Title
Projection Bias in Predicting Future Utility

Authors
Loewenstein, George
O'Donoghue, Ted
Rabin, Matthew

Publication Date
2000-06-06

Projection Bias in Predicting Future Utility

George Loewenstein
Carnegie-Mellon University

Ted O'Donoghue
Cornell University

Matthew Rabin
University of California - Berkeley

March 21, 2000

Abstract

People underappreciate how their own behavior and exogenous factors affect their future utility, and thus exaggerate the degree to which their future preferences resemble their current preferences. We present evidence which demonstrates the prevalence of such *projection bias*, and develop a formal model that draws out both descriptive and welfare implications of the bias. The model helps interpret established behavioral anomalies such as the endowment effect, and helps to explain commonly observed suboptimal patterns of behavior such as addiction and excessive pursuit of a high material standard of living. The model also suggests potentially welfare-improving policies, such as mandatory "cooling-off periods" for certain types of consumer decisions.

Keywords: Addiction, Consumption, Cooling Off, Misprediction, Projection Bias, Reference Dependence.

JEL Classification: A12, B49, D11, D91, E21

Acknowledgments: We are grateful to Erik Eyster, Chris Harris, and members of the Russell Sage Foundation Behavioral Economics Roundtable; seminar participants at Cornell University, Yale University, Harvard University, the University of Michigan, the University of Texas, the Toulouse Conference on Psychology and Economics, and the Jerome Levy Institute; and especially Colin Camerer and Drazen Prelec for very helpful discussions at the formative stages of this project. We thank Kitt Carpenter, Erik Eyster, David Huffman, and Chris Meissner for research assistance. For financial support, Loewenstein thanks the Center for the Study of Human Dimensions of Global Change at Carnegie Mellon University (NSF Grant SBR-9521914), O'Donoghue and Rabin thank the National Science Foundation (Award 9709485), and Rabin thanks the Russell Sage, MacArthur, and Sloan Foundations. This research was started while Loewenstein and Rabin were Fellows at the Center for Advanced Study in the Behavioral Sciences, supported by NSF Grant SBR-960123. They are very grateful for the Center's hospitality and the NSF's support.

Mail: George Loewenstein / Department of Social and Decision Sciences / Carnegie Mellon University /Pittsburgh, PA 15213-3890, Ted O'Donoghue / Department of Economics / Cornell University / 414 Uris Hall / Ithaca, NY 14853-7601, Matthew Rabin / Department of Economics / 549 Evans Hall #3880 / University of California, Berkeley / Berkeley, CA 94720-3880. E-mail: gl20@andrew.cmu.edu, edo1@cornell.edu, and rabin@econ.berkeley.edu. Web pages: http://sds.hss.cmu.edu/faculty/loewenstein.html, www.people.cornell.edu/pages/edo1/, and http://emlab.berkeley.edu/users/rabin/index.html.

1. Introduction

A person's current well-being is typically influenced not only by her current consumption, but also by such factors as her own past behavior, temporary fluctuations in her tastes, and changes in her environment. When a person faces an intertemporal choice in a situation in which tastes may change, she must make predictions about how changes in future *states* — all those factors besides contemporaneous consumption — will affect her future preferences. For example, when a person makes summer vacation plans during the winter, she must predict how she will feel in the summer; and when a person decides whether to try crack cocaine for the first time, she must predict how this consumption will influence her future enjoyment of activities, including consuming more crack cocaine.

In this paper, we formalize and explore the implications of a general bias in such predictions: People tend to underappreciate the effects of changes in their states, and hence falsely project their current preferences over consumption onto their future preferences. Far more than suggesting merely that people mispredict future tastes, this *projection bias* posits a systematic pattern in these mispredictions which can lead to systematic errors in dynamic-choice environments.

In Section 2, we review extensive evidence for the existence of projection bias, highlighting the diversity of situations in which projection bias occurs. Research suggests that people underappreciate short-term, transient changes in preferences, such as those induced by fluctuations in hunger or the presence of environmental cues, and slowly-developed but longer-lasting changes, such as those induced by addiction or changes in one's accustomed standard of living. Moreover, people underappreciate both endogenous changes in preferences that depend on prior choices, such as drug addiction, and exogenous changes in preferences that do not depend on prior choices, such those associated with aging.

In Section 3, we develop a formal model of projection bias. To fix ideas, consider a person with true period-τ preferences $u(c_\tau, s_\tau)$, where c_τ is her consumption in period τ and s_τ is her state in period τ. Let the person's prediction in period $t < \tau$ of her period-τ preferences be $\tilde{u}(c_\tau, s_\tau|s_t)$, where s_t is her state in period t. Our reading of the evidence suggests that the person's prediction $\tilde{u}(c_\tau, s_\tau|s_t)$ lies somewhere "in between" her true period-τ preferences $u(c_\tau, s_\tau)$ and her preferences given her current state $u(c_\tau, s_t)$. More precisely, we assume the person understands the qualitative nature of changes in her preferences — on all dimensions she correctly predicts in which

direction her preferences will move — but she underestimates the magnitudes of these changes. Formally, we assume that her predictions of the absolute utility from consumption, the marginal utility from consumption, and all higher-order derivatives of the utility function lie in between the true values and the values given her current state. To model dynamic choice given projection bias, we make the conventional assumption that the person maximizes her *perceived* intertemporal utility, $\sum_\tau \delta^\tau \tilde{u}_\tau$. Because projection bias implies that predicted utilities need not match actual utilities, however, the person's behavior need not correspond to correct intertemporal utility maximization.

Also in Section 3, we present two extended examples designed to illustrate the mechanics of our model and to show its consistency with the evidence in Section 2. The first example examines a person's decision about what to eat in the future, where projection bias implies that the person's choice depends too much on her hunger level at the time she makes the decision. The second example explores the relationship between projection bias and the endowment effect — the tendency for people to value an object more highly if they possess it than if they do not. We show that projection bias explains evidence that people fail to predict the endowment effect, and moreover that the endowment effect itself may in part be an error caused by projection bias.

In Sections 4, 5, and 6, we explore the behavioral and welfare implications of our model in a range of economic contexts. In Section 4, we consider the implications of projection bias for a person who cares not only about her current consumption but also about how her current consumption compares to her past consumption. We show how projection bias, combined with such reference-dependent utility, might lead to excessive consumption. We develop a simple two-period model in which a person chooses both consumption and leisure each period, and assume that consumption is more reference-dependent than leisure. Because the person underappreciates the extent to which increasing her current consumption makes her worse off in the future, she over-consumes early in life. Moreover, as time passes, the person will be surprised at how accustomed she has become to high levels of consumption, and so will want to work harder than she had planned in order to consume at a higher level than she had anticipated wanting to consume.

In Section 5, we consider the implications of projection bias for addiction. We consider an environment in which consumption of an addictive product has two harmful long-term consequences: It decreases a person's future well-being, and it increases the person's future desire for the addictive product. Contrary to models that view addiction as rational self-medication (Becker and Murphy, 1988), or as the unlucky outcome of a gamble based on rational uncertainty about one's own vul-

nerability (Orphanides and Zervos, 1995), our model of projection bias predicts that people too often become addicted because they underappreciate both of these deleterious effects of current consumption. Our model also predicts that people tend to overreact to transitory changes in the craving for addictive products caused by temporary factors such as a particularly stressful day at work or being in a smoke-filled room. On days when cravings are high, the person overestimates her future desire for the drug, which discourages efforts to quit. But on days where cravings are low, the person underestimates her future desire for the drug, and may engage in unrealistic attempts to quit. Hence, while projection bias over the future deleterious effects of current consumption leads a person to over-consume addictive products, projection bias over transient fluctuations in craving may lead to over-frequent, but unsuccessful, attempts to quit in moments of low craving.

In Section 6, we explore the potential for welfare-improving policies suggested by our model. In a variety of situations, people make difficult-to-reverse decisions when they are in a "hot" state that is unlikely to persist — e.g., buying automobiles when the dealer has them excited, getting married in the heat of passion, or committing suicide in the depth of depression. Because people underappreciate the degree to which intense feelings will dissipate, they may be too likely to make irreversible decisions. Imposing "cooling-off periods" — mandating that people delay for some duration before making an irreversible decision — may help correct such errors. To illustrate the potential benefits of cooling-off periods, we consider the purchase of a durable good when a dealer can exert sales hype to get customers temporarily excited about a product. In this environment, not only are people too likely to purchase the good, but also dealers who are aware that people have projection bias exert excessive sales hype. We show that a cooling-off period can alleviate both these problems, while (perhaps) imposing only a small cost on those trades that are in fact beneficial.

We feel that psychological evidence provides strong support for the existence of projection bias, and that our analysis in Sections 4, 5, and 6 highlights the potential economic importance of projection bias. We conclude in Section 7 by putting projection bias in broader economic context.

2. Evidence of Projection Bias

In this section we review a wide range of phenomena that exhibit the pattern which we refer to as projection bias.[1] We begin by discussing evidence that people underestimate the effects of transient fluctuations in their tastes — by falsely projecting their current transient preferences onto the future — and then examine evidence that people underestimate long-term changes in tastes caused by adaptation and other factors.

Several studies lend support to the folk wisdom that shopping on an empty stomach leads people to buy too much. This phenomenon can be interpreted as a manifestation of projection bias: People who are hungry act as if their future taste for food will reflect such hunger. Nisbett and Kanouse (1968) examined supermarket shoppers' shopping lists, asked them when they had last eaten, and then monitored their purchases. They found a positive correlation between over-shopping — buying more than what was on the shopping list — and hunger as measured by when the shopper last ate.[2] Gilbert, Gill and Wilson (1998) conducted a similar study in which hunger was manipulated by asking some shoppers to eat a muffin before they went shopping. Shoppers who ate muffins, and were thus less hungry, purchased a lower proportion of items that weren't on their shopping lists (34%) than those who did not eat muffins (51%).

Read and van Leeuwen (1998) obtained further evidence that people project their current hunger levels onto their future preferences. Office workers were asked to choose between healthy snacks and unhealthy snacks that they would receive in one week, either at a time when they should expect to be hungry (late in the afternoon) or satiated (immediately after lunch). Subjects were approached to make the choice either when they were hungry (late in the afternoon) or satiated (immediately after lunch). As depicted in Table 1, people who expected to be hungry the next week were more likely to opt for unhealthy snacks than those who expected to be satiated, presumably reflecting an increased taste for unhealthy snacks in the hungry state. But in addition, people who were hungry *when they made the choice* were more likely to opt for unhealthy snacks than those who were

[1] See Loewenstein and Schkade (1999) for a summary of much of the evidence presented in this section, as well as for a discussion of the psychological mechanisms that underlie projection bias.

[2] This effect was only observed for non-obese shoppers. Obese shoppers displayed the reverse pattern.

satiated, suggesting that people were projecting their current preferences onto their future selves.[3]

Table 1: Percentage of Subjects Choosing Unhealthy Snack
(from Read and van Leeuwen (1998))

		Future Hunger	
		Hungry	Satiated
Current Hunger	Hungry	78%	56%
	Satiated	42%	26%

Loewenstein, Nagin and Paternoster (1997) provide evidence of projection bias with regard to sexual arousal. Male undergraduates were randomly assigned to view sexually arousing or non-arousing photographs. Subjects were then exposed to a vivid first-person date scenario in which "their date" suddenly requested a termination of physical intimacy. Subjects reported their likelihood of behaving in a sexually aggressive fashion in this situation. Aroused subjects reported substantially higher likelihoods (70%) than nonaroused subjects (50%). That is, subjects' perceptions of their future preferences when sexually aroused depend on whether they are currently aroused.[4]

A pervasive feature of preferences that shows up in a broad array of domains is *loss aversion*: A person's preferences are typically defined with respect to some reference level (e.g., the status quo), where the person dislikes losses relative to the reference level significantly more than she likes gains. One important manifestation of loss aversion is the *endowment effect*, which refers to people's tendency to value an object more highly if they possess it than if they do not.[5]

Loewenstein and Adler (1995) demonstrate projection bias with regard to loss aversion by showing that people underestimate the magnitude of their own endowment effects. In one study, subjects who were randomly assigned to a "prediction" treatment group were shown an embossed coffee mug and then told that they would later be given one as a prize but would have the opportunity to exchange it for cash. They were then shown the form that would be used to elicit their selling price and were asked to complete it as they expected they would once they received the mug. After a

[3] The healthy snacks were apples and bananas; the unhealthy snacks were crisps, borrelnoten, Mars Bars, and Snickers Bars. We adopt the terminology healthy and unhealthy from the experimenters; whether these terms are appropriate is irrelevant to our point.

[4] These results are consistent with projection bias only if the actual likelihood of behaving in a sexually aggressive fashion is higher than 70%. Of course, the experimenters were unable to measure actual sexual behavior.

[5] See Kahneman, Knetsch and Thaler (1991) for a review of the endowment effect. Tversky and Kahneman (1991) show how the endowment effect arises naturally from assuming that people are loss-averse, and Strahilevitz and Loewenstein (1998) provide further empirical studies supporting the loss aversion account of the endowment effect.

delay, they were actually given the mug, and then asked to complete the same form eliciting selling prices. The other half of subjects were simply given mugs without first making predictions, and then they completed the form eliciting selling prices. The results, presented in Table 2, reveal a systematic underprediction of the impact of endowment on preferences: The predicted selling prices of the prediction group were substantially lower than the actual selling prices of both the prediction group and the non-prediction group.[6]

Table 2: Predicted and Actual Valuation of Mug
(from Loewenstein and Adler (1995))

Group	Condition	Number of Subjects	Prediction of Valuation	Actual Valuation
Carnegie Mellon University	Prediction	14	$3.73 (0.41)	$5.40 (0.65)
	No Prediction	13	———	$6.46 (0.54)
University of Pittsburgh	Prediction	22	$3.27 (0.48)	$4.56 (0.59)
	No Prediction	17	———	$4.98 (0.53)

(standard errors in parentheses)

The above examples, in which people underappreciate the impact of transient fluctuations in their tastes, are striking because they involve mispredictions of "changes" in tastes with which people should be entirely familiar. Virtually all humans pass back and forth between states of hunger and satiation on at least a daily basis, fluctuate between sexual arousal and non-arousal with comparable frequency, and obtain and part with objects many times over the course of their lives. Hence, even giving the rational-choice model all benefits of doubt, it is difficult to dismiss these examples of projection bias as merely rational uncertainty about the consequences of future

[6] While this study did not give participants an incentive to predict accurately, in a second study participants were told that there was a 50% chance that they would receive a mug (based on a coin flip) and were given a form eliciting selling prices that would apply if they did, in fact, obtain a mug. Selling prices were also elicited from other subjects who were simply given a mug, and choice prices were elicited from subjects who were not endowed. Again, there was a significant underprediction by non-endowed subjects of their own selling price once endowed.

states.[7]

In turning to evidence of projection bias in predicting long-term changes in tastes, we concentrate on what is probably the most important category of projection bias: the underappreciation of adaptation. There is a plethora of evidence that adaptation is a central component of human well-being (see Helson (1964), and Frederick and Loewenstein (1999) for a recent review). This literature consistently shows that people adapt to major changes in their life circumstances. But there is also a great deal of evidence that people underestimate the extent to which they will adapt to new circumstances, and hence overestimate the impact of major changes in circumstances on their long-run level of happiness.

Because research on adaptation cannot be conducted in controlled laboratory settings, it is necessarily less conclusive than research on underestimation of more short-term changes in tastes. Studies of adaptation almost always rely on comparisons of self-reported well-being across situations (e.g., before and after suffering a calamity), which raises the possibility that measured adaptation could result, in whole or in part, from changes in the way that people use the response scales (see Frederick and Loewenstein, 1999, page 308). As a result, studies may exaggerate the degree of "true" adaptation, and therefore may also exaggerate the degree to which people underestimate adaptation. In addition, studies of whether people predict adaptation often compare predictions for one group of people to actual experiences for another group of people, giving rise to selection problems. Despite these difficulties, however, we believe a number of studies do suggest that people often underappreciate their own powers of adaptation.

In a classic study, Brickman, Coates, and Janoff-Bulman (1978) interviewed people who had won lottery jackpot prizes within the last year (average winnings of $479,545) and a control group; they found virtually no difference in reported happiness of lottery non-winners and winners. They also found that lottery winners reported significantly less pleasure from each of six mundane daily

[7] Another subjective state with whose fluctuations most people should be very familiar is curiosity. Loewenstein, Prelec and Shatto (1996) showed that people who were in a non-curious state underpredicted the influence of curiosity on their own behavior. In one study, subjects attempted to answer 5 geography questions, and were given a choice between receiving the answers to the questions or getting an attractive candy bar. Subjects were first presented with a sample of 5 different geography questions and their answers. Half the subjects were asked to choose between the answers versus the candy bar *before* they attempted to answer the remaining 5 questions, while the other half attempted to answer the remaining 5 questions, and only *then* were given a choice between the answers or candy bar. Those who made the choice prior to attempting to answer the questions were significantly more likely to opt for the candy bar, as if they underestimated the force of the curiosity they would experience, than those who chose after they had attempted to answer the questions. (And a further study showed that subjects asked to predict beforehand their choice between answers and candy once they attempted to answer the questions underestimated their own subsequent likelihood of opting for the answers.)

activities. Although the paper has no data on non-winners' predictions of how they would feel if they won, the notion that lottery winners would be no happier than non-winners surely runs counter to the predictions of most people — including, presumably, those playing the lottery.

Loewenstein and Frederick (1997) compared the predictions by survey respondents of how changes in various environmental (e.g., decline in sport-fishing), social (e.g., increases in coffee shops) and personal (e.g., increases in body weight or income) factors would affect their well-being over the next decade to the reports of others about how actual changes in the last decade had affected their well-being. A clear pattern of underprediction of adaptation emerged in the data: People expected future changes to affect their well-being much more than others believed that matched changes in the past had affected their well-being.

