hat{F} \geq_{FOSD} F$ (and is sophisticated if and only if $\hat{F} = F$).*

As in the case of the general random Strotz representation, comparing naivete via overvaluations is weaker than comparing underdemand for commitment. The following corollary characterizes the implications of each.

Corollary 5. *Suppose $(\succsim_1, \lambda_1)$ and $(\succsim_2, \lambda_2)$ are naive and have RQH representations $(w, F_1, \hat{F}_1, \delta)$ and $(w, F_2, \hat{F}_2, \delta)$.*

1. *Individual 1 is more naive than individual 2 if and only if*

$$\hat{F}_1 \geq_{FOSD} \hat{F}_2 \geq_{FOSD} F_2 \geq_{FOSD} F_1,$$

   or $\hat{F}_2 = F_2$ (individual 2 is sophisticated).

2. *Individual 1 has greater probability premiums than individual 2 if and only if individual 1 has greater overvaluations than individual 2 if and only if*

$$F_1(\beta) - \hat{F}_1(\beta) \geq F_2(\beta) - \hat{F}_2(\beta), \quad \forall \beta \in [0,1].$$

The RQH representation is a member of a more general subclass of the random Strotz representation where the possible temptations are ordered by a one-dimensional parameter. We analyze this subclass, called the *uncertain intensity Strotz representation*, in Appendix B. Corollaries 4 and 5 follow directly from the results in that section.

5 Extensions and Robustness

5.1 Costly Self-Control

So far we have not considered the possibility of an agent's costly effort to resist temptations. We now turn to analyzing the robustness of our results in the presence of such costly self-control. In particular, following Gul and Pesendorfer (2001), the individual's self-control cost of choosing alternative p from the menu x is $\max_{q \in x} v(q) - v(p)$, the difference between the temptation utility of the most tempting option and that of p. The

individual maximizes her commitment utility u subject to these self-control costs, and therefore chooses the option that maximizes the compromise $u(p) + v(p)$ of commitment utility and temptation utility. The following definition permits uncertainty about the temptation utility, as in Stovall (2010), as well as possibility of incorrect beliefs about the distribution of temptation utilities.

Definition 16. *A* random Gul-Pesendorfer representation *of* $(\succsim, \lambda)$ *is a triple* $(u, \mu, \hat{\mu})$ *of a nontrivial expected-utility function* u *and nontrivial probability measures* μ *and* $\hat{\mu}$ *over* $\mathcal{V}$ *with finite-dimensional support such that the function* $U : \mathcal{K}(\Delta(C)) \to \mathbb{R}$ *defined by*

$$U(x) = \int_{\mathcal{V}} \Big[\max_{p \in x} (u(p) + v(p)) - \max_{q \in x} v(q) \Big] d\hat{\mu}(v)$$

is a utility representation of $\succsim$ *and, for all menus* x *and all measurable* $y \subset x$,

$$\lambda^x(y) = \mu(p_x^{-1}(y))$$

for some measurable selection function $p_x : \mathcal{V} \to x$ *with* $p_x(v) \in B_{u+v}(x)$ *for all* $v \in V$.

Dekel and Lipman (2012) show that under a mild continuity assumption, the menu preference alone cannot distinguish the random Gul-Pesendorfer model and the random Strotz model. However, they also find that a random Gul-Pesendorfer representation of $\succsim$ implies different ex-post choice probabilities than those implied by a random Strotz representation of $\succsim$. Analyzing whether our results can be extended to deal with self-control preferences is therefore important since the identification of naivete proposed in Sections 3 and 4 relies on a particular model of ex-ante behavior.[33]

The following theorem states that if the individual is naive and admits a random Gul-Pesendorfer representation, then the ex-ante beliefs derived from the representation are optimistic. More precisely, naivete implies that any random Gul-Pesendorfer representation predicts ex-post choices that are more virtuous than the actual ones. The intuition is the following: Self-control costs increase the attractiveness of commitment since tempting options can be undesirable ex ante even if they are not chosen ex post. Thus the definition of absolute naivete proposed in Section 4 serves as a conservative and robust test to reveal an individual's optimism even in the presence of costly self-control.

Theorem 6. *Suppose that* $(\succsim, \lambda)$ *has a random Gul-Pesendorfer representation* $(u, \mu, \hat{\mu})$, *and that the individual is naive. Then, for any* u*-upper set* $\mathcal{U}$,

$$\hat{\mu}(\{v \in \mathcal{V} : u + v \in \mathcal{U}\}) \geq \mu(\{v \in \mathcal{V} : u + v \in \mathcal{U}\}). \tag{13}$$

[33] Dekel and Lipman (2012) show that both representations can be distinguished with the additional observation of ex-post choices if we require the agent's beliefs to be correct under these representations. Otherwise, the models are indistinguishable because of possible disagreement between actual and perceived distributions.

In addition, if the individual is strictly naive, Equation (13) *is satisfied with strict inequality for some* $\mathcal{U}$.

It is important to note that the converse of Theorem 6 fails. In particular, even if $\hat{\mu} = \mu$ and the individual has correct beliefs about her future behavior, the desire to avoid self-control costs may result in $\{m(\lambda^x)\} \succ x$ for some menus, in violation of both our behavioral definitions of naivete and sophistication. Thus our behavioral definition of naivete is sufficient but *not necessary* for overoptimistic beliefs when individuals have self-control preferences.

In a companion paper Ahn, Iijima, and Sarver (2016), we explore alternative behavioral conditions that tightly characterize naivete for deterministic self-control preferences. We also show that there is an impossibility result when randomness is permitted: It is impossible to construct a behavioral definition that tightly characterizes naivete for the random Gul-Pesendorfer representation. This impossibility is closely related to the lack of tight identification in this model—the random Gul-Pesendorfer representation of ex-ante choice is not unique. That is, there are multiple different measures $\hat{\mu}$ that can be used to represent the same ex-ante menu preference, but these different measures do generate different ex-post random choices. Therefore, depending on which measure is used to represent ex-ante beliefs, the same combination of ex-ante and ex-post behavior could potentially be classified as naive (i.e., overly optimistic), sophisticated, or overly pessimistic. In light of this impossibility of a tight characterization, Theorem 6 is perhaps the best result that one can hope to obtain in the presence of random self-control costs.

5.2 Uncertainty in Normative Preferences

The random Strotz interpretation of commitment preferences relies on the assumption that normative preferences are certain ex ante. The elicitation of naivete provided in Section 4 is therefore suited to situations where long-term preferences are known and where deviations are always undesirable (e.g., temptations, addictions, memory lapses). In some situations, however, the individual might expect future shocks to her normative preferences. In that case, her menu choices trade off commitment versus flexibility and the condition $x \succ \{m(\lambda^x)\}$ does not necessarily indicate unrealistic expectations: An individual who anticipates receiving some information about her normative ranking prior to selecting an option might rationally refuse to commit to her average choice.

Identifying the flexibility-loving part from the commitment-loving component of preferences in order to detect naive anticipations requires additional assumptions. For example, in some contexts, the normative states are tied to objective contingencies that can be directly observed by the analyst (e.g., financial events, weather, health status). In this case, the identification of naivete can be performed essentially as described in

Section 4 conditional on each normative state, assuming the choice between menus can be conditioned on the realization of the state. Extensions of our analysis might also be possible without objective states. For instance, Stovall (2018) considers a model where both the normative uncertainty (over u) and the temptation uncertainty (over v) are resolved in an interim period prior to the direct experience of temptation. We conjecture that our approach could be adapted to any model such as this where uncertainty about the normative preference is resolved before temptation occurs.

Even in cases where these workarounds are not possible and our techniques are not immediately applicable, the sophistication hypothesis nonetheless imposes some necessary properties on choice data. In particular, options can be relevant ex ante only if they are chosen with some probability ex post, an axiom that Ahn and Sarver (2013) call *consequentialism*: $x \sim \text{supp}(\lambda^x)$ for all menus x is a necessary condition for the existence of a sophisticated representation. In contrast, the condition $x \succ \text{supp}(\lambda^x)$ indicates that the individual incorrectly estimates the virtue of her choices inside x.

