A MODEL OF REFERENCE-DEPENDENT PREFERENCES

Botond Kőszegi and Matthew Rabin

Abstract

We develop a model of reference-dependent preferences and loss aversion where "gain-loss utility" is derived from standard "consumption utility" and the reference point is determined endogenously by the economic environment. We assume that a person's reference point is her rational expectations held in the recent past about outcomes, which are determined in a *personal equilibrium* by the requirement that they must be consistent with optimal behavior given expectations. In deterministic environments, choices maximize consumption utility, but gain-loss utility influences behavior when there is uncertainty. Applying the model to consumer behavior, we show that willingness to pay for a good is increasing in the expected probability of purchase and in the expected prices conditional on purchase. In within-day labor-supply decisions, a worker is less likely to continue work if income earned thus far is unexpectedly high, but more likely to show up as well as continue work if expected income is high.

Keywords: Expectations, Loss Aversion, Prospect Theory, Reference-Dependent Preferences

JEL Classification: B49

I.   Introduction

How a person assesses the outcome of a choice is often determined as much by its contrast with a reference point as by intrinsic taste for the outcome itself. The most notable manifestation of such reference-dependent preferences is loss aversion: losses resonate more than same-sized gains. In this paper, we build on the essential intuitions in Kahneman and Tversky's [1979] prospect theory and subsequent models of reference dependence, but flesh out, extend, and modify these models to develop a more generally applicable theory. We illustrate such applicability by establishing some implications of loss aversion for consumer behavior and labour effort.

We present the basic framework in Section II. A person's utility depends not only on her K-dimensional consumption bundle, c, but also on a reference bundle, r. She has an intrinsic "consumption utility" $m(c)$ that corresponds to the outcome-based utility classically studied in economics. Overall utility is given by $u(c|r) \equiv m(c) + n(c|r)$, where $n(c|r)$ is "gain-loss utility." Both consumption utility and gain-loss utility are separable across dimensions, so that $m(c) \equiv \sum_k m_k(c_k)$ and $n(c|r) \equiv \sum_k n_k(c_k|r_k)$. Because the sensation of gain or loss due to a departure from the reference point seems closely related to the consumption value attached to the goods in question, we assume $n_k(c_k|r_k) \equiv \mu(m_k(c_k) - m_k(r_k))$, where $\mu(\cdot)$ satisfies the properties of Kahneman and Tversky's [1979] value function. Our model allows for both stochastic outcomes and stochastic reference points, and assumes that a stochastic outcome F is evaluated according to its expected utility, with the utility of each outcome being the average of how it feels relative to each possible realization of the reference point G: $U(F|G) = \iint u(c|r) dG(r) dF(c)$.

In addition to the widely investigated question of how people react to departures from a posited reference point, predictions of reference-dependent theories also depend crucially on the understudied issue of what the reference point is. In Section III, we propose that a person's reference point is the probabilistic beliefs she held in the recent past about outcomes. Although existing evidence is instead generally interpreted by equating the reference point with the status quo, virtually all of this evidence comes from contexts where people plausibly expect to maintain the status quo. But when expectations and the status quo are different—a common situation in economic environments— equating the reference point with expectations generally makes better predictions. Our theory, for

instance, supports the common view that the "endowment effect" found in the laboratory, whereby random owners value an object more than non-owners, is due to loss aversion—since an owner's loss of the object looms larger than a non-owner's gain of the object. But our theory makes the less common prediction that the endowment effect among such owners and non-owners with no predisposition to trade will disappear among sellers and buyers in real-world markets who *expect* to trade. Merchants do not assess intended sales as loss of inventory, but do assess failed sales as loss of money; buyers do not assess intended expenditures as losses, but do assess failures to carry out intended purchases or paying more than expected as losses. Equating the reference point with expectations rather than the status quo is also important for understanding financial risk: while an unexpected monetary windfall in the lab may be assessed as a gain, a salary of $50,000 to an employee who expected $60,000 will not be assessed as a large gain relative to status-quo wealth, but rather as a loss relative to expectations of wealth. And in non-durable consumption—where there is no object with which the person can be endowed—a status-quo-based theory cannot capture the role of reference dependence at all: it would predict, for instance, that a person who misses a concert she expected to attend would feel no differently than somebody who never expected to see the concert.

While alternative theories of expectations formation could be used, in this paper we complete our model by assuming rational expectations, formalizing (in an extreme way) the realistic assumption that people have some ability to predict their own behavior. Using the framework of Kőszegi [2005] to determine rational expectations when preferences depend on expectations, we define a "personal equilibrium" as a situation where the stochastic outcome implied by optimal behavior conditional on expectations coincides with expectations.[1]   We also define a notion of "preferred personal equilibrium," that selects the (typically unique) personal equilibrium with highest expected utility.

We show in Section III that in deterministic environments preferred personal equilibrium predicts that decisionmakers maximize consumption utility, replicating the predictions of classical reference-independent utility theory. Our analyses in Sections IV and V of consumer and labour-

---

[1] Expectations have been mentioned by many researchers as a candidate for the reference point. With the exception of Shalev's [2000] game-theoretic model, however, to our knowledge our paper is the first formalize the idea that expectations determine the reference point and to specify a rule for deriving them endogenously in any environment.

supply behavior, however, demonstrate a central implication of our theory: when there is uncertainty, a decisionmaker's preferences over consumption bundles will be influenced by her environment.

Section IV shows that a consumer's willingness to pay a given price for shoes depends on the probability with which she expected to buy them and the price she expected to pay. On the one hand, an increase in the likelihood of buying increases a consumer's sense of loss of shoes if she does not buy, creating an "attachment effect" that increases her willingness to pay. Hence, the greater the likelihood she thought prices would be low enough to induce purchase, the greater is her willingness to buy at higher prices. On the other hand, holding the probability of getting the shoes fixed, a decrease in the price a consumer expected to pay makes paying a higher price feel like more of a loss, creating a "comparison effect" that lowers her willingness to pay the high price. Hence, the lower the prices she expected among those prices that induce purchase, the lower is her willingness to buy at higher prices.

Our application in Section V to labour supply is motivated by some recent empirical research beginning with Camerer, Babcock, Loewenstein, and Thaler [1997] on flexible work hours that finds some workers seem to have a daily "target" income. We develop a model where, after earning income in the morning and learning her afternoon wage, a taxi driver decides whether to continue driving in the afternoon. In line with the empirical results of the target-income literature, our model predicts that when drivers experience *unexpectedly* high wages in the morning, for any given afternoon wage they are less likely to continue work. Yet *expected* wage increases will tend to increase both willingness to show up to work, and to drive in the afternoon once there. Our model therefore replicates the key insight of the literature that exceeding a target income might reduce effort. But in addition, it both provides a theory of what these income targets will be, and—through the fundamental distinction between unexpected and expected wages—avoids the unrealistic prediction that generically higher wages will lower effort.

Beyond improvements in specific substantive predictions, our general approach has an attractive methodological feature: because a full specification of $\mu(\cdot)$ allows us to derive both gain-loss utility and the reference point itself from consumption utility and the economic environment, it moves us

3

closer to a universally applicable, zero-degrees-of-freedom way to translate any existing reference-independent model into the corresponding reference-dependent one. Although straightforward to apply in most cases, however, our model falls short of providing a recipe for entirely formulaic application of the principles of reference-dependence. Psychological and economic judgment is needed, for instance, in choosing the appropriate notion of "recent expectations." And there are also settings where the same principles motivating our approach suggest an alternative to our reduced-form model. We discuss such shortcomings and gaps, some possible resolutions, as well as further economic applications, in Section VI.

II. Reference-Dependent Utility

We specify a person's utility for a riskless outcome as $u(c|r)$, where $c = (c_1, c_2, \ldots, c_K) \in \mathbb{R}^K$ is consumption and $r = (r_1, r_2, \ldots, r_K) \in \mathbb{R}^K$ is a "reference level" of consumption. If c is drawn according to the probability measure F, the person's utility is given by[2]

$$(1) \qquad U(F|r) = \int u(c|r)dF(c).$$

As is clearly necessary in our framework developed below, where we assume the reference point is beliefs about outcomes, we allow for the reference point itself to be stochastic. Suppose the person's reference point is the probability measure G over $\mathbb{R}^K$, and her consumption is drawn according to the probability measure F. Then, her utility is

$$(2) \qquad U(F|G) = \iint u(c|r)dG(r)dF(c).$$

This formulation captures the notion that the sense of gain or loss from a given consumption outcome derives from comparing it to all outcomes possible under the reference lottery. For example, if the reference lottery is a gamble between \$0 and \$100, an outcome of \$50 feels like a gain relative

---

[2] Despite the clear evidence that people's evaluation of prospects is not linear in probabilities, our model simplifies things by assuming preferences are linear.

to \$0, and like a loss relative to \$100, and the overall sensation is a mixture of these two feelings.[3] That a person's utility depends on a reference lottery in addition to the actual outcome is similar to several previous theories.[4] With the exception of the model in Gul [1991]—which can be interpreted as saying that the certainty equivalent of a chosen lottery becomes the decisionmaker's reference point—none of these theories provide a theory of reference-point determination, as we do below.