Gilbert, Pinel, Wilson, Blumberg, and Wheatley (1997) report several instances of what they label "immune neglect" — the tendency to underestimate one's own powers of adaptation to unfavorable events. For instance, assistant professors at the College of Liberal Arts at the University of Texas, Austin who were asked to forecast their overall well-being at various points in time following their tenure decision — conditional on the decision being favorable and unfavorable — predicted that their feelings about the decision would fully adapt in about 5 years after the decision, while those who had been assistant professors during the previous ten years, who had received either positive or negative decisions, reported actually adapting much more rapidly than that. That is, academics exaggerated the longevity of the hedonic impact of tenure: They were relatively accurate in predicting the immediate hedonic impact of getting or being denied tenure, but they extrapolated these feelings further into the future than turned out to be warranted.

Sieff, Dawes, and Loewenstein (1999) asked people who came to a clinic for an HIV test to complete a mood inventory as they thought they would complete it approximately five weeks after obtaining the test result, conditional on whether the test indicated they were HIV-positive or HIV-negative. Five weeks after receiving the test result they completed the same mood inventory. A comparison of forecasts and subsequent reported feelings revealed that people overestimated both how good they would feel after receiving a favorable result, and how bad they would feel after receiving an unfavorable result.[8]

Although the large body of research on adaptation overwhelmingly suggests dramatic adaptation

[8] Because there was a low rate of HIV-positive results among the original subject pool, the researchers also recruited, using newspaper ads, a comparison group who had received HIV-positive test results in the last 4-10 weeks. Given the noncomparability of the groups, the results for the HIV-positive test results should be treated as tentative.

to diverse circumstances, it presents a major paradox: If major changes in life-circumstances such as winning the lottery or becoming paraplegic do not produce long-term changes in well-being, then why do people exert significant amounts of effort to bring about or avoid these changes?[9] One possible explanation is that measures of well-being are flawed, as discussed earlier. Another possibility is that people care a lot about the transition periods prior to adaptation — becoming paraplegic is typically a horrible experience, even if being paraplegic is not so bad in the long-term. But the fact that people seem to exert too much effort to obtain or avoid outcomes to which they will adapt may be further evidence of projection bias in this domain, because projection bias would cause people to overestimate the duration of the transition periods, and therefore to exaggerate their aggregate impact on utility.

3. The Model and Illustrations

In this section we formulate our general model, and present two extended examples designed to illustrate the model's mechanics and its consistency with the evidence in Section 2. We assume that a person's true intertemporal preferences are given by

$$U^t = \sum_{\tau=t}^{T} \delta^\tau u(\mathbf{c}_\tau, \mathbf{s}_\tau),$$

where $u(\mathbf{c}_\tau, \mathbf{s}_\tau)$ is her instantaneous utility in period τ, $\delta \leq 1$ is her discount factor, and T is her (possibly infinite) time horizon. The vector $\mathbf{c}_\tau$ is the person's period-τ consumption vector; $\mathbf{c}_\tau$ includes all period-τ behavior relevant for current or future instantaneous utilities. The vector $\mathbf{s}_\tau$ is the person's "state" in period τ. Depending on the particular application, a person's state could be a single attribute or a vector of attributes. An individual state could be determined by past consumption (e.g., a person's addiction level), or by exogenous factors that might be internal (e.g., depression) or environmental (e.g., peer pressure). Importantly, a person cannot affect her current state; indeed, our model essentially defines the state to be all factors that affect instantaneous utility besides current consumption.[10] For analytic and notational simplicity, we assume no uncertainty in this paper; this is highly artificial, but we suspect it does not affect our qualitative results.

[9] Oswald (1997) expresses this sentiment in a paper on happiness and economic performance: "How can it be...that money buys so little well-being and yet we see individuals around us constantly striving to make more of it?"

[10] While we assume throughout the paper that the utility function itself is not a function of the date, the model could be extended by treating calendar time as a state variable.

For any period t and initial state $\mathbf{s}_t$, a "fully rational" person would choose a path of consumption $(\mathbf{c}_t, \mathbf{c}_{t+1}, ..., \mathbf{c}_T)$ to maximize true intertemporal utility U^t, taking into account how the consumption path affects the evolution of future states. In our model, a person also attempts to maximize her intertemporal utility, but she may fail to do so because she mispredicts her future instantaneous utilities.

Formally, we assume that a person understands how her behavior affects the state variables, and which exogenous factors affect her future utility, so that for any consumption plan $(\mathbf{c}_t, \mathbf{c}_{t+1}, ...)$ the person can predict exactly the future state variables $(\mathbf{s}_{t+1}, \mathbf{s}_{t+2}, ...)$. It is the impact of future state variables on her future utility that the person mispredicts.[11] Let $\tilde{u}(\mathbf{c}_\tau, \mathbf{s}_\tau | \mathbf{s}_t)$ denote the prediction of a person currently in state $\mathbf{s}_t$ of what her instantaneous utility would be from consuming $\mathbf{c}_\tau$ in state $\mathbf{s}_\tau$ in period $\tau > t$.[12] For a fully rational person, predicted utility should equal true utility — that is, $\tilde{u}(\mathbf{c}_\tau, \mathbf{s}_\tau | \mathbf{s}_t) = u(\mathbf{c}_\tau, \mathbf{s}_\tau)$. But the evidence in Section 2 suggests that for many people predicted utility is not equal to true utility. Rather, people tend to exhibit *projection bias*, which roughly speaking means predicted utility $\tilde{u}(\mathbf{c}_\tau, \mathbf{s}_\tau | \mathbf{s}_t)$ lies "in between" true utility $u(\mathbf{c}_\tau, \mathbf{s}_\tau)$ and utility in the current state $u(\mathbf{c}_\tau, \mathbf{s}_t)$.

In defining a more precise notion of "in between," we incorporate two features. First, the person understands the qualitative nature of changes in her preferences, but underestimates the magnitude of these changes. Second, the more the person's future preferences differ from her current preferences, the further her prediction is from her true future utility. To formalize these features, we introduce some notation (much of which won't be used beyond Definition 1 below). Let $\mathbf{s} \in \mathbb{R}^L$, and let s_i denote its i^{th} element. We say $\mathbf{s}$ and $\mathbf{s}'$ differ only in element j if $s_j \neq s'_j$ and $s_i = s'_i$ for all $i \neq j$. Let $\mathbf{c} \in \mathbb{R}^K$, and let c_i denote its i^{th} element. For all $n \in \{1, 2, ...\}$, define $u^n_{a_1 a_2 ... a_n}(\mathbf{c}, \mathbf{s}) \equiv \frac{\partial^n u}{\partial c_{a_1} \partial c_{a_2} ... \partial c_{a_n}}(\mathbf{c}, \mathbf{s})$, where $a_i \in \{1, 2, ..., k\}$; these are all the n^{th}-order partial derivatives of the function $u(\mathbf{c}, \mathbf{s})$ with respect to the consumption variables. Define $\tilde{u}^n_{a_1 a_2 ... a_n}(\mathbf{c}, \mathbf{s})$ as the analogs of $u^n_{a_1 a_2 ... a_n}(\mathbf{c}, \mathbf{s})$ for predicted utility, and define $u^0(\mathbf{c}, \mathbf{s}) = u(\mathbf{c}, \mathbf{s})$ and $\tilde{u}^0(\mathbf{c}, \mathbf{s}) \equiv \tilde{u}(\mathbf{c}, \mathbf{s})$. We assume that $u(\mathbf{c}, \mathbf{s})$ and $\tilde{u}(\mathbf{c}, \mathbf{s})$ are fully differentiable, so that all these items are well-defined. Finally, for any two real numbers x and y, let the set $G(x, y) \equiv [\min\{x, y\}, \max\{x, y\}]$ denote the interval between x and y.

[11] Our model is essentially equivalent to an alternative formulation of projection bias wherein people underestimate the degree to which the states will change. While our modeling choice is irrelevant for the results, we feel our formulation better reflects the underlying psychology.

[12] We assume that predicted utility, like actual utility, does not depend on the dates involved.

Definition 1. Predicted utility exhibits *projection bias* if

(1) For all $\mathbf{c}$, $\mathbf{s}$ and $\mathbf{s}'$ such that $\mathbf{s}$ and $\mathbf{s}'$ differ only in element $j \in \{1, ..., L\}$, and for all $(n, a_1, a_2, ..., a_n)$, $\tilde{u}^n_{a_1 a_2 ... a_n}(\mathbf{c}, \mathbf{s}|\mathbf{s}') \in G\left(u^n_{a_1 a_2 ... a_n}(\mathbf{c}, \mathbf{s}), u^n_{a_1 a_2 ... a_n}(\mathbf{c}, \mathbf{s}')\right)$; and

(2) For all $\mathbf{c}$, $\mathbf{s}$, $\mathbf{s}'$, and $\mathbf{s}''$ such that $\mathbf{s}$, $\mathbf{s}'$, and $\mathbf{s}''$ differ in only element $j \in \{1, ..., L\}$, and for all $(n, a_1, a_2, ..., a_n)$, $u^n_{a_1 a_2 ... a_n}(\mathbf{c}, \mathbf{s}') \in G\left(u^n_{a_1 a_2 ... a_n}(\mathbf{c}, \mathbf{s}), u^n_{a_1 a_2 ... a_n}(\mathbf{c}, \mathbf{s}'')\right)$ implies $\tilde{u}^n_{a_1 a_2 ... a_n}(\mathbf{c}, \mathbf{s}|\mathbf{s}') \in G\left(u^n_{a_1 a_2 ... a_n}(\mathbf{c}, \mathbf{s}), \tilde{u}^n_{a_1 a_2 ... a_n}(\mathbf{c}, \mathbf{s}|\mathbf{s}'')\right)$.

Condition 1 says that in addition to the predicted absolute level of utility being in between the true value and the current value, the various marginal utilities and cross-partials of all orders are also in between the true values and the current values. This condition implies that the person understands the qualitative nature of changes in her preferences, but underestimates the magnitudes of these changes. Condition 2 says that the more the person's future preferences differ from her current preferences, the further her predictions are from her true future utility. Again, the condition says that not only is this true for the predicted absolute level of utility, but it is also true for the various marginal utilities and cross-partials.

There is a particularly simple and intuitive form of projection bias that we shall often assume in this paper:

Definition 2. Predicted utility exhibits *simple projection bias* if there exists $\alpha \in [0, 1]$ such that for all $\mathbf{c}$, $\mathbf{s}$, and $\mathbf{s}'$, $\tilde{u}(\mathbf{c}, \mathbf{s}|\mathbf{s}') = (1 - \alpha)\, u(\mathbf{c}, \mathbf{s}) + \alpha\, u(\mathbf{c}, \mathbf{s}')$.

If $\alpha = 0$, the person predicts her future instantaneous utility correctly, and therefore has no projection bias. If $\alpha > 0$, the person has projection bias satisfying Definition 1, where the bigger is α the stronger is the bias. If $\alpha = 1$, the person perceives that her preferences in the future will be identical to her current preferences, independent of any changes in her state.

While we shall assume simple projection bias in several illustrations, other applications make clear that it is far too restrictive for use as a general definition. Most problematically, simple projection bias requires that the magnitude of the bias be identical for different types of states. For example, it requires that a person who is currently not thirsty and currently unaddicted to cocaine be just as bad at predicting her preferences when she is thirsty as she is at predicting her preferences when addicted to cocaine. We suspect that most non-addicts are better at imagining thirst than co-

caine craving. Definition 1 allows for such realistic manifestations of projection bias, while still imposing the general features of "betweenness."

If a person has projection bias, and her state in period t is $\mathbf{s}_t$, then she perceives her period-t intertemporal utility to be

$$\widetilde{U}^t = \sum_{\tau=t}^{T} \delta^\tau \tilde{u}(\mathbf{c}_\tau, \mathbf{s}_\tau | \mathbf{s}_t).$$

We assume that for any period t and initial state $\mathbf{s}_t$, a person with projection bias chooses a path of consumption $(\mathbf{c}_t, \mathbf{c}_{t+1}, ..., \mathbf{c}_T)$ to maximize her perceived intertemporal utility $\tilde{U}^t$, taking into account how the consumption path affects the evolution of future states. That is, she behaves exactly as a fully rational person would except that $\tilde{U}^t \neq U^t$.

To help clarify the nature of our model, and to connect it to the evidence in Section 2, we now present two extended examples that apply simple projection bias. We begin with a formalization of how current hunger influences decisions about future food choice, inspired by and paralleling the experiment by Read and van Leeuwen (1998) discussed in Section 2.

Suppose that a person can either be hungry or satiated, where state $s_t = H$ represents hunger and $s_t = S$ represents satiation. Suppose that the person can eat either fruit or chips, where $c_t = F$ represents eating fruit and $c_t = C$ represents eating chips. We can then represent instantaneous utility by

$$u(c_\tau, s_\tau) = \begin{cases} u_{FH} & \text{if } c_\tau = F \text{ and } s_\tau = H \\ u_{FS} & \text{if } c_\tau = F \text{ and } s_\tau = S \\ u_{CH} & \text{if } c_\tau = C \text{ and } s_\tau = H \\ u_{CS} & \text{if } c_\tau = C \text{ and } s_\tau = S. \end{cases}$$

Suppose the person must choose in period 0 which snack to consume in period 1. In this situation, the person's optimal behavior — what she would do if she were fully rational — is independent of the current state. A person who anticipates being hungry next week should choose chips if and only if $u_{CH} - u_{FH} > 0$, regardless of whether she is currently hungry or satiated; and a person who anticipates being satiated next week should choose chips if and only if $u_{CS} - u_{FS} > 0$, regardless of whether she is currently hungry or satiated. Since the Read and van Leeuwen experiment suggests that people's relative preference for chips is increasing in their hunger — people who expected to be hungry were more likely to choose the unhealthy snack than people who expected to be satiated — we shall proceed under the assumption $u_{CH} - u_{FH} > u_{CS} - u_{FS}$.

A person with projection bias chooses chips if and only if $\tilde{u}(C, s_1|s_0\dot{)} > \tilde{u}(F, s_1|s_0)$. For actual

behavior, unlike optimal behavior, the person's choice depends not only on her anticipated future state, but also on her current state. Of course, projection bias matters only insofar as the person's future state will be different from her current state. Hence, if either $s_0 = s_1 = H$ or $s_0 = s_1 = S$, the person behaves optimally. Suppose instead that the person is currently hungry but expects to be satiated (i.e., $s_0 = H$ and $s_1 = S$), in which case she chooses chips if and only if $\tilde{u}(C, S \mid H) > \tilde{u}(F, S \mid H)$. With a simple projection bias, $\tilde{u}(C, S \mid H) = (1-\alpha)u_{CS} + \alpha u_{CH}$ and $\tilde{u}(F, S \mid H) = (1-\alpha)u_{FS} + \alpha u_{FH}$. Hence, the person chooses chips if and only if

$$u_{CS} - u_{FS} > -\alpha\left[(u_{CH} - u_{FH}) - (u_{CS} - u_{FS})\right]$$

or

$$u_{CH} - u_{FH} > (1-\alpha)\left[(u_{CH} - u_{FH}) - (u_{CS} - u_{FS})\right].$$

Under our assumption that $u_{CH} - u_{FH} > u_{CS} - u_{FS}$, the first inequality above implies that the person is too likely to choose chips; because she projects her current hungry preferences onto her future satiated preferences, she over-estimates how much she'll want chips. It is also worth noting that the second inequality above implies that the person is less likely to choose chips than she would be if she anticipated being hungry next week. In other words, a person with projection bias does not ignore that her preferences are state-dependent — she recognizes that being satiated will diminish her preference for chips relative to fruit — she merely underappreciates the magnitude of the change.

Analogous arguments hold for the case where the person is currently satiated but expects to be hungry (i.e., $s_0 = S$ and $s_1 = H$); the person recognizes that being hungry increases her relative preference for chips over fruit, but she underappreciates the magnitude of this effect and is therefore too unlikely to choose chips. In sum, projection bias, combined with the assumption that hunger increases a person's relative preference for chips, yields conclusions that correspond to both the behavior and the intuitions from Read and van Leeuwen (1998).

Our second example explores the relationship between projection bias and the endowment-effect. Suppose a person can either own a mug, in which case her consumption is $c_t = 1$, or not own a mug, in which case her consumption is $c_t = 0$. In addition, the person's state can be either the reference point of owning a mug, $s_t = 1$, or the reference point of not owning a mug, $s_t = 0$. For simplicity, we shall assume that a person's reference point depends on only whether she enters period t owning a mug: If she enters period t owning a mug then $s_t = 1$, and if she

enters period t not owning a mug then $s_t = 0$.[13] While we assume the state in period t depends on whether the person *enters* period t owning a mug, we assume consumption in period t depends on whether the person *exits* period t owning a mug.

We assume the following instantaneous utility function:

$$u(c_t, s_t) = \begin{cases} \mu + G \cdot (1 - s_t) & \text{if } c_t = 1 \\ -L \cdot s_t & \text{if } c_t = 0. \end{cases}$$

This formulation assumes that the person gets intrinsic value μ from owning the mug in any given period. But there is also a reference-dependent component to her utility function whenever her mug status changes: If she has a mug ($c_t = 1$) when her reference level is not owning one ($s_t = 0$), then she experiences a feeling of gain, G. If she does not own a mug ($c_t = 0$) when her reference level is ownership ($s_t = 1$), then she experiences a feeling of loss, L. Loss aversion — the tendency to dislike losses more than liking gains — implies $L > G$.

Endowment-effect experiments typically compare two situations: (1) A person not endowed with a mug has the option to buy one, and (2) A person endowed with a mug has the option to sell it. The endowment effect is reflected in the finding that the selling price in situation (2) is significantly larger than the buying prices in situation (1).[14] To formalize the typical experiment in terms of our model, we suppose that buying and selling decisions occur in period 1, after which there is a second period during which the mug, if possessed, can yield benefits.[15] If the person's decision is to possess the mug, then $c_1 = c_2 = 1$; if the person's decision is to not possess the mug, then $c_1 = c_2 = 0$. Both buyers and sellers must choose between these two consumption flows, but buyers enter period 1 with $s_1 = 0$ and sellers enter period 1 with $s_1 = 1$. We assume no discounting and that the buying price P_B or selling price P_S enters as a separable and linear part of the intertemporal utility function — i.e., $U^1 = u(c_1, s_1) + u(c_2, s_2) - P$.

