Optimal After All: Information-Based Explanations Of Behavioral Decision Phenomena

by

José Fernando Camões de Mendonça Oliveira e Silva

M.B.A., Universidade Católica Portuguesa, Portugal, 1994
Lic. Computer Engineering, Universidade de Coimbra, Portugal, 1989

SUBMITTED TO THE SLOAN SCHOOL OF MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2000

©Massachusetts Institute of Technology. All rights reserved.

Author: ______________________________
Sloan School of Management
May 25, 2000

Certified by: ______________________________
Drazen Prelec
Professor of Management Science
Thesis Supervisor

Accepted by: ______________________________
Birger Wernerfelt
Professor of Management Science
Chairman, Department Committee

ARCHIVES
MASSACHUSETTS INSTITUTE
OF TECHNOLOGY
JUN 21 2000
LIBRARIES

ACKNOWLEDGEMENTS

I dedicate this thesis to my family, who have always given me nurturing support in all endeavors of my life.

The ideas in this thesis, and most of my outlook on the field, are a product of the environment of the Sloan School Marketing group. Dan Ariely, Ely Dahan, Sandy Jap, John Little, Duncan Simester, Nader Tavassoli, Glen Urban, and especially John Hauser, not only contributed to this environment, but also gave me comments and feedback on my sometimes wildly unreasonable ideas. I'd also like to thank the other students in the program, particularly Rebecca Hamilton, Dina "That's Awesome!" Mayzlin, Robert Zeithammer, Florian Zettelmeyer, and especially Dave Godes, without whom I wouldn't probably have survived my first semester here. Some people from outside the Marketing group also had a large impact on this research. Sendhil Mullainathan has suggested several avenues of exploration of the categorization ideas, and Nuno Vasconcelos and Manuela Pereira have pointed out many information theory issues.

I especially want to thank two faculty members without whom I would probably never written anything vaguely resembling this thesis: Drazen Prelec and Birger Wernerfelt. Drazen's advice and high standards turned a half-baked idea that "the payoff scale gets compressed near zero because that's where most decisions are" into a consistent explanation for several B.D.T phenomena. Birger advised me on many research ideas and opened my eyes to the world of economic thinking; all I think I know about writing elegant game-theoretic models I learned as a result of my interaction with Birger. There are few experiences comparable to the intellectual growth they helped me achieve, and I wholeheartedly thank them both for it.

I am very grateful for the financial support of the Sloan School, the Center for Innovation on Product Development and the Praxis XXI program.

Last but not least, I want to thank Sergei Rachmaninov for the escape studying his second piano concerto provided.

Abstract

Real decisions deviate from normative theory of decision making in systematic ways. Choices over risky alternatives should reflect preferences over the distributions over final states of wealth created by these alternatives. Also, choices over payoffs at different times in the future should not vary with simple passing of time. Both these implications of normative theory are violated in real decisions: framing of alternatives leads to reversal of preference, and decisions over far payoffs are less myopic than over near payoffs.

These phenomena have been described by Prospect Theory and by hyperbolic discounting. This thesis contributes an explanation of the elements of these descriptive models. It is shown that a rational agent using categorical reasoning will behave as described by this model. The loss of information in categorization is sustainable as long as the errors in decision-making due to the categorization are on average lower than the cost of thinking without the categorization.

The first essay in this thesis, distortions from encoding, looks at the effects of categorizing inputs according to their information content. The category labels are then operated upon by a concave utility function (for payoffs) or an exponential discounting (for time). This provides an explanation where the systematic error results from both the categorization and the use of a normative theory over category labels.

The second essay tests an implication of categorical reasoning over payoffs. Prospect theory does not bind the attitudes towards risk on the positive side to those on the negative side of the reference point. The categorization model does. To test this model, the essay presents a contrast with a quasi-symmetric model of value, which makes opposite predictions. Experimental validation shows the categorization model to be supported and the quasi-symmetric model to be rejected.

The third essay postulates a rational agent that makes no systematic errors other than loss of information. This agent makes decisions that are consistent with the underlying normative theories, but is constrained to either dominance seeking, or reasoning with category representatives. It is shown that the decision phenomena described by Prospect Theory and by hyperbolic discounting are explained by this categorization with generic distributions.

Contents

Chapter 1

Distortions from Encoding

1.1 Introduction

This paper shows that well-established deviations from normative decision-making models can be explained by discrete reasoning and categorization. The logic is illustrated with the prospect theory value function, and then extended to the domains of time and probability (yielding, respectively, hyperbolic discounting and the probability weighting function).

To illustrate the idea, suppose that you were to program a computer agent to process information for you. The main constraints on the agent are that it has discrete reasoning (it uses a few labels to represent all possible values in a scale) and that it is costly to reprogram it for each application.

To use discrete reasoning, the agent will need two subsystems: an *encoder* which maps the continuous dimension (say, payoffs) into the labels; and a *decoder*, which assigns meaning (say, utility) to the labels.

Suppose, for instance, that the agent is evaluating the price of personal information manager software. It will have to encode prices into labels like "Expensive", "Cheap", etc. The prices are encoded efficiently by first standardizing them (so that they are relative to, and meaningful for the category[1]). In the example the input xs are standardized to have mean zero and variance one. To do so, either the agent has information about the category under consideration and recalls it, or it can infer it from the available information[2]. Say

[1] If the agent is making decisions over cars it will use a scale where "Expensive" means something very different than what it means when the agent is evaluating t-shirts.

[2] Inference from problem data in the absence of a strong adequate norm is very common, as seen in Gourville (1998), Kahneman and Miller (1986), Prelec, Wernerfelt, and Zettelmeyer (1997), Simonson and

that in this case $\mu = -30$ (note that prices are negative payoffs) and $\sigma = 20$ and that the agent is analyzing two stores, Quick and Slow, and two models, the Einstein and the Yogi. Both stores sell both products, but there is some hassle c in obtaining them from store Slow. The prices are as follows:

	Slow	Quick
Einstein	60	90
Yogi	10	20

The standardized prices are:

	Slow	Quick
Einstein	- 3/2	- 3
Yogi	1	1/2

meaning that, for example, the Einstein at Quick is three standard deviations more expensive than the average personal information manager software. These relative prices can now be encoded.

Suppose you choose to partition the scale of payoffs into five intervals (called bins), and assign to each of these bins a label as in table 1.1. The encoder, $e(x)$, is just a system that takes a payoff x and gives it a label ℓ corresponding to the bin x fall into (so $e(x) = \ell$ iff $x \in C_\ell$).

Table 1.1: *Bins and labels for illustration.*

Bin (C_ℓ)	Label (ℓ)
$(-\infty, -0.8417]$	"Very Expensive"
$(-0.8417, -0.2533]$	"Expensive"
$(-0.2533, +0.2533]$	"Moderate"
$(+0.2533, +0.8417]$	"Cheap"
$(+0.8417, +\infty)$	"Very Cheap"

The labels associated with the four possible prices in the example are

$$\begin{aligned}
p = 10: &\quad e((-10+30)/20)) &&= e(1) &&= \text{``Very cheap''} \\
p = 20: &\quad e((-20+30)/20)) &&= e(1/2) &&= \text{``Cheap''} \\
p = 60: &\quad e((-60+30)/20)) &&= e(-3/2) &&= \text{``Very expensive''} \\
p = 90: &\quad e((-90+30)/20)) &&= e(-3) &&= \text{``Very expensive''}
\end{aligned} \tag{1.1}$$

Tversky (1992), Tversky and Kahneman (1991), Tversky and Simonson (1993), and Wernerfelt (1995).

Converting numbers into labels is only half of the problem. To create meaning, the labels need to be decoded. Since the example is about prices, the decoder will be a utility-like map $\tilde{u}(\ell)$, such that

$$\begin{aligned}\tilde{u}(\text{“Very expensive”}) < \tilde{u}(\text{“Expensive”}) < \tilde{u}(\text{“Moderate”}) < \\ < \tilde{u}(\text{“Cheap”}) < \tilde{u}(\text{“Very cheap”}).\end{aligned} \qquad (1.2)$$

Putting equations (1.1) and (1.2) together in order to determine whether to incur the hassle of shopping at Slow, the agent will conclude the following: regarding the Yogi,

$$\begin{aligned}\tilde{u}(e((-10+30)/20)) - \tilde{u}(e((-20+30)/20)) &= \\ \tilde{u}(\text{“Very cheap”}) - \tilde{u}(\text{“Cheap”}) &> 0,\end{aligned}$$

so for some cases (when $\tilde{u}(\text{“Very cheap”}) - \tilde{u}(\text{“Cheap”}) > c$) it is worth suffering the hassle and buying the Yogi at Slow. For the Einstein,

$$\begin{aligned}\tilde{u}(e((-60+30)/20)) - \tilde{u}(e((-90+30)/20)) &= \\ \tilde{u}(\text{“Very expensive”}) - \tilde{u}(\text{“Very expensive”}) &= 0,\end{aligned}$$

meaning that the agent would never advise incurring the hassle to buy the Einstein at Slow. This, *even though the price savings are much higher in that case.* Although this is a stylized version, the decision problem here is identical to that in Thaler (1980) and Tversky and Kahneman (1981). These papers show that a person is more likely to drive across town to get a \$5.00 discount on a \$15.00 calculator (68% of respondents) than a \$5.00 discount on a \$125.00 jacket (29% of respondents; calculator and jacket conditions were cross-balanced, of course)[3].

To illustrate how this encoder and a decoder might create a discrete approximation to the shape of the prospect theory value function, it is necessary to create a specific version

[3] This requires convex (rather than concave) utility: suppose the starting wealth is w and the driving cost is c; then, concavity of u implies $u(w-c-125+5) - u(w-125) > u(w-c-15+5) - u(w-15)$. Therefore, if utility were concave subjects would be more likely to drive in order to get a \$5 discount on the \$125 jacket than on the \$15 calculator, the opposite of the observation.

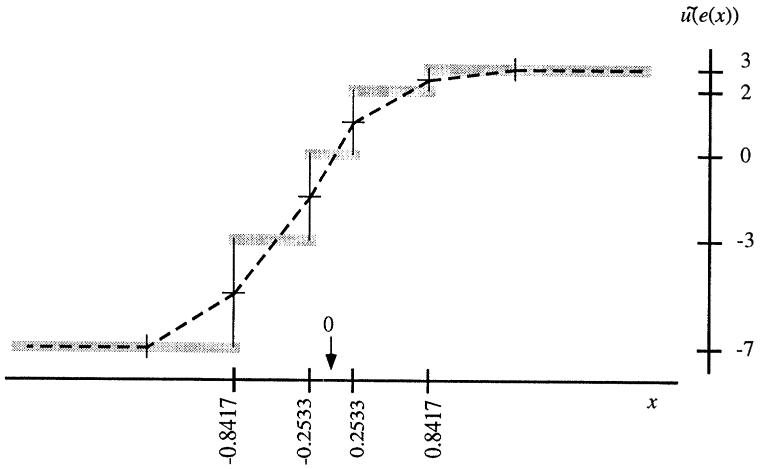

Figure 1.1: *In the example, the encoding and decoding lead to an approximation of the S-shape of v. The horizontal gray lines are $\tilde{u}(e(x))$, and they approximate the linear interpolation dashed black line.*

of equation (1.2). Using a "concave" $\tilde{u}(\cdot)$ like

$$\begin{aligned} &\tilde{u}(\text{“Very Expensive”}) = -7;\\ &\tilde{u}(\text{“Expensive”}) = -3;\\ &\tilde{u}(\text{“Moderate”}) = 0;\\ &\tilde{u}(\text{“Cheap”}) = 2;\\ &\tilde{u}(\text{“Very Cheap”}) = 3. \end{aligned} \tag{1.3}$$

we obtain the desired result (see figure 1.1 for illustration)[4].

Relation to literature: This paper uses the prospect theory value function (Kahneman and Tversky, 1979; Tversky and Kahneman, 1991; Thaler, 1985) as illustration. Deviations from normative theory (Von Neumann and Morgenstern, 1944) are well documented; see for instance (Allais, 1953; Camerer, 1995; Harless and Camerer, 1994; Kahneman and Tversky, 1996; Machina, 1987; Rabin, 1998). Cumulative prospect theory (Tversky and Kahneman, 1992; Fennema and Wakker, 1997), which includes rank dependence (Luce, 1988; Luce and

[4] Note that the values in equation (1.3) are "concave" in the sense that the difference between consecutive levels is decreasing; the second-order difference (the difference between consecutive differences) is -1; since for discrete maps differences are the equivalent of derivatives, the negative second-order difference may, with a slight abuse of language, denote concavity of $\tilde{u}$.