While preferences are reference-dependent, gains and losses are clearly not all that people care about. The sensation of gain or avoided loss from having more money does significantly affect our utility—but so does the absolute pleasure of consumption we purchase with the money. In contrast to prior formulations based on a "value function" defined solely over gains and losses, therefore, our approach makes explicit the way preferences also depend on absolute levels. We assume that overall utility has two components: $u(c|r) \equiv m(c) + n(c|r)$, where $m(c)$ is "consumption utility" typically stressed in economics, and $n(c|r)$ is "gain-loss utility."

For simplicity—and for further reasons discussed in Kőszegi and Rabin [2004]—we assume consumption utility is additively separable across dimensions: $m(c) \equiv \sum_{k=1}^{K} m_k(c_k)$, with each $m_k(\cdot)$ differentiable and strictly increasing. We also assume gain-loss utility is separable: $n(c|r) \equiv \sum_{k=1}^{K} n_k(c_k|r_k)$. Thus, in evaluating an outcome, the decisionmaker assesses gain-loss utility in each dimension separately. In combination with loss aversion, this separability is at the crux of many implications of reference-dependent utility, including the endowment effect.

Beyond saying that a person cares about both consumption and gain-loss utility, we propose a strong relationship between the two. While it surely exaggerates the tightness of the connection,

---

[3] See Larsen, McGraw, Mellers, and Cacioppo [2004] for some evidence that subjects have mixed emotions for outcomes that compare differently to different counterfactuals. Given the features below, our formula also implies that losses relative to a stochastic reference point count more than gains, so that a person who gets \$50 is more distressed by how it compares to a possible \$100 than she is pleased by how it compares to a possible \$0. We are unaware of evidence on whether this, or an alternative specification where the relief of avoiding \$0 outweighs the disappointment of not getting the \$100, is closer to reality. We believe few of the results stressed in this paper depend qualitatively on our exact formulation. We discuss this assumption in more detail in Kőszegi and Rabin [2005] in the context of risk preferences, where it is crucial.

[4] Our utility function is most closely related to Sugden's [2003]. The main difference is in the way a given consumption outcome is compared to the reference lottery. In our model, each outcome is compared to all outcomes in the support of the reference lottery. In Sugden [2003], an outcome is compared only to the outcome that would have resulted from the reference lottery in the same state. See also the axiomatic theories of Gul [1991], Masatlioglu and Ok [2005] and Sagi [2005].

our model assumes that how a person feels about gaining or losing in a dimension depends in a universal way on the changes in consumption utility associated with such gains or losses:[5]

$$n_k(c_k|r_k) \equiv \mu\left(m_k(c_k) - m_k(r_k)\right),$$

where $\mu(\cdot)$ is a "universal gain-loss function."[6]   We assume that $\mu(\cdot)$ satisfies the following properties:

A0. $\mu(x)$ is continuous for all x, twice differentiable for $x \neq 0$, and $\mu(0) = 0$.

A1. $\mu(x)$ is strictly increasing.

A2. If $y > x > 0$, then $\mu(y) + \mu(-y) < \mu(x) + \mu(-x)$.

A3. $\mu''(x) \leq 0$ for $x > 0$ and $\mu''(x) \geq 0$ for $x < 0$.

A4. $\mu'_-(0)/\mu'_+(0) \equiv \lambda > 1$, where $\mu'_+(0) \equiv \lim_{x \to 0} \mu'(|x|)$ and $\mu'_-(0) \equiv \lim_{x \to 0} \mu'(-|x|)$.

Assumptions A0-A4, first stated by Bowman, Minehart, and Rabin [1999], correspond to Kahneman and Tversky's [1979] explicit or implicit assumptions about their "value function" defined on $c - r$. Loss aversion is captured by A2 for large stakes and A4 for small stakes. Assumption A3 captures another important feature of gain-loss utility, diminishing sensitivity: the marginal change in gain-loss sensations is greater for changes that are close to one's reference level than for changes that are further away. We shall sometimes be interested in characterizing the implications of reference dependence with loss aversion but without diminishing sensitivity as a force on behavior. For doing so, we define an alternative to A3:

---

[5] In a single-dimensional model, Köbberling and Wakker [2005] also assume that the evaluation of gains and losses is related to consumption utility. Our formulation extends their insight to multiple dimensions. As one way to motivate that gain-loss utility is related to changes in consumption utility in each dimension, consider a person choosing between two gambles: a 50-50 chance of gaining a paper clip or losing a paper clip, and the comparable gamble involving $10 bills. It seems likely that she would risk losing the paper clip rather than the money, and do so because her sensation of gains and losses is smaller for a good whose consumption utility is smaller. Yet since $m(\cdot)$ is approximately linear for such small stakes, the choice depends almost entirely on the comparison of $n_k(\cdot)$ across dimensions, so that any model that does not relate gain-loss assessments to consumption utility is not equipped to provide guidance in this or related examples.

[6] Note that once $\mu(\cdot)$ is fixed, unless A3$'$ below holds affine transformations of $m(\cdot)$ will not in general result in affine transformations of our model's overall utility function. While this raises no problem in applying our model once the full utility function $u(\cdot|\cdot)$ is specified (or empirically estimated), it does mean that when deriving our model from a reference-independent model based on consumption utility alone, the specification of $\mu(\cdot)$ must be sensitive to the scaling of consumption utility.

A3′. For all $x \neq 0$, $\mu''(x) = 0$.

This utility function replicates a number of properties commonly associated with reference-dependent preferences. Proposition 1 establishes that fixing the outcome, a lower reference point makes a person happier (Part 1); and preferences exhibit a status quo bias (Parts 2 and 3):

Proposition 1 *If μ satisfies Assumptions A0-A4, then the following hold.*

1. *For all F, G, G' such that for all $k \in \{1, \ldots, K\}$, the marginals of G' first-order stochastically dominate the marginals of G in each dimension, $U(F|G) \geq U(F|G')$.*

2. *For any $c, c' \in \mathbb{R}^K$, $c \neq c'$, $u(c|c') \geq u(c'|c') \Rightarrow u(c|c) > u(c'|c)$.*

3. *Suppose μ satisfies A3′. Then, for any F, F' that do not generate the same distribution of outcomes in all dimensions, $U(F|F') \geq U(F'|F') \Rightarrow U(F|F) > U(F'|F)$.*

Parts 2 and 3 mean that if a person is willing to abandon her reference point for an alternative, then she strictly prefers the alternative if that is her reference point. Under Assumptions A0-A4, this is always true for riskless consumption bundles, but counterexamples can be constructed to show that the analogous statement for lotteries requires a more restrictive assumption such as $A3'$.

Proposition 2 establishes that in the special case where $m(\cdot)$ is linear, our utility function $u(c|r)$ exhibits the same properties as $\mu(\cdot)$:

Proposition 2 *If m is linear and μ satisfies Assumptions A0-A4, then there exists $\{v_k\}_{k=1}^K$ satisfying Assumptions A0-A4 such that, for all c and r,*

$$(3) \qquad u(c|r) - u(r|r) = \sum_{k=1}^K v_k(c_k - r_k).$$

Insofar as for local changes $m(\cdot)$ can be taken to be much closer to linear than $\mu(\cdot)$, Proposition 2 says that for small changes our utility function shares the qualitative properties of standard formulations of prospect theory.[7] This equivalence does *not* hold when the changes are large or marginal consumption utilities change quickly. This is a good thing. If, for instance, a person's

---

[7] The proof of Proposition 2 shows that, quantitatively, the degree of loss aversion observed in $u(c|r)$ is less than the degree assumed in $\mu(\cdot)$.

reference level of water is a quart below the level needed for survival, loss aversion in $\mu(\cdot)$ will not induce loss aversion in $u(c|r)$: she would be much happier about a one-quart increase in water consumption than she would be unhappy about a one-quart decrease. More importantly, when large losses in consumption or wealth are involved, diminishing marginal utility of wealth as economists conventionally conceive of it is likely to counteract the diminishing sensitivity in losses emphasized in prospect theory.[8]

III.   The Reference Point as (Endogenous) Expectations

In comparison to the extensive research on preferences over departures from posited reference points, research on the nature of reference points themselves is quite limited. While we hope that experiments and other empirical work will shed light on this topic, our model makes the extreme assumption that the reference point is fully determined by the *expectations* a person held in the *recent past*. Specifically, a person's reference point is her probabilistic beliefs about the relevant consumption outcome held between the time she first focused on the decision determining the outcome and shortly before consumption occurs.[9]   While some evidence indicates that expectations are important in determining sensations of gain and loss, our primary motivation for this assumption is that it helps unify and reconcile existing interpretations and corresponds to readily accessible intuition in many examples.[10]

The most common assumption, of course, has been that the reference point is the status quo. But in virtually all experiments interpreted as supporting this assumption, subjects plausibly expect to keep the status quo, so these studies are also consistent with the reference point being

---

[8] The tension between consumption utility and gain-loss utility in the evaluation of losses has been emphasized by prior researchers; see, e.g., Kahneman [2003] and Köbberling, Schwieren, and Wakker [2004].

[9] Our theory posits that preferences depend on *lagged* expectations, rather than expectations contemporaneous with the time of consumption. This does not assume that beliefs are slow to adjust to new information or that people are unaware of the choices that they have just made—but that preferences do not instantaneously change when beliefs do. When somebody finds out 5 minutes ahead of time that she will for sure not receive a long-expected $100, she would presumably immediately adjust her expectations to the new situation, but she will still 5 minutes later assess not getting the money as a loss.