[13] A more general formulation is $s_t = (1 - \gamma)s_{t-1} + \gamma c_{t-1}$ for some $\gamma \in (0, 1]$, where γ captures the speed of adaptation. This formulation of changing reference points used in Ryder and Heal (1973), Bowman, Minehart, and Rabin (1999), and Strahilevitz and Loewenstein (1998). Our example here is the case $\gamma = 1$, but the qualitative results hold for any $\gamma \in (0, 1]$.

[14] Many of the experiments actually compare selling prices to "choosing" prices — the amount of money a person unendowed with a mug reveals she would accept in lieu of a mug. This experimental procedure in fact corresponds more closely to our formal model. Moreover, because the money side of the transaction is identical for choosing and selling, it both allows one to ignore the legitimate concern over the role of loss aversion over money and the rather silly (but sometimes raised) concern that these experimental results might have something to do with wealth effects.

[15] Our qualitative results crucially depend on there being at least one additional period in which the mug yields benefits, since it is the period-2 benefits (or forgone benefits) which the person mispredicts. But whether there is one additional period or many additional periods is not so important for our results.

Let P_B^* and P_S^* represent the optimal buying and selling prices, and let P_B^A and P_S^A represent the actual buying and selling prices for a person who has simple projection bias. A person who enters period 1 without a mug and has the option to buy one should buy the mug if and only if $P \leq 2\mu + G \equiv P_B^*$, because her true intertemporal utility is:

$$\begin{array}{lccccccccccccc} \text{If Buy:} & u(1,0) & + & u(1,1) & - & P & = & \mu + G & + & \mu & - & P & = & 2\mu + G - P \\ \text{If Not:} & u(0,0) & + & u(0,0) & & & = & 0 & + & 0 & & & = & 0. \end{array}$$

If the person has simple projection bias of degree α, she actually buys the mug if and only if $P \leq 2\mu + G + \alpha G \equiv P_B^A$, because she perceives her intertemporal utility to be:

$$\begin{array}{lccccccccccccc} \text{If Buy:} & \tilde{u}(1,0|0) & + & \tilde{u}(1,1|0) & - & P & = & \mu + G & + & \mu + \alpha G & - & P & = & 2\mu + G + \alpha G - P \\ \text{If Not:} & \tilde{u}(0,0|0) & + & \tilde{u}(0,0|0) & & & = & 0 & + & 0 & & & = & 0. \end{array}$$

A person who enters period 1 with a mug and has the option to sell it should sell the mug if and only if $P \geq 2\mu + L \equiv P_S^*$, because her true intertemporal utility is:

$$\begin{array}{lccccccccccccc} \text{If Keep:} & u(1,1) & + & u(1,1) & & & = & \mu & + & \mu & & & = & 2\mu \\ \text{If Sell:} & u(0,1) & + & u(0,0) & + & P & = & -L & + & 0 & + & P & = & -L + P. \end{array}$$

If the person has simple projection bias of degree α, she actually sells the mug if and only if $P \geq 2\mu + L + \alpha L \equiv P_S^A$, because she perceives her intertemporal utility to be:

$$\begin{array}{lccccccccccccc} \text{If Keep:} & \tilde{u}(1,1|1) & + & \tilde{u}(1,1|1) & & & = & \mu & + & \mu & & & = & 2\mu \\ \text{If Sell:} & \tilde{u}(0,1|1) & + & \tilde{u}(0,0|1) & + & P & = & -L & + & -\alpha L & + & P & = & -L - \alpha L + P. \end{array}$$

The formulas for P_B^*, P_S^*, P_B^A, and P_S^A derived above reveal several interesting conclusions. First, $P_B^A > P_B^*$ *and* $P_S^A > P_S^*$. Projection bias leads both a potential buyer and a potential seller to over-value the mug. A potential buyer overestimates the pleasure she'll get from an object because she believes that she will continue to feel the pleasures of gain further into the future than she actually will. A potential seller overestimates the pain she'll feel from parting with an object because she believes that she will continue to feel the pain of loss further into the future than she actually will. These results are simple examples of what is probably the main class of phenomena caused by projection bias: People have a tendency to over-consume reference-dependent goods.

This over-consumption stems from an underappreciation of the temporary nature of both the pleasure they receive from gains relative to their reference point *and* the pain they feel from losses relative to their reference point. Indeed, this theme underlies many of the examples in Section 2 and our models in Sections 4, 5, and 6.

Returning to an analysis of the endowment effect itself, next notice that $P_S^* > P_B^*$ *and* $P_S^A > P_B^A$. Whether the person is fully rational or suffers from projection bias, she demands more to give up an object than she is willing to pay to get it — that is, she exhibits the endowment effect. Given our assumption that people exhibit loss aversion, the endowment effect is fully rational (Tversky and Kahneman, 1991). Because a person with projection bias understands the qualitative nature of her preferences, she exhibits the same qualitative behavior.

But projection bias increases the *magnitude* of the endowment effect — that is, $P_S^A - P_B^A > P_S^* - P_B^*$. While the person exaggerates the duration of both the sensation of loss after parting with the mug and the sensation of gain after obtaining the mug, because losses loom larger than gains, the exaggeration of losses has a greater impact. Hence, the endowment effect itself may in part be an error caused by projection bias. In other words, although the endowment effect, and loss aversion more generally, may be a manifestation of real preferences, people's behavior may be an exaggerated response to these real preferences. Insofar as people behave as if the unpleasant sensations of loss will persist for a long time, they are making an error.[16]

While the results above establish both the welfare implications of projection bias and the magnitude of its effects in buying and selling decisions, these results don't directly yield qualitative behavioral predictions about how projection bias differs in this context from rational behavior. The experiment by Loewenstein and Adler (1995) described in Section 2 identified one such difference:

[16] Indeed, Kahneman (1991, p. 143) and Tversky and Kahneman (1991) argue that the endowment effect is a bias because people's actual pain when losing an object is not commensurate with their unwillingness to part with that object. Evidence from Strahilevitz and Loewenstein (1998) also suggests this interpretation. Some subjects were endowed with mugs for several minutes, but then (under the pretext of randomization) forced to pair up with a subject who did not receive a mug and flip a coin to determine who would get to keep the mug. Shortly after this exchange of mugs, selling prices were elicited from subjects with mugs, and choice prices were elicited from those without mugs, creating prices for four groups of subjects:

(1). Began and ended with a mug: Selling Price = $5.26
(2). Began with a mug and lost it: Choice Price = $3.36
(3). Began without a mug and got one: Selling Price = $4.32
(4). Began and ended without a mug: Choice Price = $2.75

The average choice price of Group 2 subjects was higher than Group 4 subjects, showing that some of the sense of loss persists after departing with an object. But the selling price of Group 1 was higher than the choice price of Group 2, showing that subjects adapt to the loss, at least to some degree, almost immediately following the loss of the mug. The speed of adaptation suggested by this experiment seems inconsistent with the magnitudes of the endowment effect usually observed.

When subjects not endowed with mugs were asked to predict their selling prices once they owned the mug, they underestimated their selling price. Our model captures this phenomenon as well.

Suppose that a person will be given a mug in period 0 that would yield benefits for periods 1 and 2, but has the opportunity to sell it in period 1. Before the person is given the mug, however, she is asked to predict what her selling price will be in period 1. That is, the person formulates in state $s_0 = 0$ a prediction about her selling price when $s_1 = 1$.

Let $\hat{P}_S^*$ represent the predicted selling price for a person who is fully rational, and let $\hat{P}_S^A$ represent the predicted selling price for a person who has projection bias. Because a person without projection bias perceives in period 0 her true period-1 intertemporal utility, clearly $\hat{P}_S^* = P_S^*$. In contrast, if a person has simple projection bias α, she predicts that she will sell the mug if and only if $P \geq 2\mu + (1-\alpha)L + 2\alpha G \equiv \hat{P}_S^A$, because she perceives her intertemporal utility to be:

$$\begin{array}{lccccccccccc} \text{If Keep:} & \tilde{u}(1,1|0) & + & \tilde{u}(1,1|0) & & & = & \mu + \alpha G & + & \mu + \alpha G & & \\ \text{If Sell:} & \tilde{u}(0,1|0) & + & \tilde{u}(0,0|0) & + & P & = & (1-\alpha)(-L) & + & 0 & + & P. \end{array}$$

It is straightforward to see that $\hat{P}_S^A < P_S^A$; a person with projection bias underestimates her selling price. This result reflects the net effect of two countervailing biases. First, projection bias causes the individual to overestimate the persistence of the feeling of gain from keeping the mug. By itself, this error would cause her to *over*estimate her selling price. But this error is eclipsed by a second, and bigger error: She underestimates the amount of loss she will feel in giving up the mug. Hence, our model is consistent with the findings of Loewenstein and Adler (1995) that people underestimate the degree to which they will become attached to objects.

The two examples above illustrate the crucial role that "states" play in our analysis; however, our formal model does not put any restrictions on how states are defined. This is an important limitation of the model, as the implications of projection bias can differ depending on which of two specifications of preferences are employed, even when rational utility theory deems these specifications to be equivalent. For instance, suppose a person's desire to smoke is influenced by both peer pressure and the amount of smoke in the room. Our model does not *a priori* specify whether the two factors should be included as individual states, or whether the two factors should be combined into a single state, or even whether there should be two states, each of which depends on both factors (e.g., the "sum" and the "difference").

In conventional economic theory the choice of what to designate as a state in the definition of the utility function is merely a semantic point, but because projection bias is pinned to particular

states, in our model such designations matter a great deal. The only "psychology-free" way to define states in a way that fully pins down projection bias is to allow only one state; however, this is too restrictive. Returning to the example above, if we want to allow a person to underappreciate the effects of peer pressure to a different degree than she underappreciates the effects of the amount of smoke in the room, we must define each factor to be an individual state. Hence, the starting point for each of our applications is to propose a domain in which a person might exhibit projection bias, and to specify assumptions about what are the natural "states" to consider. The psychology of the particular application, therefore, determines the relevant states in what we hope and believe is a non-arbitrary and non-post hoc way. Again returning to our example above, choosing to label environmental smoke and peer pressure as the two states rather than the sum and the difference between the two factors is an assumption that we suspect readers can intuit as being appropriate.

The endowment-effect example also illustrates how projection bias can lead to *dynamic inconsistency* — planning to behave a certain way in the future, but later, in the absence of new information, revising this plan. Formally, a person is dynamically inconsistent if $(\mathbf{c}_t, \mathbf{c}_{t+1}, ..., \mathbf{c}_T)$ maximizes $\tilde{U}^t$ and yet after following $(\mathbf{c}_t, \mathbf{c}_{t+1}, ..., \mathbf{c}_{\tau-1})$, $(\mathbf{c}_\tau, \mathbf{c}_{\tau+1}, ..., \mathbf{c}_T)$ does not maximize $\tilde{U}^\tau$. Dynamic inconsistency can arise because perceived preferences may be time-inconsistent: Even though the person's true intertemporal preferences are time-consistent, projection bias implies that the person's perceived period-τ intertemporal utility may turn out to be different from that predicted in period t.

In models of self-control problems of the sort we discuss in the conclusion, a person's *true* preferences are time-*in*consistent, and it therefore makes sense to assume that the person is aware of, but disapproves of, her own future preferences. In our model, in contrast, the time inconsistency in perceived preferences derives solely from misprediction of future utilities, so it would make little sense to assume that the person is completely aware of this inconsistency. We assume throughout the paper that the person is unaware of the time inconsistency — that is, at all times the person perceives her preferences to be time-consistent.[17]

Of course, a person need not be dynamically inconsistent even when she does not anticipate

[17] Given the logic of our model, it is inherent that a person is unaware of her *current* misprediction; but the person could have an awareness of her *future* propensity to mispredict. A person could, for instance, be aware of her general propensity to over-shop when hungry, while still committing the error on a case-by-case basis. The coexistence of day-to-day mispredictions with a "meta-awareness" of these mispredictions is similar to the discussion in O'Donoghue and Rabin (1999*b*) of how people can simultaneously be aware of their general tendency to procrastinate and yet still procrastinate on a case-by-case basis. A model of "sophisticated projection bias" could plausibly better describe behavior in some circumstances, but we choose our current formulation as a simpler starting point.

changes in her preferences. Proposition 3.1, in fact, provides sufficient conditions for dynamic consistency despite projection bias:

Proposition 3.1. A person is dynamically consistent if for all $\mathbf{s}_t$, $\mathbf{s}_\tau$, $\mathbf{c}_\tau$, and $\mathbf{c}'_\tau$, $\tilde{u}(\mathbf{c}_\tau, \mathbf{s}_\tau|\mathbf{s}_t) - \tilde{u}(\mathbf{c}'_\tau, \mathbf{s}_\tau|\mathbf{s}_t) = u(\mathbf{c}_\tau, \mathbf{s}_\tau) - u(\mathbf{c}'_\tau, \mathbf{s}_\tau)$, and for all $\mathbf{c}_\tau$, $\mathbf{s}_\tau$, $\mathbf{s}'_\tau$, $\mathbf{s}_t$, and $\mathbf{s}'_t$, $\tilde{u}(\mathbf{c}_\tau, \mathbf{s}_\tau|\mathbf{s}_t) - \tilde{u}(\mathbf{c}_\tau, \mathbf{s}'_\tau|\mathbf{s}_t) = \tilde{u}(\mathbf{c}_\tau, \mathbf{s}_\tau|\mathbf{s}'_t) - \tilde{u}(\mathbf{c}_\tau, \mathbf{s}'_\tau|\mathbf{s}'_t)$.

Proposition 3.1 says that as long as projection bias does not cause a person either to misperceive the relative merits of any two consumption bundles, or to misperceive the relative impact on preferences of any two future states, then the person is dynamically consistent.[18] Though certain themes will be repeated throughout the remainder of the paper, beyond Proposition 3.1 we reach no general formal conclusions about the implications of projection bias. Instead, we turn now to three extended applications of our model.[19]

4. Over-Consumption

In this section we explore how projection bias combined with reference-dependent utility might lead to excessive pursuit of a high material standard of living.[20] We consider a simple model of consumption-leisure decisions. In period τ, a person chooses consumption, c_τ, and leisure, l_τ; that is, the vector of choice variables is (c_τ, l_τ). In principle, both consumption and leisure could be reference-dependent. But since realistically, we believe, consumption is more reference-dependent than leisure, we shall assume for simplicity that leisure is not reference-dependent.[21] Let r^c_τ denote the reference level for consumption in period τ — that is, the state variable is $s_\tau = r^c_\tau$. We assume the instantaneous utility function takes the following form:

[18] The second condition necessarily holds for a simple projection bias.

[19] Each of our applications employs a simple model that abstracts away from some of the richer features of the respective environments. The introduction of richer features would have similar qualitative effects in our model as in a rational-choice model, although our model allows that people might over- or under-react to such features relative to what is optimal.

[20] Many models over the years posit habit persistence in consumption of the sort we consider here. See, e.g., Duesenberry (1952), Ryder and Heal (1973), and Bowman, Minehart, and Rabin (1999).

[21] Our main points in this section depend on merely consumption being more reference-dependent than leisure, and not on leisure being reference-independent. In general, the degree of reference dependence for a good depends on both how much the person cares about deviations from her reference level and how quickly the person's reference level adjusts. Frank (1999, Chapter 6) provides diverse support for the assumption that leisure is less reference-dependent than consumption.

$$u\left((c_\tau, l_\tau), s_\tau\right) = v(c_\tau) + w(l_\tau) + R\left(c_\tau - r^c_\tau\right).$$

The functions $v(c_\tau)$ and $w(l_\tau)$ are the direct utilities from consumption and leisure, where we assume $v', w' > 0$ and $v'', w'' < 0$.[22] $R\left(c_\tau - r^c_\tau\right)$ is the reference-dependent utility from consumption, which captures how a person cares about gains and losses relative to the reference point. We assume that R is strictly increasing and weakly concave: $R' > 0$ and $R'' \leq 0$.[23]

We suppose that the reference level for consumption evolves according to

$$r^c_\tau = (1-\gamma) r^c_{\tau-1} + \gamma c_{\tau-1}.$$

The parameter γ represents how quickly the person adapts to changes in consumption. If $\gamma = 1$, then each period's reference level is equal to the prior period's choice, meaning that the person adapts very quickly to changes. If $\gamma = 0$, in contrast, then the person's reference level never changes, and consumption is effectively reference-independent.

For simplicity, we consider a two-period model where there is no discounting, and the person can borrow or save at 0% interest.[24] We normalize units of time and income such that in each period the person is endowed with one unit of time to divide between labor and leisure, and each unit of time allocated to labor yields one unit of consumption purchasing power. Hence, the person's budget constraint is[25]

$$[c_1 + c_2] + [l_1 + l_2] \leq 2.$$

Finally, we assume that the initial reference point is $r^c_1 = 0$. Let $(c^*_1, c^*_2, l^*_1, l^*_2)$ denote optimal

[22] Our assumption that the utility from consumption and leisure are separable is of course potentially restrictive, since the utility of spending on consumption is likely to depend on how much leisure time a person has to enjoy that consumption. We assume separability for convenience, but do not know how relaxing it would change results.

[23] Based on behavioral evidence, Kahneman and Tversky (1979) propose as part of their prospect theory that $R''(x) < 0$ for $x > 0$ while $R''(x) > 0$ for $x < 0$, which implies people have a decreasing sensitivity to both losses and gains. We focus on the case where only gains are relevant — because we assume the initial reference point is small and there is no uncertainty — so it does not matter that we assume global concavity. Bowman, Minehart, and Rabin (1999) consider a rational-choice model of reference-dependent consumption where income is uncertain and therefore people might experience losses. They assume diminishing sensitivity to losses, and show that it implies people are prone to resist an immediate reduction in consumption in response to unexpected decreases in income. Since their results are strongest when a person *believes* that future reference levels of consumption will adjust slowly, projection bias would likely enhance their results.

[24] Our main results hold even if the person cannot borrow or save.