Fishburn, 1991), solves some problems with the original prospect theory; its treatment of prospects by rank order is explained in this paper by the standardization and encoding.

Another important stream of literature is that of bounded rationality and cost of thinking (Gingerenzer and Goldstein, 1996; Simon, 1955; Shugan, 1980; Rubinstein, 1998). Models of semi-orders and similarity (Beja and Gilboa, 1992; Bridges, 1983; Gilboa and Lapson, 1995; Rubinstein, 1988; Tversky, 1977) are also related to the model in this paper: fat indifference curves may be an outcome of categorization of stimuli into bins, but the approximation makes indifference intransitive.

A final stream of literature related to this paper is that of categorization of numerical stimuli (Holyoak and Alker, 1976; Lakoff, 1987; Macmillan, Kaplan, and Creelman, 1977; Olson and Budescu, 1997; González-Vallejo, Erev, and Wallsten, 1994; Viswanathan and Childers, 1996, 1997; Wallsten, Budescu, and Zwick, 1993) and approximate reasoning (Armstrong, Gleitman, and Gleitman, 1983; Basu, 1984; Huttenlocher and Hedges, 1994; Zadeh, 1965, 1978, 1982).

Structure of paper: The next section presents the rationale behind the axioms of the formal model, which is developed section 1.3. Section 1.4 has the extensions to time and probability. The last section summarizes the paper.

1.2 Rationale

In this section I will illustrate the logic behind the formal model. I present the formal assumptions in an informal way and discuss how they result from two assumptions on reasoning.

The first central assumption is that reasoning is discrete[5]. To have discrete reasoning over continuous quantities it is necessary to encode these quantities into some discrete set of labels, and then decode, or attach meaning to these labels.

The second main assumption is that it is costly to set up a pair of encoder and decoder specially designed for each problem. In appendix A.1 there is an illustration of the effects of small changes in parameters of the decision problem and their effects on encoding and decoding. The conclusion is that optimality requires creation of encoder/decoder pairs from basic principles for every problem. Since it is costly to create problem-specific encoder/decoder pairs, in this paper I assume that a general purpose system will leave the decision-maker better off on average: for each problem, the general-purpose encoder/decoder performs worse than a problem-specific encoder/decoder; the paper assumes that the average losses in performance are smaller than the setup cost, hence on average the general-purpose encoder/decoder should perform better.

If the encoder is general-purpose, then data entering it should be standardized: bins and labels for cars and for chewing gum can be reused as long as the variance and the mean are taken care of: by standardizing prices chewing gum that costs twice the standard deviation more than the average price is labeled "Very expensive", exactly as a car than costs twice the standard deviation more than the average car, even if the price of the former is \$1.50 and the price of the latter is \$40,000.00.

Once the data is standardized, it can be encoded by a general-purpose set of bins. These bins are the result of a optimization process: they are chosen to maximize the information transmitted; a result in information theory states that this being the case, the optimal encoder is one in which each bin has equal probability of being selected[6]. In the case of

[5] Discrete, and essentially symbolic, reasoning over numeric quantities is a result of cognitive limitations; see for example Miller (1956). Given a linguistic reasoning ability — see Pinker (1994), Pinker and Bloom (1992), or Deacon (1997), for instance — it makes sense for a capacity-constrained decision-maker to reuse that ability for new problems, like numerical inference.

[6] The result itself is a classic result in information theory, see for instance (Gersho and Gray, 1992); Parducci (1965) and Krumhansl (1978) offer empirical evidence that decision-makers follow this encoding criterion.

continuous variables under consideration, these bins are intervals in the real line. In order to determine what these intervals are, it is necessary to assume a probability distribution.

Empirically, the distribution that seems more appropriate is one that has high probability density near the mean, and low density away from the mean. This shape can also be derived by two other, more theoretical, criteria.

If the distribution is the most general-purpose possible over $\mathbb{R}$, a reasonable argument would be that it should require the minimal set of assumptions. A distribution $f_X(x)$ requires the minimal set of assumptions if it maximizes entropy. Entropy of a distribution with p.d.f. $f_X(\cdot)$ is defined as

$$H(f_X) = -\int_{-\infty}^{+\infty} f_X(x) \log f_X(x) dx.$$

Using calculus of variations to maximize this functional under the constraints that the mean and variance exist, and that $f_X \geq 0$ and integrates to one (and these constraints are necessary, otherwise the problem is not identified), results in $f_X(x)$ being a Gaussian Normal. Hence, the Gaussian Normal is the distribution over $\mathbb{R}$ that requires the least information to be assumed about its characteristics. Using the least-assumption criterion then yields a distribution which is empirically sound.

A final argument in favor a distribution that is roughly shaped like a Normal is that of problem filtering. Suppose the decision-maker has some capability for choosing decision problems, or at least to avoid some. Then the decision-maker would invest some effort to get rid of very negative payoffs: these are in a region of the utility curve where the slope is very steep, leading to large potential losses. On the other hand, the decision-maker would not invest much effort to increase the likelihood of very high payoffs, since at the high payoff levels the utility is flat and returns to effort are very small. Under these circumstances, the resulting distribution of payoffs is roughly compatible with the two previous criteria.

Using a distribution that looks like a Normal leads to bins that are narrow near zero and wide away from zero, since they have to include the same probability and the density is decreasing away from zero. Bins like those in table 1.1 are obtained from a Normal distribution.

The second part of the encoder/decoder pair is the decoder. The decoder is designed by using the normative axioms which are appropriate for the dimension encoded; only, since the decoder acts on labels, this is as if the decision-maker had an information barrier, which stops her from realizing the distortion created by the encoding.

It is important to note that it is the sharing of the encoder across several types of dimensions (payoffs, times, temperatures, height, etc.) that make the optimization objective for the encoder independent of the utility function. If the encoder were designed specifically for a specific payoff distribution without standardization, then it would make sense to use bins that were narrower on the losses side and wider on the gains side (relative to the equiprobable ones) in order to account for the fact that errors on the losses side are more costly in utility terms (since the utility function is concave) than errors on the gains side. (To be precise, errors are ever more relevant to the left and ever less relevant to the right.)

Note also that even though the utility-based decoder may be specific for payoffs, the standardization prior to encoding destroys information about the concavity, which makes matching the decoder to the encoder impossible. Standardizing leads to loss of information about the concavity of u: let x_1, x_2 be two payoffs from different problems such that $\mu_1 \neq \mu_2$ and $\sigma_1 \neq \sigma_2$; then, in general the concavity of $z_1 \doteq (x_1 - \mu_1)/\sigma_1$ will be different from the concavity of $z_2 \doteq (x_2 - \mu_1)/\sigma_2$ for any value of $z_1 = z_2$.

To review the chain of inference: (i) discrete reasoning requires encoding and decoding; (ii) cost of setup leads to use of general-purpose encoder/decoder pairs; (iii) in order to reuse the encoder efficiently, inputs are standardized; (iv) the encoder is optimized for minimum error in representing the inputs; (v) the distribution of inputs is shaped like a Gaussian Normal; (vi) standardization implies that information about the concavity of u is lost; (vii) thus, the decoder is derived from the normative axioms. All these steps are stated as assumptions in the formal model, although they are consequences of the more basic cost of setup and discrete reasoning assumptions. The formal derivation in the next section is illustrated in figure 1.2.

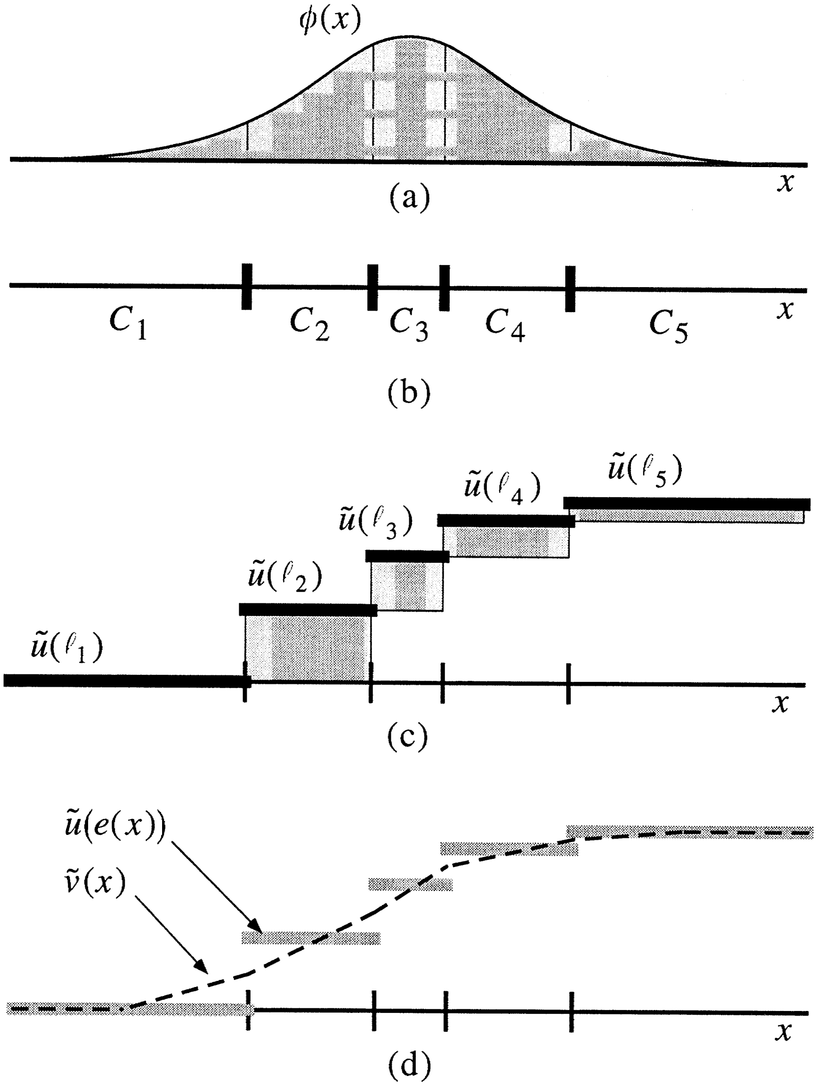

Figure 1.2: *Derivation of the approximation to the value function with five labels. Panel (a) is the probability density function $\phi(x)$ for a Gaussian Normal $\mathcal{N}(0,1)$, with the quintiles marked by the vertical dashed lines. Panel (b) shows where the cutoff points for a 5-label encoder for distribution in panel (a) are located; bins $C_1, \ldots, C_5$ are equiprobable, hence they correspond to the quintiles. Note that the size of the bins increases with distance to zero. Panel (c) shows what happens when a decoder $\tilde{u}(\ell)$ with decreasing first-order differences is applied to the labels of the encoder in panel (b). Panel (d) shows that the outcome of (c) approximates a s-shaped function $\tilde{v}(x)$ with steeper slopes on the negative side, hence the shape of $v(\cdot)$.*

1.3 Model

I start by defining the constructs of discrete reasoning, and the criterion of optimality.

Definition 1 (Building blocks for discrete reasoning). *Define the following:*

1. *Reasoning is discrete: the number of bins (and labels) is N, finite.*

2. *A bin C_ℓ is a subset of $\mathbb{R}$. Denote a set of N bins forming a partition of $\mathbb{R}$ by $\mathcal{C}$.*

3. *The set of labels is $\mathcal{L} = \{-L, \dots, -1, 0$ (if N is odd), $1, \dots, L\} \subset \mathbb{Z}$ where $L = N/2$ if N is even and $L = (N-1)/2$ if N is odd*[7].

4. *Encoding is a map $e_{\mathcal{C}}(x) : \mathbb{R} \mapsto \mathcal{L}$ such that $e_{\mathcal{C}}(x) = \ell$ if and only if $x \in C_\ell$.*

5. *A decoder is a map $\tilde{u}_{\mathcal{C}}(\ell) : \mathcal{L} \mapsto \mathbb{R}$. Holding $\mathcal{C}$ fixed, a decoder is completely defined by a set of N numbers $\tilde{u}_{-L}, \dots, \tilde{u}_L$.*

Definition 2 (Optimality). *Given an appropriate norm, $||\cdot||$*[8]*, for a random variable $X \sim \mathcal{D}$ with p.d.f. $F_X(x)$, and a utility function $u(x)$,*

1. *The loss from using a encoder/decoder pair $(e_{\mathcal{C}}, \tilde{u})$ is defined as*

$$L(\mathcal{D}, u, e_{\mathcal{C}}, \tilde{u}) \doteq \int_{\mathbb{R}} ||u(x) - \tilde{u}(e_{\mathcal{C}}(x))|| dF_X(x).$$

2. *The optimal encoder/decoder pair is obtained by finding the maximum entropy encoding bins which minimize losses,*

$$\min_{\mathcal{C}} \left\{ L(\mathcal{D}, u, e_{\mathcal{C}}, \tilde{u}) \right\} \tag{1.4}$$

subject to the constraint of optimal decoding,

$$\tilde{u}(e_{\mathcal{C}}(x)) = E_z \left[u(z) | z \in \text{support}(e_{\mathcal{C}}(x)) \right] \tag{1.5}$$

From this optimality criterion, the following important observations can be derived. The first is a simplification of the choice of bins: not all possible partitions of $\mathbb{R}$ need to be considered.