[10] For examples of some of the more direct evidence of expectations-based counterfactuals affecting reactions to outcomes, see Mellers, Schwartz, and Ritov [1999], Breiter, Aharon, Kahneman, Dale, and Shizgal [2001], and Medvec, Madey, and Gilovich [1995].

expectations. For instance, procedures that have generated the classic endowment effect might plausibly have induced a disposition of subjects to believe that their current ownership status is indicative of their ensuing ownership status, so that our expectations-based theory makes the same prediction as a status-quo-based theory. Indeed, our theory may be useful for understanding instances where the endowment effect has *not* been found. One interpretation of the rare exceptions to laboratory findings of the effect, such as Plott and Zeiler [2005], is that they have successfully de-coupled subjects' expectations from their initial ownership status. Similarly, the field experiment by List [2003], which replicates the effect for inexperienced sports card collectors but finds that experienced collectors show a much smaller, insignificant effect, is consistent with our theory if more experienced traders come to expect a high probability of parting with items they have just acquired. And the important limits on the endowment effect noted by Tversky and Kahneman [1991] and Novemsky and Kahneman [2005]—that budgeted spending by buyers and successful reduction of inventory by sellers are not coded as losses—are clearly also predicted by our model: parties in market settings *expect* to exchange money for goods.[11]

Consider also an instance of reference dependence commonly discussed in economics: employees' aversion to wage cuts. A wise inconsistency has pervaded the application of the reference-point-as-status-quo perspective to the the laboratory vs. the labour market: a decrease in salary is not a reduction in the status-quo level of wealth—it is a reduction from the expected rate of increase in wealth. While good judgment and obfuscatory language can be used to variously deem aversion to losses in current wealth (when we reject unexpected gambles) *vs.* aversion to losses in increases in current wealth (when we are bothered by wage cuts) *vs.* aversion to losses in increases in the rate of increase of current wealth (when we are bothered by not getting an expected pay raise) as the relevant notion of loss aversion, our model not only accomodates all these scenarios—but predicts which is the appropriate notion as a function of the environment.

Finally, a status-quo theory of the reference point is especially unsatisfying when applied to the many economic activities—such as food, entertainment, and travel—that involve fleeting con-

---

[11] Indeed, while researchers such as Novemsky and Kahneman [2005] seem to frame these examples as determining some "boundaries of loss aversion," we view them more narrowly as determining the boundaries of the endowment effect. As we demonstrate in Section IV, loss aversion *does* have important implications in markets—but the endowment effect is not among them.

sumption opportunities and no ownership of physical assets. If a person expects to undergo a painful dental procedure, finding out that it is not necessary after all may feel like a gain. Yet there is no meaningful way in which her status-quo endowment of dental procedures is different from somebody's who never expected the procedure, so irrespective of expectations a status-quo theory would always predict the same gain-loss utility of zero from this experience.

Our model of how utility depends on expectations could be combined with any theory of how these expectations are formed. But as a disciplined and largely realistic first pass, we assume that expectations are fully rational. To illustrate with an example analyzed in more detail in Section IV, suppose a consumer had long known that she would have the opportunity to buy shoes, and faced with price uncertainty, had formed plans whether to buy at each price. If given the reference point based on her expectation to carry through with these plans, there is some price where she would in fact prefer *not* to carry through with her plans, our theory says that she should not have expected these plans in the first place. More generally, our notion of *personal equilibrium* assumes that a person correctly predicts both the environment she faces—here, the market distribution of prices—and her own reaction to this environment—here, her behavior in reaction to market prices—and taking the reference point generated by these expectations as given, in each contingency maximizes expected utility.

Formally, suppose the decisionmaker has probabilistic beliefs described by the distribution Q over $\mathbb{R}$ capturing a distribution over possible choice sets $\{D_l\}_{l\in\mathbb{R}}$ she might face, where each $D_l \subset \Delta(\mathbb{R}^K)$. In the first and weaker of two solution concepts we consider, rational expectations is the only restriction we impose:

Definition 1 *A selection* $\{F_l \in D_l\}_{l\in\mathbb{R}}$ *is a* <u>personal equilibrium</u> *(PE) if for all* $l \in \mathbb{R}$ *and* $F_l' \in D_l$, $U\left(F_l \,\middle|\, \int F_l dQ(l)\right) \geq U\left(F_l' \,\middle|\, \int F_l dQ(l)\right).$

If the person expects to choose F_l from choice set D_l, then given her expectations over possible choice sets she expects the distribution of outcomes $\int F_l dQ(l)$. Definition 1 says that with those expectations as her reference point, she should indeed be willing to choose F_l from choice set D_l.

Of existing models, our notion of PE is most related to the notion of "loss-aversion equilibrium" that Shalev [2000] defined for multiplayer games as a Nash equilibrium fixing each player's reference

point, where the reference point is equal to the player's (implicitly defined) reference-dependent expected utility. Although Shalev does not himself pursue this direction, reformulating his notion of loss-aversion equilibrium using our utility function and applying it to individual decisionmaking corresponds to PE.[12]

As we shall illustrate in Section IV, there may be multiple PE: it may be that if a person expects to buy shoes at a particular deterministic price, she prefers to buy them, and if she expects not to buy, she prefers not to buy them. Since the decisionmaker has a single utility function, she can always rank the PE outcomes in terms of ex-ante expected utility. Insofar as a person is free to make any plan so long as she will follow it through, therefore, she will choose her favorite plan:

Definition 2 *A selection $\{F_l \in D_l\}_{l \in \mathbb{R}}$ is a <u>preferred personal equilibrium</u> (PPE) if it is a PE, and $U\left(\int F_l dQ(l) \,\middle|\, \int F_l dQ(l)\right) \geq U\left(\int F_l' dQ(l) \,\middle|\, \int F_l' dQ(l)\right)$ for all PE selections $\{F_l' \in D_l\}_{l \in \mathbb{R}}$.*

To see how the solution concepts work, consider the shoe shopper from above. Her choice set at the time of the purchase decision is the decision of whether to buy at the (ex-ante possibly unknown) price she faces. Her reference point is the probabilistic distribution over money outlays and shoe acquisitions determined by her planned behavior in each choice set—her plans whether to buy at each possible price—combined with the distribution over possible choice sets—her beliefs about the prices she might face. PE requires her planned behavior in each choice set to be optimal given this reference point, and PPE requires her to choose the PE plan with highest *ex-ante* expected utility.[13] ,[14]

---

[12] Proposition 3 below is also related to Shalev's [2000] Proposition 2 showing that pure-strategy Nash equilibria are always myopic loss-aversion equilibria. Insofar as they provide a framework for having utility depend on beliefs, personal equilibrium is also very closely related to Geanakoplos, Pearce, and Stacchetti's [1989] notion of psychological Nash equilibrium in games. But since in a psychological game a player's utility can only depend on her *own* (possibly higher-order) beliefs about *other* players' strategies, it seems impossible to reformulate our model in that framework.

[13] The above glosses over a key way that our model is under-specified. A premise of our model is that preferences depend on expectations formed after the decisionmaker started focusing on a decision. The specification of Q should correspond to this timing, and is therefore an important interpretational matter in any application. As an illustration, if the shoe shopper had been thinking about her possible purchase for a long time, her expectations from before she knew the price will affect her preferences, and the appropriate Q is the lottery representing her probabilistic beliefs over prices. But if she only considered the possible purchase upon seeing shoes at the store, Q should be the deterministic lottery corresponding to the relevant price at the time.

[14] Because PE is simply an application to our environment of the more general definition in Kőszegi [2005], Theorem

While PE and PPE may in general exhibit different properties, our main results and intuitions in this paper hold for both concepts. Except where we wish to make explicit how the concepts differ, we have therefore chosen examples where there is a unique PE, so that the set of PE and PPE coincide.

Before considering specific contexts, we note a simple and striking feature of our general model. When a loss-averse decisionmaker's choice set is deterministic and all choices in it are deterministic, the predictions of PPE are identical to those of a model based solely on consumption utility:

Proposition 3 *Suppose Q is a lottery putting probability 1 on a choice set consisting of all convex combinations of a set D^* of deterministic outcomes. If A3′ holds, then a lottery is a PPE if and only if it puts probability 1 on an outcome that is a solution to $argmax_{c' \in D^*} m(c')$.*

Any deterministic consumption level c that maximizes consumption utility is a PE: if the decisionmaker had expected c, another outcome would both lower consumption utility and lead to sensations of loss in dimensions where it falls short of c. Furthermore, notice that ex-ante expected gain-loss utility is zero for deterministic outcomes, and because losses from comparing an outcome to a counterfactual loom larger than the symmetric gains, it is negative for non-deterministic lotteries. Hence, c yields strictly greater ex-ante expected utility than any deterministic outcome that does not maximize consumption utility, and also than any non-deterministic lottery. This means both that c is a PPE, and that the latter possibilities cannot be.

Because $u(c|c) = m(c)$, Proposition 3 implies that in deterministic environments, not only does loss aversion not affect behavior, it also does not affect welfare. These statements are not true for PE: as we illustrate in Section IV, if a person comes to anticipate choosing an option that does not maximize consumption utility, she may carry it out to avoid the sensation of loss associated with switching to another option. But more importantly, as many of our results in Sections IV and V highlight, in environments with uncertainty even applying PPE reference dependence can hugely influence behavior.