[25] Technically, the budget constraint also requires $c_1, c_2, l_1, l_2 \geq 0$ and $l_1, l_2 \leq 1$. Our analysis here assumes an interior solution.

behavior, which is derived from the following choice problem:

$$\begin{aligned} \max_{(c_1,c_2,l_1,l_2)} U^1 &= [v(c_1) + w(l_1) + R(c_1)] \\ &+ [v(c_2) + w(l_2) + R(c_2 - \gamma c_1)] \end{aligned}$$

such that $c_1 + c_2 + l_1 + l_2 \leq 2.$

Optimal behavior in this model exhibits a simple qualitative feature:

Lemma 4.1. $c_1^* < c_2^*$ and $l_1^* = l_2^*$.

Lemma 4.1 states that the person should increase her consumption over time, while working the same amount each period. Since there is no discounting, if consumption were not reference-dependent, then the person would consume the same amount each period. But since the person will become accustomed to her consumption level, the more she consumes now, the less she will enjoy future consumption. As a result, it is optimal to have a consumption stream that increases over time. Since leisure is not reference-dependent, it should be the same each period.

Now consider how a person with simple projection bias behaves. Let $(c_1^A, c_2^A, l_1^A, l_2^A)$ denote the person's actual behavior, and let $(\hat{c}_2^A, \hat{l}_2^A)$ denote the person's plan in period 1 for period-2 behavior. In period 1, the person chooses c_1^A, l_1^A, $\hat{c}_2^A$, and $\hat{l}_2^A$ to solve the following problem:

$$\begin{aligned} \max_{(c_1,c_2,l_1,l_2)} \tilde{U}^1 &= [v(c_1) + w(l_1) + R(c_1)] \\ &+ (1-\alpha)[v(c_2) + w(l_2) + R(c_2 - \gamma c_1)] \\ &+ \alpha [v(c_2) + w(l_2) + R(c_2)] \end{aligned}$$

such that $c_1 + c_2 + l_1 + l_2 \leq 2.$

Whereas in period 1 the person chooses her *planned* period-2 behavior, she chooses her *actual* period-2 behavior in period 2, after c_1^A and l_1^A have been chosen and carried out. Hence, c_2^A and l_2^A solve the following problem:

$$\max_{(c_2,l_2)} \tilde{U}^2 = \left[v(c_2) + w(l_2) + R\left(c_2 - \left[\gamma c_1^A\right]\right)\right]$$

such that $c_2 + l_2 \leq 2 - \left[c_1^A + l_1^A\right].$

How do actual and planned behavior compare to optimal behavior? Because the person understands the qualitative nature of her preferences, her planned behavior has the same qualitative features as her optimal behavior — that is, the person plans to increase her consumption over time, while planning to work the same amount each period.

Lemma 4.2. $c_1^A < \hat{c}_2^A$ and $l_1^A = \hat{l}_2^A$.

Although the person plans to increase her consumption over time, as is optimal, she may do so in a suboptimal way. To distinguish the different errors that the person might make, we separate the cases where the reference function $R(\cdot)$ is linear from those where it is strictly concave.

Proposition 4.1. Suppose the reference function $R(\cdot)$ is linear. Then

(1) $c_1^A > c_1^*$, $\hat{c}_2^A < c_2^*$, and $l_1^A = \hat{l}_2^A < l_1^* = l_2^*$, and

(2) $c_2^A = \hat{c}_2^A$ and $l_2^A = \hat{l}_2^A$.

Part 1 of Proposition 4.1 establishes that the person over-consumes in period 1, and over-works in both periods to pay for this over-consumption. Intuitively, projection bias leads the person to underappreciate how much she will become accustomed to high consumption, and therefore to underestimate how much increasing her current consumption will reduce her future well-being. As a result, the person over-indulges in the consumption activity.[26] Part 2 states that when the reference function is linear, the person behaves in period 2 exactly as she planned to behave — that is, she is dynamically consistent. When the reference function is linear, the reference point affects only the person's absolute utility, and not her marginal utility of consumption. Hence, the utility function satisfies the conditions for Proposition 3.1.

Proposition 4.1 illustrates how projection bias leads a person to over-consume while young, because she underappreciates how much she will become accustomed to her consumption level. By assuming the reference function $R(\cdot)$ is linear, however, it does not capture a second effect of

[26] The extension of this logic to more than two periods says that in all periods except the last, actual consumption is larger than optimal *conditional on past consumption*. That is, in a T-period model, actual consumption in period $t < T$ may or may not be larger than first-best optimal consumption in period t, but it will be larger than is optimal given $(c_1^A, ..., c_{t-1}^A, l_1^A, ..., l_{t-1}^A)$.

projection bias in this environment: dynamic inconsistency. Proposition 4.2 shows what happens when R is strictly concave, so that a person exhibits decreasing sensitivity to gains relative to her reference point:

Proposition 4.2. Suppose the reference function $R(\cdot)$ is strictly concave. Then

(1) $c_1^A > c_1^*$; and

(2) $c_2^A > \hat{c}_2^A$ and $l_2^A < \hat{l}_2^A$.

Part 1 of Proposition 4.2 establishes that, just as with a linear reference function and for the same reason, the person over-consumes in period 1 when the reference function is strictly concave. Unlike when the reference function is linear, however, over-consumption in period 1 need not be associated with working more than is optimal; depending on functional forms, it could be that $l_1^A = \hat{l}_2^A < l_1^* = l_2^*$, as in the linear case, but it could also be that $l_1^A = \hat{l}_2^A \geq l_1^* = l_2^*$. The person might under-work when R is strictly concave because in addition to *over* estimating the marginal utility of current consumption, the person also *under* estimates the marginal utility of future consumption, and therefore may pay for excess current consumption by planning to forego future consumption rather than current and future leisure.[27]

Part 2 establishes that the concavity of R introduces a dynamic-inconsistency effect that is absent when R is linear. When period 2 arrives, the person discovers the full effects of having a larger reference point. When R is linear, this unfortunate discovery does not affect the person's desire to consume. But when R is concave, the person discovers that the marginal utility of consumption has increased. Hence, she decides to "ramp up" her consumption, which in turn requires that she work more than she had planned. Indeed, if there were more periods, this process would continually repeat itself: Each period the person will be surprised at how accustomed she has become to her consumption level, and as a result she will revise upward her planned consumption and therefore revise downward her planned leisure.

Hence, our analysis shows that under the plausible assumption that consumption is more highly reference-dependent than leisure, people with projection bias tend to over-consume, while working more than they should and than they expected to support this consumption. This analysis indeed

[27] E.g., if $\alpha = 1$ and $\gamma = 1$, one can show that $l_1^A = \hat{l}_2^A < l_1^* = l_2^*$ for $v(x) = R(x) = 2x^{1/2}$ whereas $l_1^A = \hat{l}_2^A > l_1^* = l_2^*$ for $v(x) = R(x) = -x^{-1}$.

parallels the arguments of many previous researchers, such as Scitovsky (1976) and Frank (1999), who have argued that people spend too much time and energy generating wealth and too little time on leisure activities, and that people enjoy increases in their standard of living less than they think they will.

5. Addiction

An important domain in which state-dependent preferences play a role is the consumption of harmful addictive products. As with all products, the temptation to consume addictive products can vary over time due to factors such as age, environmental cues, traumatic events, peer pressure, and so forth. But the very essence of harmful addictive products is that a person's preferences depend on her own past behavior. In this section, we explore the implications of projection bias for the consumption of harmful addictive products.[28]

We consider a simple model of addiction with two periods, $t \in \{1, 2\}$, and in each period a person can either "hit" (choose $c_t = 1$) or "refrain" (choose $c_t = 0$). Our results generalize to a multi-period model, but the arguments are more complicated. We also simplify the model by assuming that the addictive product is free, so that the decision to hit is based solely on whether the current benefits of hitting outweigh the perceived long-run costs.

There are two ways in which past consumption of a harmful addictive product can affect current preferences. First, addictive products involve *negative internalities*: Past consumption decreases a person's current level of utility.[29] Negative internalities capture the adverse effects of harmful addictive products on a person's day-to-day life, including health problems, decreased job perfor-

[28] Based on earlier reasearch on habit formation (e.g., Pollak (1970)), Becker and Murphy (1988) formulate a model of instantaneous utilities for addictive products, and characterize steady-state levels of addiction at which a person would rationally not be motivated to become unaddicted. Using a similar model of instantaneous utilities, Orphanides and Zervos (1995) explore whether a person might rationally choose to become addicted when she is uncertain about the effects of addiction. Although Orphanides and Zervos allow only fully rational agents in their formal model, they in fact discuss the possibility that there may be a systematic propensity for people to underestimate the addictiveness of products. This posited departure from rational-choice theory is similar to our formal model where people underappreciate addictiveness. In this section, we study a variant of the model developed by O'Donoghue and Rabin (1999*c*, 1999*d*), who modify and simplify the Becker-Murphy model of instantaneous utilities to explore the role of self-control problems for addictive behavior. The model is also closely related to Loewenstein's (1999) "visceral account of addiction" which incorporates a specific kind of projection bias that he labels a "cold/hot intrapersonal empathy gap." His paper also presents diverse evidence from the literature on drug addiction supporting the assumptions we make in this section.

[29] We borrow the term "internalities" from Herrnstein *et al.* (1993), who define an internality to be a "within-person externality".

mance, strained personal relationships, and so forth. "Tolerance" for an addictive product — the need to consume increasing amounts to obtain the same effect — also falls under the rubric of negative internalities. Second, addictive products involve *habit formation*: Past consumption increases a person's current marginal utility for the product, which means that her "craving" for the addictive product will be stronger the more she has consumed in the past. The combination of negative internalities and habit formation creates the trap of addiction: As a person consumes more and more of an addictive product, she receives less and less pleasure from this consumption, and yet she finds it more and more difficult to refrain.

To incorporate the effects of negative internalities and habit formation, we suppose that all effects of past consumption can be incorporated into a single statistic, k_t, which we shall refer to as a person's *addiction level* in period t. In our two-period model, we assume that $k_1 \geq 0$ is exogenous (i.e., k_1 reflects consumption prior to period 1), and that $k_2 = \gamma k_1 + c_1$ for some $\gamma \in [0, 1)$.[30] In addition to past consumption, instantaneous utility in period t can depend on exogenous factors such as fluctuations in environmental cues, peer pressure, etc. We let x_1 and x_2 denote the exogenously determined component of the instantaneous utility from hitting in periods 1 and 2, respectively. To fit within the framework of this paper, we shall treat x_1 and x_2 as deterministic, although the results easily generalize to the case where they are stochastic.

The person's instantaneous utility function depends on current consumption c_t and the current vector of state variables $\mathbf{s}_t = (k_t, x_t)$. We use the following simple formulation:

$$u(c_t, \mathbf{s}_t) = \begin{cases} x_t - \rho k_t & \text{if } c_t = 1 \\ -(\rho + \sigma) k_t & \text{if } c_t = 0. \end{cases}$$

In this formulation, the parameter $\rho > 0$ represents the negative internalities. Whether or not a person chooses to hit in period t, her instantaneous utility is reduced by amount ρk_t, reflecting that past consumption has reduced her current well-being. The parameter $\sigma > 0$ represents habit formation. A person's desire to consume in period t is $u(1, \mathbf{s}_t) - u(0, \mathbf{s}_t) = x_t + \sigma k_t$. Hence, the more a person has consumed in the past, the bigger is k_t, and therefore the larger is the current desire to consume. The current desire to consume also depends on exogenous factors as captured

[30] We assume $k_1 \leq 1/(1 - \gamma)$, which guarantees that hitting in period 1 makes the person more addicted in period 2. The more general form used by Becker and Murphy (1988) and O'Donoghue and Rabin (1999*c*, 1999*d*) is $k_t = \gamma k_{t-1} + c_{t-1}$.

by x_t.[31]

Of course, a person hits in period t only if the current desire to consume is larger than the perceived future cost of this consumption, which depends on projection bias. In this model, there are three dimensions along which a person's preferences are state-dependent — negative internalities, habit formation, and exogenous factors — and projection bias could operate differently along these three dimensions. Indeed, the different implications of projection bias along these different dimensions is the focus of our analysis. We use the following formulation of projection bias:

$$\tilde{u}\left(c_2,(k_2,x_2)|(k_1,x_1)\right)=\begin{cases}\left[(1-\alpha_x)x_2+\alpha_x x_1\right] & - & \rho\left[(1-\alpha_\rho)k_2+\alpha_\rho k_1\right] & \text{if } c_2=1\\ -\ \rho\left[(1-\alpha_\rho)k_2+\alpha_\rho k_1\right] & - & \sigma\left[(1-\alpha_\sigma)k_2+\alpha_\sigma k_1\right] & \text{if } c_2=0.\end{cases}$$

In this formulation, α_ρ represents how much a person underappreciates the effects of negative internalities, α_σ represents how much a person underappreciates the effects of habit formation, and α_x represents how much a person underappreciates the effects of exogenous factors.[32]

Given exogenous parameters $k_1 \geq 0$, $\sigma, \rho > 0$, $\gamma \in [0,1]$, $x_1, x_2 \in \mathbb{R}$, and $\alpha_\rho, \alpha_\sigma, \alpha_x \in [0,1]$, we let c_1^* and c_2^* denote the person's optimal consumption in periods 1 and 2, and we let c_1^A and c_2^A denote the person's actual consumption in periods 1 and 2. We shall also be interested in how a person would behave in period 2 as a function of her period-1 behavior. Let $C_2^*(c_1)$ be optimal behavior in period 2 conditional on period-1 behavior, let $C_2^A(c_1)$ be actual behavior in period 2 conditional on period-1 behavior, and let $\hat{C}_2^A(c_1)$ be perceived optimal behavior in period 2 conditional on period-1 behavior. Clearly, $c_2^* = C_2^*(c_1^*)$ and $c_2^A = C_2^A(c_1^A)$; but it need not be the case that $c_2^A = \hat{C}_2^A(c_1^A)$ because the person might be dynamically inconsistent.[33]

Because period 2 is the last period, the person hits in period 2 if and only if the temptation to hit is non-negative.[34] Moreover, habit formation implies the temptation to consume in period 2 will be larger if the person hits in period 1 than if she refrains in period 1. Because this intuition holds for both the person's true future preferences or her perceived future preferences, a fully rational person is more likely to hit in period 2 if she hits in period 1, and a person with projection bias perceives

[31] Our assumption that exogenous factors influence the utility of hitting rather than the utility from refraining is purely for notational convenience — the two assumptions are formally equivalent.

[32] Note that this formulation does not fall under the rubric of simple projection bias, and our analysis in this section illustrates why that assumption is too restrictive.

[33] It will always be the case that $C_2^A(c_1) = C_2^*(c_1)$ — that is, a person always behaves optimally in period 2 conditional on her behavior in period 1 (which may not have been optimal). In a multi-period model, this is always true in the final period, but need not be true in any other period since, as with period 1 in this model, a person's behavior depends on her perceived future preferences.

[34] For simplicity, we assume throughout that a person hits when indifferent.

herself as more likely to hit in period 2 if she hits in period 1. This logic is summarized in Lemma 5.1.

Lemma 5.1. For all $k_1 \geq 0$, $\sigma, \rho > 0$, $\gamma \in [0,1]$, $x_1, x_2 \in \mathbb{R}$, and $\alpha_\rho, \alpha_\sigma, \alpha_x \in [0,1]$:

(1) $C_2^*(1) \geq C_2^*(0)$, and

(2) $\hat{C}_2^A(1) \geq \hat{C}_2^A(0)$.

In period 1, the person cares about the current temptation to hit, but also cares about the cost to future well-being from hitting now. For optimal behavior, this trade-off takes the following form: It is optimal to hit in period 1 (i.e., $c_1^* = 1$) if and only if

$$u\left(1,(k_1,x_1)\right) - u\left(0,(k_1,x_1)\right) \geq \delta\left[u\left(C_2^*(0),(\gamma k_1,x_2)\right) - u\left(C_2^*(1),(\gamma k_1+1,x_2)\right)\right]$$

or

$$x_1 + \sigma k_1 \geq \delta\left[u\left(C_2^*(0),(\gamma k_1,x_2)\right) - u\left(C_2^*(1),(\gamma k_1+1,x_2)\right)\right].$$

For optimal behavior, the temptation to hit in period 1 is $x_1 + \sigma k_1$. The future cost of hitting is the person's period-2 utility following restraint in period 1, which is $u\left(C_2^*(0),(\gamma k_1,x_2)\right)$, minus her period-2 utility following hitting in period 1, which is $u\left(C_2^*(1),(\gamma k_1+1,x_2)\right)$.

For a person suffering from projection bias, the perceived trade-off of current temptation vs. future cost takes a similar form: The person hits in period 1 if and only if

$$u\left(1,(k_1,x_1)\right) - u\left(0,(k_1,x_1)\right) \geq$$
$$\delta\left[\tilde{u}\left(\hat{C}_2^A(0),(\gamma k_1,x_2)\middle|\,(k_1,x_1)\right) - \tilde{u}\left(\hat{C}_2^A(1),(\gamma k_1+1,x_2)\middle|\,(k_1,x_1)\right)\right]$$

or

$$x_1 + \sigma k_1 \geq \delta\left[\tilde{u}\left(\hat{C}_2^A(0),(\gamma k_1,x_2)\middle|\,(k_1,x_1)\right) - \tilde{u}\left(\hat{C}_2^A(1),(\gamma k_1+1,x_2)\middle|\,(k_1,x_1)\right)\right].$$

For actual behavior, the temptation to consume in period 1 is $x_1 + \sigma k_1$, which is of course identical to that for optimal behavior — projection bias does not affect current instantaneous utilities. The perceived future cost of hitting is her perceived period-2 utility following restraint in period 1, which is $\tilde{u}\left(\hat{C}_2^A(0),(\gamma k_1,x_2)\middle|\,(k_1,x_1)\right)$, minus her perceived period-2 utility following hitting in period 1, which is $\tilde{u}\left(\hat{C}_2^A(1),(\gamma k_1+1,x_2)\middle|\,(k_1,x_1)\right)$.