[7] $\mathbb{Z}$ is used for notational convenience: it already includes a successor function and an order relation. Zero is excluded when N is even, so in these cases $\text{succ}(-1) \doteq 1$.

[8] For this paper's purposes, the norm will be the mean squared error; in other words $||x - y|| \doteq (x - y)^2$. This is usual in quantizing problems; it makes for more tractable problems than most alternatives, including the absolute value.

Observation 1. *Each bin is an interval. Also, the encoder is fully defined by a set of cutoff points, $k_{-L+1}, \dots, k_{-1}$, $k_0 = 0$ (if N is even), $k_1, \dots, k_{L-1}$.*

(Proof in the appendix.) The second is the observation that if the optimal encoder/decoder pair is used, the result is unbiased; hence, the systematic biases that this paper aims at explaining require information loss:

Observation 2 (Unbiasedness of perfect information encoding). *If $e_C(x)$ and $\tilde{u}(\ell)$ solve equations (1.4) and (1.5) then $E[\tilde{u}(e(x))] = u(x)$.*

This observation follows from equation (1.5); a more general case is proved in Gersho and Gray (1992), Lemma 6.2.2.

The reason for information loss is that it is costly to optimize for each individual problem (illustrated in the appendix, section A.1). (Notation: define e_G and $\tilde{u}_G$ to be general-purpose encoder and decoder.) That leads to the following assumption.

Assumption 1. *For a set of distributions $\mathbb{D}$, with a probability measure $F_{\mathcal{D}}(\mathcal{D}_i)$ defined over its elements $\mathcal{D}_i$, setup cost S dominates losses:*

$$\int_{\mathcal{D}_i \in \mathbb{D}} \left[L(\mathcal{D}_i, u, e_G, \tilde{u}_G) - L\Big(\mathcal{D}_i, u, e_{C^*(\mathcal{D}_i, u)}, \tilde{u}^*(\mathcal{D}_i, u)\Big)\right] dF_{\mathcal{D}}(\mathcal{D}_i) < S \tag{1.6}$$

An immediate consequence of this assumption is that for $\mathbb{D}$ a general-purpose encoder/decoder pair is better than solving the optimization problem for each decision. I will now postulate the characteristics of such general purpose pair.

Assumption 2 (Characterization of $e_G(\cdot)$). *For the encoder,*

1. *$e_G(x)$ is designed for $E[x] = 0$, $Var[x] = 1$.*

2. *$e_G(x)$ minimizes $E[||x - \tilde{x}(\ell)||]$ where $\tilde{x}(\ell) = E[x | x \in C_\ell]$.*

(From now on, I will drop the G subscript on $e(\cdot)$.) Given point 2 of this assumption, I can derive the following observation

Observation 3. *The optimal $e(x)$ is such that $\Pr(x \in C_\ell) = 1/N$ for all ℓ.*

This is a consequence of maximum entropy, applied to assumption 2, point 2.

In order to design the encoder bins, it is necessary to characterize the space of distributions $\mathbb{D}$.

Assumption 3 (Characterization of $\mathbb{D}$). *For all $\mathcal{D} \in \mathbb{D}$ and $Z \sim \mathcal{D}$ described by a p.d.f. $f_Z(z)$, we have:*

1. *$f_Z(z)$ is symmetric, smooth except possibly in a set of measure zero, and decreasing in $|z - E[Z]|$*

2. *$E[Z] < \infty$, $Var[Z] < \infty$.*

This characterization, plus observations 1 and 3, leads to:

Observation 4. *For each $\mathcal{D} \in \mathbb{D}$, the optimal $\mathcal{C}$ has* $\text{support}(C_\ell)$ *increasing in $|\ell|$ for all ℓ.*

(Proof in appendix.)

The second part of the encoder/decoder pair is the decoder, and it follows the normative axioms for utility. These imply:

Assumption 4 (Characterization of $\tilde{u}$). *Fix the following characteristics of $\tilde{u}$:*

1. *$\tilde{u}(\ell) < \tilde{u}(\ell + 1)$*

2. *$\tilde{u}(\ell + 2) - \tilde{u}(\ell + 1) < \tilde{u}(\ell + 1) - \tilde{u}(\ell)$*

3. *Without loss of generality,* $\max\{\tilde{u}(\ell)\} - \min\{\tilde{u}(\ell)\}$ *is constant across N.*

There is also a condition on the concavity of $\tilde{u}$,

Assumption 5 (Not too concave condition). *For all $\ell \leq e(0)$,*

$$\frac{k_{\ell+3} - k_{\ell+2}}{k_{\ell+2} - k_{\ell+1}} \leq \frac{\tilde{u}(\ell + 3) - \tilde{u}(\ell + 1)}{\tilde{u}(\ell + 2) - \tilde{u}(\ell)} \tag{1.7}$$

Putting these together, I derive the main result.

Result 1 (Main result). *Under assumptions 2, 3, and 4*

1. *Fix a number of labels N. There exists a partition $\mathcal{C}^*$ of $\mathbb{R}$ with N intervals which forms a generic encoder for $\mathbb{D}$, such that* $\text{support}(C_\ell)$ *is increasing with distance of C_ℓ to 0, for all $C_\ell \in \mathcal{C}^*$.*

2. *If condition (1.7) holds, then for each $\epsilon > 0$ there exists a N and a function $\tilde{v}(\cdot)$ such that $\forall_x |\tilde{u}(e(x)) - \tilde{v}(x)| < \epsilon$, such that: $\tilde{v}$ is concave over gains, convex over losses, and $\tilde{v}(z) - \tilde{v}(0) \leq |\tilde{v}(-z) - \tilde{v}(0)|$ for all $z \geq 0$.*

(Proof in the appendix.)

There are a few important points to note about this result:

One of the more obvious points in the result is the not too concave condition (equation 1.7). This condition is mostly a technical condition with not much importance: for instance, if the second order difference is constant at $-\eta$, increasing N leads to a ever smaller part of the η for which $\tilde{u}$ is increasing does not satisfy equation (1.7). This is shown in appendix A.3.

Another obvious point is the approximation. The model of bins as sets does not capture realistically the meaning of linguistic categories. Therefore, instead of considering more complex models of bins — namely bins with probabilistic boundaries or bins as fuzzy sets — I prefer to use the approximation. In the appendix there is the complete description of the linear approximation.

Regarding the number of labels, there is another interpretation of the result. If the decision-maker uses successive refinements, the main result still holds: note that partitioning of a bin, into say two sub-bins, leads to a new encoder where the bins are narrow near zero and become wider away from zero. So a decision-maker using a successive refinement would also have a s-shaped valuation of payoffs. Comparative statics of the number of bins is in appendix A.4.

Another important point is that $\tilde{u}$ need not be strictly concave. In fact, it might just be the case that there are only two levels of first-order difference, one for losses (larger) and one for gains. The semantics of such $\tilde{u}$ would be that the decision-maker has a preference for gains (hence changes in the loss side are more penalized), and a simple algorithm to enforce it.

One last important point is that the bins need not be exactly symmetric. Using an encoder with bins that are narrower on the losses side than on the gains side is better than a symmetric encoder and one could assume that the decision-makers might use some form of shading on the cutoff points. Since a small degree of concavity for the utility inside a category C_ℓ would lead to a small change in the position of the $k_{\ell-1}$ and k_ℓ, those small changes would not defeat the first part of result 1: the s-shape would now be asymmetric even if the utility map had a constant first-order difference, a stronger result. Only high degrees of concavity would lead to bins whose width does not increase with distance to zero, and that is confounded by the number of labels.

1.4 Extensions and predictions

I will now explore the implications of applying the main idea in the reconstruction of the prospect theory value function to other numerical domains in decision-making: time and probability. The main point is that deviations from the normative model can be explained from optimal encoding (in the information-theory sense) followed by decoding which applies the normative axioms to the labels.

1.4.1 Extension to time: from exponential decoding to hyperbolic discounting

Let A and B be streams of utility starting at time T. If a decision-maker chooses A over B at T, then she should always prefer A over B for any time before T. (Otherwise, non-stream-related passing of time would create reversal of preferences and inconsistent behavior.) For example, if an apple now is preferred to an orange tomorrow, then an apple in one year should be preferred to an orange in one year and a day. The normative appeal of exponential discounting of time is derived from this property: the exponential is the only function $\delta(t)$ for which $A \succsim \delta(\tau)B \Leftrightarrow \delta(t)A \succsim \delta(t+\tau)B$ for all $t \geq 0$. In other words, the exponential is the only discounting that guarantees that the simple passing of time will not lead to reversal of preferences.

In reality, however, people who prefer apples today to oranges tomorrow, when prompted to make the choice one year in advance, may well reverse their choices. This behavior leads to time-inconsistent choices by itself, since the decision-maker forced to make decisions in advance will find herself in conflict with her future preferences, and to reversal of stated preferences at the time they are revealed.

Hyperbolic discounting (Ainslie, 1975, 1991; Azfar, 1999; Kirby and Marakovic, 1995; Laibson, 1997; Prelec, 1989; Prelec and Lowenstein, 1991) can explain time-inconsistent behavior. For suitable values of the parameters a hyperbolic discount function underweights differences in (with respect to the exponential) distant outcomes and overweights differences in near outcomes. This may lead to preference reversals[9].

The process used for explaining the value function can be used to explain the form of

[9] It bears mentioning that visceral factors are a different explanation of time-inconsistent preferences. If the decision-makers preferences are variable with time, then, obviously, apparently inconsistent behavior may occur; more so when the decision-maker is unable to forecast the effects on preferences. This is a different and somewhat orthogonal argument to that in the paper.

hyperbolic discounting: standardization ("*long* moments" vs "*short* years"), followed by encoding, followed by negative exponential decoding.

Since time can be thought of as unbounded in both directions (past and future), the appropriate distribution could again be the Normal; after all, the most relevant events are in most cases those in the near future and those in the immediate past. However, a different distribution could be used for analysis of the future only. For a distribution over $\mathbb{R}^+$, the three criteria used before converge: Empirically, most events considered by decision makers tend to be in the near future or immediate; since the future is uncertain, incentive compatibility leads decision-makers to prefer immediate payoffs, ceteris paribus; and, the solution to maximization of entropy over $\mathbb{R}^+$ conditional on the existence of the mean yields the exponential p.d.f., $f_T(t) = \exp(-t)$. So, a specific encoder for positive time would have narrow bins near zero, and bins becoming wider and wider as they moved towards $+\infty$.

If either of these encoding schemes is used to process time (which is a number, after all), this distortion can be explained in the following way[10]: assuming today is the reference point, the difference in labels between today (ℓ_0) and tomorrow (ℓ_1) is much higher than between a year from now (ℓ_2) and a year and a day from now (ℓ_3), since categories are larger with distance to today. (It is even possible that $\ell_2 = \ell_3$.) Therefore, the discounting, if computed using the time labels, as opposed to the original time data, will create the reversal of preferences. This reasoning can be extended to yield the form of hyperbolic discounting.

Hyperbolic discounting, or an approximation thereof, can result from applying a normative discount to the labels of time: these labels, when taken as cardinal, represent a concave map of time; the discount factor associated with t is $e^{-r\,e(t)}$ rather than e^{-rt}. This leads to a function that decreases faster than the exponential at first (because near 0 categories are very small) and then decreases slower than the exponential (because away from zero categories are large). These are the generic characteristics of hyperbolic discounting (see figure 1.3 for an illustration).

Observation 5 (Hyperbolic discounting). *The composition of negative exponential decoding with a encoder with bins increasing in size away from zero creates an approximation to hyperbolic discounting.*

If the encoder for $\mathbb{R}$ instead of the encoder for $\mathbb{R}^+$ is used, then the reversal of preferences with respect to the future should be mirrored with respect to the past. This has to be tested (as intertemporal reversals are) without getting memory into play.