---

1 of that paper establishes that, if $\int D_l dQ(l)$ is convex and compact, a PE exists. Since the set of PE is closed and $U(F|F)$ is continuous in F, PPE also exists.

IV. Shopping

In this section we use our model to study how a consumer's valuation for a good is endogenously determined by market conditions and her own anticipated behavior.

Suppose there are two dimensions of choice, with $m(c) = c_1 + c_2$, where $c_1 \in \{0, 1\}$ reflects whether or not a consumer has a pair of shoes, and $c_2 \in \mathbb{R}$ is her dollar wealth. This means we can think of her "intrinsic value" for shoes as \$1. To isolate the consequences of loss aversion, we assume that $\mu(\cdot)$ satisfies A3': $\mu(x) = \eta x$ for $x > 0$ and $\mu(x) = \eta\lambda x$ for $x \leq 0$. In this formulation, $\eta > 0$ is the weight the consumer attaches to gain-loss utility, and $\lambda > 1$ is her "coefficient of loss aversion." We normalize her initial endowment to $(0, 0)$.

Whatever the consumer expected, she assesses a price p paid for shoes as some combination of loss and foregone gain. Adding this gain-loss sensation to her consumption value for money, her disutility from spending on the shoes is between $(1 + \eta)p$—her disutility if she had expected to pay p or more—and $(1 + \eta\lambda)p$—her disutility if she had expected to pay nothing. By a similar argument, the consumer's total utility from getting the shoes is between $1 + \eta$ and $1 + \eta\lambda$. Hence, the expectations most conducive to buying induce a disutility of $(1+\eta)p$ from spending money and a utility of $1 + \eta\lambda$ from getting the shoes, so that no matter her expectations the consumer would never buy for prices $p > p_{max} \equiv (1 + \eta\lambda)/(1 + \eta)$. Conversely, even given the expectations least conducive to buying, the consumer buys for all prices $p < p_{min} \equiv (1 + \eta)/(1 + \eta\lambda)$.

The above implies that if shoes are available at a certain price p, then for $p > p_{max}$ there is a unique PE in which the consumer does not buy, and for $p < p_{min}$ there is a unique PE in which she buys. In addition, with a deterministic price, expecting to buy in fact creates the preferences most favorable for buying, and expecting not to buy creates the preferences least favorable for buying. Hence, for $p \in [p_{min}, p_{max}]$, buying for sure and not buying for sure are both pure-strategy PE.[15] Intuitively, if the consumer had expected to get the shoes, she would feel a loss of shoes if she did not buy, and her aversion to this loss leads her to buy even at relatively high prices. In contrast, if she had expected not to buy, buying would result in a loss of money, and her aversion to this

---

[15] In the interior of this range, there is also a unique mixed-strategy equilibrium, where the consumer buys with probability $q = [(1+\eta\lambda)p - (1+\eta)]/[\eta(\lambda-1)(p+1)] \in [0, 1]$, and is indifferent between buying and not buying. This equilibrium is unstable by analogy to conventional notions of instability.

loss leads her not to buy. But as an application of Proposition 3, the consumer's PPE is typically unique: she buys if and only if $p < 1$.[16]

Reference dependence leads to more interesting behavior when prices are uncertain. In this case, changes in the price distribution can significantly affect a consumer's willingness to pay a particular price.[17] To examine the consumer's willingness to pay without worrying about how her behavior feeds back into her expectations, we suppose she expects the price $p_L < p_{min}$ with probability q_L and the price $p_H > p_{max}$ with probability $q_H = 1 - q_L$, and consider the "out-of-equilibrium" question of whether she buys at the intermediate price p_M. Our calculations are limiting cases of environments where p_M occurs with a small probability.[18]

Given the assumptions above, the consumer will buy the shoes at price p_L, but not at price p_H. Hence, she expects to consume shoes and spend p_L with probability q_L, and not to consume shoes or spend money with probability q_H. As a result, her utility from buying at price p_M is

$$1 - p_M$$
$$+ q_H(\eta - \eta\lambda p_M)$$
$$(4) \qquad - q_L\eta\lambda(p_M - p_L).$$

The first line is the consumption utility from buying. The second line is gain-loss utility from comparing buying to not buying, which is assessed as a gain of 1 pair of shoes and a loss of p_M dollars. The third line is gain-loss utility from comparing buying at p_M to buying at the expected purchase price p_L, which leads to no gain or loss in shoes and a loss of $p_M - p_L$ dollars.

The consumer's utility from not buying is

$$(5) \qquad q_L(\eta p_L - \eta\lambda).$$

---

[16] To avoid cumbersome presentation, for the remainder of the paper a statement of the form "if and only if $x < y$" will be used to mean "if $x < y$ and only if $x \leq y$."

[17] One implication of this distribution-dependence is that different mechanisms that are traditionally considered equivalent, "incentive compatible" ways of eliciting preferences should yield different answers. For instance, the Becker-DeGroot-Marschak [1964] value-elicitation procedure commonly used in experimental work under the presumption that any induced or subjective expectation by subjects should yield the same—intrinsic—value, according to our model would yield different results depending on expectations. Hence, our model tells us that not only does this procedure not work in practice—it does not work in theory, either.

[18] In Kőszegi and Rabin [2004], we show how to solve for the set of personal equilibria for any price distribution faced by the consumer, and how the principles we establish with our three-price examples extend to such distributions. We also show that sufficient price uncertainty implies that the PE is unique.

Relative to the expectation to buy, not buying is a gain of p_L dollars and a loss of 1 pair of shoes.

We consider consumer behavior in several different situations by comparing these expected utilities from buying and not buying. First, suppose that the good may be available for free: $p_L = 0$. Then, using Expressions 4 and 5, the consumer buys the shoes at p_M if and only if

$$(6) \qquad\qquad p_M < 1 - (1 - q_L) \cdot \frac{\eta(\lambda - 1)}{1 + \eta\lambda}.$$

The right-hand side of the above inequality is increasing in q_L. Since the consumer expects to spend nothing no matter what, an increase in q_L only increases her expectation to get the shoes and hence the loss she will feel if she does not buy. This "attachment effect" increases her willingness to pay.

Now suppose that $p_L \geq 0$ and $q_L = 1$. Then, comparing Expressions 4 and 5, the consumer buys the shoes at price p_M if and only if

$$(7) \qquad\qquad p_M < 1 + p_L \cdot \frac{\eta(\lambda - 1)}{1 + \eta\lambda}.$$

The right-hand side of this inequality is increasing in p_L. Hence, in violation of the law of demand, a decrease in the price distribution decreases the consumer's willingness to pay for the shoes. If there is a possibility of acquiring the shoes at a low price, by comparison the consumer considers paying the higher price p_M to be a loss. The greater the difference between p_M and p_L, the greater is the sense of loss, and the more this "comparison effect" decreases her willingness to pay.

This violation of the law of demand occurs in our model solely because of the feedback of the environment into expectations: for any fixed expectations, the consumer's demand is downward-sloping. But because lowering the prices at which she had expected to buy changes her view of paying higher prices, she might be led not to buy at those prices.[19]

The above illustrations of the attachment and comparison effects can be generalized to broader principles regarding the consumer's reaction to different types of price decreases. If prices drop

---

[19] The logic behind the attachment and comparison effects can shed light on non-equilibrium situations in which buyers believe that the price is lower than it actually is. For example, car dealers sometimes use the strategy of "throwing a lowball," promising a very low price and attempting to raise it once the consumer gets used to the expectation of buying—using the attachment effect to sell at a price that would otherwise be too high. Our model says this strategy may backfire if the high price looks too awful compared to the expected price, but car dealers presumably make it their business to be well calibrated about which combinations of expected and actual prices are most successful in the situations they face.

from above the consumer's reservation price to below it, this boosts demand not only by the direct effect of lower prices, but also because the reservation price itself increases: since the consumer now expects to buy with higher probability, she becomes more attached to the idea of having the good. If prices drop from a level at which she would have bought anyway, then the price decrease intensifies the sense of loss a consumer feels from comparing buying at a high price to other possible purchase prices, reducing her demand at higher prices.

While Proposition 3 says that (if applying PPE) consumption utility fully governs consumer behavior in deterministic environments, the above results show that the same is very much not true in stochastic ones. As Inequality 6 indicates, there may be a unique PE in which the consumer does not buy the shoes for some prices below their intrinsic value; and as Inequality 7 shows, she may be induced to buy even for some prices above their intrinsic value.

Price uncertainty may in fact induce a shopper to behave in a way that does not maximize her overall utility among strategies available to her. As an example, suppose that she faces equiprobable prices of zero or one-half. It is easy to check that for any $\eta > 0$ and $\lambda > 0$ the unique PE is to buy at both prices. Yet as $\lambda \to \infty$, the disutility from paying one-half as compared to nothing approaches infinity, with other parts of the consumer's PE utility, as well as her utility from expecting and carrying through a plan never to buy, remaining the same. Hence, for a sufficiently high λ, the consumer is worse off than if she were unable to buy the shoes.