Projection bias can cause the person to mispredict the future cost of hitting in two ways. First, there is a direct effect: projection bias causes the person to perceive incorrect period-2 instantaneous utilities (i.e., $\tilde{u}$ differs from u). Second, there is an indirect effect: projection bias causes the person to mispredict her own future behavior (i.e., $\hat{C}_2^A(c_1)$ possibly differs from $C_2^*(c_1)$).

It is straightforward to show that optimal and actual period-1 behavior both take the form of a "cutoff rule," where the person hits if and only if her current addiction level is above some threshold:

Lemma 5.2. For all $\sigma, \rho > 0$, $\gamma \in [0,1]$, $x_1, x_2 \in \mathbb{R}$, and $\alpha_\rho, \alpha_\sigma, \alpha_x \in [0,1]$:

(1) There exists $\bar{k}^* \geq 0$ such that $c_1^* = 1$ if and only if $k_1 \geq \bar{k}^*$, and

(2) There exists $\bar{k}^A \geq 0$ such that $c_1^A = 1$ if and only if $k_1 \geq \bar{k}^A$.

The key result behind Lemma 5.2 is that the future cost of hitting is decreasing in k_1 — the more addicted a person currently is, the less a current hit will hurt her in the future. This feature of our model is true both for the actual future cost of hitting and for the perceived future cost of hitting. Since the temptation to hit is increasing in her addiction level, clearly the person will follow a cutoff rule.

While Lemmas 5.1 and 5.2 pertain to any combination of α_ρ, α_σ, and α_x, we now attempt to distinguish the implications of projection bias on these three dimensions. We begin by assuming that the person fully understands exogenous factors, and consider the effects of an underappreciation of negative internalities and habit formation — that is, we consider $\alpha_x = 0$ and $\alpha_\rho, \alpha_\sigma \geq 0$. Proposition 5.1 establishes that projection bias on either of these dimensions leads to over-consumption of addictive products.[35]

Proposition 5.1. Suppose $\alpha_x = 0$. If either $\alpha_\rho > 0$ or $\alpha_\sigma > 0$, then for all $k_1 \geq 0$, $\sigma, \rho > 0$, $\gamma \in [0,1]$, and $x_1, x_2 \in \mathbb{R}$:

(1) $c_1^A \geq c_1^*$, and

(2) If $c_1^A = c_1^*$, then $c_2^A = c_2^*$; if $c_1^A > c_1^*$, then $c_2^A \geq c_2^*$.

[35] Since there always exist examples where the projection bias does not affect behavior, this and other results are stated with weak inequalities. However, in each case there also exist examples where the inequalities are strict.

The intuition for these over-consumption results is simple: An underappreciation for either negative internalities or habit formation leads the person to underestimate the cost to future well-being of hitting now. If a person underappreciates negative internalities, then she does not realize how much hitting today will hurt her tomorrow; and if a person underappreciates habit formation, then she does not realize how much hitting today will hurt her tomorrow if she plans to refrain tomorrow. Part 2 of Proposition 5.1 emphasizes that over-consumption in period 1 can lead to over-consumption in period 2. If a person over-consumes in period 1, then she will be more addicted in period 2 than she would have been had she behaved optimally in period 1, and higher addiction levels make a person more prone to consume.

Although Proposition 5.1 establishes that an underappreciation for negative internalities and an underappreciation for habit formation both lead to over-consumption, there are differences between the two errors. Proposition 5.2 shows that these biases have different implications for dynamic inconsistency:

Proposition 5.2. Suppose $\alpha_x = 0$. For all $k_1 \geq 0$, $\sigma, \rho > 0$, $\gamma \in [0,1]$, and $x_1, x_2 \in \mathbb{R}$, for any $\alpha_\rho \in [0,1]$:

(1) If $\alpha_\sigma = 0$, then $\hat{C}_2^A(c_1) = C_2^A(c_1)$ for $c_1 \in \{0,1\}$, and

(2) If $\alpha_\sigma > 0$, then $\hat{C}_2^A(1) \leq C_2^A(1)$ and $\hat{C}_2^A(0) \geq C_2^A(0)$.

Part 1 of Proposition 5.2 establishes that if a person fully appreciates both habit formation and exogenous variability, then she will be dynamically consistent; an underappreciation of negative internalities does not itself lead to dynamic inconsistency. If a person underappreciates the adverse effects of addictive products on her day-to-day life, then she may be surprised (and upset) at her overall level of well-being next year. But because she is not surprised by her desire to hit, she will behave according to her original plans. Hence, when over-consumption is driven by an underappreciation of negative internalities, it is merely because hitting looks less costly than it really is. Part 2 of Proposition 5.2 establishes that whenever a person underappreciates habit formation, however, she will be dynamically *in*consistent: She underestimates the likelihood of hitting tomorrow following hitting today, and also underestimates the likelihood of restraint tomorrow following restraint today.

These over-consumption results hold for any addiction level. Hence, while one implication is that an initially unaddicted person might develop a harmful addiction when she shouldn't, a second implication is that an initially addicted person might choose to stay addicted when she shouldn't. Addicts who should quit may feel that it is not worth becoming unaddicted because they underappreciate the degree to which their persistently strong craving is linked to their addiction.

We now assume that the person fully understands negative internalities and habit formation, and consider the effects of an underappreciation of exogenous factors — that is, we consider $\alpha_\rho = \alpha_\sigma = 0$ and $\alpha_x > 0$. Proposition 5.3 establishes that whether a person is over-optimistic or over-pessimistic about future behavior depends on whether the exogenous temptation will be larger or smaller in the future than it is currently:

Proposition 5.3. Suppose $\alpha_\rho = \alpha_\sigma = 0$. For all $k_1 \geq 0$, $\sigma, \rho > 0$, $\gamma \in [0, 1]$, and $x_1, x_2 \in \mathbb{R}$, for any $\alpha_x \in (0, 1]$:

(1) If $x_1 > x_2$, then $\hat{C}_2^A(c_1) \geq C_2^A(c_1)$ for $c_1 \in \{0, 1\}$, and

(2) If $x_1 < x_2$, then $\hat{C}_2^A(c_1) \leq C_2^A(c_1)$ for $c_1 \in \{0, 1\}$.

Part 1 of Proposition 5.3 says that if the exogenous temptation is higher now than it will be in the future, the person overestimates the likelihood that she will hit in period 2 whether she is predicting her behavior following hitting or refraining. The intuition is simple: If the person projects today's large exogenous temptation onto tomorrow's small exogenous temptation, then she perceives the utility from hitting tomorrow to be larger than it actually is, and therefore is over-pessimistic about hitting in the future. Part 2 says that the reverse result holds if the exogenous temptation now is smaller than it will be in the future; the intuition is analogous.

Consider next period-1 behavior. Suppose $x_1 > x_2$. The direct effect of mispredicting future utility is that the person projects today's large exogenous temptation onto tomorrow's small exogenous temptation, and therefore perceives the utility from hitting tomorrow to be larger than it actually is. Since hitting tomorrow is more likely following hitting today (Lemma 5.1), this direct effect makes a person too likely to hit today. Moreover, the indirect effect is that the person is over-pessimistic about the likelihood of future restraint, which further reduces the perceived future cost of hitting. Hence, both effects imply that the person is too likely to hit in period 1. An analogous logic applies when $x_1 < x_2$. This logic is summarized in Proposition 5.4:

Proposition 5.4. Suppose $\alpha_\rho = \alpha_\sigma = 0$. For all $k_1 \geq 0$, $x_1, x_2 \in \mathbb{R}$, $\sigma, \rho > 0$, and $\gamma \in [0, 1]$, for any $\alpha_x \in (0, 1]$:

(1) If $x_1 > x_2$, then $c_1^A \geq c_1^*$ and $c_2^A \geq c_2^*$, and

(2) If $x_1 < x_2$, then $c_1^A \leq c_1^*$ and $c_2^A \leq c_2^*$.

Extrapolating from our model, our results suggest that people may over-react to exogenous variability in the desire to consume — they are too likely to give in to large temptations, and too unlikely to give in to small temptations. When people have a strong desire to consume, they wrongly predict that the strong desire will persist into the future and therefore that they will consume a lot in the future, and as a result decide they might as well consume now. Analogously, when people have a weak desire to consume, they wrongly predict that the desire will remain weak in the future and therefore that they will consume very little in the future, and as a result they decide they should refrain now.

6. Cooling Off

In a variety of situations, people make irreversible (or difficult-to-reverse) decisions when they are in a "hot" state that is unlikely to persist — people buy automobiles when the dealer has them excited, get married in the heat of passion, and commit suicide when depression is particularly intense. Since the current state of mind affects well-being and will persist to some degree, responding to it is not *per se* a mistake. But projection bias leads people to underappreciate the degree to which intense feelings will dissipate, and hence people may be too likely to make an irreversible or otherwise costly decision. This suggests a useful policy prescription: Impose "cooling-off periods" that force people to delay before making irreversible decisions.

In this section, we use our model of projection bias to explore the role of cooling-off periods. We concentrate on one particular economic application: the purchase of a durable good, such as an automobile, when sellers can exert effort to get consumers "excited" about a purchase. We present a simple, extremely stylized model that identifies some basic intuitions that should hold more generally.

Suppose that a person goes to a dealer to possibly purchase a durable good. The dealer quotes a price P, and also engages in "sales hype" ϕ. After she observes P and experiences ϕ, the person

decides whether to buy the product. If she does not buy the product, then her intertemporal utility is zero. If she buys the product, she will enjoy the benefits of ownership, but must forego the future benefits she could have financed with wealth P.

The person derives ownership benefits from two sources, the intrinsic usefulness of the product and the excitement of ownership. Each period the product provides intrinsic value μ. The dealer's sales hype creates excitement level ϕ, but this excitement decays over time at rate $\gamma \in (0, 1)$; since hype occurs in period 1, the person's excitement level in period τ is $\gamma^{\tau-1}\phi$. Hence, the person's instantaneous utility function is $u(c_\tau, s_\tau) = c_\tau \cdot (\mu + s_\tau)$, where consumption is $c_\tau = 1$ if the person owns the good in period τ and $c_\tau = 0$ otherwise, and the state is $s_\tau = \gamma^{\tau-1}\phi$. The person has discount rate $\delta < 1$; for much of the analysis below, it is more natural and useful to interpret δ as also reflecting the probability that the product purchased will break or be lost.

We consider a monopoly dealer who faces a single cohort of potential customers. In period 1, all consumers enter the store, and each customer either buys the product in period 1 or never buys the product. Consumers differ in their intrinsic valuation of the product μ, where μ is distributed uniformly on interval $[0, X]$. The dealer cannot observe μ, and therefore must offer the same P and ϕ to all customers. But we assume that all customers have the same simple projection bias α, and that the dealer knows α. The dealer faces constant marginal cost of production $d \geq 0$. In addition, sales hype ϕ requires cost $C(\phi)$, where $C' > 0$ and $C'' > 0$.

We begin by assuming that contracts are irreversible.[36] If a consumer chooses to buy the product in period 1, then she must pay the price P and take possession immediately (i.e., $c_\tau = 1$ for all $\tau \geq 1$). In such an environment, the person should buy the product if and only if the discounted intrinsic value plus excitement of ownership is sufficiently large to cover the price — that is, if and only if

$$\frac{1}{1-\delta}\mu + \frac{1}{1-\delta\gamma}\phi - P \geq 0.$$

Projection bias leads the person to exaggerate the stream of benefits she will get from the excitement of ownership, and she buys the product if and only if

$$\frac{1}{1-\delta}\mu + \left[(1-\alpha)\frac{1}{1-\delta\gamma} + \alpha\frac{1}{1-\delta}\right]\phi - P \geq 0.$$

Both optimal behavior and actual behavior can be characterized by a cutoff rule: There exist

[36] A more realistic assumption would be that the parties could write reversible contracts if they wished. But given our (extreme) assumptions, customers see no benefits of flexibility, and the dealer, because he is aware of the customer's projection bias, strictly prefers an irreversible contract to a reversible contract.

$\bar{\mu}^*(P,\phi)$ and $\bar{\mu}^A(P,\phi)$ such that people should buy the product if and only if $\mu \geq \bar{\mu}^*(P,\phi)$, but people actually buy the product if and only if $\mu \geq \bar{\mu}^A(P,\phi)$. Lemma 6.1 compares how people actually behave to how they should behave:

Lemma 6.1. For any P and ϕ, $\overline{\mu}^*(P,\phi) - \overline{\mu}^A(P,\phi) = \alpha\frac{\delta(1-\gamma)}{1-\delta\gamma}\phi$.

Anyone with $\mu \in [\overline{\mu}^A(P,\phi), \overline{\mu}^*(P,\phi))$ makes an irreversible purchase that they shouldn't; Lemma 6.1 therefore establishes that if the dealer engages in any sales hype at all ($\phi > 0$), too many people make the irreversible purchase. This result is a simple implication of the fact that projection bias leads a person to exaggerate the stream of benefits she will get from the excitement of ownership. Lemma 6.1 also yields some simple comparative statics. First, the quicker a person cools off — the smaller is γ — the more likely she is to buy when she shouldn't. The quicker a person cools off, the more her true future preferences differ from her current preferences, and therefore the bigger is the error caused by projection bias. Second, the more patient is a person or the longer lasting is the product being purchased— the larger is δ — the more likely she is to buy when she shouldn't. Projection bias makes the person perceive that her future utility from the product will be larger than it actually is; hence the more she cares about the distant future benefits of the product, the bigger the errors she makes. Third, the bigger is the projection bias — the bigger is α — the more likely the person is to buy when she shouldn't.

While these comparative statics reflect some important intuitions, and parallel some of our earlier examples, we are also interested in the fact that ϕ is endogenous — it is chosen by the dealer. If people buy whenever $\mu \geq \bar{\mu}(P,\phi)$, then, recalling that μ is distributed uniformly on $[0,X]$, the dealer faces demand function $D(\bar{\mu}(P,\phi)) = X - \bar{\mu}(P,\phi)$. Given constant production cost $d \geq 0$ and cost of sales hype $C(\phi)$, the dealer's payoff function is[37]

$$(P-d)\cdot D(\bar{\mu}(P,\phi)) \quad - \quad C(\phi).$$

If people were fully rational, then the dealer would face demand $D(\bar{\mu}^*(P,\phi))$; we let (P^*,ϕ^*) denote what optimal dealer behavior would be if people were fully rational. But the dealer actually faces demand $D(\bar{\mu}^A(P,\phi))$; we let (P^A,ϕ^A) denote optimal dealer behavior when the dealer is

[37] We assume X is sufficiently large and C is sufficiently convex to guarantee an interior solution. In particular, $X > d$, $C'(0) = 0$, and $C''(\phi) > [(1-\alpha)(1-\delta)/(1-\delta\gamma)+\alpha]^2/[2(1-\delta)]$ for all ϕ guarantees an interior solution for both optimal and actual behavior.

aware that people have projection bias. Proposition 6.1 describes how projection bias affects dealer behavior:

Proposition 6.1. (1) $P^A > P^*$ and $\phi^A > \phi^*$, and

(2) $\bar{\mu}^*(P^A, \phi^A) - \bar{\mu}^A(P^A, \phi^A) > \bar{\mu}^*(P^*, \phi^*) - \bar{\mu}^A(P^*, \phi^*)$.

Part 1 of Proposition 6.1 states that a monopoly dealer who is aware that people have projection bias distorts upwards both sales hype and the price. Sales hype causes a person with projection bias to overvalue the product. If dealers are aware that people have projection bias, they (correctly) perceive a larger marginal return to sales hype, and therefore choose more sales hype than they would if people were fully rational. The increased price further takes advantage of the increased value, both real and perceived, created by sales hype.[38]

Part 2 of Proposition 6.1 emphasizes that the dealer takes advantage of people: Because those people with $\mu \in [\bar{\mu}^A(P, \phi), \bar{\mu}^*(P, \phi))$ buy the product when they shouldn't, Part 2 establishes that the dealer reacts to projection bias in a way that increases the number of incorrect purchases. We should also note that the dealer's reaction changes the identity of those who make incorrect purchases, in ways that depend on the relative magnitudes of the increases in sales hype and price.[39]

We now consider imposing a cooling-off period that forces people to delay for some duration before making the irreversible decision of buying the product. For ease of exposition, we focus on a mandatory one-period delay: Consumers who choose to buy the product in period 1 cannot take possession until period 2, and they have the option to cancel the contract in period 2 rather than take possession. Formally, we assume that if a consumer chooses to buy the product in period 1, then she pays the price P in period 1. In period 2, the person either takes possession, in which case $c_\tau = 1$ for all $\tau \geq 2$, or cancels the contract and gets a full refund.

[38] That the increased demand created by sales hype increases rather than decreases the profit-maximizing price of course depends on our assumption about the distribution of tastes, which yields a linear demand function.

[39] While the dealer's reaction increases the number of customers who suffer negative utility, it could be that total consumer surplus increases. Intuitively, if the dealer's response involves a sufficiently large increase in sales hype, then all customers who would have made the purchase at (P^*, ϕ^*) will be better off despite the higher price. Since the increased sales hype increases the product's value while the increased price decreases the product's value, the net effect of the dealer's reaction could be to increase or decrease the product's value. In the event that the product's value decreases, total consumer surplus necessarily decreases. In the event that the product's value increases, total consumer surplus might increase or decrease depending on the relative magnitudes of the two effects — increased product value for all, but more incorrect purchases.

Under such a cooling-off period, the person should buy the product in period 1 if and only if

$$\frac{\delta}{1-\delta}\mu + \frac{\delta\gamma}{1-\delta\gamma}\phi - P \geq 0.$$

Because the cooling-off period delays possession by one period, the person loses period-1 ownership benefits $(\mu + \phi)$; otherwise, this condition is identical to that under binding contracts. This inequality defines the cutoff $\bar{\mu}^{**}(P, \phi)$ for optimal behavior.