[10] This is related to the idea of similarity in (Rubinstein, 1988)

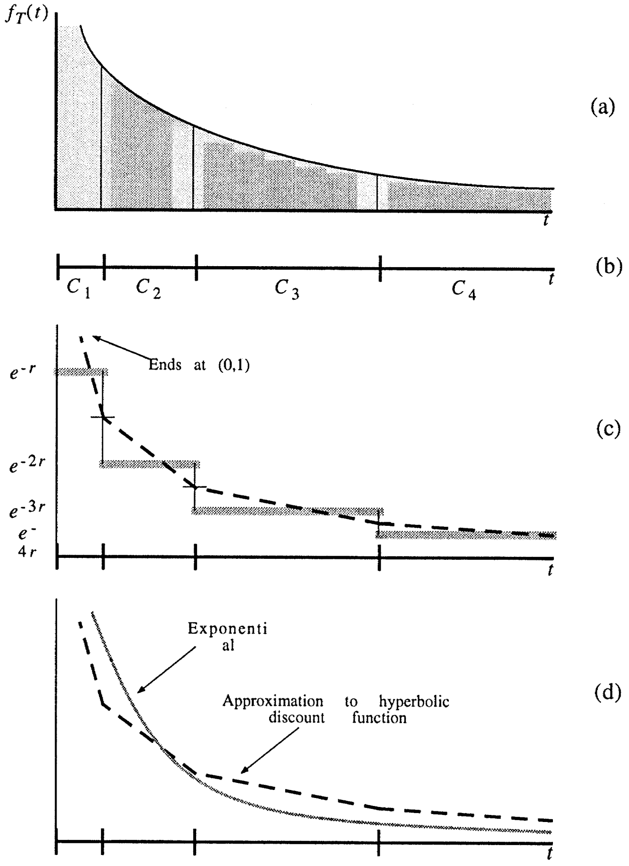

Figure 1.3: *Derivation of the approximation to hyperbolic discounting. Panel (a) depicts the probability density function for a possible distribution for t. Panel (b) is the optimal decoder for the distribution in panel (a). Panel (c) depicts the effect of exponential discounting decoding applied to the labels of the categories in panel (b). The dashed line is a linear interpolation of the discrete categories. Panel (d) compares the linear interpolation with the exponential discounting function. The interpolation over-discounts near events and under-discounts far-away events, approximating the behavior of a hyperbolic discount function.*

1.4.2 Distortion of probabilities: encoding effects and the probability weighting function

The probability weighting function is the second element of Prospect Theory (Kahneman and Tversky, 1979). The value function by itself implies risk seeking over all losses and risk aversion over all gains, both regardless of the probabilities involved. This cannot explain why people buy insurance (being risk averse over losses when the probability of a loss is very small) and why thy play Powerball (being risk seeking over gains when the probability of a gain is very small). The probability weighting function $\pi(p) : [0,1] \mapsto [0,1]$ solves that problem: decision-makers choose based on $v(x) \cdot \pi(p)$, not on $v(x) \cdot p$. The four types of behavior (risk seeking or avoidance over losses or gains) come from the interaction of $v(\cdot)$ and $\pi(\cdot)$. The probability weighting function is s-shaped (concave first, then convex), regressive (starts being above the diagonal and ends from below). An axiomatization and thorough analysis of $\pi(\cdot)$ can be found in Prelec (1998); Wu and Gonzalez (1996, 1999) measure the values of $\pi(\cdot)$.

The process used for the other two numerical dimensions of decision-making can be used to explain distortions of probability. First note that the decoder is trivial: since probabilities are used by themselves (payoffs are mapped into a value function, time is processed by discounting), the decoder is a linear function[11].

Standardization occurs here too: consider the difference between "rarely occurring headaches" (maybe once a month) versus "rarely occurring blindness" (maybe in one out of a million people) in a paper with this title (Fischer and Jungermann, 1996). So the process of interpretation of probabilities is sensitive to the norm on likelihood.

The remaining element is the encoder. To create the encoder, it is necessary to determine the distribution of probability values, $F_P(p)$. Note that in this case the support is bounded ($p \in [0,1]$) as opposed to both previous cases. Finding the distribution will require balancing the idea of maximum entropy with both the empirical and incentive-compatible distributions.

Probabilities have bounded support, and maximum entropy over a bounded support is attained with a uniform distribution. However, the distribution of probabilities faced by a decision-maker is anything but uniform: most facts in life are either sure or nearly so, few are completely uncertain. Risk-aversion and some control over the choice of probabilities by the

[11] González-Vallejo et al. (1994) show that presenting probabilities as numbers or as words leads to changes in behavior. This could be explained by the verbal presentation bypassing the encoder.

decision maker can help explain why and derive a distribution for probability values. Since people are averse to uncertainty, they will prefer to choose problems that have probabilities near the extremes (either sure or impossible). This leads to a distribution as depicted in figure 1.4a.

Encoding a distribution $f_P(p)$ as depicted in figure 1.4A using a equiprobable partition (derived from the additively separable errors) creates categories with sizes decreasing away from 1/2 (as shown in figure 1.4b). When a linear metric is applied to these categories, we obtain the shape of the probability weighting function (illustrated in figure 1.4c).

Observation 6. *If the distribution of probability values is increasing from the point 1/2 to the extremes, then optimal encoding generates the shape of the probability weighting function.*

1.5 Summary

In this paper I have shown that some well documented deviations from normative models involving numeric quantities can be explained by a model of discrete reasoning where there is information loss between encoding and decoding.

Future work is necessary on the areas of reference point formation, inference from data, and empirical validation.

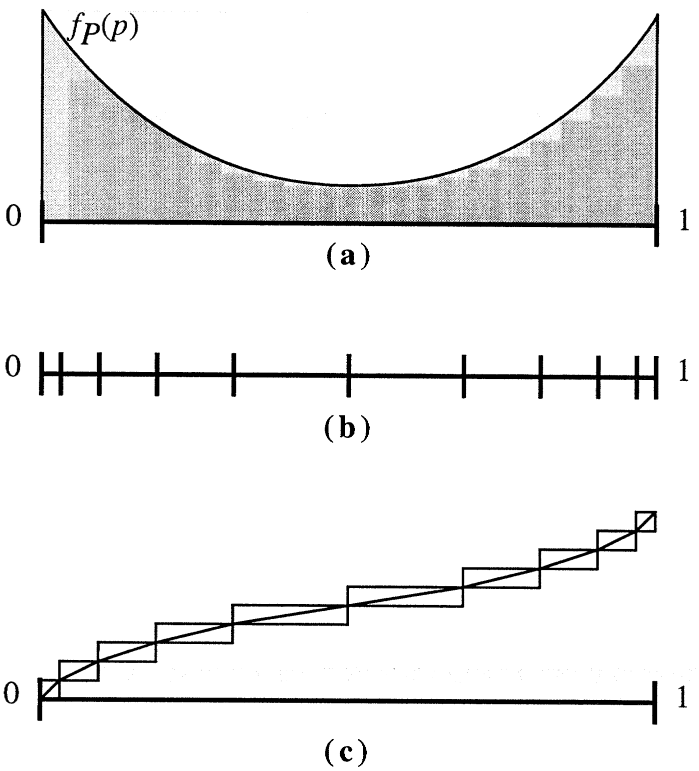

Figure 1.4: *Derivation of the probability weighting function. Panel (a) is the distribution of probability values when the decision-maker avoids risk. Panel (b) is optimal encoding designed for the distribution in panel (a). Panel (c) shows that using a linear decoding function on the labels generated by encoding in panel (b) leads to the probability weighting function.*

Chapter 2

Effects of Payoff Range on Risk-Taking

2.1 Introduction: What problem do we look into?

In this paper we extend the knowledge of decision-maker attitudes towards risk in decision-making, by looking into possible theories for the prospect theory value function. In particular, how we can use attitudes towards risk in the gains domain to predict attitudes towards risk in the losses domain, and vice-versa. To do that, we look into the nature of decision-making under risk, proposing two simple theories for a descriptive model (the prospect theory value function) and testing these with an experiment. Therefore, the paper makes both an empirical contribution (the predictive connection between attitudes towards risk over losses and over gains) and a theoretical contribution (separating two theories for the prospect theory value function).

The normative theory of decision-making under risk is Von Neumann and Morgenstern (1944) expected utility theory. Expected utility theory states that a normative agent, given several choice alternatives, $C_1, C_2, \ldots$, each of which creates a distribution, $f_1(w), f_2(w), \ldots$, over final states of wealth, should choose the alternative that maximizes *expected utility* defined as $\int u(w) f_i(w) dw$, where u is a cardinal representation of preferences over wealth states, called the *utility function.* It is customary to assume that u is a concave function, leading to a behavior described as risk-aversion: when faced with the choice between a sure payoff s and a lottery L that has expected value s, the decision-maker should choose s over L, denoted $s \succ L$. This concavity can be derived under some assumptions from maximization of utility over consumption bundles, where w enters only through the income constraint[1].

Empirical evidence does not support expected utility theory. For instance, expected utility theory predicts that the way the problem is posed has no effect on the final choice. In particular, neither the complexity of the alternatives, nor the way they are framed, should make a difference, since the decision relies only on the final distribution over wealth states $f_i(w)$ and this does not change with complexity or framing. So, making a choice between (1 w.p. 1/2; (1 w.p. 1/2; 0 w.p. 1/2) w.p. 1/2)[2] and 1 should yield the same result as choosing between (1 w.p. 3/4; 0 w.p. 1/4) and 1, since the final distributions over wealth are identical — complexity shouldn't matter. Also, making it a choice between $1 + 1$ and $1 +$ (2 w.p. 1/2; 0 w.p. 1/2) or between 2 and (3 w.p. 1/2; 1 w.p. 1/2) — that is framing the choice using the first 1 as a reference point, versus using implicitly zero as a reference point (in the second choice) — should not make any difference in the final result. It has been shown thoroughly (see for instance Allais (1953), Machina (1987), Thaler (1980), Tversky and Kahneman (1974); reviews in Camerer (1995) and Rabin (1998) give a more complete view of the field) that both complexity and framing lead to preference reversals. These problems in decision making are more than curiosities: decision-making under uncertainty is pervasive in management, medicine, politics, to name a few; see also a comment in Kahneman and Tversky (1996). A particularly enlightening example is a choice between two alternative medical treatments in Tversky and Kahneman (1981). When the same choice is framed in the gains domain (600 people will die if untreated, choose between either saving 200 or taking a 1/3 chance of saving 600), the conservative (risk-averse) option is chosen; when the choice is framed in the losses domain (600 people will die if untreated, choose between a method that lets 400 die and a method that has a 1/3 chance nobody will die), decision-makers become gamblers and choose to take the risk. Given that the decisions are equivalent in terms of final distribution over states of the world, the decision-makers' choices are inconsistent.

Since expected utility theory fails to describe the observed phenomena, Kahneman and Tversky (1979) propose an alternative theory, prospect theory. Prospect theory describes

[1] A rational agent will use the first dollar to buy the best possible consumption alternative, the second dollar to buy the second best alternative, and so on, leading to decreasing marginal utility for wealth; these preferences for bundles coupled with a income constraint can be described by a concave utility function over wealth.

[2] The notation "w.p." means "with probability".

agent choices over risky options (which these authors call prospects) through two functions: an s-shaped value function over payoffs, $v(x)$, and a probability weighting function, $\pi(p)$. The value function defines a measure of preference over changes relative to a reference point, not over final states of wealth. This makes the choices dependent on the formulation of the alternatives. If a reference point is established at "current wealth minus 600", and the alternatives are then framed as gains of 200 and 400, the result is potentially different from one where the reference point is established as "current wealth" and the alternatives are framed as losses of 400 and 200, although the final wealth states are equivalent. Also, given the s-shape of $v(\cdot)$, agents are risk-averse over gains (they prefer a sure gain $s_i > 0$ over a lottery yielding the same expected value, say $(s_i/p_i$ w.p. $p_i; 0$ w.p. $1-p_i))$ and risk-seeking over losses (they prefer a lottery over losses $(-s_i/p_i$ w.p. $p_i; 0$ w.p. $1-p_i)$ over its equivalent sure loss $-s_i < 0)$. There is also loss aversion in $v(\cdot)$, in that $|v(-x)| > v(x), x > 0$; this characteristic is not important for the topic of this paper, but is very important in reality and serves as a first face validity test for both the theoretical development and the empirical results.[3]

In this paper, we extend our knowledge the prospect theory value function by looking into how the form of the concavity and convexity are related. The main objective of the paper is to investigate if the behavior on each side is symmetric or asymmetric with respect to range of payoffs (we are interested in knowing how the propensity for risk taking varies with the size of payoffs). We will be looking at simple choices of the form

$$\text{choose between } s \text{ and } L_{s,p} \doteq (s/p \text{ w.p. } p; 0 \text{ w.p. } (1-p)). \tag{2.1}$$

A strict interpretation of prospect theory would state that for gains, $s > 0$, the sure outcome would always be preferred, $s \succ L_{s,p}$, since the concavity of v makes the arc, $v(s)$ be above the chord, $pv(s/p) + (1-p)v(0)$[4]; for losses the lottery would always be preferred, since in a convex function any chord is above the arc to which it corresponds. However, this is too strict an interpretation, and does not match empirical data: for example, in the supracited medical treatment decision, 72% of respondents chose the sure option in the gains framing and 78% chose the lottery in the losses framing; these numbers are significantly different

[3] The second element of prospect theory, the probability weighting function is necessary to model reversal of attitudes towards risk at extreme probability values; it will not be considered in this paper, since we will be mostly interested in median probability values (away from extremes). For more on the probability weighting function see Prelec (1998).