Intuitively, since the consumer values the shoes, she buys whenever the price is very low. The attachment to the good induced by realizing that she will do this, however, changes her attitudes toward the purchase decision. If the price turns out to be higher, she must choose between a loss of money and a loss of shoes. While buying is her best response to her expectations, it is still worse than if she could have avoided the risk of loss by avoiding the expectation of getting the shoes in the first place. More generally, because the consumer does not internalize the effect of her ex-post behavior on ex-ante expectations, the strategy that maximizes ex-ante expected utility is often not a PE.

16

V. Driving

A literature has recently emerged studying the relationship between labor supply and wages of workers with flexible daily work schedules.[20]   While strongly disagreeing about the extent of this behavior and whether it is irrational, Camerer, Babcock, Loewenstein, and Thaler [1997] and Farber [2004, 2005] analyzing New York city taxi drivers all find a negative relationship between earnings early in the day and duration of work later in the day. Studies analyzing participation decisions as a function of expected wages, on the other hand, find a positive relationship between earnings and effort: Oettinger [1999] finds that stadium vendors are more likely to go to work on days when their wage can be expected to be higher, and Fehr and Goette [2002] show bicycle messengers sign up for more shifts when their commision is experimentally increased.

The negative relationship between earnings early and work later in the day found in this literature can be thought of in terms of drivers having a daily "target income." To our knowledge, all existing research assumes that the target is exogenous, and none specifies its determinants. A person's target, however, is likely to be endogenous to her situation. If a driver generally makes around $200 a day, her target is probably close to $200; if she generally makes $300, it will be $300. If there are days (such as holidays) where she *predictably* makes more money than average, her target on those days will be higher; and on days she does not drive, it is presumably zero.

In this section we analyze labour-supply decisions with such endogenous targets. We find that a driver's response to wage changes depends on whether those changes were predictable: a) she is more likely to go to work on days her wage is *predictably* higher, and b) once at work she is likely to work more at a given realized wage when those predicted wages are higher, but c) she may drive less when her wages are *un*predictably high.

By endogenizing the target as the driver's expectations, our theory contributes to the literature in a few ways. Most fundamentally, feature (b) above reverses the problematic implication of fixed-income-target models that predictably higher wages lower effort. In addition, because the model predicts how targets vary between drivers and for a driver from day to day, it may qualify the interpretation of existing empirical studies of the target-income hypothesis and aid substantially in

---

[20] See Goette, Huffman, and Fehr [2004] and Farber [2004] for useful reviews of this literature.

designing new tests. Finding that the reference-dependent model best fits the data if the target is allowed to vary across drivers and days, for instance, Farber [2004] concludes that such variation— implicitly equated with indeterminacy—undermines the usefulness of the concept of target income. But assuming expectations reflect the empirical patterns of average income, our model provides a way of allowing realistic variation in targets without introducing degrees of freedom. Finally, while this literature seems to assume that reference-dependence is limited to income, our model assumes effort is also reference-dependent. Although we follow the literature in referring to the phenomenon as income-targeting, in fact we obtain a negative relationship between unexpectedly high income and effort without a priori assuming that only income is reference dependent.

Formally, a taxi driver decides each day whether to go to work in the morning, and, if she does, whether to continue driving in the afternoon. Denote by e^m her morning participation decision, where $e^m = 1$ if she drives and $e^m = 0$ if she does not. Define her afternoon driving decision, $e^a \in \{0, 1\}$, similarly. Driving in the morning yields income I^m, and driving in the afternoon brings in I^a. I^m and I^a are independent random variables. If the driver works in the morning, she learns her afternoon wage I^a before deciding whether to drive in the afternoon. The driver has daily loss-averse preferences over her income and driving time. Her consumption utility is linear in each of the two dimensions, with $m_1(I^m + I^a) = I^m + I^a$ and $m_2(e^m + e^a) = f \cdot (e^m + e^a)$, where f is the per-unit consumption-utility cost of effort. Assume again that $\mu(x) = \eta x$ for $x > 0$ and $\mu(x) = \eta \lambda x$ for $x \le 0$. We solve for properties of PPE.[21]

As in the previous section, to develop intuition we consider an "out-of-equilibrium" situation. The driver expects the morning income to be w_E^m with probability one, and the realized morning income is w_R^m. Although when expected income changes, typically so does realized income, to highlight the conceptual distinction between expected and surprise wage changes we perform comparative statics separately with respect to w_E^m and w_R^m. The driver expects the afternoon wage to be $w_H^a > \frac{1+\eta\lambda}{1+\eta} \cdot f \equiv w_{max}$ or $w_L^a < \frac{1+\eta}{1+\eta\lambda} \equiv w_{min}$ with probabilities q_L and $q_H = 1 - q_L$, respectively, and the realized afternoon wage is the intermediate value w_R^a. These assumptions imply that if the driver works in the morning, then no matter her expectation she will work in the afternoon if the

---

[21] Appropriately restated, the results we stress hold for PE as well, with more complicated characterizations due to the multiplicity of equilibria.

18

wage is w_H^a, and not work if the wage is w_L^a. To be able to demonstrate all the effects of interest within this simple model, we assume that at midday the driver is short but within reach of her expected morning income: $w_E^m - w_R^a < w_R^m \leq w_E^m$.

Since our main interest is in the driver's behavior *if* she works, we first investigate the properties of PPE, labelled "participation PPE," in which she shows up; we turn below to whether showing up is indeed a PPE. With the above distributions, the cabbie expects to work in the afternoon with probability q_H, and to earn w_E^m if she drives in the morning only, and $w_E^m + w_H^a$ if she drives all day. Her utility if she continues to work is then

$$
w_R^m + w_R^a - 2f
$$
$$
+ \quad q_L(\eta(w_R^m + w_R^a - w_E^m) - \eta\lambda f)
$$
$$
- \quad q_H\eta\lambda((w_E^m + w_H^a) - (w_R^m + w_R^a)).
$$

The first line is the consumption utility from working all day. The second line is gain-loss utility from comparing driving all day to driving only in the morning, which leads to a gain of $w_R^m + w_R^a - w_E^m$ in income and a loss of f in work effort. Finally, the third line is the loss from comparing working all day to her expectation to work all day, which feels like a loss because the driver earns $w_R^m + w_R^a$ instead of $w_E^m + w_H^a$. The utility from not working is

$$
w_R^m - f + q_H(\eta f - \eta\lambda(w_E^m + w_H^a - w_R^m)) - q_L\eta\lambda(w_E^m - w_R^m).
$$

Hence, the driver continues to work so long as

$$
(8) \qquad w_R^a \geq \frac{[1 + \eta + q_L\eta(\lambda - 1)] \cdot f - q_L\eta(\lambda - 1)(w_E^m - w_R^m)}{1 + \eta + q_H\eta(\lambda - 1)}.
$$

Finally, if the expected income level is realized in the morning ($w_R^m = w_E^m$), this inequality becomes

$$
(9) \qquad w_R^a \geq \frac{1 + \eta + q_L\eta(\lambda - 1)}{1 + \eta + q_H\eta(\lambda - 1)} \cdot f.
$$

Notice that this last inequality does not depend on the expected morning wage. With an expectations-based income target, a fully anticipated increase in the morning wage leaves the driver on average equally far from her target in the middle of the day—and hence does not affect her willingness to

19

continue work. By contrast, the driver is more likely to drive even at a moderate afternoon wage if the probability q_H of a high afternoon wage is increased. Intuitively, an increase in the expected afternoon wage increases the driver's target income as well as her expectation to work, making lower income more painful and work less painful. Unlike the prediction of reduced work in fixed-income-target models, these results imply that (depending on the form of increase) anticipated wage increases either increase work effort or leave it unaffected.

Examining how a divergence between expected and realized incomes affects the driver's work effort, however, shows that our model does capture the core intuition in the target-income findings of Camerer, Babcock, Loewenstein, and Thaler [1997]. Since the right-hand side of Inequality 8 is increasing in w_R^m, when the driver's morning income is lower she requires a lower afternoon income to induce her to keep working. Intuitively, if a driver earns less money than she expected early in the day, she will be more willing to work later, so as to mitigate the losses from falling short of her target. Finally, Inequality 8 shows one more sense in which increases in expected income increase the driver's willingness to work: an increase in the predicted *morning* wage w_E^m has a positive effect on the *afternoon* work decision for any given morning wage.[22]

Based on the above, it is easy to derive the driver's participation decision. Since the afternoon work decision is independent of w_E^m, an increase in w_E^m merely shifts the entire income distribution from participation to the right. Hence, showing up is more likely to be part of a PPE. This result is consistent with the findings of Oettinger [1999] and Fehr and Goette [2002] that the elasticity of participation with respect to *predictable* wage variation is positive.

Although these intuitions do not fully extend to all stochastically-dominant increases in the income distribution, they do generalize to simple shifts in it.[23] Suppose that the driver's morning income has two additive components, a constant w^m and a continuously distributed "surprise" part $\tilde{w}^m \sim F^m$ with bounded support. Her afternoon income is the continuously distributed

---

[22] It bears emphasizing that only an increase in the realized *morning* income has a negative effect on willingness to drive in the afternoon. Because an increase in the realized afternoon income w_R^a does not affect the driver's reference point, our model predicts (as does virtually any model) that such an increase increases the driver's willingness to work. Hence, in cases when both w_R^m and w_R^a increase—when surprises in the wage are positively correlated over the course of a day—the overall effect on work effort is ambiguous.