A person with projection bias chooses to buy the product in period 1 if and only if

$$\frac{\delta}{1-\delta}\mu + \left[(1-\alpha)\frac{\delta\gamma}{1-\delta\gamma} + \alpha\frac{\delta}{1-\delta}\right]\phi - P \geq 0.$$

Once again, this condition is identical to that under binding contracts except for losing the period-1 ownership benefits $(\mu + \phi)$. But since the person will have cooled off somewhat before period 2, she might cancel the contract in period 2. The person's state in period 2 will be $s_2 = \gamma\phi$, and she will project this smaller excitement level onto her future preferences. She will therefore take possession in period 2 if and only if[40]

$$\frac{1}{1-\delta}\mu + \left[(1-\alpha)\frac{1}{1-\delta\gamma} + \alpha\frac{1}{1-\delta}\right]\gamma\phi - \left(\frac{1}{\delta}P\right) \geq 0.$$

Our focus shall be on whether the person actually takes possession; this latter inequality defines the cutoff $\bar{\mu}^{AA}(P, \phi)$ such that a person actually completes the contract if and only if $\mu \geq \bar{\mu}^{AA}(P, \phi)$.

Lemma 6.2 describes how a one-period cooling-off period affects behavior (holding dealer behavior fixed):

Lemma 6.2. For any P and ϕ,

(1) $\bar{\mu}^{**}(P, \phi) > \bar{\mu}^{*}(P, \phi)$ and $\bar{\mu}^{AA}(P, \phi) > \bar{\mu}^{A}(P, \phi)$, and

(2) $\bar{\mu}^{**}(P, \phi) - \bar{\mu}^{AA}(P, \phi) < \bar{\mu}^{*}(P, \phi) - \bar{\mu}^{A}(P, \phi)$.

Lemma 6.2 reflects the basic trade-off associated with a mandatory cooling-off period. Part 1 reflects the costs of a cooling-off period: Fewer people buy the product. By delaying possession, people must give up both the intrinsic value and ownership excitement that they would have received in period 1. But part 2 reflects the potential benefits of a cooling-off period: Fewer people buy the product when they shouldn't. A cooling-off period is exactly that — people must take time

[40] That the price P is multiplied by the factor $1/\delta$ reflects that the price paid in period 1 is the period-1 discounted value of forgone future benefits. Since in period 2 those future benefits are one period closer, the period-2 discounted value is P/δ.

to cool off and move out of the extreme state. As they do, their current preferences become more similar to their future preferences, and they are therefore less prone to make the mistake.

Lemma 6.2 describes the benefits of a cooling-off period holding dealer behavior fixed. How does the dealer react to this law? Under the cooling-off period, the dealer faces demand $D(\bar{\mu}^{AA}(P,\phi))$; we let (P^{AA},ϕ^{AA}) denote optimal dealer behavior. Proposition 6.2 describes how the dealer responds to the cooling-off period:

Proposition 6.2. (1) $\phi^{AA} < \phi^{A}$ and $P^{AA} < P^{A}$, and

(2) $\bar{\mu}^{**}(P^{AA},\phi^{AA}) - \bar{\mu}^{AA}(P^{AA},\phi^{AA}) < \bar{\mu}^{**}(P^{A},\phi^{A}) - \bar{\mu}^{AA}(P^{A},\phi^{A}) < \bar{\mu}^{*}(P^{A},\phi^{A}) - \bar{\mu}^{A}(P^{A},\phi^{A})$.

Part 1 of Proposition 6.2 establishes that the cooling-off period reduces the dealer's incentive to engage in excessive sales hype. Intuitively, the marginal return to sales hype is reduced by the cooling-off period because the purchase decision is effectively made after the person's excitement level decays some. Note that less sales hype also leads to a lower price. Part 2 establishes that fewer people buy the product when they shouldn't under the cooling-off period. The second inequality follows from Lemma 6.2 — holding dealer behavior fixed at (P^{A},ϕ^{A}), the cooling-off period makes people less likely to buy when they shouldn't. The first inequality reflects that the effect of the cooling-off period on dealer behavior — making the dealer reduce sales hype — further enhances the basic effect.[41]

Our analysis above does not establish whether a one-period cooling-off law is on net beneficial. It merely lays out the benefits — fewer people make incorrect purchases — and costs — forgone short-run benefits. But the logic described above makes clear when cooling-off periods are likely to be particularly beneficial: when cooling off occurs relatively quickly — i.e., when γ is small — and when short-term ownership benefits are relatively small — i.e., when $\mu + \phi$ is small. Of course, in some environments the costs would be small, as when there is a natural delay before a good can be enjoyed, or when a customer could return the product in good-as-new condition. In other situations, the costs could be substantial.

[41] In fact, our analysis may understate the degree to which firms will reduce sales hype in response to a cooling-off law because we assume that cancelled contracts are costless to the dealer. If cancelled contracts were costly for the dealer, it would reduce sales hype even further to reduce the number of cancelled contracts. Of course, the cancelled-contract costs would be another social cost of the cooling-off law.

Moreover, cooling-off periods may be attractive because they can have large benefits for those with projection bias, while imposing relatively small costs on those who are fully rational. In particular, for α large and γ small, binding contracts may be very damaging, and short cooling-off periods may (nearly) fix the problem. At the same time, for a fully rational person, a mandated cooling-off period is costly only to the extent of the short delay before the benefits are initiated.

Cooling-off periods of the sort described above are potentially useful for any difficult-to-reverse decisions that people tend to make under the influence of states of mind that are unlikely to persist. As discussed at the outset, marriage and suicide also meet these criteria. In the case of marriage, people often decide to commit themselves when experiencing intense feelings that will not last. The failure to appreciate the transience of these feelings not only results in marriages that probably shouldn't take place, but may also lessen the perceived need for steps, such as prenuptial agreements, that could lessen the negative impacts of divorce. Many states and religions, in fact, place barriers to impulsive marriage, including long engagement periods, waiting periods for marriage certificates, or medical tests that take time. Such restrictions suggest that there may be some awareness of projection bias and of the benefits of cooling off. In the case of suicide, people often commit suicide in the midst of extreme depression or hopelessness. There is a common assertion in writings on depression (see, e.g., Solomon 1998, p. 49) that depressed people are unable to imagine that they will feel better — i.e., not depressed — in the future, suggesting that both clinical and situational depression and situational hopelessness are subject to projection bias. While explicit cooling-off periods for suicide would be awkward and difficult to enforce, de facto cooling-off periods are imposed legally and socially when the norm is to intervene in observed cases of attempted suicide, or to pay special attention to suicide attempts by people in bad situations.

7. Discussion and Conclusion

Our goal in this paper has been to improve the realism of formal economic models of intertemporal choice by incorporating under the rubric of projection bias a range of related psychological phenomena of how people mispredict their preferences. Our analysis in Sections 4, 5, and 6 highlights the potential economic importance of projection bias. Rather than choose applications that would highlight qualitative differences between our model and the rational-choice model, we have chosen applications that demonstrate two different categories of projection-bias implications.

First, projection bias makes sounder *quantitative* predictions. For instance, acquiring a harmful addiction is consistent in principle with both the rational-choice model and the projection-bias model. But under reasonable assumptions about how the pleasure from consumption compares to the magnitude of future harm, the rational-choice model requires absurd levels of impatience whereas our projection-bias model could predict the same behavior with more reasonable levels of impatience.

Second, projection bias improves welfare analysis. Researchers outside the field of economics take it for granted that people sometimes behave in suboptimal ways. While many economists likely share this opinion, they often view such issues as outside the domain of economics. By introducing into formal models a behaviorally-based, precise articulation of a systematic error, our model of projection bias helps facilitate a principled analysis of the ways in which people behave suboptimally in economic environments, and thus allows more accurate welfare conclusions.

Many of these issues also apply to another psychological phenomenon that has received increasing attention from economists: time-inconsistent discounting and self-control problems. It is worth briefly distinguishing the two phenomena, particularly since both can give rise to dynamic inconsistency. Recent formal models of self-control problems are based on a long tradition in psychology — and in common sense — that people have a time-inconsistent taste for immediate gratification.[42] A person with self-control problems may or may not mispredict her future behavior: If the person is sophisticated and fully understands her future self-control problems, she will be dynamically consistent; but if she is naive and underestimates her future self-control problems, she will systematically over-estimate the likelihood of behaving herself in the future.

Insofar as projection bias and naivete about self-control problems both involve misprediction of future intertemporal utility functions, they are of course related. They may also share psychological underpinnings, because both involve a failure by people to appreciate how their future feelings will differ from their current feelings. But the two sources of dynamic inconsistency are distinguished by the object of the person's misprediction: Whereas projection bias is a misprediction of future instantaneous utilities, naivete about self-control problems is a misprediction of the relative weights she will attach to future instantaneous utilities.

In many contexts, it is difficult to disentangle whether mispredictions of future behavior are

[42] See Ainslie (1992), Loewenstein and Prelec (1992), and Thaler (1991) for reviews of the evidence on self-control problems. Such time-inconsistent preferences were first formally studied in economics by Strotz (1956), and more recently by Laibson (1994,1997), O'Donoghue and Rabin (1999*a*), and others.

coming from projection bias or naivete about self-control problems. Consider, for instance, the tendency to underpredict the propensity to yield to sexual desire in the future. People may go on dates planning to refrain from unsafe sexual contact, but then in the heat of the moment behave in ways that they had not predicted. Both projection bias and naivete about self-control problems can play a role in this phenomenon. Projection bias generates the misprediction because the person does not fully appreciate in a cool state how tempting sex will be in a hot state. Naivete about self-control problems generates the misprediction because the person does not appreciate how prone she will be to pursue activities that yield her immediate gratification.

But projection bias and naivete about self-control problems often generate distinct predictions. Indeed, several of the mispredictions discussed in this paper can clearly come only from projection bias and not from naivete about self-control problems. For instance, neither Loewenstein, Nagin, and Paternoster's (1997) finding that males' predictions about behaving in a sexually aggressive fashion depended on their current arousal nor Read and van Leeuwen's (1998) finding that people's choices of snacks depended on their current hunger can be explained by naivete about self-control problems; projection bias is clearly implicated.[43]

While we hope we have already made our case for the broad applicability of projection bias, we now briefly speculate about three additional economic applications. Social-comparison theory, which studies the ways a person cares about her status relative to a comparison group, is a major topic in psychological research, and has recently received more attention among economists. Because people often make decisions, such as switching jobs, that cause their comparison groups to change, we suspect that projection bias in the form of underappreciating the effects of a change in comparison groups may be important in this context.

As a concrete example inspired by Frank (1985), consider a person's decision whether to switch jobs. Suppose the jobs differ in two types of status they offer the person. They provide her with "general status" — the prestige accorded to the job by society at large — and "local status" — how her status compares to the average of those in her comparison group. When deciding whether to switch jobs, the person should optimally weigh the trade-off between general status and local status.

[43] A second result in Read and van Leeuwen (1998) seems to implicate self-control problems and not projection bias: Subjects were asked at the time the snacks were delivered whether they would like to switch, and whereas very few people (5%) who chose an unhealthy snack changed their minds and requested a healthy snack, a large fraction (71%) of subjects who chose a healthy snack changed their minds and requested an unhealthy snack. This asymmetry in switching behavior is exactly what we would expect to see if people have self-control problems wherein they prefer to eat healthy snacks *in the future* while at the same time they prefer to eat the unhealthy snacks *now*.

But if projection bias leads the person to underappreciate the effects of a changing comparison group, and if the comparison group affects local status but not general status, then the person will under-pursue local status and therefore over-pursue general status. As a result, the person will be too likely to choose a high-status job whose comparison group has high average status. Graduate students may, for example, be too likely to take a job at Harvard because they believe the short-term, local-status effect of prestige among their current comparison group (fellow graduate students) will persist. They do not fully appreciate that once they get to Harvard they won't impress their new comparison group (colleagues at Harvard) with the fact that they have a job at Harvard.

A broader, but more speculative, economic application of projection bias is to diminishing marginal utility. Traditional consumer theory assumes that recent consumption of a product reduces the marginal utility from further consumption. Eating a second pint of ice cream yields less pleasure than the first, and watching a Johnny Depp movie for the 30^{th} time generates less pleasure than watching it for merely the 3^{rd} time. While consumer theory usually suppresses the temporal nature of diminishing marginal utility, it may be important when a person suffers from projection bias. Most people understand satiation: We all realize that eating the second pint of ice cream will be less satisfying than the first. But from anecdotal evidence and intuition — and from extrapolating the hunger findings discussed earlier — we suspect that projection bias leads people to underappreciate these effects. That is, we suspect that people tend to project their current marginal utility of consumption onto their future marginal utility, and hence act as if our utility function for consumption of a good is less concave than it actually is.

If people extrapolate marginal utilities in this way, then they will be prone to over-purchase activities they currently don't engage in. People may plan overly long vacations, believing the ninth day lying on the beach will be nearly as enjoyable as the first; and professionals who have little time for reading or traveling may falsely anticipate the blissfulness of spending their retirement years with non-stop reading and traveling. Firms may, of course, take advantage of such mispredictions, just as they exploited over-heated consumers in Section 6, by selling long-term contracts for a series of purchases. For instance, we suspect that initial purchase agreements for book-of-the-month clubs and yearly passes to gyms are an attempt to take advantage of people who are currently enthusiastic about reading or working out at the gym. On a shorter time-scale, we suspect that restaurants take advantage of projection bias by offering all-you-can-eat meals to hungry diners who don't realize that they will become satiated relatively quickly.

Our final economic application involves extending projection bias to the interpersonal domain: Perhaps people make the same types of mistakes in predicting other people's preferences and behavior that they make in predicting their own behavior. Indeed, Van Boven, Dunning and Loewenstein (1999) provide evidence of such interpersonal projection bias and illustrate some potential economic implications. In one experiment, the usual endowment effect was replicated by eliciting selling prices from subjects endowed with coffee mugs and buying prices from subjects not endowed. But sellers were then asked to estimate how much buyers would pay and buyers to estimate how much sellers would charge, with all subjects rewarded for accurate predictions. Consistent with interpersonal projection bias, sellers over-estimated buying prices, and buyers underestimated selling prices.

In further experiments, some subjects ("sellers") were given mugs and privately asked their minimum selling prices. Other subjects ("buyers' agents") were given $10 to purchase a mug on behalf of a buyer, and told to make a single offer to one seller which would be either accepted or rejected based on the seller's stated minimum selling price. Buyers' agents were told that if their offer was accepted they could keep the difference between the $10 and the amount that they bought the mug for. If their offer was rejected they would get nothing. In one (representative) experiment, the mean offer was $5.54, but the mean minimum selling price was $6.98; only 29% of offers were accepted, and the average earnings of buyers' agents was $.85. At the profit-maximizing offer of $7.00, 66% of offers would have been accepted, and average earnings would have been $2.00. Further experiments showed that the buyers' agents' interpersonal error resulted from an intrapersonal error — from their tendency to underestimate the price at which they themselves would sell the mug.[44]

By demonstrating that people can actually lose money by failing to appreciate projection bias, and hence have incentives to overcome the bias, this last application raises the question of whether projection bias disappears with experience. That projection bias operates on states such as hunger with which people should have ample experience suggests that projection bias does not disappear. A possible explanation for the persistence of projection bias is Loewenstein's (1996) hypothesis that projection bias applies retrospectively. Just as a person may not fully attend to all relevant utility changes when making decisions, she may not fully attend to all relevant utility changes

[44] These experiments also provide some insight into how projection bias persists even in settings with feedback: Buyers' agents interpreted failures to transact as greed on the part of sellers, only slowly learned from experience that their offers were too low, and even then failed to generalize the lesson to other goods.

when reflecting on past behavior. Such a retrospective projection bias may explain the tendency to forget how one actually behaved — to think that one ate less than one actually ate, displayed less anger than one did, etc. — and to view the elements of past behavior that one does recall as inexplicable and flukish. The failure to recall one's behavior and to view it as flukish may prevent learning from experience.

At the same time, there is reason to believe that people are aware of projection bias on some meta-level: A person might be aware that when she shops tomorrow she will suffer from projection bias, and yet when tomorrow arrives she will be unaware that she currently suffers from the bias.[45] Indeed, that "never shop on an empty stomach" is such a part of folk wisdom suggests people are aware of projection bias. In addition, we suspect that many rules people develop are designed to deal with moment-by-moment projection bias. For instance, in the context of our cooling-off model, people might develop rules such as never buy a car the first time you visit a dealer. Such rules may be quite important — and also may be further evidence that people suffer from projection bias. But as a careful analysis of such rules is beyond the scope of this paper, we leave the analysis for future research.

As models that reflect the reality of both short-term fluctuations and long-term changes in preferences become more widespread in economics, economists must seriously address the question of whether people accurately predict how their preferences will change. Much as there has been a growing recognition among economists that behavioral and welfare economics will be improved by developing models that incorporate self-control problems, we hope our analysis and examples illustrate the potential benefits for both behavioral and welfare economics of incorporating mispredictions of utilities in general, and projection bias in particular, into formal economic analysis.

[45] As we discuss in Section 3, the very essence of projection bias requires that at any moment when the person suffers from projection bias she must be unaware of that fact.