[4] We include $v(0)$ in the equation just for clarity, since as a notation simplification we assume $v(0) = 0$.

from chance, but they are also significantly different from 100%. Other factors, which are not accounted for in the measurement (and prospect theory) make the decision more stochastic. Therefore, we will talk about a propensity for risk-taking[5].

The propensity for risk taking, denoted $r(s,p)$ is a measure of preference for $L_{s,p}$ over s: higher $r(s,p)$ means a higher preference for $L_{s,p}$. We can measure it either as a percentage of subjects choosing the lottery, when only binary choice data is available, or the difference between subjects' preference rating for the lottery and for the safe payoff. In the medical treatment example, the results for the gains frame could be summarized in our notation by $r(200, 1/3) = .72$. The s-shape of v implies $r(-s,p) > r(s,p)$ for $s > 0$ (and this is only an accurate description of actual behavior for a non-extreme p; extreme values of p will bring into play the distortion of probability values created by the probability weighting function). But there is no implication from $v(\cdot)$ that given $s_1 > s_2 > 0$ either $r(s_1,p) > r(s_2,p)$ or $r(s_1,p) < r(s_2,p)$, nor any information in the loss domain. Also, there is no link between the behavior with respect to losses and the behavior with respect to gains[6]. Finding a theory that explains the shape of v and that can make predictions about such relationship is the objective of this paper.

In simple terms, what this paper sets out to do is to use a theory-building approach to find out whether our knowledge of, say, $r(50, 1/2) > r(500, 1/2)$ can be used to predict that $r(-50, 1/2) < r(-500, 1/2)$ or that $r(-50, 1/2) > r(-500, 1/2)$. This will be done in two steps: first, the development of theories for where the shape of $v(\cdot)$ comes from (in the next two sections); and second, an empirical test of the theories (in section four). Section five summarizes the paper and establishes connections to the literatures on categorization and processing of numerical information. The appendix goes over the technical details of the second theory, information compression.

[5] For instance, if we consider an additive effect of the unmeasured effects, ϵ, the decision then is to choose the sure payoff, s, if $v(s) + \epsilon > pv(s/p)$, and the lottery otherwise. Assuming that $\epsilon \sim N(0, \sigma^2)$ for illustration, the propensity for risk-taking will be increasing in the convexity of $v(s)$, since $v(s) - pv(s/p)$ is increasing in the concavity; the choice of L will be decreasing in this difference and increasing in the variance of ϵ, σ^2.

[6] To be precise, the s-shape only implies the relationship between $r(\cdot)$ for losses and $r(\cdot)$ for gains presented; all other relationships require further assumptions on curvature.

2.2 An elegant solution: decreasing sensitivity and a factor for losses

One parsimonious way to model $v(\cdot)$ is to choose a continuous, increasing, and concave function $g(x) : \mathbb{R}_0^+ \mapsto \mathbb{R}_0^+$, with $g(0) = 0$, to model decreasing sensitivity; then supplement that function with a penalty factor for losses $\lambda > 1$, creating the following model for $v(\cdot)$

$$\hat{v}(x) = \begin{cases} g(x) & \text{If } x \geq 0 \\ -\lambda\, g(-x) & \text{If } x < 0 \end{cases} . \tag{2.2}$$

An interpretation of this approach is that agents have decreasing sensitivity in perception over absolute magnitudes, and then they penalize losses separately. So, the difference between 5 and 6 is perceived as being in some sense larger than the difference between 500 and 501, and that is independent of sign: the difference between -5 and -6 is also larger than that between -500 and -501; in fact, with the model in equation 2.2, the differences between -5 and -6 and 5 and 6 are equal in magnitude, as are those between -500 and -501 and between 500 and 501. The idea of decreasing sensitivity is not new to perception, being an integral part of several theories of perception (Weber's and Fechner's among others — see Herrnstein and Boring (1965) pages 64-75 for instance). Kahneman and Tversky (1979) also propose some form of decreasing sensitivity over the absolute value of magnitude:

> "Many sensory and perceptual dimensions share the property that the psychological response is a concave function of the magnitude of physical change. [...] We propose that this principle applies in particular to the evaluation of monetary changes." (Page 278.)

In this formulation the penalty for losses is a primitive of the theory. Agents are assumed to dislike losses more than they like gains of equivalent magnitude, and do so at the "loss / gain" degree, as if they separated out the sign information component from the absolute magnitude information component. Kahneman and Tversky refer to this behavior (*nota bene*: in general, not to the constant penalty factor $-\lambda$ formulation in equation (2.2)) as a "reflection effect" (*op. cit.* page 268).

The elegance of the decreasing sensitivity model can be thought of as the result a cost-of-thinking reduction argument, with specialization along the (sign, magnitude) dimensions: an interpretation of this model is that there is separate coding for gains versus losses (one piece of information) and for magnitude. In a high cost-of-thinking situation (say with

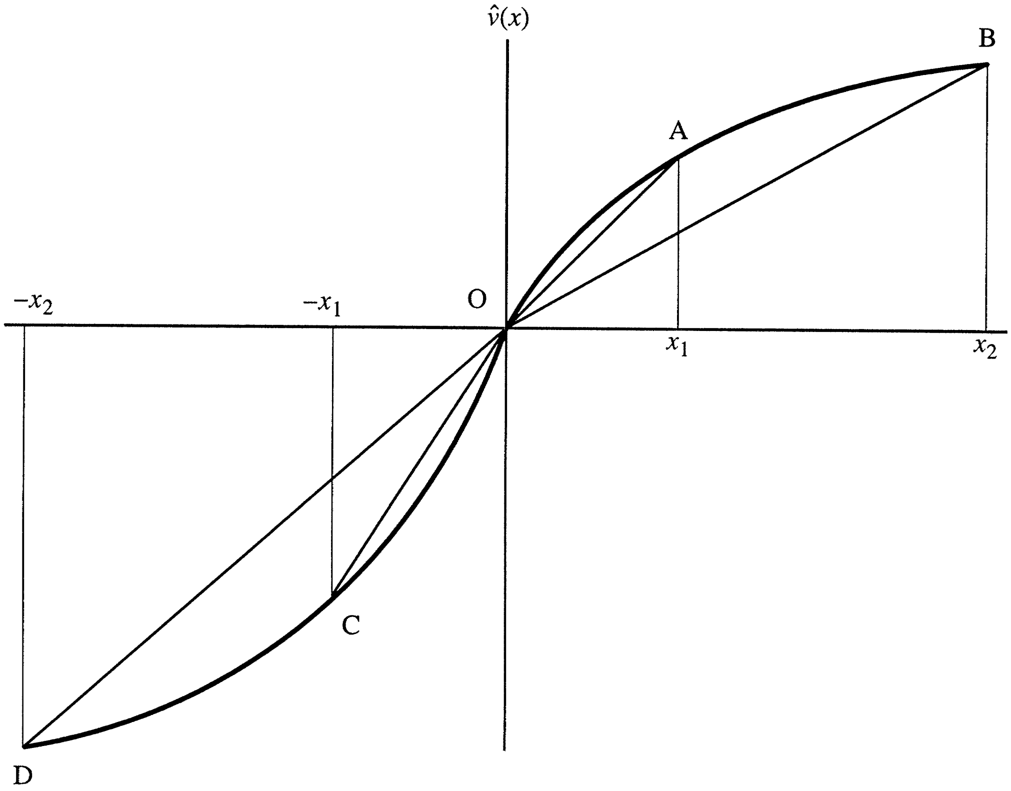

Figure 2.1: QUASI-SYMMETRIC $\hat{v}(\cdot)$ IMPLIES SYMMETRIC ATTITUDES TOWARDS RISK. In this illustration, the increase in range over gains leads to increasing risk-aversion: the distance between the arc OB and the line OB is higher than that between the arc OA and the line OA. The symmetric nature of $\hat{v}$ leads to the mirror image in the negative domain with OC and OD, making increasing range lead to increasing risk-seeking over losses.

constrained time) an agent will use the most important bit of information, the sign, to decide; as cost decreases, the agent will use magnitude information.

The model in equation (2.2) implies a link between the behavior over losses and the behavior over gains: it implies they are symmetric, in the sense that under this functional form, for any payoffs of the same sign s_1, s_2, and a non-extreme probability p[7],

$$\operatorname{sign}\big(r(s_1,p) - r(s_2,p)\big) = -\operatorname{sign}\big(r(-s_1,p) - r(-s_2,p)\big); \tag{2.3}$$

[7] In this paper, we will assume that probabilities are not extreme, meaning that the effect of the value function is not overwhelmed by the effect of the probability weighting function. This is coherent with our experiment, and with the decisions that we want to analyze. Plus, the result should hold (albeit with reverse main effects) for extreme probability values, since prospect theory assumes independence of $v(\cdot)$ and $\pi(\cdot)$.

in other words, if $r(50,p) > r(500,p)$, using $\hat{v}$ as a model for v implies $r(-50,p) < r(-500,p)$. (This is illustrated in figure 2.1.) It does not imply either separately, just the symmetry, so it could be that $r(50,p) < r(500,p)$ (which would imply $r(-50,p) > r(-500,p)$). In choice terms, this means that if more people prefer a gain of 500 over $L_{500,1/2} = (1000$ w.p. $1/2; 0$ w.p. $1/2)$ than a gain of 5 over $L_{5,1/2} = (10$ w.p. $1/2; 0$ w.p. $1/2)$, the model predicts that more people will prefer a loss of -500 over $L_{-500,1/2} = (-1000$ w.p. $1/2; 0$ w.p. $1/2)$ than a loss of -5 over $L_{-5,1/2} = (-10$ w.p. $1/2; 0$ w.p. $1/2)$. This symmetric behavior can be shown more formally by deriving the Arrow-Pratt coefficient of absolute risk aversion for $\hat{v}$:

$$-\frac{\hat{v}''}{\hat{v}'} = \begin{cases} -g''/g' & \text{for gains} \\ g''/g' & \text{for losses} \end{cases} . \tag{2.4}$$

Note that the symmetry is independent of the curvature of $g(\cdot)$, so even if that changed with range, the symmetry of the result would hold.

At this point we already have one hypothesis that extends $v(\cdot)$. Using the generic form of $v(\cdot)$, one cannot use the fact of, say, $r(5,p) > r(500,p)$ to predict whether $r(-5,p) > r(-500,p)$ or $r(-5,p) < r(-500,p)$. Using the more structured $\hat{v}(\cdot)$, we can use the symmetry to relate behavior over losses to behavior over gains, and would predict $r(-5,p) < r(-500,p)$ in the example. However, this is just one way to think about the origin of $v(\cdot)$. Another possibility follows.

2.3 An alternative model: information compression plus concave utility over internal representation

To build an alternative hypothesis, we start with a different model of what goes on in the decision-making process. The idea here is that the agent acts as if she had a concave utility function, $\tilde{u}$; this function models preferences defined over internal representations of the payoffs. The internal representations are categorical. There is information loss in that the preference structure does not take into account the way the internal representations were created. In particular, the preferences are defined over category labels, without regard for the categories underlying the labels other than through their order. The internal representation (the categorization) is created by an independent process, information compression: the input payoffs x are mapped into these representations by a function $e(x)$. Interesting perspectives on the importance of symbolic reasoning and creation and manipulation of internal representations can be found in (Lakoff, 1987; Deacon, 1997; Margolis and Laurence, 1999).

To build a decision-making model with information compression implies some assumptions about the allocation of effort. We will assume several things, essentially (and in sequence, since the former imply the latter):

- The information-compression process is costly to set-up, and should be shared across different decision instances. This is essentially a cost-of-thinking argument. Solving the problem of setting up an information compression process is a complex task, therefore, if said information compression process, $e(x)$, can be shared across applications (decision instances in our case) there are savings in setup effort; there are also losses from not designing the ideal information-compression scheme for each decision, but this cost is assumed to be less than that of setting up a specific system. This is a main assumption of the model, that the system is shared across decision instances due to cost of setup being higher than expected cost of using a generic information compression process instead of a instance-specific one. This assumption implies the next several ones.