As an illustration, suppose that an individual is considering buying a membership that gives her free access to the gym. Let x denote the option set that includes any number of gym visits, and let $p \in x$ denote zero visits. Observing that she values the membership ex ante ($x \succ \{p\}$) but that she attends the gym with probability zero ex post ($\lambda^x(\{p\}) = 1$) is sufficient to conclude that she had unrealistic expectations regarding her gym attendance.[34] Relatedly, suppose that the individual can self-impose a penalty for smoking. Her initial choice set is $\{p^1, p^2\}$ (smoking or not) but she can replace p^1 by a contract p^3 according to which smoking results in the payment of a penalty. Observing that she selects the contract ($\{p^3, p^2\} \succ \{p^1, p^2\}$) but continues smoking with probability one despite the penalty ($\lambda^{\{p^3,p^2\}}(\{p^3\}) = 1$) is sufficient to conclude that her menu choice was led by naive anticipations.[35]

[34] Note that all of the options in x are in fact pairs consisting of a number of visits together with the expense of the gym membership. Letting q denote zero visits without the paying for the membership, the choice to join the gym corresponds to the preference $x \succ \{q\}$. Since $\{q\} \succ \{p\}$ by dominance (the individual prefers not to pay the cost of the membership without going), we have $x \succ \{p\} = \text{supp}(\lambda^x)$.

[35] This argument implicitly assumes that the individual weakly prefers having the option to quit to being forced to smoke ($\{p^1, p^2\} \succsim \{p^1\}$) and that the individual prefers not to pay the penalty all else equal ($\{p^1\} \succ \{p^3\}$). Thus $\{p^3, p^2\} \succ \{p^1, p^2\} \succsim \{p^1\} \succ \{p^3\} = \text{supp}(\lambda^{\{p^3,p^2\}})$.

A A Comparative from Dekel and Lipman (2012)

In this section, we summarize a relevant result from Dekel and Lipman (2012) that will play a central role in our proofs of Theorems 3, 4, and 5. Recall the definition of comparative temptation aversion: Individual 2 is *more temptation averse* than individual 1 if, for all menus x and lotteries p,

$$\{p\} \succ_1 x \implies \{p\} \succ_2 x.$$

Theorem 7 (Dekel and Lipman (2012))**.** *Suppose $\succsim_1$ and $\succsim_2$ have random Strotz representations (u, μ_1) and (u, μ_2). Then $\succsim_2$ is more temptation averse than $\succsim_1$ if and only if $\mu_1 \gg_u \mu_2$.*

Dekel and Lipman (2012) consider only a finite prize space C in their paper. In the Supplemental Appendix, we prove that their result can be extended to any compact metric space C and any random Strotz representation (with finite-dimensional support) defined on that space.[36] This extension to compact spaces is not merely a technical exercise, as it is critical for many of the applications of our results, such as dynamic consumption problems where $C = [a, b]^{\mathbb{N}}$.

B Uncertain Intensity Random Strotz

In this section, we highlight a useful special case of the random Strotz representation where the uncertainty over future behavior is only over the magnitude of the future temptation, and not in its basic direction. For example, the individual may know that she will crave sweet snacks (but not salty snacks) ex post, but is uncertain of how strong her craving for sweets will be. This uncertain intensity Strotz representation encompasses the random quasi-hyperbolic discounting model studied in Section 4.4 where the individual is uncertain of the intensity of her present bias, and the corollaries presented below provide a bridge between our main theorems and the results in that section.

Two expected-utility functions u and v are *independent* if they are nontrivial and it is not the case that $v \approx u$ or $v \approx -u$.

Definition 17. *An* uncertain intensity Strotz representation *of $(\succsim, \lambda)$ is a quadruple $(u, v, F, \hat{F})$ of two independent expected-utility functions u, v and two cumulative distribution functions $F, \hat{F}$ on $[0, 1]$ such that the function $U : \mathcal{K}(\Delta(C)) \to \mathbb{R}$ defined by*

$$U(x) = \int_0^1 \max_{p \in B_{\alpha u + (1-\alpha)v}(x)} u(p)\, d\hat{F}(\alpha)$$

[36] Although Definition 9 and Theorem 7 impose the restriction that the measure μ in the random Strotz representation must have finite-dimensional support, our proof in the Supplemental Appendix shows that the "if" direction in Theorem 7 is true even without the finite-dimensional support assumption. It remains an open question whether the "only if" direction can be extended to probably measures with arbitrary support. However, we view the exploration of additional generalizations of this results as a purely technical question. As we discussed in Section 4, we are not aware of any application of the random Strotz model that does not have finite-dimensional support.

is a utility representation of $\succsim$ *and, for all menus* x *and all measurable* $y \subset x$,

$$\lambda^x(y) = F(p_x^{-1}(y))$$

for some measurable selection function $p_x : [0,1] \to x$ *with* $p_x(\alpha) \in B_u(B_{\alpha u+(1-\alpha)v}(x))$ *for all* $\alpha \in [0,1]$.

For the case of an uncertain intensity Strotz representation, the direction of the temptation is known to be v, but the magnitude of that temptation relative to the virtuous utility u is uncertain. A naive individual underestimates the influence of v, and this bias is reflected in her belief $\hat{F}$ over the intensities in $[0,1]$ putting more likelihood on larger weighting of u (hence lower weighting of v) than F.

Corollary 6. *Suppose* $(\succsim, \lambda)$ *has a uncertain intensity Strotz representation* $(u, v, F, \hat{F})$. *Then the individual is naive if and only if* $\hat{F} \geq_{FOSD} F$ *(and is sophisticated if and only if* $\hat{F} = F$*).*

Corollary 7. *Suppose* $(\succsim_1, \lambda_1)$ *and* $(\succsim_2, \lambda_2)$ *are naive and have uncertain intensity Strotz representations* $(u, v, F_1, \hat{F}_1)$ *and* $(u, v, F_2, \hat{F}_2)$.

1. *Individual 1 is more naive than individual 2 if and only if*

$$\hat{F}_1 \geq_{FOSD} \hat{F}_2 \geq_{FOSD} F_2 \geq_{FOSD} F_1,$$

   or $\hat{F}_2 = F_2$ *(individual 2 is sophisticated).*

2. *Individual 1 has greater probability premiums than individual 2 if and only if individual 1 has greater overvaluations than individual 2 if and only if*

$$F_1(\alpha) - \hat{F}_1(\alpha) \geq F_2(\alpha) - \hat{F}_2(\alpha), \quad \forall \alpha \in [0,1].$$

C Proofs

C.1 Proof of Theorem 3

Suppose the random choice rule λ has a random Strotz representation (u, μ). Consider the hypothetical sophisticated ex-ante preference $\succsim^*$ that is also be represented by (u, μ). The following lemma shows how this hypothetical preference can be determined from λ and u.

Lemma 1. *Suppose* λ *has a random Strotz representation* (u, μ). *Then for any menu* x,

$$u(m(\lambda^x)) = \int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\mu(v).$$

In particular, if we define a binary relation $\succsim^*$ *on* $\mathcal{K}(\Delta(C))$ *by*

$$x \succsim^* y \iff u(m(\lambda^x)) \geq u(m(\lambda^y)),$$

then (u, μ) *is a random Strotz representation for* $\succsim^*$.

Proof. If (u, μ) represents λ then by definition there exists, for all menus x, a measurable selection function $p_x : \mathcal{V} \to x$ with $p_x(v) \in B_u(B_v(x))$ such that

$$\lambda^x(y) = \mu(p_x^{-1}(y))$$

for all measurable $y \subset x$. Thus λ^x is the distribution on x induced by the random variable p_x defined on the measure space $(\mathcal{V}, \mu)$. Therefore, the standard change of variables formula together with the linearity and continuity of u imply

$$\begin{aligned}\int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\mu(v) &= \int_{\mathcal{V}} u(p_x(v))\, d\mu(v) \\ &= \int_x u(p)\, d\lambda^x(p) = u\bigg(\int_x p\, d\lambda^x(p)\bigg) = u(m(\lambda^x)),\end{aligned}$$

as desired. ■

Turning now to the proof of Theorem 3, fix a random Strotz representation $(u, \mu, \hat{\mu})$ for $(\succsim, \lambda)$, and define $\succsim^*$ as in Lemma 1. To establish sufficiency, suppose the individual is naive. Then for all menus x and lotteries p,

$$\begin{aligned}\{p\} \succ x &\implies \{p\} \succ \{m(\lambda^x)\} \qquad \text{(naivete)} \\ &\implies u\big(m\big(\lambda^{\{p\}}\big)\big) = u(p) > u(m(\lambda^x)) \\ &\implies \{p\} \succ^* x.\end{aligned}$$

Thus $\succsim^*$ is more temptation averse than $\succsim$. Since (u, μ) represents $\succsim^*$ by Lemma 1, Theorem 7 implies that $\hat{\mu} \gg_u \mu$. If the individual is sophisticated, then a similar argument shows that the converse also holds: $\succsim$ is also more temptation averse than $\succsim^*$ (in particular, $\succsim = \succsim^*$) and hence $\mu \gg_u \hat{\mu}$ also holds, i.e., $\hat{\mu} \approx \mu$.