Appendix: Proofs

Proof of Proposition 3.1: Let $\mathbf{C} \equiv (\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_T)$ and suppose $\mathbf{C}$ induces states $(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_T)$. Let $\mathbf{C}' \equiv (\mathbf{c}'_1, \mathbf{c}'_2, ..., \mathbf{c}'_T)$ and suppose $\mathbf{C}'$ induces states $(\mathbf{s}'_1, \mathbf{s}'_2, ..., \mathbf{s}'_T)$. Then define $\tilde{V}^t(\mathbf{C}, \mathbf{s}_t) \equiv \sum_{\tau=t}^{T} \delta^\tau \tilde{u}(\mathbf{c}_\tau, \mathbf{s}_\tau | \mathbf{s}_t)$, which is the continuation utility perceived in period t from following consumption path $\mathbf{C}$ in all periods. Suppose $\mathbf{C}$ is the optimal consumption path perceived in period 1 — that is, $\mathbf{C}$ maximizes $\tilde{V}^1(\mathbf{C}, \mathbf{s}_1)$. The person is dynamically consistent if for all $\mathbf{C}'$ such that for some $\bar{\tau} > 1$, $\mathbf{c}'_\tau = \mathbf{c}_\tau$ for all $\tau < \bar{\tau}$ (and therefore $\mathbf{s}'_\tau = \mathbf{s}_\tau$ for all $\tau \leq \bar{\tau}$), $\tilde{V}^t(\mathbf{C}, \mathbf{s}_t) \geq \tilde{V}^t(\mathbf{C}', \mathbf{s}_t)$ for all $t \leq \bar{\tau}$. We now prove that the conditions in Proposition 3.1 are sufficient for the person to be dynamically consistent. For all t, $\tilde{V}^t(\mathbf{C}, \mathbf{s}_t) - \tilde{V}^t(\mathbf{C}', \mathbf{s}_t) = \sum_{\tau=t}^{T} \delta^\tau \left[\tilde{u}(\mathbf{c}_\tau, \mathbf{s}_\tau | \mathbf{s}_t) - \tilde{u}(\mathbf{c}'_\tau, \mathbf{s}'_\tau | \mathbf{s}_t)\right]$. Since $\mathbf{c}'_\tau = \mathbf{c}_\tau$ for all $\tau < \bar{\tau}$, which implies $\mathbf{s}'_\tau = \mathbf{s}_\tau$ for all $\tau \leq \bar{\tau}$, $\tilde{V}^t(\mathbf{C}, \mathbf{s}_t) - \tilde{V}^t(\mathbf{C}', \mathbf{s}_t) = \sum_{\tau=\bar{\tau}}^{T} \delta^\tau \left[\tilde{u}(\mathbf{c}_\tau, \mathbf{s}_\tau | \mathbf{s}_t) - \tilde{u}(\mathbf{c}'_\tau, \mathbf{s}'_\tau | \mathbf{s}_t)\right]$ for all $t \leq \bar{\tau}$. Note that $\tilde{u}(\mathbf{c}_\tau, \mathbf{s}_\tau | \mathbf{s}_t) - \tilde{u}(\mathbf{c}'_\tau, \mathbf{s}'_\tau | \mathbf{s}_t) = \tilde{u}(\mathbf{c}_\tau, \mathbf{s}_\tau | \mathbf{s}_t) - \tilde{u}(\mathbf{c}'_\tau, \mathbf{s}_\tau | \mathbf{s}_t) + \tilde{u}(\mathbf{c}'_\tau, \mathbf{s}_\tau | \mathbf{s}_t) - \tilde{u}(\mathbf{c}'_\tau, \mathbf{s}'_\tau | \mathbf{s}_t)$. The first condition in Proposition 3.1 implies $\tilde{u}(\mathbf{c}_\tau, \mathbf{s}_\tau | \mathbf{s}_t) - \tilde{u}(\mathbf{c}'_\tau, \mathbf{s}_\tau | \mathbf{s}_t)$ is independent of $\mathbf{s}_t$. The second condition of the Proposition 3.1 implies $\tilde{u}(\mathbf{c}'_\tau, \mathbf{s}_\tau | \mathbf{s}_t) - \tilde{u}(\mathbf{c}'_\tau, \mathbf{s}'_\tau | \mathbf{s}_t)$ is independent of $\mathbf{s}_t$. Hence, for all $t < t' \leq \bar{\tau}$, $\tilde{V}^t(\mathbf{C}, \mathbf{s}_t) - \tilde{V}^t(\mathbf{C}', \mathbf{s}_t) = \tilde{V}^{t'}(\mathbf{C}, \mathbf{s}_{t'}) - \tilde{V}^{t'}(\mathbf{C}', \mathbf{s}_{t'})$. Finally, because $\mathbf{C}$ maximizes $\tilde{V}^1(\mathbf{C}, \mathbf{s}_1)$, we know $\tilde{V}^1(\mathbf{C}, \mathbf{s}_1) - \tilde{V}^1(\mathbf{C}', \mathbf{s}_1) \geq 0$, and the result follows. Q.E.D.

Proof of Lemma 4.1: The first-order conditions for optimal behavior are

$$\begin{aligned} v'(c_1^*) + R'(c_1^*) - \gamma R'(c_2^* - \gamma c_1^*) &= \lambda^* \\ v'(c_2^*) + R'(c_2^* - \gamma c_1^*) &= \lambda^* \\ w'(l_1^*) &= \lambda^* = w'(l_2^*) \\ c_1^* + c_2^* + l_1^* + l_2^* &= 2 \end{aligned}$$

where λ^* is the Lagrange multiplier. Given $v'' < 0$, $w'' < 0$, and $R'' \leq 0$, there is a unique maximum, and we assume an interior solution. It is clear that $l_1^* = l_2^*$. We prove $c_1^* < c_2^*$ by contradiction. Suppose $c_1^* \geq c_2^*$. Then $c_1^* > c_2^* - \gamma c_1^*$, in which case $R'' \leq 0$ implies $R'(c_1^*) \leq R'(c_2^* - \gamma c_1^*)$, and therefore $R' > 0$ implies $R'(c_1^*) - \gamma R'(c_2^* - \gamma c_1^*) < R'(c_2^* - \gamma c_1^*)$. But $R'(c_1^*) - \gamma R'(c_2^* - \gamma c_1^*) < R'(c_2^* - \gamma c_1^*)$ and $v'' < 0$ implies $c_1^* < c_2^*$, a contradiction. Q.E.D.

Proof of Lemma 4.2: The first-order conditions for planned behavior are

$$v'(c_1^A) + R'(c_1^A) - (1-\alpha)\gamma R'(\hat{c}_2^A - \gamma c_1^A) = \lambda^A$$

$$\begin{aligned} v'(\hat{c}_2^A) + (1-\alpha)R'(\hat{c}_2^A - \gamma c_1^A) + \alpha R'(\hat{c}_2^A) &= \lambda^A \\ w'(l_1^A) &= \lambda^A = w'(\hat{l}_2^A) \\ c_1^A + \hat{c}_2^A + l_1^A + \hat{l}_2^A &= 2 \end{aligned}$$

where λ^A is the Lagrange multiplier. Given $v'' < 0$, $w'' < 0$, and $R'' \leq 0$, there is a unique maximum, and we assume an interior solution. It is clear that $l_1^A = \hat{l}_2^A$. We prove $c_1^A < \hat{c}_2^A$ by contradiction. If $c_1^A \geq \hat{c}_2^A > \hat{c}_2^A - \gamma c_1^A$, then $R'' \leq 0$ implies $R'(c_1^A) \leq R'(\hat{c}_2^A) \leq R'(\hat{c}_2^A - \gamma c_1^A)$, and therefore $R'(c_1^A) \leq (1-\alpha)R'(\hat{c}_2^A - \gamma c_1^A) + \alpha R'(\hat{c}_2^A)$. Then $R' > 0$ implies $R'(c_1^A) - (1-\alpha)\gamma R'(\hat{c}_2^A - \gamma c_1^A) < (1-\alpha)R'(\hat{c}_2^A - \gamma c_1^A) + \alpha R'(\hat{c}_2^A)$, but then $v'' < 0$ implies $c_1^A < \hat{c}_2^A$, a contradiction. Q.E.D.

Proof of Proposition 4.1: (1) Letting $R'(x) = \bar{R}$ for all x, the first-order conditions become

Optimal Behavior	Planned Behavior
$v'(c_1^*) + [1-\gamma]\bar{R} = \lambda^*$	$v'(c_1^A) + [1-(1-\alpha)\gamma]\bar{R} = \lambda^A$
$v'(c_2^*) + \bar{R} = \lambda^*$	$v'(\hat{c}_2^A) + \bar{R} = \lambda^A$
$w'(l_1^*) = \lambda^* = w'(l_2^*)$	$w'(l_1^A) = \lambda^A = w'(\hat{l}_2^A)$
$c_1^* + c_2^* + l_1^* + l_2^* = 2$	$c_1^A + \hat{c}_2^A + l_1^A + \hat{l}_2^A = 2.$

Because $[1-(1-\alpha)\gamma] > [1-\gamma]$, $\lambda^A \leq \lambda^*$ would imply $c_1^A > c_1^*$, $\hat{c}_2^A \geq c_2^*$, and $l_1^A = \hat{l}_2^A \geq l_1^* = l_2^*$, which would violate the budget constraint. Hence, it must be that $\lambda^A > \lambda^*$, which implies $\hat{c}_2^A < c_2^*$, and $l_1^A = \hat{l}_2^A < l_1^* = l_2^*$. The budget constraint then implies $c_1^A > c_1^*$.

(2) The first-order conditions for actual period-2 behavior are

$$\begin{aligned} v'(c_2^A) + \bar{R} &= \lambda^{AA} \\ w'(l_2^A) &= \lambda^{AA} \\ c_2^A + l_2^A &= 2 - \left[c_1^A + l_1^A\right] = \hat{c}_2^A + \hat{l}_2^A \end{aligned}$$

where λ^{AA} is the Lagrange multiplier. Clearly $c_2^A = \hat{c}_2^A$ and $l_2^A = \hat{l}_2^A$ are the solutions to these equations (with $\lambda^{AA} = \lambda^A$). Q.E.D.

Proof of Proposition 4.2: (1) The first-order conditions for optimal and planned behavior are as in the proofs of Lemmas 4.1 and 4.2. We prove that $c_1^A > c_1^*$ whether $\lambda^A \leq \lambda^*$ or $\lambda^A > \lambda^*$. First suppose $\lambda^A \leq \lambda^*$, which implies $l_1^A = \hat{l}_2^A \geq l_1^* = l_2^*$ and therefore $c_1^A + \hat{c}_2^A \leq c_1^* + c_2^*$. Because the first-order conditions for optimal behavior imply $v'(c_1^*) + R'(c_1^*) + \gamma v'(c_2^*) = (1+\gamma)\lambda^*$ and the first-order conditions for planned behavior imply $v'(c_1^A) + R'(c_1^A) + \gamma v'(\hat{c}_2^A) + \gamma\alpha R'(\hat{c}_2^A) = (1+\gamma)\lambda^A$, $\lambda^A \leq \lambda^*$ and $R' > 0$ imply either $c_1^A > c_1^*$ or $\hat{c}_2^A > c_2^*$ (but not both since $c_1^A + \hat{c}_2^A \leq c_1^* + c_2^*$). But

$\hat{c}_2^A > c_2^*$ and $c_1^A < c_1^*$ implies

$$\lambda^A = v'(c_1^A) + R'(c_1^A) - (1-\alpha)\gamma R'(\hat{c}_2^A - \gamma c_1^A) > v'(c_1^*) + R'(c_1^*) - \gamma R'(c_2^* - \gamma c_1^*) = \lambda^*,$$

which contradicts $\lambda^A \leq \lambda^*$. Hence, if $\lambda^A \leq \lambda^*$ then $c_1^A > c_1^*$ and $\hat{c}_2^A < c_2^*$.

Next suppose $\lambda^A > \lambda^*$, which implies $l_1^A = \hat{l}_2^A < l_1^* = l_2^*$ and therefore $c_1^A + \hat{c}_2^A > c_1^* + c_2^*$. Then $c_1^A \leq c_1^*$ requires $\hat{c}_2^A > c_2^*$ which implies

$$\lambda^A = v'(\hat{c}_2^A) + (1-\alpha)R'(\hat{c}_2^A - \gamma c_1^A) + \alpha R'(\hat{c}_2^A) < v'(c_2^*) + R'(c_2^* - \gamma c_1^*) = \lambda^*,$$

which contradicts $\lambda^A > \lambda^*$. Hence, if $\lambda^A \leq \lambda^*$ then $c_1^A > c_1^*$ (and the relationship between $\hat{c}_2^A$ and c_2^* is ambiguous).

(2) The first-order conditions for actual period-2 behavior are

$$\begin{aligned} v'(c_2^A) + R'(c_2^A - \gamma c_1^A) &= \lambda^{AA} \\ w'(l_2^A) &= \lambda^{AA} \\ c_2^A + l_2^A &= 2 - \left[c_1^A + l_1^A\right] = \hat{c}_2^A + \hat{l}_2^A \end{aligned}$$

where λ^{AA} is the Lagrange multiplier. Because $R'(c_2 - \gamma c_1^A)\dot{} \geq (1-\alpha)R'(c_2 - \gamma c_1^A)\dot{} + \alpha R'(c_2)$ for any c_2, $\lambda^{AA} \leq \lambda^A$ would imply $c_2^A > \hat{c}_2^A$ and $l_2^A \geq \hat{l}_2^A$, which would violate the budget constraint. Hence, it must be that $\lambda^{AA} > \lambda^A$, which implies $l_2^A < \hat{l}_2^A$. The budget constraint then implies $c_2^A > \hat{c}_2^A$. Q.E.D.

Proof of Lemma 5.1: (1) If the person hits in period 1, then $k_2 = \gamma k_1 + 1$; hence $C_2^*(1) = 1$ if and only if $u(1,(\gamma k_1 + 1, x_2)) \geq u(0,(\gamma k_1 + 1, x_2))$ or $x_2 \geq -\sigma(\gamma k_1 + 1)$. If the person refrains in period 1, then $k_2 = \gamma k_1$; hence $C_2^*(0) = 1$ if and only if $u(1,(\gamma k_1, x_2)) \geq u(0,(\gamma k_1, x_2))$ or $x_2 \geq -\sigma(\gamma k_1)$. Hence if $C_2^*(0) = 1$ then $C_2^*(1) = 1$.

(2) If the person hits in period 1, then $k_2 = \gamma k_1 + 1$; hence $\hat{C}_2^A(1) = 1$ if and only if $\tilde{u}(1,(\gamma k_1 + 1, x_2)|(k_1, x_1)) \geq \tilde{u}(0,(\gamma k_1 + 1, x_2)|(k_1, x_1))$ or $(1-\alpha_x)x_2 + \alpha_x x_1 \geq -\sigma[(1-\alpha_\sigma)(\gamma k_1 + 1) + \alpha_\sigma(k_1)]$. If the person refrains in period 1, then $k_2 = \gamma k_1$; hence $\hat{C}_2^A(0) = 1$ if and only if $\tilde{u}(1,(\gamma k_1, x_2)|(k_1, x_1)) \geq \tilde{u}(0,(\gamma k_1, x_2)|(k_1, x_1))$ or $(1-\alpha_x)x_2 + \alpha_x x_1 \geq -\sigma[(1-\alpha_\sigma)(\gamma k_1) + \alpha_\sigma(k_1)]$. Hence if $\hat{C}_2^A(0) = 1$ then $\hat{C}_2^A(1) = 1$. Q.E.D.

Proof of Lemma 5.2: (1) Define $V^2(k) \equiv \max_{c\in\{0,1\}} u(c,(k, x_2))$, in which case it is optimal to hit in period 1 if and only if $x_1 + \sigma k \geq \delta[V^2(\gamma k) - V^2(\gamma k + 1)]$. $V^2(k)$ is the upper envelope of the functions $u(0,(k, x_2))$ and $u(1,(k, x_2))$, and since $u(0,(k, x_2))$ and $u(1,(k, x_2))$ are decreasing

linear functions of k, $V^2(k)$ is weakly convex, and therefore $[V^2(\gamma k) - V^2(\gamma k + 1)]$ is weakly decreasing in k. Given $x_1 + \sigma k$ is increasing in k, the result follows.

(2) Define $\tilde{V}^2(k) \equiv \max_{c \in \{0,1\}} \tilde{u}(c, (k, x_2)|(k_1, x_1))$, in which case the person hits in period 1 if and only if $x_1 + \sigma k \geq \delta[\tilde{V}^2(\gamma k) - \tilde{V}^2(\gamma k + 1)]$. Since $\tilde{V}^2(k)$ is the upper envelope of the decreasing linear functions $\tilde{u}(0, (k, x_2)|(k_1, x_1))$ and $\tilde{u}(1, (k, x_2)|(k_1, x_1))$, it is weakly convex, and therefore $[\tilde{V}^2(\gamma k) - \tilde{V}^2(\gamma k + 1)]$ is weakly decreasing in k. Given $x_1 + \sigma k$ is increasing in k, the result follows. Q.E.D.

Proof of Proposition 5.1: (1) We prove $V^2(\gamma k_1) - V^2(\gamma k_1 + 1) \geq \tilde{V}^2(\gamma k_1) - \tilde{V}^2(\gamma k_1 + 1)$ for any $k_1 \geq 0$, from which the result follows. First note that for any k_1 and k_2, $V^2(k_2) - \tilde{V}^2(k_2) =$

$$\begin{aligned} & \max\{x_2 - \rho k_2, -\rho k_2 - \sigma k_2\} \\ & - \max\{x_2 - \rho[(1-\alpha_\rho)k_2 + \alpha_\rho k_1], -\rho[(1-\alpha_\rho)k_2 + \alpha_\rho k_1] - \sigma[(1-\alpha_\sigma)k_2 + \alpha_\sigma k_1]\} \\ = \; & \rho\alpha_\rho(k_1 - k_2) + \max\{x_2, -\sigma k_2\} - \max\{x_2, -\sigma[(1-\alpha_\sigma)k_2 + \alpha_\sigma k_1]\}. \end{aligned}$$

If $k_2 = \gamma k_1 \leq k_1$, then $\rho\alpha_\rho(k_1 - k_2) \geq 0$ and $\max\{x_2, -\sigma k_2\} \geq \max\{x_2, -\sigma[(1-\alpha_\sigma)k_2 + \alpha_\sigma k_1]\}$, and therefore $V^2(\gamma k_1) - \tilde{V}^2(\gamma k_1) \geq 0$. If $k_2 = \gamma k_1 + 1 \geq k_1$, then $\rho\alpha_\rho(k_1 - k_2) \leq 0$ and $\max\{x_2, -\sigma k_2\} \leq \max\{x_2, -\sigma[(1-\alpha_\sigma)k_2 + \alpha_\sigma k_1]\}$, and therefore $V^2(\gamma k_1 + 1) - \tilde{V}^2(\gamma k_1 + 1) \leq 0$. Then $V^2(\gamma k_1) - \tilde{V}^2(\gamma k_1) \geq 0 \geq V^2(\gamma k_1 + 1) - \tilde{V}^2(\gamma k_1 + 1)$ implies $V^2(\gamma k_1) - V^2(\gamma k_1 + 1) \geq \tilde{V}^2(\gamma k_1) - \tilde{V}^2(\gamma k_1 + 1)$.