- Since the process is shared across different decisions, the payoffs are standardized prior to compression. The reason for that is as follows: given that information compression will generate a fixed partition of the payoff space, varying the scale of the payoffs would lead to large variation in the discrimination ability of the partition; by standardizing the inputs, the partition will always have the same discriminant ability. For instance consider a partition of the dimension *height* into five categories, say ranging from "very short" to "very tall". Suppose that one categorization, denoted $\mathcal{C}$, is optimal for people. If $\mathcal{C}$ is used to categorize, or encode, the height of buildings, it will be too fine in the lower end, and too coarse in the high end; all buildings will be "very tall", probably. On the other hand, if $\mathcal{C}$ is used to categorize the height of champagne flutes, then it will be too coarse in the lower end and too fine in the high end (most champagne flutes will be encoded as "very short"). Standardizing the input essentially creates partitions that are measured in units of standard deviation and solves this problem. Norm theory (Kahneman and Miller, 1986) is one way to think of standardization in reality, because it implies that objects bring with themselves notions of dimensionality/scale. For the rest of this section standardization will be ignored, since that simplifies notation and does not change anything substantive.

- Information compression should minimize the loss of information between inputs and outputs. (The detailed form of this criterion is in the appendix; the distance measure is the Kullback-Leibler information loss distance — it measures the loss of information across $e(x)$ from the viewpoint (distribution) of the inputs — and will lead to partitioning where the categories are equally likely.) This criterion, at least in its abstract form, is intuitively obvious: given that information loss is bad, in the sense that it cannot lead to better decisions and usually leads to worse decisions, we want to minimize it. The specific form of the criterion requires more technical details, hence is left to the appendix.

- The information compression process should assume distributions that have minimal information content, in order to maximize the generic nature of the process. In other words this criterion means that the design should make the fewest possible assumptions about the distribution. Minimal information content means that each realized value should be maximally surprising on average. For instance, if the distribution were degenerate, say $X = x_0$, constant, each realization would be completely unsurprising and the information content of the realization would be minimal - whereas the information content of the functional form of the distribution would be maximal - the opposite of what a generic system should have. (The detailed form is also in the appendix; solving this under some required constraints leads to a Normal distribution.)

As is shown in the appendix, under these assumptions about the generic nature of information compression and the most informative inputs, $e(x)$ creates a symmetric s-shaped map from $\mathbb{R}$ to $[0, 1]$. (See figure 2.2 for illustration.) These assumptions also require problem-specific data to enter the utility function, since there is loss of information about the concavity of $u(0)$ due to the standardization ($e(0)$ varies withe the standardization; plus, since there is a scaling of the payoffs, the curvature of $\tilde{u}(\cdot)$ would only be preserved if it had the CARA form, $\tilde{u}(e) = -\exp(-re)$). Therefore, given the internal representations $e = e(x)$, the final result is $\tilde{v}(x; \eta) \doteq \tilde{u}(e(x); \eta)$, where η is a problem-specific parameter determining the degree of risk-aversion adequate for $\tilde{u}(\cdot)$ in the problem at hand.

In order to make predictions regarding the effect of payoff range on risk-taking behavior in this model, we need to assume a relationship between range and η. A strong assumption

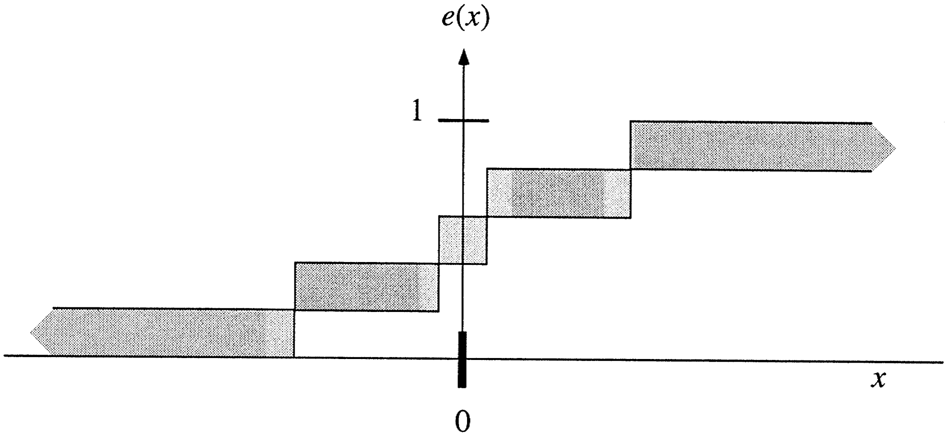

Figure 2.2: ENCODING CREATES AN S-SHAPE. In this illustration, a five category encoder maps the real line $x \in \mathbb{R}$ into five labels. Using the convention in the appendix, the labels are the c.d.f. values for elements of the category. The result is decreasing sensitivity, which, given the choice of labels, is visualized by a discrete s-shape.

would be that η increases with increasing payoff range[8]. A weaker assumption, and one that is sufficient for the purpose of developing a testable hypothesis, is that η is monotonic in range. (This assumption can be tested independently from the main hypothesis.)

Under the monotonicity assumption, the effects of range on risk attitudes should be in the same direction over losses and gains, meaning that if an increase in payoff range leads to increasing risk-taking over losses it should also lead to increasing risk-taking over gains, and mutatis mutandis for decreasing risk-taking. (Note how this is the opposite of the result in equation (2.3).) This follows from the Arrow-Pratt coefficient of risk-aversion for $\tilde{v}(x;\eta)$, which is

$$-\left(\frac{\partial^2 \tilde{v}}{\partial x^2}\right) / \left(\frac{\partial \tilde{v}}{\partial x}\right) = -\left(\frac{d^2 e}{d x^2}\right) / \left(\frac{d e}{d x}\right) - \frac{d e}{d x}\left(\frac{\partial^2 \tilde{u}}{\partial e^2}\right) / \left(\frac{\partial \tilde{u}}{\partial e}\right). \tag{2.5}$$

Note that if the increase in range leads to some change in $\tilde{u}''/\tilde{u}'$ (meaning the partial derivatives above), then the effect on $\tilde{v}''/\tilde{v}'$ (same proviso) is independent of the decision being over losses or over gains, since $de/dx > 0, \forall x$; the main effect of losses versus gains is in the first term on the right hand side of equation (2.5), where the second derivative of e changes sign between losses and gains. We haven't assumed any directional effect of range

[8] As we shall see, this assumption would be supported by data.

on $\tilde{u}''/\tilde{u}'$, only that it is monotonic, but the data will show that the reasonable assumption (increasing range increases risk-aversion) holds.

With this monotonicity assumption on $\tilde{u}$, it follows from equation (2.5) that under this model, $\tilde{v}$, the following condition holds

$$\operatorname{sign}(r(C_i) - r(C_j)) = \operatorname{sign}(r(-C_i) - r(-C_j)). \tag{2.6}$$

In other words, under the information-compression model, if people are more likely to choose $L_{5,p}$ over \$5 than $L_{500,p}$ over \$500, then they are more likely to choose $L_{-5,p}$ over -\$5 than $L_{-500,p}$ over -\$500. These are opposite predictions to those of equation (2.3).

In summary, we now have two different simple theories (decreasing sensitivity with a penalty factor for losses and information compression with a concave utility over the internal representation) of the prospect theory value function; these theories make opposite testable predictions. Under $\hat{v}$, the quasi-symmetric model of decreasing sensitivity,

$$\begin{aligned} r(5,p) > r(500,p) \quad &\text{implies} \quad r(-5,p) < r(-500,p) \quad \text{and} \\ r(5,p) < r(500,p) \quad &\text{implies} \quad r(-5,p) > r(-500,p); \end{aligned}$$

under $\tilde{v}$, the information compression model,

$$\begin{aligned} r(5,p) > r(500,p) \quad &\text{implies} \quad r(-5,p) > r(-500,p) \quad \text{and} \\ r(5,p) < r(500,p) \quad &\text{implies} \quad r(-5,p) < r(-500,p). \end{aligned}$$

We will now test these predictions.

2.4 Empirical test

The contrasting results in equations (2.3) and (2.6) were tested with a survey containing choices between sure payoffs and lotteries. Each question in the survey is a choice between a sure payoff s_i and a lottery of the form used so far, L_{s_i,p_j}, as in equation (2.1). For this experiment, there were four payoff conditions $s_i \in \{5, 70, 900, 4000\}$, three probability conditions $p_j \in \{1/2, 1/3, 1/5\}$ and a losses versus gains condition. The probability values are high enough that there are no significant effects of the probability weighting function. Also, even if there were effects of the probability weighting function, we could tease these out by testing within probability condition — where the effects of the probability weighting function are held constant. Besides, there is no reason to assume any interaction between the concavity of $g(\cdot)$, for $\hat{v}(\cdot)$, or of $\tilde{u}(\cdot)$, for $\tilde{v}(\cdot)$, and the probability values, therefore a within probability condition test would be enough to separate the two theories. A pre-test has shown the range \$5 – \$4000 sufficient to have an effect on risk attitudes.

Payoff and probability manipulations are run within subjects; losses versus gains is run between subjects. The reason for running losses/gains between subjects is that running it within subjects might bias the experiment in favor of the $\tilde{v}(\cdot)$ model: if a subject had a choice between -\$5 or -\$10 with probability 1/2 in the loss condition and chose response $r(-5, 1/2) = 4$, then when answering the equivalent condition over gains (choosing between \$5 and \$10 w.p. 1/2) she could use the previous answer as an heuristic and, purely by inference, answer $r(5, 1/2) = 4$; this would bias the result in favor of $\tilde{v}(\cdot)$, since effects over losses and over gains would be in the same direction through inference alone: if inference leads to $r(5, 1/2) = r(-5, 1/2)$ and $r(4000, 1/2) = r(-4000, 1/2)$, then the directional effects are immediately equal, $\text{sign}(r(4000, 1/2) - r(5, 1/2)) = \text{sign}(r(-4000, 1/2) - r(-5, 1/2))$, without any effect of $\tilde{v}(\cdot)$. In other words, even if $\hat{v}(\cdot)$ were the true theory, inference alone would make the results support $\tilde{v}(\cdot)$. Running the condition between subjects gets rid of this problem. However, since there are unmeasured individual effects, comparing conditions between subjects lowers the significance measure for a fixed number of subjects. This lowered significance acts against both models, $\hat{v}(\cdot)$ and $\tilde{v}(\cdot)$: as we are evaluating a contrast the decrease in significance does not create a bias. Since the outcome of the survey was still significant, this was not a problem in our experiment. We also make repeated measures of the same underlying models — considering each probability value a repeated measure —, which help establish the validity of the theory.

Each questionnaire contains twelve questions (six choices over losses and six over gains). Table 2.1 shows the distribution of conditions per questionnaire. The order of options, left and right side, and up and down positioning on the page were counterbalanced across questionnaires. Since we have a small number of subjects, direct tests for the effect of order are not feasible (too few subjects per cell if divided by question order), but any effects of order — given the counterbalance — would present themselves as a stochastic disturbance and lower the significance of the measured effects; in no way would counterbalanced questions lead to order biasing the result towards one model or the other. Thirty-two MBA students responded to the questionnaires at the end of a class session.

The response scale was a six-point preference scale anchored on “strongly prefer A” and “strongly prefer B”, where A and B were the references for the options. We coded strong

Table 2.1: DISTRIBUTION OF CONDITIONS PER QUESTIONNAIRE. The notation G:1, L:2 means that the question is asked as a choice over gains on questionnaire 1 and as a choice over losses on questionnaire 2.

	s_i			
p_j	5	70	900	4000
1/2	G:1, L:2	G:2, L:1	G:1, L:2	G:2, L:1
1/3	G:2, L:1	G:1, L:2	G:2, L:1	G:1, L:2
1/5	G:1, L:2	G:2, L:1	G:1, L:2	G:2, L:1

preference for the sure option as a 1 and strong preference for the lottery as a 6, so higher values reflect higher $r(\cdot)$. Given the absence of a middle point in the scale, a test for use of the two adjacent choices (3 and 4) was done. The responses were strongly bimodal for each of the conditions and pooled data, therefore no effect of middle category absence was found. For further reassurance, the effect of middle category absence — if existent — would be in the form or randomization between 3 and 4, thus acting against the significance of both models ($\hat{v}(\cdot)$ and $\tilde{v}(\cdot)$); the result would be non-significant effects of range on preferences *tout court.*

Table 2.2 contains the means and standard errors thereof for the twenty-four conditions. The trends are apparent: increasing range decreases risk-taking behavior both for losses and for gains. Analysis of variance on pooled data shows that there is a significant effect of losses versus gains ($t_{383} = 8.45; p = 0.0039$) as one would expect from the s-shape of $v(\cdot)$, which is preserved in both models. This observation does not support or defeat either of the models in this paper, but is a confirmation that the experiment meets past empirical evidence. The analysis of pooled data also shows that both losses ($F_{3,189} = 5.3; p = 0.0016$) and gains ($F_{3,189} = 14.14; p < 0.0001$) have significant main effects and these are both decreasing with payoff range. (Individual cell differences are in general too small for significance with this few subjects.) These results show behavior that is consistent with $\tilde{v}$ (the results conform with the predictions from equation 2.6) but not with $\hat{v}$ (the results are the opposite of those predicted by equation 2.4).