[23] Specifically, it is Parts 2 and 3 of Proposition 4 below regarding the driver's participation decision that does not extend to all shifts. If, for instance, the variance in income increases significantly along with an increase in the mean, the driver's expected utility in a PE involving participation may decrease due to the variance.

random variable $I^a \sim F^a[\underline{w}, \overline{w}]$, where $\underline{w} < w_{min}$ and $\overline{w} > w_{max}$.[24]   Proposition 4 establishes some properties of PPE as a function of the driver's beliefs about income described by the triple (w^m, F^m, F^a).

Proposition 4 *The following hold for any $\eta > 0$ and $\lambda > 1$.*

1. *For any participation PPE, there is a function $g : \mathbb{R} \to [w_{min}, w_{max}]$ such that for any $\tilde{w}^m$, the driver continues work if and only if $I^a \geq g(\tilde{w}^m)$. The function $g(\cdot)$ is increasing, and is strictly increasing at $\tilde{w}^m$ if $F^m(\tilde{w}^m + g(\tilde{w}^m)) - F^m(\tilde{w}^m) > 0$.*

2. *If $g(\cdot)$ describes a participation PPE for (w^m, F^m, F^a), it also describes a participation PPE for any $(w^{m\prime}, F^m, F^a)$ such that $w^{m\prime} \geq w^m$.*

3. *For any F^m and F^a, there is a constant $\underline{w}^m$ such that participation is part of a PPE for (w^m, F^m, F^a) if and only if $w^m \geq \underline{w}^m$, and non-participation is a PPE for (w^m, F^m, F^a) if and only if $w^m \leq \underline{w}^m$.*

In the proposition, $g(\tilde{w}^m)$ is the driver's "reservation wage" for driving in the afternoon as a function of the surprise component of the morning wage. Part 1 shows that $g(\cdot)$ is increasing, so that the driver has negative labor supply elasticity with respect to surprises in the morning wage. Part 2 says that once the morning wage is sufficiently high to induce participation, it is only surprises relative to expectations that determine the driver's afternoon work decision. This means that an increase in the expected morning wage does not affect the afternoon work decision if it is matched by an increase in the actual wage, but makes the driver more likely to continue work if it is not borne out in the actual wage (which therefore represents a worse surprise). Part 3 shows that an increase in the expected morning wage makes the driver more likely to show up to work.

VI.   Conclusion

By directly constructing reference-dependent utility from consumption utility and assuming that the reference point is endogenously determined as rational expectations about outcomes, our theory

---

[24] These assumptions ensure that if the driver works in the morning, then for any I^m there is a positive probability that she continues work, and a positive probability that she does not.

provides an algorithm for translating a "classical" reference-independent model into the corresponding reference-dependent one. This helps render it highly portable and generally applicable. While we illustrated some implications of our framework for consumer and labour-supply decisions, there are many other domains where it can be applied. In the context of risk preferences, Kőszegi and Rabin [2005] and Sydnor [2005] show that an expectations-based theory may help explain very different reactions consumers have to insuring anticipated risks versus reacting to unanticipated risks. In auction theory, since expectations of winning an auction affect the desire to do so, different auction designs predicted in current theory to be revenue equivalent may in fact generate different revenues because of different expectations induced. In intertemporal choice, Stone [2005] shows how expectations-based preferences can generate behavior that can be mistaken for present-biased preferences in the sense of Laibson [1997] and O'Donoghue and Rabin [1999] or temptation disutility in the sense of Gul and Pesendorfer [2001], since surprising oneself with immediate consumption tends to be more pleasurable than inherently unsurprising planned future consumption. In principal-agent models, performance-contingent pay may not only directly motivate the agent to work harder in pursuit of higher income, but also indirectly motivate her by changing her expected income and effort. In bargaining, early posturing might be used not only to influence the other's beliefs about achievable outcomes, but also to influence her preferences over outcomes by this change in beliefs.[25]   And while firms are presumably not directly subject to reference-dependent utilities, Heidhues and Kőszegi [2005, 2006] show that their responses to consumer loss aversion have important implications for wage and price setting in imperfectly competitive markets.

Several features, however, leave our framework short of providing a formulaic way to apply reference dependence to all economic contexts. Our model takes as given the set of consumption dimensions over which gain-loss utility is separately defined. While in most applications it is appropriate to identify these dimensions with the physical consumption dimensions, with this approach our theory (or any theory) of reference dependence would in some cases make bad predictions.[26]

---

[25] For related intuitions outside of our personal-equilibrium framework, see DeMeza and Webb [2003] on principal-agent theory and Compte and Jehiel [2003] on bargaining theory.

[26] Consider, for instance, a decisionmaker making choices over Tropicana and Florida's Natural premium orange juices, two separate consumption goods she enjoys but can barely distinguish. Would she be willing to trade six ounces of Tropicana juice she had expected to consume for eight ounces of Florida's Natural juice? If the trade were coded as a loss in the Tropicana dimension and a gain in the Florida's Natural dimension, under usual parametrizations

Kőszegi and Rabin [2004] argue that gain-loss utility should be defined over "hedonic" dimensions of consumption that people experience as psychologically distinct; in some situations, judgment is needed to identify these dimensions. And as noted in Section III, even more judgment is required in determining the moment of first focus.

There are also contexts to which this paper's model does not apply, but to which alternative models using the same psychological principles would apply. When considerable time passes between a decision and consumption, by the time consumption takes place the reference point—which we assume is recent expectations—will presumably reflect the decision made. In the purchase of extended warranties, for instance, whether or not a good breaks is typically realized long after the decision of whether to buy the warranty. Kőszegi and Rabin [2005] consider an approach to incorporating such possibilities into a rational-expectations model.

While the utility-maximization approach in this paper suggests that people maximize reference-dependent preferences corresponding to true experienced well-being, there are reasons to doubt that this is so. Substantial evidence indicates that people "narrowly bracket": they do not fully integrate decisions at hand with other decisions and events. Ignoring the way gains and losses from different decisions cancel each other out leads to overweighting of gains and losses. People may also over-attend to losses and gains because they underestimate how quickly they will adapt to these changes.[27]   On both of these accounts, the nature and scope of reference-dependent choices seems to reflect mistakes our fully rational model does not capture.

APPENDIX: PROOFS

Proof of Proposition 1.

1. Obvious.

2. Suppose not; that is, suppose $u(c, c') \geq u(c', c')$ and $u(c, c) \leq u(c', c)$. Adding these and

---

she would not. If the person does not view the juices she consumes as substantively different experiences, however, she would presumably accept the trade.

[27] See Kahneman and Tversky [1979], Thaler [1985], Kahneman and Lovallo [1993], Benartzi and Thaler [1995], and Read, Loewenstein, and Rabin [1999] on the role of narrow bracketing in loss aversion, and Kahneman [1991] and Loewenstein, O'Donoghue, and Rabin [2003] on the role of underestimating adaptation.

using the definition of u implies that

(10) $$m(c) + m(c') + n(c|c') + n(c'|c) \geq m(c) + m(c') + n(c|c) + n(c'|c').$$

Eliminating $m(c) + m(c')$ from both sides and using the definition of $n(\cdot)$, this reduces to

(11) $$\sum_{k=1}^{K} \left[\mu(m_k(c_k) - m_k(c_k')) + \mu(m_k(c_k) - m_k(c_k')) \right] \geq 0.$$

By A2, and using that $c \neq c'$, this is a contradiction.

3. We prove that for any $F, F' \in \Delta(\mathbb{R}^K)$,

$$U(F|F) + U(F'|F') > U(F|F') + U(F'|F).$$

This is obviously sufficient to establish the claim by contradiction. Let the marginals of F and F' on dimension k (that is, the distribution of outcomes in dimension k generated by F and F') be F_k and F_k', respectively. Noticing that expected consumption utilities are the same on the two sides, it is sufficient to prove that

$$\iint \mu(m_k(c_k) - m_k(r_k)) dF_k(c_k) dF_k(r_k) + \iint \mu(m_k(c_k) - m_k(r_k)) dF_k'(c_k) dF_k'(r_k)$$
$$\geq \iint \mu(m_k(c_k) - m_k(r_k)) dF_k(c_k) dF_k'(r_k) + \iint \mu(m_k(c_k) - m_k(r_k)) dF_k'(c_k) dF_k(r_k)$$

for all k, and the inequality is strict for any k for which $F_k \neq F_k'$. The inequality obviously holds with equality when $F_k = F_k'$, so we establish that it holds strictly when $F_k \neq F_k'$.