(2) Because period 2 is the final period, $C_2^A(c_1) = C_2^*(c_1)$ for $c_1 \in \{0, 1\}$. Given $c_2^* = C_2^*(c_1^*)$ and $c_2^A = C_2^A(c_1^A) = C_2^*(c_1^A)$, it immediately follows that $c_1^A = c_1^*$ implies $c_2^A = c_2^*$. That $c_1^A > c_1^*$ implies $c_2^A \geq c_2^*$ follows from Lemma 5.1. Q.E.D.

Proof of Proposition 5.2: Because period 2 is the final period, $C_2^A(c_1) = C_2^*(c_1)$ for $c_1 \in \{0, 1\}$. The proof of part 1 of Lemma 5.1 then implies that $C_2^A(1) = 1$ if and only if $x_2 \geq -\sigma(\gamma k_1 + 1)$, and $C_2^A(0) = 1$ if and only if $x_2 \geq -\sigma(\gamma k_1)$. Given $\alpha_x = 0$, the proof of part 2 of Lemma 5.1 implies $\hat{C}_2^A(1) = 1$ if and only if $x_2 \geq -\sigma[(1-\alpha_\sigma)(\gamma k_1 + 1) + \alpha_\sigma(k_1)]$, and $\hat{C}_2^A(0) = 1$ if and only if $x_2 \geq -\sigma[(1-\alpha_\sigma)(\gamma k_1) + \alpha_\sigma(k_1)]$. Clearly $\alpha_\sigma = 0$ implies $\hat{C}_2^A(c_1) = C_2^A(c_1)$ for $c_1 \in \{0, 1\}$. If $\alpha_\sigma > 0$, then $[(1-\alpha_\sigma)(\gamma k_1 + 1) + \alpha_\sigma(k_1)] \leq \gamma k_1 + 1$ and $[(1-\alpha_\sigma)(\gamma k_1) + \alpha_\sigma(k_1)] \geq \gamma k_1$, and it follows that $\hat{C}_2^A(1) \leq C_2^A(1)$ and $\hat{C}_2^A(0) \geq C_2^A(0)$. Q.E.D.

Proof of Proposition 5.3: Again, $C_2^A(1) = 1$ if and only if $x_2 \geq -\sigma(\gamma k_1 + 1)$, and $C_2^A(0) = 1$ if

and only if $x_2 \geq -\sigma(\gamma k_1)$. Given $\alpha_\sigma = 0$, the proof of part 2 of Lemma 5.1 implies $\hat{C}_2^A(1) = 1$ if and only if $(1-\alpha_x)x_2 + \alpha_x x_1 \geq -\sigma(\gamma k_1 + 1)$, and $\hat{C}_2^A(0) = 1$ if and only if $(1-\alpha_x)x_2 + \alpha_x x_1 \geq -\sigma(\gamma k_1)$. Then $x_1 > x_2$ implies $(1-\alpha_x)x_2 + \alpha_x x_1 > x_2$, and therefore $\hat{C}_2^A(c_1) \geq C_2^A(c_1)$ for $c_1 \in \{0,1\}$. Similarly, $x_1 < x_2$ implies $(1-\alpha_x)x_2 + \alpha_x x_1 < x_2$, and therefore $\hat{C}_2^A(c_1) \leq C_2^A(c_1)$ for $c_1 \in \{0,1\}$. Q.E.D.

Proof of Proposition 5.4: First note that for any x_1, x_2, and k_1,

$$\begin{aligned} V^2(\gamma k_1) - V^2(\gamma k_1 + 1) &= \max\{x_2, -\sigma\gamma k_1\} - \max\{x_2, -\sigma(\gamma k_1 + 1)\} \qquad \text{and} \\ \tilde{V}^2(\gamma k_1) - \tilde{V}^2(\gamma k_1 + 1) &= \max\{[(1-\alpha_x)x_2 + \alpha_x x_1], -\sigma\gamma k_1\} - \\ &\quad \max\{[(1-\alpha_x)x_2 + \alpha_x x_1], -\sigma(\gamma k_1 + 1)\}. \end{aligned}$$

Next note that $\max\{X, -\sigma\gamma k_1\} - \max\{X, -\sigma(\gamma k_1 + 1)\}$ is weakly decreasing in X because

$$\max\{X, -\sigma\gamma k_1\} - \max\{X, -\sigma(\gamma k_1 + 1)\} = \begin{cases} \sigma & \text{if } X \leq -\sigma(\gamma k_1 + 1) \\ -\sigma\gamma k_1 - X & \text{if } X \in [-\sigma(\gamma k_1 + 1), -\sigma\gamma k_1] \\ 0 & \text{if } X \geq -\sigma\gamma k_1. \end{cases}$$

Hence, if $x_2 > x_1$ then $[(1-\alpha_x)x_2 + \alpha_x x_1] < x_2$, in which case $\tilde{V}^2(\gamma k_1) - \tilde{V}^2(\gamma k_1 + 1) \leq V^2(\gamma k_1) - V^2(\gamma k_1 + 1)$ for all $k_1 \geq 0$ and therefore $c_1^A \geq c_1^*$. Similarly, if $x_2 < x_1$ then $[(1-\alpha_x)x_2 + \alpha_x x_1] > x_2$, in which case $\tilde{V}^2(\gamma k_1) - \tilde{V}^2(\gamma k_1 + 1) \geq V^2(\gamma k_1) - V^2(\gamma k_1 + 1)$ for all $k_1 \geq 0$ and therefore $c_1^A \leq c_1^*$. Finally, since $c_2^* = C_2^*(c_1^*)$ and $c_2^A = C_2^A(c_1^A) = C_2^*(c_1^A)$, and since $C_2^*(1) \geq C_2^*(0)$ by Lemma 5.1, $c_1^A \geq c_1^*$ implies $c_2^A \geq c_2^*$ and $c_1^A \leq c_1^*$ implies $c_2^A \leq c_2^*$. Q.E.D.

Proof of Lemma 6.1: From the inequalities in the text, we can derive

$$\begin{aligned} \bar{\mu}^*(P,\phi) &= (1-\delta)P - [(1-\delta)/(1-\delta\gamma)]\phi \qquad \text{and} \\ \bar{\mu}^A(P,\phi) &= (1-\delta)P - [(1-\alpha)(1-\delta)/(1-\delta\gamma) + \alpha]\phi. \end{aligned}$$

It follows that $\bar{\mu}^*(P,\phi) - \bar{\mu}^A(P,\phi) = [\alpha\delta(1-\gamma)/(1-\delta\gamma)]\phi$. Q.E.D.

Proof of Proposition 6.1: (1) First consider the general problem: If $\tilde{D}(P,\phi) = X - AP + B\phi$ and the dealer maximizes $(P-d)\cdot\tilde{D}(P,\phi) - C(\phi)$, then the optimal $(\tilde{P},\tilde{\phi})$ satisfy

$$\begin{aligned} \tilde{P} &= [X + B\tilde{\phi} + Ad]/2A \\ C'(\tilde{\phi}) &= B\tilde{P} - Bd, \end{aligned}$$

where $X > Ad$, $C'(0) = 0$, and $C''(\phi) > B^2/(2A)$ for all ϕ guarantees a unique interior solution. Combining these equations yields

$$C'(\tilde{\phi}) = [B(X - Ad)]/2A + [B^2/(2A)]\tilde{\phi}.$$

It is straightforward to show $\tilde{\phi}$ is decreasing in A and increasing in B, and therefore $\tilde{P}$ is decreasing in A and increasing in B.

For optimal behavior, $A^* = 1 - \delta$ and $B^* = (1-\delta)/(1-\delta\gamma)$; for actual behavior $A^A = 1 - \delta$ and $B^A = (1-\alpha)(1-\delta)/(1-\delta\gamma) + \alpha$. The assumptions $X > d(1-\delta)/\delta$, $C'(0) = 0$, and $C''(\phi) > [(1-\alpha)(1-\delta)/(1-\delta\gamma) + \alpha]^2/[2(1-\delta)]$ for all ϕ guarantee a unique interior solution for both optimal and actual behavior; and $A^A = A^*$ and $B^A > B^*$ imply $\phi^A > \phi^*$ and $P^A > P^*$.

(2) The result follows from $\phi^A > \phi^*$ and Lemma 6.1. Q.E.D.

Proof of Lemma 6.2: (1) From the inequalities in the text, we can derive

$$\begin{aligned} \bar{\mu}^{**}(P,\phi) &= [(1-\delta)/\delta]P - [(1-\delta)/(1-\delta\gamma)]\gamma\phi \qquad \text{and} \\ \bar{\mu}^{AA}(P,\phi) &= [(1-\delta)/\delta]P - [(1-\alpha)(1-\delta)/(1-\delta\gamma) + \alpha]\gamma\phi. \end{aligned}$$

It follows that $\bar{\mu}^{**}(P,\phi) - \bar{\mu}^{*}(P,\phi) = [(1-\delta)^2/\delta]P + [(1-\delta)/(1-\delta\gamma)](1-\gamma)\phi > 0$ and $\bar{\mu}^{AA}(P,\phi) - \bar{\mu}^{A}(P,\phi) = [(1-\delta)^2/\delta]P + [(1-\alpha)(1-\delta)/(1-\delta\gamma) + \alpha](1-\gamma)\phi > 0$.

(2) $\bar{\mu}^{**}(P,\phi) - \bar{\mu}^{AA}(P,\phi) = [\alpha\delta(1-\gamma)/(1-\delta\gamma)]\gamma\phi < [\alpha\delta(1-\gamma)/(1-\delta\gamma)]\phi$, and the result follows from Lemma 6.1. Q.E.D.

Proof of Proposition 6.2: (1) The dealer's problem falls within the general form in the proof of Proposition 6.1, where $A^{AA} = (1-\delta)/\delta$ and $B^{AA} = [(1-\alpha)(1-\delta)/(1-\delta\gamma) + \alpha]\gamma$. The assumptions $X > d(1-\delta)/\delta$, $C'(0) = 0$, and $C''(\phi) > [(1-\alpha)(1-\delta)/(1-\delta\gamma) + \alpha]^2/[2(1-\delta)]$ for all ϕ guarantee a unique interior solution; and $A^{AA} > A^A$ and $B^{AA} < B^A$ imply $\phi^{AA} < \phi^A$ and $P^{AA} < P^A$.

(2) $\bar{\mu}^{**}(P,\phi) - \bar{\mu}^{AA}(P,\phi) = [\alpha\delta(1-\gamma)/(1-\delta\gamma)]\gamma\phi$ is increasing in ϕ, so $\phi^{AA} < \phi^A$ implies $\bar{\mu}^{**}(P^{AA},\phi^{AA}) - \bar{\mu}^{AA}(P^{AA},\phi^{AA}) < \bar{\mu}^{**}(P^{A},\phi^{A}) - \bar{\mu}^{AA}(P^{A},\phi^{A})$. The second inequality follows from Lemma 6.2. Q.E.D.

References

Ainslie, George (1992). *Picoeconomics: The strategic interaction of successive motivational states within the person*. New York: Cambridge University Press.

Becker, G. and Murphy, K. (1988). "A Theory of Rational Addiction." *Journal of Political Economy*, **96**, pp. 675-700.

Bowman, D., Minehart, D., and Rabin, M. (1999). "Loss Aversion in a Consumption-Savings Model." *Journal of Economic Behavior and Organization*, **38**, pp. 155-178.

Brickman, P., Coates, D. and Janoff-Bulman, R. (1978). "Lottery winners and accident victims: is happiness relative?" *Journal of Personality and Social Psychology*, **36**, pp. 917-27.

Duesenberry, J. S. (1952). *Income, Saving and the Theory of Consumer Behavior*. Cambridge, MA: Harvard University Press.

Frank, Robert (1985). *Choosing the Right Pond: Human Behavior and the Quest for Status*. Oxford University Press.

Frank, Robert (1999). *Luxury Fever*. New York: Free Press.

Frederick, S. and Loewenstein, G. (1999). "Hedonic Adaptation," in Daniel Kahneman, Edward Diener, and Norbert Schwarz, eds., *Well-Being: The Foundations of Hedonic Psychology*. New York: Russell Sage Foundation Press, pp. 302-329.

Gilbert, Daniel T., Gill, Michael J. and Wilson, Timothy D. (1998). "How do we know what we will like? The informational basis of affective forecasting." Working paper, Harvard University Department of Psychology.

Gilbert, Daniel T., Pinel, E. C., Wilson, Timothy D., Blumberg, S. J., and Wheatley, T. (1997). "Immune neglect: A source of durability bias in affective forecasting." *Journal of Personality and Social Psychology*, **75**(3), pp. 617-638.

Helson, Harry (1964). *Adaptation-level theory: An experimental and systematic approach to behavior*. New York: Harper and Row.

Herrnstein, R.J., Loewenstein, G., Prelec, D., and Vaughan, W., Jr. (1993). "Utility Maximization and Melioration: Internalities in Individual Choice." *Journal of Behavioral Decision Making*, **6**, pp. 149-185.

Kahneman, Daniel (1991). "Judgement and Decision-Making: A Personal View." *Psychological Science*, 2(3), pp. 142-145.

Kahneman, Daniel, Knetsch, Jack L., and Thaler, Richard H. (1991). "The Endowment Effect, Loss Aversion, and Status Quo Bias: Anomalies." *Journal of Economic Perspectives*, **5**(Winter), pp. 193-206.

Kahneman, Daniel and Tversky, Amos (1979). "Prospect Theory: An Analysis of Decision Under Risk." *Econometrica*, **47**, pp. 263-291.

Laibson, David (1994). "Essays on Hyperbolic Discounting." Dissertation, MIT Department of Economics.

Laibson, David (1997). "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics*, **112**(2), pp. 443-477.

Loewenstein, George (1996). "Out of Control: Visceral Influences on Behavior." *Organizational Behavior and Human Decision Processes*, **65**, pp. 272-92.

Loewenstein, George (1999). "A visceral account of addiction," in Jon Elster and Ole-Jørgen Skog, eds., *Getting Hooked: Rationality and Addiction.* Cambridge, England: Cambridge University Press, pp. 235-264.

Loewenstein, George and Adler, D. (1995). "A Bias in the Prediction of Tastes." *Economic Journal*, **105**, pp. 929-37.

Loewenstein, George and Frederick, S. (1997). "Predicting Reactions to Environmental Change," in M. Bazerman, D. Messick, A. Tenbrunsel, and K. Wade-Benzoni, eds., *Environment, Ethics, and Behavior*. San Francisco, CA: The New lexington Press, pp. 52-72.

Loewenstein, G., Nagin, D. and Paternoster, R. (1997). "The effect of sexual arousal on predictions of sexual forcefulness." *Journal of Crime and Delinquency*, **34**, pp. 443-73.

Loewenstein, G., Prelec, D. and Shatto (1996). "Hot/cold intrapersonal empathy gaps and the prediction of curiosity." Working paper, Carnegie Mellon University.

Loewenstein, G. and Schkade, D. (1999). "Wouldn't it be Nice? Predicting Future Feelings," in Daniel Kahneman, Edward Diener, and Norbert Schwarz, eds., *Well-Being: The Foundations of Hedonic Psychology*. New York: Russell Sage Foundation Press, pp. 85-105.

Nisbett, Richard E. and Kanouse, D.E. (1968). "Obesity, hunger, and supermarket shopping behavior." *Proceedings. American Psychological Association Annual Convention*, **3**, pp. 683-84.

O'Donoghue, Ted and Rabin, Matthew (1999*a*). "Doing It Now or Later." *American Economic Review*, **89**, pp. 103-124.

O'Donoghue, Ted and Rabin, Matthew (1999*b*), "Incentives for Procrastinators." *Quarterly Journal of Economics*, **114**(3), pp.769-816

O'Donoghue, Ted and Rabin, Matthew (1999*c*). "Addiction and Present-Biased Preferences." Mimeo, Cornell University and U.C. Berkeley.

O'Donoghue, Ted and Rabin, Matthew (1999*d*). "Addiction and Self Control," in Jon Elster, ed., *Addiction: Entries and Exits*. New York: Russell Sage Foundation, pp. 196-206.

Orphanides, A., and Zervos, D. (1995). "Rational addiction with learning and regret." *Journal of Political Economy*, **103**, pp. 739-758.

Oswald, Andrew J. (1997). "Happiness and economic performance." *Economic Journal*, **107**, pp. 1815-1831

Pollak, R. A. (1970). "Habit Formation and Dynamic Demand Functions." *Journal of Political Economy*, **78**(4), pp. 745-763.

Read, Daniel, and van Leeuwen, Barbara (1998). "Predicting hunger. The effects of appetite and delay on choice." *Organizational Behavior and Human Decision Processes*, **76**, pp. 189-205.

Ryder, H. E. and Heal, G. M. (1973). "Optimal Growth with Intermporally Dependent Preferences." *Review of Economic Studies*, **40**, pp. 1-33.

Scitovsky, T. (1976). *The Joyless Economy: An Inquiry into Human Satisfaction*. Oxford University Press.

Sieff, E.M., Dawes, R.M., and Loewenstein, George F. (1999). "Anticipated versus actual responses to HIV test results." *American Journal of Psychology*, **112**(2), pp. 297-311.

Solomon, Andrew (1998). "Personal History: Anatomy of Melancholy." *New Yorker*, LXXIII, Jan 12, pp. 46-61.

Strahilevitz, Michal and Loewenstein, George (1998). "The effect of ownership history on the valuation of objects." *Journal of Consumer Research*, **25**, pp. 276-289.

Strotz, R. (1956). "Myopia and Inconsistency in Dynamic Utility Maximization." *Review of Economic Studies*, **23**, pp. 165-180.

Thaler, R. H. (1991). "Some Empirical Evidence on Dynamic Inconsistency." *Quasi Rational Economics*, Russell Sage Foundation, pp. 127-133.

Tversky, Amos, and Kahneman, Daniel (1991). "Loss Aversion in Riskless Choice: A Reference-Dependent Model." *Quarterly Journal of Economics*, **106**(4), pp. 1039-61.

VanBoven, Leaf, Dunning, David, and Loewenstein, George (1999). "Trading places: egocentric empathy gaps between owners and buyers." Working paper, Department of Psychology, Cornell University.