To establish necessity, suppose $\hat{\mu} \gg_u \mu$. By Theorem 7, $\succsim^*$ is more temptation averse than $\succsim$. By contrapositive, this is equivalent to the condition

$$x \succsim^* \{p\} \implies x \succsim \{p\}.$$

Note that for any menu x, if we take $p = m(\lambda^x)$ then

$$u(m(\lambda^x)) = u(p) = u\big(m\big(\lambda^{\{p\}}\big)\big)$$

and hence $x \sim^* \{p\} = \{m(\lambda^x)\}$. Since $\succsim^*$ is more temptation averse than $\succsim$, this implies $x \succsim \{m(\lambda^x)\}$. Thus the individual is naive. If we also have $\mu \gg_u \hat{\mu}$ then another application of Theorem 7 implies the condition above can be strengthened to $x \succsim^* \{p\} \iff x \succsim \{p\}$. In this case, $x \sim^* \{m(\lambda^x)\}$ implies $x \sim \{m(\lambda^x)\}$ and hence the individual is sophisticated.

C.2 Proof of Theorem 4

Our proof of this theorem is based on two lemmas. This first lemma shows that once we exclude the case where individual 2 is sophisticated, our condition for more naive in Definition 12 is equivalent to the analogous condition but with weak preferences rather than strict.

Lemma 2. *Suppose* $(\succsim_1, \lambda_1)$ *and* $(\succsim_2, \lambda_2)$ *are naive and have random Strotz representations* $(u, \mu_1, \hat{\mu}_1)$ *and* $(u, \mu_2, \hat{\mu}_2)$*. Then individual 1 is more naive than individual 2 if and only if either*

$$x \succsim_2 \{p\} \succsim_2 \{m(\lambda_2^x)\} \implies x \succsim_1 \{p\} \succsim_1 \{m(\lambda_1^x)\}, \tag{14}$$

or individual 2 is sophisticated.

Proof. The "if" direction is straightforward and is therefore omitted. To prove the "only if" direction, suppose individual 1 is more naive than individual 2. We will show that if individual 2 is strictly naive (i.e., naive but not sophisticated), then Equation (14) must hold. First, note that since $(\succsim_i, \lambda_i)$ has a random Strotz representation $(u, \mu_i, \hat{\mu}_i)$, we have that for any menus x and y and any $\alpha \in [0, 1]$,

$$u(m(\lambda_i^{\alpha x + (1-\alpha)y})) = \alpha u(m(\lambda_i^x)) + (1-\alpha) u(m(\lambda_i^y)). \tag{15}$$

This assertion is easy to verify by appealing to Lemma 1 together with the linearity of the random Strotz value function for menus. Second, under the assumption that individual 2 is strictly naive, there exists some menu y such that $y \succ_2 \{m(\lambda_2^y)\}$. Fix any lottery q such that $y \succ_2 \{q\} \succ_2 \{m(\lambda_2^y)\}$. Then, for any menu x and lottery p,

$$\begin{aligned}
&x \succsim_2 \{p\} \succsim_2 \{m(\lambda_2^x)\} \\
&\qquad \implies \alpha x + (1-\alpha) y \succ_2 \{\alpha p + (1-\alpha) q\} \succ_2 \{m(\lambda_2^{\alpha x + (1-\alpha)y})\}, \quad \forall \alpha \in (0,1) \\
&\qquad \implies \alpha x + (1-\alpha) y \succ_1 \{\alpha p + (1-\alpha) q\} \succ_1 \{m(\lambda_1^{\alpha x + (1-\alpha)y})\}, \quad \forall \alpha \in (0,1) \\
&\qquad \implies x \succsim_1 \{p\} \succsim_1 \{m(\lambda_1^x)\}.
\end{aligned}$$

The first implication follows from the linearity of the random Strotz value function for menus together with Equation (15). The second implication follows because individual 1 is more naive than individual 2. The final implication follows by taking the limit as $\alpha \to 1$, given that the random Strotz representation is continuous in the mixture operation. Thus we have shown that Equation (14) holds. ■

The following lemma decomposes the condition in Lemma 2 into two more basic conditions. The comparative of being more temptation averse is defined in the main text. The comparative of being more virtuous is defined for the first time in this lemma. Intuitively, individual 2 is more virtuous than individual 1 if she makes "better" choices (as measured by her commitment preference) from every menu than individual 1.

Lemma 3. *Suppose* $(\succsim_1, \lambda_1)$ *and* $(\succsim_2, \lambda_2)$ *are naive, and suppose* $\succsim_1$ *and* $\succsim_2$ *share the same commitment preference, i.e.,* $\{p\} \succsim_1 \{q\} \iff \{p\} \succsim_2 \{q\}$ *for all lotteries* $p, q \in \Delta(C)$*. Then Equation* (14) *is satisfied if and only if both of the following hold:*

1. *Individual 2 is more temptation averse than individual 1:* $\{p\} \succ_1 x \implies \{p\} \succ_2 x$.
2. *Individual 2 is more virtuous than individual 1:* $\{p\} \succ_2 \{m(\lambda_2^x)\} \implies \{p\} \succ_1 \{m(\lambda_1^x)\}$.

Proof. Equation (14) *implies condition (1):* Fix any menu x and lottery p such that $\{p\} \succ_1 x$. By Equation (14), we cannot have $x \succsim_2 \{p\} \succsim_2 \{m(\lambda_2^x)\}$. Thus either $\{p\} \succ_2 x$ or $\{m(\lambda_2^x)\} \succ_2 \{p\}$. To rule out the second possibility, note that since individual 2 is naive, we must have $x \succsim_2 \{m(\lambda_2^x)\}$. By Equation (14), this implies $x \succsim_1 \{m(\lambda_2^x)\} \succsim_1 \{m(\lambda_1^x)\}$. Therefore, $\{p\} \succ_1 \{m(\lambda_2^x)\}$, and hence $\{p\} \succ_2 \{m(\lambda_2^x)\}$ since individuals 1 and 2 have the same commitment preference. Thus the only possibility is $\{p\} \succ_2 x$, as desired.

Equation (14) *implies condition (2):* Fix any menu x and lottery p such that $\{p\} \succ_2 \{m(\lambda_2^x)\}$. Since individual 2 in naive, $x \succsim_2 \{m(\lambda_2^x)\}$. By Equation (14), this implies $x \succsim_1 \{m(\lambda_2^x)\} \succsim_1 \{m(\lambda_1^x)\}$. Individuals 1 and 2 share the same commitment preference, and therefore $\{p\} \succ_1 \{m(\lambda_2^x)\} \succsim_1 \{m(\lambda_1^x)\}$, as desired.

Conditions (1) and (2) together imply Equation (14)*:* If individual 2 is more virtuous than individual 1, then we must have $\{m(\lambda_2^x)\} \succsim_2 \{m(\lambda_1^x)\}$. Otherwise, taking $p = m(\lambda_1^x)$ in condition (2) would lead to a contradiction. Therefore, since the individuals share the same commitment preference, $\{p\} \succsim_2 \{m(\lambda_2^x)\} \implies \{p\} \succsim_1 \{m(\lambda_1^x)\}$ for any lottery p. Combining this with the contrapositive of condition (1), it follows directly that Equation (14) holds. ■

We are now ready to prove Theorem 4. If individual 2 is sophisticated, then the asserted equivalence holds trivially. Therefore, suppose that individual 2 is strictly naive. By Theorem 3, $\hat{\mu}_2 \gg_u \mu_2$. Also, since individual 2 is strictly naive, Lemmas 2 and 3 imply that individual 1 is more naive than individual 2 if and only if 2 is both more temptation averse and more virtuous than 1. By Theorem 7, individual 2 is more temptation averse than individual 1 if and only if $\hat{\mu}_1 \gg_u \hat{\mu}_2$. The proof is therefore completed if we can show that individual 2 is more virtuous than individual 1 if and only if $\mu_2 \gg_u \mu_1$. To see that this is true, define $\succsim_1^*$ and $\succsim_2^*$ as in Lemma 1 for λ_1 and λ_2, respectively. Then (u, μ_1) and (u, μ_2) represent $\succsim_1^*$ and $\succsim_2^*$. Note that for all menus x and lotteries p,

$$\{p\} \succ_i \{m(\lambda_i^x)\} \iff u(p) > u(m(\lambda_i^x)) \iff \{p\} \succ_i^* x, \quad i = 1, 2.$$

Therefore, individual 2 is more virtuous than individual 1 if and only if $\succsim_1^*$ is more temptation averse than $\succsim_2^*$. By Theorem 7, this is true if and only if $\mu_2 \gg_u \mu_1$.