Since μ satisfies A3′, there is an $\alpha > 0$ such that for any $x \in \mathbb{R}$ we have $\mu(x) + \mu(-x) = -\alpha|x|$. Using this and dividing by $-\frac{1}{2}\alpha$, the above becomes

$$\iint |m_k(c_k) - m_k(r_k)| dF_k(c_k) dF_k(r_k) + \iint |m_k(c_k) - m_k(r_k)| dF_k'(c_k) dF_k'(r_k)$$
(12) $$< \; 2 \cdot \iint |m_k(c_k) - m_k(r_k)| dF_k(c_k) dF_k'(r_k).$$

For reals x, a, b, let $x \in ((a, b))$ denote that x is between a and b (i.e. that $x \in (a, b)$ if $a < b$ and $x \in (b, a)$ if $b < a$). Also, let $I(\cdot)$ denote the indicator function. Then, the above inequality can be rewritten as

$$\iiint I[x \in ((m_k(c_k), m_k(r_k)))] dx dF_k(c_k) dF_k(r_k) + \iiint I[x \in ((m_k(c_k), m_k(r_k)))] dx dF_k'(c_k) dF_k'(r_k)$$
$$< \; 2 \cdot \iiint I[x \in ((m_k(c_k), m_k(r_k)))] dx dF_k(c_k) dF_k'(r_k).$$

24

Reversing the order of integration, the above becomes

$$\int \Pr_{\substack{c_k \sim F_k \\ r_k \sim F_k}} [x \in ((m_k(c_k), m_k(r_k)))] dx + \int \Pr_{\substack{c_k \sim F'_k \\ r_k \sim F'_k}} [x \in ((m_k(c_k), m_k(r_k)))] dx$$

(13) $$< \ 2 \cdot \int \Pr_{\substack{c_k \sim F_k \\ r_k \sim F'_k}} [x \in ((m_k(c_k), m_k(r_k)))] dx.$$

We prove that the above is true weakly point-by-point, and strictly on a set of positive measure. Let $F_k(m_k^{-1}(x)) = p(x)$ and $F'_k(m_k^{-1}(x)) = p'(x)$. Notice that since $F_k \neq F'_k$, there is a set of positive measure such that $p(x) \neq p'(x)$. The probability that x is on a line segment of two points $m_k(c_k)$ and $m_k(r_k)$, where c_k and r_k are chosen independently according to F_k is $2p(x)(1 - p(x))$. Similarly, the probability that it is between two such points when c_k and r_k are chosen according to F'_k is $2p'(x)(1 - p'(x))$. And the probability that it is between two such points when c_k and r_k are chosen according to F_k and F'_k, respectively, is $p(x)(1 - p'(x)) + p'(x)(1 - p(x))$. It is sufficient to prove that

$$p(x)(1 - p(x)) + p'(x)(1 - p'(x)) \leq p(x)(1 - p'(x)) + p'(x)(1 - p(x))$$

and that the inequality is strict for a set of positive measure. This is true since $(p(x) - p'(x))^2 \geq 0$ and the inequality is strict whenever $p(x) \neq p'(x)$. $\square$

Proof of Proposition 2. Let $v_k(x) = m_k(x) - m_k(0) + \mu(m_k(x) - m_k(0))$. Since m_k is linear for each k,

$$
\begin{aligned}
u(c|r) - u(r|r) &= \sum_{k=1}^{K} [m_k(c_k) - m_k(r_k) + \mu(m_k(c_k) - m_k(r_k))] \\
&= \sum_{k=1}^{K} [m_k(c_k - r_k) - m(0) + \mu(m_k(c_k - r_k) - m_k(0))] \\
&= \sum_{k=1}^{K} v_k(c_k - r_k).
\end{aligned}
$$

25

A0 and A1 are obviously satisfied. Notice that for any $y > x > 0$,

$$v_k(-y) + v_k(y) = \mu(m_k(y) - m_k(0)) + \mu(-(m_k(y) - m_k(0)))$$

$$< \mu(m_k(x) - m_k(0)) + \mu(-(m_k(x) - m_k(0))) = v_k(-x) + v_k(x)$$

since m_k is increasing and μ satisfies A2. Thus, v_k satisfies A2. A3 is obvious, given the linearity of m_k. Next,

$$\lim_{x \searrow 0} \frac{v_k'(-x)}{v_k'(x)} = \frac{m_k'(0) + \mu_-'(0)m_k'(0)}{m_k'(0) + \mu_+'(0)m_k'(0)} = \frac{1 + \mu_-'(0)}{1 + \mu_+'(0)} < \lambda.$$

This completes the proof. $\square$

Proof of Proposition 3. Without loss of generality, let (as in Sections IV and V) $\mu(x) = \eta x$ for $x > 0$ and $\mu(x) = \eta \lambda x$ for $x \leq 0$. Suppose $c \in \text{argmax}_{c' \in D^*} m(c')$. Then, by definition, for all $c' \in D^*$

$$\sum_{k:m_k(c_k') > m_k(c_k)} \left[m_k(c_k') - m_k(c_k) \right] \leq \sum_{k:m_k(c_k') < m_k(c_k)} \left[m_k(c_k) - m_k(c_k') \right].$$

Therefore, using that $m(c') \leq m(c)$,

$$
\begin{aligned}
u(c'|c) &= m(c') + \sum_k \mu \left(m_k(c_k') - m_k(c_k) \right) \leq m(c) + \sum_k \mu \left(m_k(c_k') - m_k(c_k) \right) \\
&= m(c) + \eta \sum_{k:m_k(c_k') > m_k(c_k)} \left[m_k(c_k') - m_k(c_k) \right] - \eta \lambda \sum_{k:m_k(c_k') < m_k(c_k)} \left[m_k(c_k) - m_k(c_k') \right] \\
&\leq m(c) = u(c|c),
\end{aligned}
$$

which implies that c is a PE choice.

To show that c is a PPE, it is sufficient to prove that it is preferred ex ante to all feasible choices within the decisionmaker's choice set (not only PE choices). We show that for any distribution F over D^*, we have $U(F|F) \leq \int m(c')dF(c')$ $(\leq m(c))$.

We prove dimension by dimension. The gain-loss utility part of $U(F|F)$ in dimension k is

(14) $$\frac{1}{2} \iint \left[\mu(m_k(c_k) - m_k(r_k)) + \mu(m_k(r_k) - m_k(c_k)) \right] dF_k(c_k) dF_k(r_k).$$

By assumption A2, this is non-positive.

Finally, to show that a lottery F that is non-deterministic or does not maximize $m(\cdot)$ is not a PPE, we show that c is strictly preferred to these outcomes ex ante. This is obvious for deterministic outcomes that do not maximize $m(\cdot)$. And for a non-deterministic F_k, Expression (14) is strictly negative, establishing the claim for that case as well. $\square$

Proof of Proposition 4.

1. Suppose the driver expects to work in the afternoon with probability q, and her daily income *net of* w^m to be distributed according to $H(\cdot)$. If the driver's realized morning income net of w^m is $\tilde{w}^m$, then her utility from working in the afternoon for wage I^a is

(15)
$$w^m + \tilde{w}^m + I^a - 2f + \eta \int_{-\infty}^{\tilde{w}^m + I^a} [(\tilde{w}^m + I^a) - w] dH(w) - \eta\lambda \int_{\tilde{w}^m + I^a}^{\infty} [w - (\tilde{w}^m + I^a)] dH(w) - (1-q)\eta\lambda f,$$

while her utility from not working is

(16)
$$w^m + \tilde{w}^m - f + \eta \int_{-\infty}^{\tilde{w}^m} [\tilde{w}^m - w] dH(w) - \eta\lambda \int_{\tilde{w}^m}^{\infty} [w - \tilde{w}^m] dH(w) + q\eta f.$$

Hence, she works if

(17)
$$(1+\eta)I^a + \eta(\lambda-1)I^a(1 - H(\tilde{w}^m + I^a)) + \eta(\lambda-1) \int_{\tilde{w}^m}^{\tilde{w}^m + I^a} (w - \tilde{w}^m) dH(w) \geq [1+\eta+(1-q)\eta(\lambda-1)]f.$$

Notice that the left-hand side of Inequality 17 is greater than or equal to $(1+\eta)I^a$, and the right-hand side is less than or equal to $(1+\eta\lambda) \cdot f$. Hence, for any $I^a > w_{max}$, the left-hand side is greater. Conversely, the left-hand side is less than or equal to $[1+\eta+\eta(\lambda-1)H(\tilde{w}^m)] \cdot I^a \leq (1+\eta\lambda) \cdot I^a$, while the right-hand side is greater than or equal to $(1+\eta) \cdot f$. Hence, for any $I^a < w_{min}$, the left-hand side is lower. Since the left-hand side of the inequality is strictly increasing and continuous in I^a while the right-hand side is constant in I^a, there is a unique $I^a \in [w_{min}, w_{max}]$ where the two sides are equal. Call this value $g(\tilde{w}^m)$. Clearly, the driver works in the afternoon whenever $I^a > g(\tilde{w}^m)$, but not if $I^a < g(\tilde{w}^m)$.

The left-hand side of Inequality 17 is differentiable in $\tilde{w}^m$, with the derivative being $-\eta(\lambda - 1)[H(\tilde{w}^m + I^a) - H(\tilde{w}^m)] \leq 0$. Hence, $g(\cdot)$ is non-decreasing, and whenever $H(\tilde{w}^m + g(\tilde{w}^m)) -$

$H(\tilde{w}^m) > 0$, it is strictly increasing at $\tilde{w}^m$. Since $Prob[I^a < w_{min}] > 0$, for any realized morning wage the probability that the driver does not work in the afternoon is positive. Thus, $H(\tilde{w}^m + g(\tilde{w}^m)) - H(\tilde{w}^m) > 0$ if $F^m(\tilde{w}^m + g(\tilde{w}^m)) - F^m(\tilde{w}^m) > 0$. This completes the proof.