Within each probability value the results replicated the main finding, although the number of subjects was too small for all effects to be significant. For instance, the four cells for $p = 1/5$, loss condition, show a significant decreasing trend ($F_{3,61} = 2.67, p = 0.06$), as

Table 2.2: Results of Experiment. The numbers are means of preference ratings for the risky option, a measure of $r(\cdot)$; higher numbers reflect higher preference for risk. Standard error in parenthesis.

		Gains		
p_j		s_i		
	$5	$70	$900	$4000
1/2	4.75 (0.44)	3.88 (0.44)	2.56 (0.38)	2.31 (0.39)
1/3	4.37 (0.48)	3.81 (0.51)	2.31 (0.28)	2.32 (0.36)
1/5	4.31 (0.48)	2.81 (0.48)	3.00 (0.40)	2.81 (0.40)
Pooled	4.48 (0.27)	3.50 (0.28)	2.63 (0.21)	2.48 (0.22)

		Losses		
p_j		s_i		
	$5	$70	$900	$4000
1/2	4.69 (0.37)	4.25 (0.37)	4.06 (0.33)	3.88 (0.34)
1/3	4.31 (0.36)	3.13 (0.34)	3.50 (0.34)	3.19 (0.36)
1/5	4.25 (0.40)	3.44 (0.38)	3.31 (0.41)	2.88 (0.30)
Pooled	4.41 (0.22)	3.60 (0.22)	3.63 (0.21)	3.31 (0.20)

do those for $p = 1/5$, gains condition, albeit only marginally ($F_{3,61} = 2.36, p = 0.08$).

2.5 Conclusions

In this paper we have compared two simple models of the prospect theory value function. Empirical data support an explanation based on information compression and preferences defined over the internal representation with information loss. The paper's contributions are thus:

- THEORETICAL. In this paper we propose that the prospect theory value function may be the result of information compression of payoffs into internal representations which are then subjected to a preference structure that can be characterized as by a concave function over cardinalized categories; this theory is more consistent with empirical results than an alternative theory that is based on decreasing sensitivity an a factor penalization for losses as primitives.

- EMPIRICAL. We have explored the effect of payoff range on the propensity for risk-taking. This extends the basic prospect theory value function by making more specialized claims regarding choices with significantly different payoff ranges.

Following this work there are some avenues for exploration. Theoretical issues include the influence of stocasticity and fuzziness on the categorization model, the relationship between categorization and cost of thinking and bounded rationality, and information processing implications of arithmetical manipulation. In this sense, the paper is part of the literatures on categorization (Macmillan et al., 1977; Lakoff, 1987; Rubinstein, 1988; Deacon, 1997; Margolis and Laurence, 1999), on bounded rationality (Gingerenzer and Goldstein, 1996; Rubinstein, 1998; Kaufman, 1999), and on numerical information processing (see, for instance, Parducci (1965) and Holyoak and Alker (1976)).

Empirical issues to explore further include finding moderators for the effects of range on the Arrow-Pratt coefficient of risk-aversion for $\tilde{u}(\cdot)$, on the implications of categorization themselves, and on the mechanisms of standardization. This links this paper to the empirical side of the behavioral decision theory literature and also to the estimation of categorization models.

Chapter 3

Categorical Rational Agents: The Reason For Behavioral Decision Phenomena?

3.1 Sypnosis

The decision-making and psychology literatures document several well-established phenomena that deviate from the normative model of decision-making. In this paper we focus on phenomena described by prospect theory and hyperbolic discounting. These phenomena include reversal of preferences with changes in the way a problem is framed, inconsistent attitudes towards risk with median versus extremely low probabilities, and inconsistency in choices in inter-temporal decision making. We propose a theory that explains how an agent, using normative decision-making rules and categorical reasoning, but without making systematic errors in setup of a reasoning scheme, will act in accordance with the behavioral descriptions. This paper proposes that an agent using categories compatible with concave utility functions and exponential discounting will behave as described by prospect theory and hyperbolic discounting. (The categorization used for payoffs is derived from minimization of the expected error in utility; the one for time is derived from minimization of the error in discounting.) The deviations are caused by the loss of information in the categorization process. If on average these errors in utility and discounting are smaller than the cost of setting up a specific categorization for a problem, an agent using a generic categorization will do better than one that creates specific categorizations for each decision problem and incurs a set-up cost. A generic categorization requires a distribution which maximizes the information revealed by each realization of the payoffs; using entropy as the metric for information, we use a normal as the distribution for payoffs and an exponential for time. With a concave utility function and a Normal distribution, the optimal categorization will yield a partition of the payoff space that has narrow categories near zero and wide categories away from zero; these categories will also be asymmetric, to account for the concavity of the utility function. With these categories, there will be decreasing sensitivity with distance to zero and given the asymmetry, the categorization also creates loss aversion. Regarding time, the exponential distribution coupled with exponential discounting lead to decreasing sensitivity over increasing time; this creates effects similar to those of hyperbolic discounting.

3.2 Introduction

In this paper we offer an explanation for behavior that deviates from the normative model of decision-making, based on categorical reasoning. We analyze behavioral phenomena of decision-making under risk, and inter-temporal decision making. This section gives an overview of the phenomena, and summarizes the logic of the paper.

Notation

We will be looking at payoffs, time and probabilities. We will use x, y, z, w to represent payoffs (with indices if necessary), t_i or τ_i (both with or without indices) for times, and p_i or π_i for probabilities. We use the notation $(x_1, p_1; x_2, p_2; \dots; x_M, p_M)$ to mean the lottery that pays x_1 with probability p_1, x_2 with probability p_2, and so on till x_M with probability p_M. If $x_M = 0$, we will abbreviate and leave that out; this makes simple lotteries like $(x, p; 0, 1-p)$ be represented by (x, p) alone. We use the notation $(x \otimes t)$ to mean a payoff of x at time t.

3.2.1 Decision-Making Under Uncertainty and Prospect Theory

The normative model for decision making under uncertainty is Von Neumann and Morgenstern (1944)'s expected utility theory. This theory is based on axioms describing how a decision-maker should choose between sets of uncertain alternatives in a way that avoids inconsistency. As a result of these axioms, a decision-maker's preferences can be described by a utility function over wealth, $u(w)$, which is unique up to an affine transformation. Suppose that the choice is between two alternatives i and j which create probability distributions over the space of wealth $F_i(w)$ and $F_j(w)$, respectively; according to expected utility theory, the agent should choose i over j if

$$\int u(w)\,dF_i(w) > \int u(w)\,dF_j(w). \tag{3.1}$$

The utility function is usually assumed concave to model *ceteris paribus* preference for less risk.

The axioms and, by implication, the expected utility form in equation (3.1), imply that a rational agent will make consistent choices independently of the way the problem is framed; e.g., a choice between

$$L_1 \doteq (-20, 1/2)$$

and $x_1 = -10$ should lead to the same choice as one between

$$L_2 \doteq -20 + (20, 1/2)$$

and $x_2 = -10$. So, if the agent chooses $x_1 \succ L_1$, she should choose $x_2 \succ L_2$. The reason for this consistency is that L_1 and L_2 create equal distributions over final states of wealth, and are therefore equivalent for decision-making purposes.

However, actual decision-making does not conform with the normative theory. The literatures of decision-making, economics, and psychology (Allais, 1953; Camerer, 1995; Harless and Camerer, 1994; Kahneman and Tversky, 1979, 1996; Machina, 1987; Rabin, 1998; Thaler, 1985; Tversky and Kahneman, 1991) document instances of preference reversal with different framing of the same problem. E.g.: Tversky and Kahneman (1981) present data on choices which violate expected utility. Decision-makers were told that a disease would kill 600 people if nothing was done, and were asked to choose between treatment methods for said disease. By manipulating the reference point, and framing choices as either gains (denoted here by x_1^G and x_2^G) or losses (denoted here by x_2^L and x_1^L) with respect to that reference point, the experimenters obtain results where $x_1^G \succ x_2^G$ and $x_2^L \succ x_1^L$. Namely, decision-makers preferred a method that would save 400 people for sure (x_1^G) — and, by implication, would leave 200 to die — over one that would save 600 with probability 2/3 (x_2^G), but preferred a method that could kill 600 with probability 1/3 (x_2^L) over one that would kill 200 for sure (x_1^L). Since the methods are pairwise equivalent (x_1^G is equivalent to

x_1^L and x_2^G is equivalent to x_2^L, because they create the same final outcome distributions), the result is inconsistent. In summary, people are risk averse over gains and risk-seeking over losses, even if the final distributions are equivalent across the gain/loss manipulation.

Kahneman and Tversky (1979) propose an alternative model of decision-making under risk, *prospect theory.* Prospect theory accommodates the inconsistency just shown using a value function $v(x)$ where the argument is a payoff, a *change* in the wealth level with respect to a reference point. The value function is concave over gains, convex over losses, and is steeper on the loss side than on the gains side, creating loss aversion,

$$x > 0 \quad \Rightarrow \quad v(x) - v(0) < v(0) - v(-x).$$

However, these attitudes towards risk only hold when the probabilities are not too extreme. In cases where there is a small probability of a large gain (as in a state lottery ticket), the same agents who were risk-averse over gains with more median probabilities become risk-seeking. Conversely, with a small probability of a large loss, the agents who were risk-seeking over loss alternatives with median probabilities become risk-averse; this is the case with most insurance, for instance. Prospect theory models this by having a probability weighting function $w(\pi)$ that overweights probabilities near zero. So, when $\pi \approx 0$, $w(\pi) >> \pi$ leading to $w(\pi)|v(x)| >> \pi|v(x)|$, and whereas concavity over gains (resp. convexity over losses) makes $\pi v(x/\pi) < v(x)$ (resp. $\pi|v(x/\pi)| > |v(x)|$), the effect of $w(\pi) >> \pi$ overwhelms these smaller differences.

Another point, already implicit in the formulation of $v(x)$, is the separation of payoffs from the aggregate wealth. This may lead to other inconsistencies depending on whether a set of payoffs is analyzed element by element or only as a summary statement. Since $v(\cdot)$ is non-linear, the analysis element by element tends to exacerbate the effects of payoffs. This is illustrated more thoroughly in Thaler (1985).

Since the value function is centered on a reference point, and the reference point is outside the scope of the theory, there are several effects that movement of said reference create. Framing, as shown above, leads to inconsistent decisions across different versions of the same problem. Dynamics of the reference point lead to inconsistencies across time, namely differences in acquisition and selling price for a good, and self-control problems. An example of the former is the endowment effect: subjects given one of two similarly valued objects value the one they are assigned more than the other. Adaptation, the idea that the reference point shifts to the status quo, explains this: denote the value function with explicit dependency on the reference point x_0 by $v(x; x_0)$ and the starting wealth by w. Then, the value of acquiring an object x is $v(x;w) - v(0;w)$; the value of keeping x after acquisition and integration is $v(0, w+x) - v(-x; w+x)$. Since the value function has loss aversion $v(0, w+x) - v(-x; w+x) > v(x;w) - v(0;w)$, leading to different bidding and asking prices. Note that utility would predict that the asking price is such that $u(p) = u(w+x) - u(w)$, and the same constraint applies to the bidding price. (It is the reference point dependency, and not the asymmetric s-shape of v that matters here; a concave $u(\cdot)$ also exhibits loss aversion.)

3.2.2 Inter-Temporal Decision-Making and Hyperbolic Discounting.

In order to make decisions over choice alternatives involving payoffs that occur at different instants in time, it is necessary to discount the utility of payoffs, so that the immediate utility equivalent of a payoff $(x \otimes t)$ is given by $D(t)\, u(x)$, where $D(t)$ is a decreasing convex function of t, called a discount factor with $D(0) = 1$ by definition.