C.3 Proof of Theorem 5

(1) ⇔ (2): Let U_i denote the value function from the representation $(u, \hat{\mu}_i)$ for the ex-ante preference $\succsim_i$ for $i = 1, 2$. Also, recall from Lemma 1 that if λ_i has random Strotz representation

(u, μ_i), then

$$u(m(\lambda_i^x)) = \int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\mu_i(v).$$

Thus $OV_i(x) = U_i(x) - u(m(\lambda_i^x))$. Now, fix any lotteries p, q with $\{q\} \succ_i \{p\}$ for $i = 1, 2$. For each menu x, define

$$\begin{aligned} A_i^x &\equiv \big\{\alpha \in [0,1] : (1-\alpha)x + \alpha\{p\} \succsim_i (1-\alpha)\{m(\lambda_i^x)\} + \alpha\{q\}\big\} \\ &= \big\{\alpha \in [0,1] : (1-\alpha)U_i(x) + \alpha u(p) \geq (1-\alpha)u(m(\lambda_i^x)) + \alpha u(q)\big\} \\ &= \big\{\alpha \in [0,1] : (1-\alpha)OV_i(x) \geq \alpha(u(q) - u(p))\big\}. \end{aligned}$$

By definition, $P_i(x; p, q) = \sup A_i^x$. Note that A_i^x is a closed interval. Moreover, since both individuals are naive, we have $x \succsim_i m(\lambda_i^x)$ and therefore $0 \in A_i^x$. Also, $1 \notin A_i^x$ since $\{q\} \succ_i \{p\}$. This implies

$$\alpha = P_i(x; p, q) \iff (1-\alpha)OV_i(x) = \alpha(u(q) - u(p)).$$

Therefore, $OV_1(x) \geq OV_2(x)$ if and only if $P_1(x; p, q) \geq P_2(x; p, q)$.

(2) $\Leftrightarrow$ (3): For any menu x,

$$\begin{aligned} &OV_1(x) \geq OV_2(x) \\ &\iff \int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\hat{\mu}_1(v) - \int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\mu_1(v) \\ &\qquad\qquad \geq \int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\hat{\mu}_2(v) - \int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\mu_2(v) \\ &\iff \int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\Big(\frac{1}{2}\hat{\mu}_1 + \frac{1}{2}\mu_2\Big)(v) \geq \int_{\mathcal{V}} \max_{p \in B_v(x)} u(p)\, d\Big(\frac{1}{2}\hat{\mu}_2 + \frac{1}{2}\mu_1\Big)(v). \end{aligned}$$

If this is true of all menus x, then the (hypothetical) preference represented by the random Strotz representation $(u, \frac{1}{2}\hat{\mu}_2 + \frac{1}{2}\mu_1)$ is more temptation averse than the preference represented by $(u, \frac{1}{2}\hat{\mu}_1 + \frac{1}{2}\mu_2)$. Thus by Theorem 7, $OV_1(x) \geq OV_2(x)$ for all x if and only if $\frac{1}{2}\hat{\mu}_1 + \frac{1}{2}\mu_2 \gg_u \frac{1}{2}\hat{\mu}_2 + \frac{1}{2}\mu_1$ or, equivalently,

$$\frac{1}{2}\hat{\mu}_1(\mathcal{U}) + \frac{1}{2}\mu_2(\mathcal{U}) \geq \frac{1}{2}\hat{\mu}_2(\mathcal{U}) + \frac{1}{2}\mu_1(\mathcal{U})$$

for every u-upper set $\mathcal{U}$. Rearranging terms, this is precisely condition (3).

C.4 Proof of Theorem 6

Define the function $\sigma : \mathcal{V} \to \mathcal{V}$ by $\sigma(v) = u + v$, and define the measures $\hat{\nu}$ and ν on $\mathcal{V}$ by $\hat{\nu}(E) = \hat{\mu}(\sigma^{-1}(E))$ and $\nu(E) = \mu(\sigma^{-1}(E))$ for any measurable set E. Observe that for any

menu x,

$$\begin{aligned}
\int_{\mathcal{V}} \min_{p\in B_v(x)} u(p)d\hat{\nu}(v) &= \int_{\mathcal{V}} \min_{p\in B_{u+v}(x)} u(p)d\hat{\mu}(v) \qquad \text{(change of variables)} \\
&\geq \int_{\mathcal{V}} \Big[\max_{p\in x}(u(p)+v(p)) - \max_{q\in x} v(q)\Big] d\hat{\mu}(v) \\
&= U(x) \\
&\geq u(m(\lambda^x)) \qquad \text{(naivete)} \\
&= \int_{\mathcal{V}} u(p_x(v))d\mu(v) \\
&\geq \int_{\mathcal{V}} \min_{p\in B_{u+v}(x)} u(p)d\mu(v) \\
&= \int_{\mathcal{V}} \min_{p\in B_v(x)} u(p)d\nu(v). \qquad \text{(change of variables)}
\end{aligned} \tag{16}$$

Thus,

$$\int_{\mathcal{V}} \min_{p\in B_v(x)} u(p)d\hat{\nu}(v) \geq \int_{\mathcal{V}} \min_{p\in B_v(x)} u(p)d\nu(v),$$

which we rewrite as

$$\int_{\mathcal{V}} \max_{p\in B_v(x)} [-u(p)]d\hat{\nu}(v) \leq \int_{\mathcal{V}} \max_{p\in B_v(x)} [-u(p)]d\nu(v). \tag{17}$$

Consider the binary relations $\succsim^{\hat{\nu}}$ and $\succsim^{\nu}$ defined by their Random Strotz representations $(-u, \hat{\nu})$ and $(-u, \nu)$, respectively. Equation (17) shows that $\succsim^{\hat{\nu}}$ is more temptation-averse than $\succsim^{\nu}$. Theorem 7 applies since $\hat{\nu}$ and ν have finite-dimensional supports, and implies that $\nu \gg_{-u} \hat{\nu}$.

Consider a u-upper set $\mathcal{U}$, and $v \in \mathcal{V}\setminus\mathcal{U}$, $v' \in \mathcal{V}$ such that $v' \gg_{-u} v$. It is easy to show that this latter condition is equivalent to $v \gg_u v'$. Suppose that $v' \in \mathcal{U}$. Since $\mathcal{U}$ is a u-upper set, the condition $v \gg_u v'$ implies $v \in \mathcal{U}$, which is a contradiction. Hence, $v' \in \mathcal{V}\setminus\mathcal{U}$ for any v' such that $v' \gg_{-u} v$. This shows that $\mathcal{V}\setminus\mathcal{U}$ is a $(-u)$-upper set, and therefore $\nu(\mathcal{V}\setminus\mathcal{U}) \geq \hat{\nu}(\mathcal{V}\setminus\mathcal{U})$, or equivalently $\hat{\nu}(\mathcal{U}) \geq \nu(\mathcal{U})$.

We therefore have

$$\hat{\mu}(\{v \in \mathcal{V} : u+v \in \mathcal{U}\}) = \hat{\nu}(\mathcal{U}) \geq \nu(\mathcal{U}) = \mu(\{v \in \mathcal{V} : u+v \in \mathcal{U}\}). \tag{18}$$

To complete the proof, we show that Equation (18) is strict for some $\mathcal{U}$ if the individual is strictly naive. Suppose, by contradiction, that Equation (18) is satisfied as an equality for all u-upper sets. The arguments above imply that $\hat{\nu}(\mathcal{U}) = \nu(\mathcal{U})$ for any $(-u)$-upper set $\mathcal{U}$, i.e., by Theorem 7 that $\succsim^{\hat{\nu}}$ is more temptation-averse than $\succsim^{\nu}$ and vice versa. This implies that Equation (17) is satisfied as an equality for all x, and therefore the system in Equation (16) only contains equalities. In particular, $U(x) = u(m(\lambda^x))$ for all x, i.e., the individual is sophisticated.