2. Let a "compulsory-work PPE" refer to a PPE in the decision problem in which the driver must work in the morning. Notice that any participation PPE is a compulsory-work PPE, and in a compulsory-work PPE the Expressions 15 and 16 describe the driver's utilities from working and not working in the afternoon. Now since w^m additively enters both the driver's utility from working in the afternoon (Expression 15) and her utility from not working in the afternoon (Expression 16), if $H(\cdot)$ and q describe a compulsory-work PPE for some (w^m, F^m, F^a), they describe a compulsory-work PPE for any $(w^{m\prime}, F^m, F^a)$: with the same expectations of behavior as a function of $\tilde{w}^m$, the same behavior remains optimal.

Suppose a compulsory-work PPE described by $H(\cdot)$ and q is a PPE for some (w^m, F^m, F^a). Then, by the above argument $H(\cdot)$ and q also describe a compulsory-work PPE for any $(w^{m\prime}, F^m, F^a)$ such that $w^{m\prime} \geq w^m$. Furthermore, since (relative to the former compulsory-work PPE) in the latter compulsory-work PPE the distribution of income is just shifted to the right by a constant, increasing the driver's expected utility by the same constant, this compulsory-work PPE is also a PPE.

3. Notice that for any F^m and F^a, the set of w^m such that a compulsory-work PPE for (w^m, F^m, F^a) is a PPE for (w^m, F^m, F^a) is closed. And clearly a compulsory-work PPE is a PPE for a sufficiently large w^m, and not a PPE for a sufficiently small w^m. Combining these facts with Part 2, the set of w^m for which a compulsory-work PPE is a PPE is of the form $[\underline{w}^m, \infty)$ for some $\underline{w}^m \in \mathbb{R}$. Since PPE exists and has non-participation whenever it is not a participation PPE, this means that non-participation is a PPE for any $w^m < \underline{w}^m$. Since the set of w^m for which non-participation is a PPE is also closed, non-participation is a PPE for $w^m = \underline{w}^m$. Finally, since an increase in w^m shifts the distribution of income from participation to the right while leaving income from non-participation unchanged, non-participation cannot be a PPE for any $w^m > \underline{w}^m$. $\square$

Department of Economics
University of California, Berkeley

References

Becker, Gordon M.; DeGroot, Morris H. and Marschak, Jacob. "Measuring Utility by a Single-Response Sequential Method." Behavioral Science, July 1964, 9, pp. 226-232.

Benartzi, Shlomo and Thaler, Richard H. "Myopic Loss Aversion and the Equity Premium Puzzle." Quarterly Journal of Economics, February 1995; 110(1), pp. 73-92.

Bowman, David; Minehart, Deborah and Rabin, Matthew. "Loss Aversion in a Consumption-Savings Model." Journal of Economic Behavior and Organization, February 1999, 38(2), pp. 155-78.

Breiter, Hans C.; Aharon, Itzhak; Kahneman, Daniel; Dale, Anders and Shizgal, Peter. "Functional Imaging of Neural Responses to Expectancy and Experience of Monetary Gains and Losses." Neuron, May 2001, 30, pp. 619-639.

Camerer, Colin; Babcock, Linda; Loewenstein, George and Thaler, Richard. "Labor Supply of New York City Cabdrivers: One Day at a Time." Quarterly Journal of Economics, August 1997; 112(2), pp. 407-441.

Compte, Olivier and Jehiel, Philippe. "Bargaining with Reference Dependent Preferences." Working Paper, March 2003.

DeMeza, David and Webb, David. "Principal Agent Problems with Prospect Theory Preferences: An Application to Executive Stock Options," Working Paper, London School of Economics, 2003.

Farber, Henry S. "Is Tomorrow Another Day? The Labor Supply of New York Cab Drivers." Journal of Political Economy, 2005; 113(1), pp. 46-82.

Farber, Henry S. "Reference-Dependent Preferences and Labor Supply: The Case of New York City Taxi Drivers." Princeton University Industrial Relations Working Paper #497, 2004.

Geanakoplos, John; Pearce, David and Stacchetti, Ennio. "Psychological Games and Sequential Rationality." Games and Economic Behavior, 1989, 1(1), pp. 60-79.

Goette, Lorenz, Fehr, Ernst, and Huffman, David. "Loss Aversion and Labor Supply." Journal of the European Economic Association, April-May 2004, 2(2-3), pp. 216-228.

Gul, Faruk. "A Theory of Disappointment Aversion." Econometrica, 1991, 59(3), pp. 667-686.

Gul, Faruk and Pesendorfer, Wolfgang. "Temptation and Self-Control." Econometrica, November 2001, 69(6), pp. 1403-35.

Heidhues, Paul and Kőszegi, Botond. "The Impact of Consumer Loss Aversion on Pricing," Working Paper, 2005.

Heidhues, Paul and Kőszegi, Botond. "Competition and Price Variation When Consumers Are Loss Averse," Working Paper, 2006.

Kahneman, Daniel. "A Perspective on Judgment and Choice; Mapping Bounded Rationality," American Psychologist, 2003, 58(9), pp. 697-720.

Kahneman, Daniel; Knetsch, Jack L. and Thaler, Richard H. "Experimental Tests of the Endowment Effect and the Coase Theorem." Journal of Political Economy, December 1990, 98(6), pp. 1325-48.

Kahneman, Daniel; Knetsch, Jack L. and Thaler, Richard H. "The Endowment Effect, Loss Aversion, and Status Quo Bias: Anomalies." Journal of Economic Perspectives, Winter 1991, 5(1), pp. 193-206.

Kahneman, D. and Lovallo, D. (1993), "Timid choices and bold forecasts. A cognitive perspective on risk taking." Management Science, 39, pp. 17-31.

Kahneman, Daniel and Tversky, Amos. "Prospect Theory: An Analysis of Decision under Risk." Econometrica, March 1979, 47(2), pp. 263-91.

Köbberling, Veronika, Schwieren, Christiane, and Wakker, Peter P. "Prospect-Theory's Diminishing Sensitivity Versus Economic's Intrinsic Utility of Money: How the Introduction of the Euro Can Be Used to Disentangle the Two Empirically." Working Paper, December 2004.

Köbberling, Veronika and Wakker, Peter P. "An Index of Loss Aversion." Journal of Economic Theory, 2005, 122, pp. 119-131.

Kőszegi, Botond. "Utility from Anticipation and Personal Equilibrium." Working Paper, 2005.

Kőszegi, Botond and Rabin, Matthew. "A Model of Reference-Dependent Preferences," UC Berkeley, Department of Economics, Working Paper #1061, 2004.

Kőszegi, Botond and Rabin, Matthew. "Reference-Dependent Risk Preferences," Working Paper, 2005.

Laibson, David. "Golden Eggs and Hyperbolic Discounting." Quarterly Journal of Economics, May 1997, 112(2), pp. 443-77.

Larsen, Jeff T.; McGraw, A. Peter; Mellers, Barbara A. and Cacioppo, John T. "The Agony of Victory and Thrill of Defeat: Mixed emotional reactions to disappointing wins and relieving losses." Psychological Science, May 2004, 15(5), pp. 325-330.

List, John. "Does Market Experience Eliminate Market Anomalies?" Quarterly Journal of Economics, February 2003, 118(1), pp. 41-71.

Loewenstein, George; O'Donoghue, Ted and Rabin, Matthew. "Projection Bias in Predicting Future Utility." Quarterly Journal of Economics, November 2003, 118(4), pp. 1209-1248.

Masatlioglu, Y. and Ok, Efe. "Rational Choice with Status-Quo Bias." Journal of Economic Theory, March 2005; 121(1), pp. 1-29

Medvec, Victoria Husted, Madey, Scott F. and Gilovich, Thomas. "When Less Is More: Counterfactual Thinking and Satisfaction Among Olympic Medalists." Journal of Personality and Social Psychology, October 1995, 69(4), pp. 603-610.

Mellers, Barbara; Schwartz, Alan and Ritov, Ilana. "Emotion-Based Choice." Journal of Experimental Psychology: General, September 1999, 128(3), pp. 332-45.

Novemsky, Nathan and Kahneman, Daniel. "The Boundaries of Loss Aversion." Journal of Marking Research, May 2005, 42, pp. 119-128.

O'Donoghue, Ted and Rabin, Matthew. "Doing It Now or Later." American Economic Review, March 1999, 89(1), pp. 103-24.

Plott, Charles and Zeiler, Kathryn. "The Willingness to Pay/Willingness to Accept Gap, the 'Endowment Effect,' Subject Misconceptions and Experimental Procedures for Eliciting Valuations." American Economic Review, June 2005, 95(3), pp. 530-545.

Read, Daniel, Loewenstein, George, and Rabin, Matthew, "Choice Bracketing," Journal of Risk and Uncertainty 19 (1999), pp. 171-197.

Sagi, Jacob. "Anchored Preference Relations," forthcoming, Journal of Economic Theory, 2005.

Shalev, Jonathan. "Loss Aversion Equilibrium," International Journal of Game Theory, 2000, 29, pp. 269-87.

Stone, Rebecca. "Behavioural Implications of Loss Aversion," Working Paper, University College London, 2005.

Sugden, Robert. "Reference-Dependent Subjective Expected Utility." Journal of Economic Theory, August 2003, 111(2), pp. 172-191.

Sydnor, Justin. "Sweating the Small Stuff: The Demand for Low Deductibles in Homeowners Insurance." Working Paper, UC Berkeley, 2005.

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