C.5 Proof of Corollary 6

Lemma 4. *Suppose u and v are independent expected-utility functions, and define a function $g : [0,1] \to \mathcal{V}$ by $g(\alpha) = \alpha u + (1-\alpha)v$.*

1. *Take any cumulative distribution functions F and $\hat{F}$ on $[0,1]$, and define probability measures μ and $\hat{\mu}$ on $\mathcal{V}$ by $\mu \equiv F \circ g^{-1}$ and $\hat{\mu} \equiv \hat{F} \circ g^{-1}$.[37] If $(u, v, F, \hat{F})$ is an uncertain intensity Strotz representation of a preference $(\succsim, \lambda)$, then $(u, \mu, \hat{\mu})$ is a random Strotz representation of $(\succsim, \lambda)$.*

2. *Take any cumulative distribution functions F_1 and F_2 on $[0,1]$, and define probability measures μ_1 and μ_2 on $\mathcal{V}$ by $\mu_i \equiv F_i \circ g^{-1}$. Then $\mu_1 \gg_u \mu_2$ if and only if $F_1 \geq_{FOSD} F_2$.*

Proof. (1): Note that by assumption $\succsim$ is represented by

$$U(x) = \int_0^1 \max\{u(p) : p \in B_{g(\alpha)}(x)\}\, d\hat{F}(\alpha).$$

By the standard change of variables formula, this implies

$$\begin{aligned} U(x) &= \int_{\mathcal{V}} \max\{u(p) : p \in B_{\tilde{v}}(x)\}\, d(\hat{F} \circ g^{-1})(\tilde{v}) \\ &= \int_{\mathcal{V}} \max\{u(p) : p \in B_{\tilde{v}}(x)\}\, d\hat{\mu}(\tilde{v}), \end{aligned}$$

and hence $(u, \hat{\mu})$ is a random Strotz representation of $\succsim$.

Note also that by assumption there exists, for each menu x, a measurable selection function $p_x : [0,1] \to x$ with $p_x(\alpha) \in B_u(B_{g(\alpha)}(x))$ for all $\alpha \in [0,1]$ such that

$$\lambda^x(y) = F(p_x^{-1}(y))$$

for all measurable $y \subset x$. Take any measurable selection function $\tilde{p}_x : \mathcal{V} \to x$ with $\tilde{p}_x(\tilde{v}) \in B_u(B_{\tilde{v}}(x))$ for all $\tilde{v} \in \mathcal{V}$ that also satisfies $p_x(\alpha) = \tilde{p}_x(g(\alpha))$ for all $\alpha \in [0,1]$.[38] Therefore, for any measurable $y \subset x$,

$$\lambda^x(y) = F(g^{-1}(\tilde{p}_x^{-1}(y))) = \mu(\tilde{p}_x^{-1}(y)),$$

and hence (u, μ) is a random Strotz representation of λ.

[37] We are abusing notation slightly and using F to also denote the probability measure on $[0,1]$ that has F as its distribution function. That is, for any measurable set $A \subset [0,1]$, we write $F(A)$ to denote $\int_A dF(\alpha)$. Thus $\mu(E) = \int_{\{\alpha' : g(\alpha') \in E\}} dF(\alpha)$ for any measurable $E \subset \mathcal{V}$.

[38] To see that such a selection function $\tilde{p}_x$ exists, fix any measurable selection function $\hat{p}_x : \mathcal{V} \to x$ with $\hat{p}_x(\tilde{v}) \in B_u(B_{\tilde{v}}(x))$ for all $\tilde{v} \in \mathcal{V}$. Let $\bar{\mathcal{V}} = g([0,1]) \subset \mathcal{V}$. When the codomain of g is restricted to $\bar{\mathcal{V}}$, i.e., $g : [0,1] \to \bar{\mathcal{V}}$, this function is a bijection. Now define $\tilde{p}_x(\tilde{v}) = p_x(g^{-1}(\tilde{v}))$ for $\tilde{v} \in \bar{\mathcal{V}}$ and $\tilde{p}_x(\tilde{v}) = \hat{p}_x(\tilde{v})$ for $\tilde{v} \notin \bar{\mathcal{V}}$.

(2): Suppose $\mu_i \equiv F_i \circ g^{-1}$ for $i = 1, 2$ and $\mu_1 \gg_u \mu_2$. Fix any $\alpha \in [0, 1]$, and let $\mathcal{U} = \{v' \in \mathcal{V} : v' \gg_u \alpha u + (1-\alpha)v\}$. By construction, $\mathcal{U}$ is a u-upper set, so $\mu_1(\mathcal{U}) \geq \mu_2(\mathcal{U})$. In addition, $g^{-1}(\mathcal{U}) = [\alpha, 1]$. Therefore,

$$F_1([\alpha, 1]) = \mu_1(\mathcal{U}) \geq \mu_2(\mathcal{U}) = F_2([\alpha, 1]).$$

Since this is true for all $\alpha \in [0, 1]$, $F_1 \geq_{FOSD} F_2$.

Conversely, suppose $F_1 \geq_{FOSD} F_2$. Fix any u-upper set $\mathcal{U}$. Note that for any $0 \leq \alpha \leq \alpha' \leq 1$, we have $g(\alpha') \gg_u g(\alpha)$ and hence

$$g(\alpha) \in \mathcal{U} \implies g(\alpha') \in \mathcal{U}.$$

This implies that the set $g^{-1}(\mathcal{U})$ is an interval from some $\alpha^* \in [0, 1]$ to 1.[39] Therefore,

$$\mu_1(\mathcal{U}) = F_1(g^{-1}(\mathcal{U})) \geq F_2(g^{-1}(\mathcal{U})) = \mu_2(\mathcal{U}).$$

Since this is true for all u-upper sets, $\mu_1 \gg_u \mu_2$. ■

Turning now to the proof of Corollary 6, suppose $(\succsim, \lambda)$ has an uncertain intensity Strotz representation $(u, v, F, \hat{F})$. Define g as in Lemma 4 for u and v, define measures $\mu \equiv F \circ g^{-1}$ and $\hat{\mu} \equiv \hat{F} \circ g^{-1}$ on $\mathcal{V}$. By part 1 of Lemma 4, $(u, \mu, \hat{\mu})$ is a random Strotz representation for $(\succsim, \lambda)$. Therefore, by Theorem 3 together with part 2 of Lemma 4, the individual is naive if and only if $\hat{F} \geq_{FOSD} F$ (and is sophisticated if and only if $\hat{F} = F$).

C.6 Proof of Corollary 7

Suppose $(\succsim_1, \lambda_1)$ and $(\succsim_2, \lambda_2)$ are naive and have uncertain intensity Strotz representations $(u, v, F_1, \hat{F}_1)$ and $(u, v, F_2, \hat{F}_2)$. Define g as in Lemma 4 for u and v, define measures $\mu_i \equiv F_i \circ g^{-1}$ and $\hat{\mu}_i \equiv \hat{F}_i \circ g^{-1}$ on $\mathcal{V}$. By part 1 of Lemma 4, $(u, \mu_i, \hat{\mu}_i)$ is a random Strotz representation for $(\succsim_i, \lambda_i)$ for $i = 1, 2$. The result follows from applications of Theorems 4 and 5, respectively, together with part 2 of Lemma 4.

C.7 Proof of Corollaries 4 and 5

A maximally present-biased preference only values immediate consumption in period 1 and ignores all subsequent consumption, which is equivalent to the extreme case where $\beta = 0$: $v_0(c) = w(c_1)$. Any convex combination of the virtuous utility u and maximally present-biased v_0 can be rewritten as the following familiar formula:

$$\beta u(c) + (1-\beta)v_0(c) = w(c_1) + \beta \sum_{t=2}^{\infty} \delta^{t-1} w(c_t) = v_\beta(c).$$

[39] That is, it is equal to either $(\alpha^*, 1]$ or $[\alpha^*, 1]$, where $\alpha^* = \inf g^{-1}(\mathcal{U})$.

Therefore, uncertainty about the present-bias parameter β simply parameterizes uncertainty about the intensity of u relative to v_0, and β is the relative weighting of exponential discounting versus extreme impatience. Thus an RQH representation $(w, F, \hat{F}, \delta)$ can equivalently be expressed as an uncertain intensity Strotz representation $(u, v_0, F, \hat{F})$. With this observation, the results follow directly from Corollaries 6 and 7 in Appendix B.

References

Ahn, D. S. (2008): "Ambiguity Without a State Space," *Review of Economic Studies*, 75, 3–28.

Ahn, D. S., R. Iijima, and T. Sarver (2016): "Naivete about Temptation and Self-Control: Foundations for Naive Quasi-Hyperbolic Discounting," Working paper.

Ahn, D. S., and T. Sarver (2013): "Preference for Flexibility and Random Choice," *Econometrica*, 81, 341–361.

Ali, S. N. (2011): "Learning Self-Control," *Quarterly Journal of Economics*, 126, 857–893.

Aliprantis, C., and K. Border (2006): *Infinite Dimensional Analysis*, 3rd edition. Berlin, Germany: Springer-Verlag.

Augenblick, N., M. Niederle, and C. Sprenger (2015): "Working Over Time: Dynamic Inconsistency in Real Effort Tasks," *Quarlery Journal of Economics*, 130, 1067–1115.

Augenblick, N., and M. Rabin (2015): "An Experiment on Time Preference and Misprediction in Unpleasant Tasks," Working paper, Haas School of Business and Harvard University.

Bousquet, L. (2017): "Measuring Time Preferences and Anticipation: A Lab Experiment," Working paper.

Chatterjee, K., and R. V. Krishna (2009): "A "Dual Self" Representation for Stochastic Temptation," *American Economic Journal: Microeconomics*, 1, 148–167.

DellaVigna, S. (2009): "Psychology and Economics: Evidence from the Field," *Journal of Economic Literature*, 47, 315–372.

DellaVigna, S., and U. Malmendier (2004): "Contract Design and Self-Control: Theory and Evidence," *Quarterly Journal of Economics*, 119, 353–402.

DellaVigna, S., and U. Malmendier (2006): "Paying Not to Go to the Gym," *American Economic Review*, 96, 694–719.

Dekel, E., and B. L. Lipman (2010): "Costly Self-Control and Random Self-Indulgence," Working paper.

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

Duflo, E., M. Kremer, and J. Robinson (2011): "Nudging Farmers to Use Fertilizer: Evidence from Kenya," *American Economic Review*, 101, 2350–2390.

Eliaz, K., and R. Spiegler (2006): "Contracting with Diversely Naive Agents," *Review of Economic Studies*, 72, 689–714.

Epstein, L. G. (1999): "A Definition of Uncertainty Aversion," *Review of Economic Studies*, 66, 579-608.

Giné, X., D. Karlan, and J. Zinman (2010): "Put your Money where your Butt is: a Commitment Contract for Smoking Cessation," *American Economic Journal: Applied Economics*, 2, 213-235.

Ghirardato, P., and M. Marinacci (2002): "Ambiguity Made Precise: A Comparative Foundation," *Journal of Economic Theory*, 102, 251-289.

Grant, S., A. Kajii, and B. Polak (2000): "Preference for Information and Dynamic Consistency," *Theory and Decision*, 48, 263–286.

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

Gul, F., and W. Pesendorfer (2004): "Self-Control and the Theory of Consumption," *Econometrica*, 72, 119–158.

Gul, F., and W. Pesendorfer (2005): "The Revealed Preference Theory of Changing Tastes," *Review of Economic Studies*, 72, 429–448.

Heidhues, P., and B. Kőszegi (2009): "Futile Attempts at Self-Control," *Journal of the European Economic Association*, 7, 423–434.

Heidhues, P., and B. Kőszegi (2010): "Exploiting Naivete about Self-Control in the Credit Market," *American Economic Review*, 100, 2279–2303.

Kaur, S., M. Kremer, and S. Mullainathn (2015): "Self-Control at Work," *Journal of Political Economy*, 123, 1227–1277.

Kopylov, I. (2012): "Perfectionism and Choice," *Econometrica*, 80, 1819–1943.

Kőszegi, B. (2014): "Behavioral Contract Theory," *Journal of Economic Literature*, 52, 1075–1118.

Kreps, D., and E. Porteus (1978): "Temporal Resolution of Uncertainty and Dynamic Choice Theory," *Econometrica*, 46, 185–200.

Le Yaouanq, Y. (2015): "Anticipating Preference Reversal," Toulouse School of Economics Working Paper No. TSE-585.

Lipman, B. L., and W. Pesendorfer (2013): "Temptation," in Acemoglu, Arellano, and Dekel, eds., *Advances in Economics and Econometrics: Tenth World Congress*, Volume 1, Cambridge University Press.

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

O'Donoghue, T., and M. Rabin (1999): "Doing It Now or Later," *American Economic Review*, 89, 103–124

O'Donoghue, T., and M. Rabin (2001): "Choice and Procrastination," *Quarterly Journal of Economics*, 116, 121–160.

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

Shui, H., and L. M. Ausubel (2005): "Time Inconsistency in the Credit Card Market," Working paper, University of Maryland.

Siniscalchi, M. (2011): "Dynamic Choice Under Ambiguity," *Theoretical Economics*, 6, 379–421.

Spiegler, R. (2011): *Bounded Rationality and Industrial Organization*. New York, NY: Oxford University Press.

Stovall, J. (2010): "Multiple Temptations," *Econometrica*, 78, 349-376.

Stovall, J. (2018): "Temptation with Uncertain Normative Preferences," *Theoretical Economics*, 13, 145–174.

Supplementary Appendix for
Behavioral Characterizations of naivete for Time-Inconsistent Preferences

Abstract

Theorem 7 from Appendix A of the main paper is an extension of the characterization of comparative temptation aversion from Dekel and Lipman (2012): While their result required a finite consumption space, our extension applies to any random Strotz representation defined on any compact and metrizable consumption space C, provided the measure in the representation has finite-dimensional support. As discussed in the paper, this extension is important for a number of applications, including dynamic consumption decisions where C is a set of infinite consumption streams. In this supplement, we provide a proof of Theorem 7.

S.1 Proof of Theorem 7

S.1.1 Sufficiency: more temptation averse $\Longrightarrow$ less u-aligned

The following is the relevant result from Dekel and Lipman (2012), which they proved for the case of finite C.

Theorem S.1 (Dekel and Lipman (2012))**.** *Suppose C has finite cardinality. Suppose $\succsim_1$ and $\succsim_2$ have random Strotz representations (u, μ_1) and (u, μ_2). Then $\succsim_2$ is more temptation averse than $\succsim_1$ if and only if $\mu_1 \gg_u \mu_2$.*

Proof. Theorem 4 in Dekel and Lipman (2012) establishes the equivalence of $\succsim_2$ being more temptation averse than $\succsim_1$ and another condition on the representations that they refer to as conditional dominance. However, they also establish that $\mu_1 \gg_u \mu_2$ as an intermediate step in their proof.[40] The equivalence asserted in Theorem S.1 is also stated explicitly in Theorem 4 of their working paper, Dekel and Lipman (2010).[41] ■

[40] To show that $\succsim_2$ being more temptation averse that $\succsim_1$ implies $\mu_1 \gg_u \mu_2$, the relevant results in Dekel and Lipman (2012) are the following: Lemma 3 shows that a partial order vC_uv' used in their paper is equivalent to our order $v \gg_u v'$ (ignoring their normalization of utility functions). Lemmas 4, 5, and 6 and the arguments on page 1296 show that for any set W that is closed under C_u (is a u-upper set in our terminology), $\mu_1(W) \geq \mu_2(W)$.

[41] Dekel and Lipman (2010) impose a normalization on the set of utility functions used in their result. However, by the uniqueness properties of the random Strotz representation established in Theorem 3 of Dekel and Lipman (2012), the probability of any u-upper set is the same for any random Strotz representation of the same preference. Therefore, their normalization of utilities is inconsequential for the result.

To prove the sufficiency part of Theorem 7, we now show that the sufficiency direction in Theorem S.1 can be extended to any compact and metrizable space C and any random Strotz representations (u, μ_1) and (u, μ_2) defined on that space, subject to our restriction that each μ_i has finite-dimensional support. Our approach is to show that the relationship between μ_1 and μ_2, specifically $\mu_1 \gg_u \mu_2$, can be inferred from looking at the restriction of the representations and preferences to a carefully chosen finite consumption space $C^* \subset C$.

The following preliminary result will be useful in the sequel. Recall that $\mathcal{V}$ denotes the set of all continuous functions $v : C \to \mathbb{R}$, i.e., the set of all expected-utility functions.

Lemma S.1. *Suppose the set $\{v_1, \ldots, v_n\} \subset \mathcal{V}$ is linearly independent. Then there exists a finite subset $C^* \subset C$ such that the set $\{v_1^*, \ldots, v_n^*\}$ is linearly independent, where $v_i^* = v_i|_{C^*}$ is the restriction of the function v_i to C^*.*

Proof. Suppose to the contrary that for every finite $B \subset C$, the collection $\{v_1|_B, \ldots, v_n|_B\}$ is linearly dependent. Then for any finite $B \subset C$, the set $A_B \subset \mathbb{R}^n$ defined by

$$A_B = \{\alpha \in \mathbb{R}^n : \|\alpha\| = 1 \text{ and } \alpha_1 v_1(c) + \cdots + \alpha_n v_n(c) = 0 \ \forall c \in B\}$$

is nonempty. Note that A_B is also a closed subset of the unit ball in $\mathbb{R}^n$, which is itself compact because n is finite. Let $\mathcal{B}$ denote the set of all nonempty finite subsets of C. For any $B_1, \ldots, B_k \in \mathcal{B}$, we have

$$A_{B_1} \cap \cdots \cap A_{B_k} = A_{B_1 \cup \cdots \cup B_k} \neq \emptyset,$$

since $B_1 \cup \cdots \cup B_k$ is finite and hence also in $\mathcal{B}$. Thus the collection $\{A_B\}_{B \in \mathcal{B}}$ has the finite intersection property. Since these sets are closed subsets of a compact set, this implies $\bigcap_{B \in \mathcal{B}} A_B \neq \emptyset$. However, since

$$\bigcap_{B \in \mathcal{B}} A_B = \{\alpha \in \mathbb{R}^n : \|\alpha\| = 1 \text{ and } \alpha_1 v_1(c) + \cdots + \alpha_n v_n(c) = 0 \ \forall c \in C\},$$

this implies the set $\{v_1, \ldots, v_n\}$ is linearly dependent, a contradiction. ■

Since μ_1 and μ_2 have finite-dimensional support, there exists a finite set of expected-utility functions $\{v_1, \ldots, v_n\} \subset \mathcal{V}$ such that $\text{supp}(\mu_i) \subset \text{span}(\{v_1, \ldots, v_n\})$ for $i = 1, 2$. Consider the set of function $\{u, \mathbf{1}, v_1, \ldots, v_n\}$, where $\mathbf{1}$ denotes the constant function with $\mathbf{1}(c) = 1$ for all $c \in C$. Without loss of generality, assume that this set of functions is linearly independent. Otherwise, we can sequentially remove the functions v_i until we obtain a linearly independent set.[42] To simplify notation in what follows, let $\mathcal{V}_s \equiv \text{span}(\{u, \mathbf{1}, v_1, \ldots, v_n\}) \subset \mathcal{V}$. Thus $\mu_1(\mathcal{V}_s) = \mu_2(\mathcal{V}_s) = 1$.

[42]Note that the set $\{u, \mathbf{1}\}$ must be linearly independent since u assumed to be nontrivial (i.e., not

Take C^* as in Lemma S.1 for the set $\{u, \mathbf{1}, v_1, \ldots, v_n\}$. Let $\mathcal{V}^*$ denote the set of all continuous real-valued functions on C^* and let $\mathcal{V}_s^* \equiv \text{span}(\{u^*, \mathbf{1}^*, v_1^*, \ldots, v_n^*\}) \subset \mathcal{V}^*$, where $u^* = u|_{C^*}$, $\mathbf{1}^* = \mathbf{1}|_{C^*}$, and $v_i^* = v_i|_{C^*}$. Note that each of the functions $u^*, v_1^*, \ldots, v_n^*$ must be nontrivial (i.e., not constant) since function $\mathbf{1}^*$ together with these functions forms a linearly independent set.

Lemma S.2. *Define a function $g : \mathcal{V}_s \to \mathcal{V}_s^*$ by $g(v) = v|_{C^*}$, and define a measure μ_i^* on $\mathcal{V}^*$ by $\mu_i^*(E) = \mu_i(g^{-1}(E))$ for any measurable set $E \subset \mathcal{V}^*$ for $i = 1, 2$.*[43]

1. *The function g is a homeomorphism. That is, g is bijection and both g and its inverse function g^{-1} are continuous.*
2. *For any measurable set $E \subset \mathcal{V}$, $\mu_i(E) = \mu_i^*(g(E \cap \mathcal{V}_s))$.*
3. *For any proper u-upper set $\mathcal{U}$ in $\mathcal{V}$ (i.e., $\mathcal{U} \subsetneq \mathcal{V}$), the set $\mathcal{U}^* = g(\mathcal{U} \cap \mathcal{V}_s)$ is a u^*-upper set in $\mathcal{V}^*$.*
4. *Let $\succsim_i^*$ denote the restriction of $\succsim_i$ to sets of lotteries with support in C^*, which we can identify with the set $\mathcal{K}(\Delta(C^*))$. Then (u^*, μ_i^*) is a random Strotz representation for $\succsim_i^*$ for $i = 1, 2$.*

Proof. (1): This is a standard application of the fundamental theorem of linear algebra for finite-dimensional vector spaces. Note that g is a linear function from the linear space $\mathcal{V}_s$ with basis vectors $\{u, \mathbf{1}, v_1, \ldots, v_n\}$ to the linear space $\mathcal{V}_s^*$ with basis vectors $\{u^*, \mathbf{1}^*, v_1^*, \ldots, v_n^*\}$. Since g maps each basis vector for $\mathcal{V}_s$ to the corresponding basis vector for $\mathcal{V}_s^*$ and the number of basis vectors is the same for each space, g is a bijection. Since any linear function between finite-dimensional spaces is continuous, both g and g^{-1} are continuous.[44]

(2): Fix any measurable set $E \subset \mathcal{V}$. Then

$$\mu_i(E) = \mu_i(E \cap \mathcal{V}_s) = \mu_i(g^{-1}(g(E \cap \mathcal{V}_s))) = \mu_i^*(g(E \cap \mathcal{V}_s)),$$

constant). Moreover, if $\text{span}\{u, \mathbf{1}\} = \text{span}\{u, \mathbf{1}, v_1, \ldots, v_n\}$, then the support of the measures in the random Strotz representations (u, μ_i) must assign all probability to the set of affine transformations of u. In this case, the representations reduce to time-consistent expected-utility maximization, and we have $\mu_1 \approx \mu_2$. Except in this trivial case, the linearly independent set of expected-utility functions whose span contains the support of μ_i must contain u, $\mathbf{1}$, and at least some of the v_i functions.

[43] In the definition of μ_i^*, we are implicitly treating g as a function from $\mathcal{V}_s$ into $\mathcal{V}^*$. We could equivalently define μ_i^* by $\mu_i^*(E) = \mu_i(g^{-1}(E \cap \mathcal{V}_s^*))$.

[44] A more detailed argument is as follows: Define $h : \mathbb{R}^{n+2} \to \mathcal{V}_s$ by $h(\alpha) = \alpha_1 v_1 + \cdots + \alpha_n v_n + \alpha_{n+1} u + \alpha_{n+2}\mathbf{1}$ and define $h^* : \mathbb{R}^{n+2} \to \mathcal{V}_s^*$ by $h^*(\alpha) = \alpha_1 v_1^* + \cdots + \alpha_n v_n^* + \alpha_{n+1} u^* + \alpha_{n+2}\mathbf{1}^*$. By the linear independence of these sets of functions, both h and h^* are bijections. It is trivial that both functions are continuous, and by Aliprantis and Border (2006, Corollary 5.24) both h^{-1} and h^{*-1} are also continuous. Note that $g = h^* \circ h^{-1}$ and $g^{-1} = h \circ h^{*-1}$, and hence these functions are continuous.

where the first equality follows from $\mu_i(\mathcal{V}_s) = 1$, the second follows from $g^{-1}(g(E \cap \mathcal{V}_s)) = E \cap \mathcal{V}_s$ (which holds because g is a bijection), and the third follows from the definition of μ_i^*.

(3): First observe that for any $v, v' \in \mathcal{V}_s$,

$$\begin{aligned} v \approx v' &\iff v = av' + b\mathbf{1} \text{ for some } a > 0, b \in \mathbb{R} \\ &\iff g(v) = ag(v') + b\mathbf{1} \text{ for some } a > 0, b \in \mathbb{R} \\ &\iff g(v) \approx g(v'). \end{aligned} \tag{S.1}$$

Now fix any proper u-upper set $\mathcal{U}$ in $\mathcal{V}$, and let $\mathcal{U}^* = g(\mathcal{U} \cap \mathcal{V}_s)$. To see that $\mathcal{U}^*$ is a u^*-upper set, fix any $v^* \in \mathcal{U}^*$ and $v^{*\prime} \in \mathcal{V}^*$ with $v^{*\prime} \gg_{u^*} v^*$. We need to show that $v^{*\prime} \in \mathcal{U}^*$. Let $v = g^{-1}(v^*) \in \mathcal{U} \cap \mathcal{V}_s$. Note that we cannot have $v^* \approx -u^*$, as this would imply by Equation (S.1) that $v \approx g^{-1}(-u^*) = -u$, which would in turn imply by the definition of a u-upper set that $\mathcal{U} = \mathcal{V}$, contradicting our assumption that $\mathcal{U}$ is a proper subset of $\mathcal{V}$. Therefore, there exists some $\alpha \in [0, 1]$ such that

$$v^{*\prime} \approx \alpha u^* + (1 - \alpha)v^*.$$

Thus there exist $a > 0$ and $b \in \mathbb{R}$ such that

$$v^{*\prime} = a\alpha u^* + a(1 - \alpha)v^* + b\mathbf{1}^*.$$

Let

$$v' = a\alpha u + a(1 - \alpha)v + b\mathbf{1}.$$

Clearly $v' \in \mathcal{V}_s$. Moreover, since $v' \gg_u v$ we have $v' \in \mathcal{U}$. Thus $v' \in \mathcal{U} \cap \mathcal{V}_s$, which implies $v^{*\prime} = g(v') \in \mathcal{U}^*$.

(4): We can treat a lottery $p \in \Delta(C^*)$ as a measure defined only on the space C^*, or we treat this as a lottery in $\Delta(C)$ that assigns probability zero to the set $C \setminus C^*$. Thus we will abuse notation slightly and evaluate the lotteries $p \in \Delta(C^*)$ using both functions in $\mathcal{V}^*$ and functions in $\mathcal{V}$. Note that for any $v \in \mathcal{V}_s$, $v(p) = v^*(p)$ for $v^* = g(v) \in \mathcal{V}_s^*$.

Therefore, for any $x \in \mathcal{K}(\Delta(C^*))$,

$$\begin{aligned}
U_i^*(x) &= \int_{\mathcal{V}^*} \max_{p \in B_{v^*}(x)} u^*(p)\, d\mu_i^*(v^*) \\
&= \int_{\mathcal{V}_s^*} \max_{p \in B_{v^*}(x)} u^*(p)\, d(\mu_i \circ g^{-1})(v^*) && \text{(definition of } \mu_i^*) \\
&= \int_{\mathcal{V}_s} \max_{p \in B_{g(v)}(x)} u^*(p)\, d\mu_i(v) && \text{(change of variables)} \\
&= \int_{\mathcal{V}_s} \max_{p \in B_v(x)} u(p)\, d\mu_i(v) \\
&= U_i(x).
\end{aligned}$$

Thus U_i^* is the restriction of U_i to $\mathcal{K}(\Delta(C^*))$. Also, note that μ_i^* is nontrivial (i.e., assigns probability zero to the set of constant functions) since

$$\mu_i^*(\{\alpha \mathbf{1}^* : \alpha \in \mathbb{R}\}) = \mu_i(g^{-1}(\{\alpha \mathbf{1}^* : \alpha \in \mathbb{R}\})) = \mu_i(\{\alpha \mathbf{1} : \alpha \in \mathbb{R}\}) = 0,$$

by the nontriviality of μ_i. Hence (u^*, μ_i^*) is a random Strotz representation of $\succsim_i^*$. ■

We now prove that $\mu_1 \gg_u \mu_2$. By assumption, $\succsim_2$ is more temptation averse than $\succsim_1$. Thus for any menu x and lottery p, $\{p\} \succ_1 x$ implies $\{p\} \succ_2 x$. This implies a fortiori that the same condition must hold for lotteries and menus of lotteries with support in C^*, and hence $\succsim_2^*$ is more temptation averse than $\succsim_1^*$, where $\succsim_i^*$ is defined as in part 4 of Lemma S.2. Since C^* is finite and (u^*, μ_i^*) represents $\succsim_i^*$ for $i = 1, 2$, Theorem S.1 implies that $\mu_1^* \gg_{u^*} \mu_2^*$.

Now fix any u-upper set $\mathcal{U}$ in $\mathcal{V}$. If $\mathcal{U} = \mathcal{V}$, then trivially $\mu_1(\mathcal{U}) = \mu_2(\mathcal{U}) = 1$. Otherwise, by part 3 of Lemma S.2, $g(\mathcal{U} \cap \mathcal{V}_s)$ is a u^*-upper set in $\mathcal{V}^*$ and therefore

$$\mu_1(\mathcal{U}) = \mu_1^*(g(\mathcal{U} \cap \mathcal{V}_s)) \geq \mu_2^*(g(\mathcal{U} \cap \mathcal{V}_s)) = \mu_2(\mathcal{U}),$$

where the equalities follow from part 2 of Lemma S.2 and the inequality follows from $\mu_1^* \gg_{u^*} \mu_2^*$. Since this is true for any u-upper set $\mathcal{U}$, conclude that $\mu_1 \gg_u \mu_2$.

S.1.2 Necessity: less u-aligned $\Longrightarrow$ more temptation averse

In this section we prove that the more temptation averse comparative is implied by $\mu_1 \gg_u \mu_2$. It is worth noting that the proof of this direction does not rely on the assumption that these measures have finite-dimensional support.

The following preliminary result will be useful.

Lemma S.3. *Let u, v, v' be expected-utility functions defined on $\Delta(C)$, and suppose $v \gg_u v'$. Then for any menu x,*

$$\max_{p \in B_v(x)} u(p) \geq \max_{q \in B_{v'}(x)} u(q).$$

Proof. If $v' \approx -u$, then for any menu x,

$$\max_{q \in B_{v'}(x)} u(q) = \min_{q \in x} u(q) \leq u(p), \quad \forall p \in x.$$

In particular,

$$\max_{q \in B_{v'}(x)} u(q) \leq \max_{p \in B_v(x)} u(p).$$

If we do not have $v' \approx -u$, then $v \gg_u v'$ implies $v \approx \alpha u + (1-\alpha)v'$ for some $\alpha \in [0,1]$. First, consider $\alpha = 0$. In this case, $v \approx v'$. Therefore $B_v(x) = B_{v'}(x)$, which implies

$$\max_{p \in B_v(x)} u(p) = \max_{q \in B_{v'}(x)} u(q).$$

Finally, consider the case of $\alpha > 0$. Note that for any menu x and any $p \in B_v(x)$ and $q \in B_{v'}(x)$,

$$\alpha u(p) + (1-\alpha)v'(p) \geq \alpha u(q) + (1-\alpha)v'(q) \quad \text{and} \quad v'(q) \geq v'(p).$$

Since $\alpha > 0$, these inequalities imply $u(p) \geq u(q)$. Therefore,

$$\max_{p \in B_v(x)} u(p) \geq \max_{q \in B_{v'}(x)} u(q),$$

as claimed. ■

Suppose (u, μ_1) and (u, μ_2) are random Strotz representations of $\succsim_1$ and $\succsim_2$, and suppose $\mu_1 \gg_u \mu_2$. Fix any menu x, and let $[a, b] = u(x)$. Define $f_x : \mathcal{V} \to [a, b]$ by

$$f_x(v) = \max_{p \in B_v(x)} u(p).$$

By Lemma S.3, $v \gg_u v'$ implies $f_x(v) \geq f_x(v')$. Therefore, for any $\alpha \in [a, b]$ and $v \gg_u v'$,

$$v' \in f_x^{-1}([\alpha, b]) \iff f_x(v') \geq \alpha \implies f_x(v) \geq \alpha \iff v \in f_x^{-1}([\alpha, b]).$$

Thus $f_x^{-1}([\alpha, b])$ is a u-upper set. Therefore,

$$\mu_1(f_x^{-1}([\alpha, b])) \geq \mu_2(f_x^{-1}([\alpha, b])).$$

Define distributions $\eta_i^x \equiv \mu_i \circ f_x^{-1}$ on $[a, b]$ for $i = 1, 2$. By the preceding arguments, η_1^x first-order stochastically dominates η_2^x. Therefore, by the change of variables formula,

$$U_1(x) = \int_{\mathcal{V}} f_x(v)\, d\mu_1(v) = \int_a^b \alpha\, d\eta_1^x(\alpha) \geq \int_a^b \alpha\, d\eta_2^x(\alpha) = \int_{\mathcal{V}} f_x(v)\, d\mu_2(v) = U_2(x).$$

Since this is true for every x, and using the fact that $U_1(\{p\}) = U_2(\{p\})$ for any lottery p, it follows immediately that $\succsim_2$ is more temptation averse than $\succsim_1$.