Normative theory requires discounting to be exponential, such that the current utility of a payoff x at time t, given a discount rate ρ is $u(x) \exp(-\rho t)$. Exponential discounting guarantees that there is no inconsistency in choice due to the simple passing of time: if payoff x_1 at time t_1 is preferred to payoff x_2 at time t_2, then for any $\tau > 0$, x_1 at time $t_1 + \tau$ is preferred to x_2 at $t_2 + \tau$:

$$u(x_1) \exp(-\rho\, t_1) > u(x_2) \exp(-\rho\, t_2) \tag{3.2}$$

implies, for all ρ,

$$u(x_1) \exp(-\rho\, (t_1 + \tau)) > u(x_2) \exp(-\rho\, (t_2 + \tau)) \tag{3.3}$$

This property of time-invariance is important because it avoids inconsistencies in short-term versus long-term decisions. However, these inconsistencies, and other, more obvious, cases of inter-temporal inconsistencies have been documented in the decision-making literature. For instance, (Ainslie, 1975, 1991) document cases when the choice reverses with the passing of time. Subjects being offered a choice between one chocolate bar today versus two chocolate bars tomorrow choose the single, immediate, bar; the same subjects, when offered a choice between one chocolate bar in a week versus two chocolate bars in eight days, chose overwhelmingly the two chocolate bars. This is a clear violation of the implications in equations (3.2) and (3.3).

An alternative form of discounting, hyperbolic discounting, where

$$D(t) = \frac{1}{(1 + \alpha t)^{\beta/\alpha}},$$

can be used to explain the inter-temporal preference reversals. Assuming that $\alpha = \beta = 1$ for simplicity, we consider the choice between two payoffs x_1 and x_2. Say $u(x_1) = 1$ and $u(x_2) = 2$. Then $(x_1 \otimes 0) \succ (x_2 \otimes 2)$, but $(x_2 \otimes 4) \succ (x_1 \otimes 2)$, that is, a shift of two time units reverses preferences. Note that hyperbolic discounting over-discounts small differences in time for small t, relative to the exponential discounting form, and under-discounts small differences in time for large t; this can be seen as the hyperbolic discounting form having decreasing sensitivity over time (relative to the exponential discounting form)[1].

The absence of inter-temporal consistency is a possible explanation for the persistence of addictions even in decision-makers who recognize the addiction and state their intention to stop the addictive behavior, say B. For example, consider a decision-maker choosing whether to drop B today or tomorrow; as shown in the preceding paragraph, there are combinations of discount functional forms (hyperbolic) and utilities that make it possible that doing B today is preferred to the utility stream on not having B in the future (say starting tomorrow), while at the same decision time the stream of utility starting in two days is preferred to doing B tomorrow. Such a reasoning leads to always postponing the decision to drop the behavior, and consequentially never dropping it albeit it makes sense from a discounted utility point to drop it at some future time.

3.2.3 Preview Of The Rest Of The Paper

The objective of this paper is to develop an explanation for the two classes of phenomena just described, under the requirement that such explanation must include the rationality of the decision-making agent as a core assumption.

In the next section we will introduce the idea of categorical reasoning. In categorical reasoning there is loss of information, because the continuous numerical quantities like payoffs, probabilities, and times are categorized into a few distinct possibilities; within each

[1] Note also that if decision-maker X has hyperbolic discounting, another decision-maker Y can extract surplus from X without providing any service. Suppose X has a call option over A at time τ; then, at time 0, Y offers to sell X an upgrade from A at time τ to B at time $t + \tau$, which X prefers. At time τ, Y offers to sell X another upgrade, from B in t time units to A now (more precisely, at that time, τ) – in fact selling back to X her original call option. In both cases an hyperbolic discounting X will have some gain from accepting the "upgrade", therefore allowing Y to extract a surplus. In the above numerical example, Y could charge up to $1/15$ utils (X's utils) for changing the call option on C_1 at time 2 to a call on C_2 at time 4, and at time 2 charge up to $1/3$ utils exchange the options with X. (In this setup Y never has to worry about X executing the option over B, so there is no risk, nor float, in issuing it.) X would accept in both cases and Y would receive the value of $2/5$ X's utils.

possibility all values are indistinguishable. For this to be optimal, there must be some cost to precision, which is an assumption of the paper. If that cost of precision is higher than the expected value of the errors in decision made by agents using categorical reasoning, then losing information is optimal. The section describes two types of categorical reasoning: ordinal, where only the order of the categories is used, and cardinal, in which a category is represented by a value which can be operated upon as if it was the value of the input to the categorization (so if the input is a time, the representative value can be used as the representative time for the category into which the input fell).

Section 3.4 describes how the categories are derived endogenously by a rational agent. In order to do so, two more assumptions are added to the paper: first, that the categorization is generic, in the sense that it is not built from scratch for each problem (supported by the same argument of cost of thinking, here manifested as a cost of setup); second, that in order to make efficient use of the generic categorization the agent will standardize the inputs. The parameters for the standardization can either be recalled from memory, in cases when there is a norm for the problem at hand, or can be inferred from problem data.

Using the criterion that categorization should lead to unbiased representative values while minimizing the errors made by the decision-maker, we derive some results guiding the relative size of categories. With a concave utility function we choose categories that have a lower probability of occurrence when they contain low payoffs than when they contain high payoffs. (A concave utility function has steeper slopes for low payoffs than for high payoffs. Therefore, losing precision at low levels is more costly than at high levels. This leads to the optimality of choosing higher error probability when the payoffs are high.)

In order to develop the categorization, we require a distribution for the inputs. Using the criterion that we should assume the least about the functional form of such a distribution, we minimize a measure of information and obtain distributions for the inputs. The distribution for payoffs is a Normal, leading to decreasing sensitivity in categorization as the payoffs are further from zero; the distribution for time is an exponential, leading to decreasing sensitivity with increasing time.

Section 3.5 shows how an agent using the categorization derived in section 3.4 and the categorical decision rules derived in section 3.3 would behave as described by prospect theory and hyperbolic discounting. The last section then summarizes the points of the paper.

3.2.4 Relation to the literature

The work in this paper extends work on prospect theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1991; Thaler, 1985) and hyperbolic discounting.

Cumulative prospect theory (Tversky and Kahneman, 1992; Fennema and Wakker, 1997), which includes rank dependence (Luce, 1988; Luce and Fishburn, 1991), solves some problems with the original prospect theory; its treatment of prospects by rank order is explained in this paper by the standardization and encoding.

Another important stream of literature is that of bounded rationality and cost of thinking (Gingerenzer and Goldstein, 1996; Simon, 1955; Shugan, 1980; Rubinstein, 1998). Models of semi-orders and similarity (Beja and Gilboa, 1992; Bridges, 1983; Gilboa and Lapson, 1995; Rubinstein, 1988; Tversky, 1977) are also related to the model in this paper: fat indifference curves may be an outcome of categorization of stimuli into bins, but the approximation makes indifference intransitive: if there were no standardization, the indifference would be transitive, but with the standardization the categories may change enough that indifference becomes intransitive.

A final stream of literature related to this paper is that of categorization of numerical stimuli (Holyoak and Alker, 1976; Lakoff, 1987; Macmillan et al., 1977; Olson and Budescu, 1997; González-Vallejo et al., 1994; Viswanathan and Childers, 1996, 1997; Wallsten et al., 1993) and approximate reasoning (Armstrong et al., 1983; Basu, 1984; Huttenlocher and Hedges, 1994). We will use a model of sets as categories, and justify that after the development of the endogenous categorization model.

The contribution of the paper to each stream of literature is discussed in the conclusions, section 3.7.

3.3 Categorical choice

In this section we look at how an agent makes choices using an internal representation of payoffs, probabilities, and time, when that representation is categorical. The main issue with categorical representation is that there is information loss: if x and y are in the same category, then for decision-making purposes, the agent cannot distinguish x from y and her actions will reflect that. We will abstract from the creation of the categories (postponed till section 3.4) and assume that the categorization is in place; the objective of this section is to determine how, given that categorization, the rational agent should make decisions. The agent will treat the categorization as optimal, and that will be a consistent assumption by design of the categorization criteria.

In a categorical reasoning framework there are several possible levels of representation of a payoff x. For illustration we will consider a categorization with three different categories. Each category needs three components to be completely defined:

- A *label*, or name for the category. These are arbitrary objects, the placeholders for the information that an input has fallen into a specific category. For instance, for height we could choose "short", "average", "tall", as category labels – but they would have no meaning except when attached to a category. Since we will be interested in categories that have an underlying order (given by the order of the elements they represent) and are ordinal in nature[2], we will choose integer numbers as the labels and use the order of the labels to match the order of the categories. For the example, $\ell = \{-1, 0, 1\}$. (Another advantage of using $\ell \in \mathbb{Z}$ is that successor and predecessor operation become simple $+1$ and -1 operations.)

- The support, or *bin*, of the category: the values of x that belong in each category. Using C_ℓ to represent the category bin, we can choose, for the example, $C_{-1} = (-\infty, -1), C_0 = [-1, 2]$, and $C_1 = (2, +\infty)$. Note that the choice of a simple interval for each category is not required by categorization per se. We will impose an a-priori constraint that the C_ℓ have to be a partition of $\mathbb{R}$, so that no x is in more that one category and every x is in some category, but these partitions need not be sets of

[2] For example, in height, we know that "tall" refers to more height than "short"; however, we cannot say that any member of the former category is say three times taller than any of the latter; for cardinal relationships we have to define a representative element and only that representative element can be compared in a cardinal way. An ordinal relation is preserved by any monotonic increasing transformation.

simple intervals. Formally, the set of bins, $\mathcal{C} = \{C_\ell\}$, is in the set of all partitions of $\mathbb{R}$, or $\mathcal{C} \in \mathcal{P}(\mathbb{R})$.

- A *representative value*, r_ℓ, for the category. This is a value that is used in a cardinal way (as opposed to the labels which can at most be ordinal) by the agent using the categories. These values are unique up to a linear transformation. The choice of these values has to do with the criteria used for choice of $\mathcal{C}$. For the illustration we will use $r_{-1} = -2, r_0 = 0$, and $r_1 = 3$.

To summarize:

Definition 3. *A categorization for payoffs is a triple* (L, C, R) *where* $L = \{1, \dots, N\}$, $C \subset 2^{\mathbb{R}}$ *with* $\#C = N$, $R \in \mathbb{R}^N$.

Given the three elements per category, the label ℓ, the bin C_ℓ, and the representative value r_ℓ, the process of categorization has two steps: *encoding*, denoted $e(\cdot)$, the choice of a label for a payoff; and *decoding*, the output of the representative value for the category indicated by the label.

Definition 4. $e(x) = \ell$ *if and only if* $x \in C_\ell$.

Suppose $x = 3/2$; encoding will yield $e(1/2) = 1$; then, decoding will be $r_{e(x)} = r_1 = 3$.

For any given categorization, there are three different levels of precision in reasoning. At the highest precision level, the agent uses the payoffs as given, with no categorization: if x and y are two different payoffs, the agent uses x and y for decision-making. A second level involves categorization using the cardinal values: the agent uses $r_{e(x)}$ and $r_{e(y)}$ as inputs to the decision process. This leads to loss of information, since if $e(x) = e(y)$ the agent will not be able to distinguish x and y. At a third, and least precise, level, the agent will use only the category labels for decision-making: $e(x)$ and $e(y)$. In this case there is no cardinal value, so the agent must use decision rules defined over ordinal data. These will be presented in section 3.3.2.

There is a significant difference between using $r_{e(x)}$ and using $e(x)$. Consider the case of a risk-neutral agent choosing between two lotteries:

$$L_1 \doteq (-2, 1/2\,; 5, 1/2)$$

and

$$L_2 \doteq (1, 2/3).$$

(Assume for now that only payoffs are categorized, and probabilities are used unchanged; this is just to keep this example simple, not the way decision rules are derived.) Using the purely ordinal labels, there is no way to use the cardinal information in the probabilities. Since $e(-2) < e(0) = e(1) < e(5)$ does not create any dominance either for L_1 or for L_2, the purely ordinal agent cannot decide which lottery to choose. When the representative values for the categories are used, the agent can now compare expected values since the probabilities and the representative values are cardinal. The agent will choose L_2 since

$$1/2\, r_{e(-2)} + 1/2\, r_{e(5)} = 1/2 > 0 = 1/3\, r_{e(0)} + 2/3\, r_{e(1)}$$

Note that if the negative payoff in L_1 had been -100, the decision would still be for L_1, although a risk-neutral agent using the payoffs instead of the representative values would choose L_2. This is an example of the main problem of categorical reasoning: loss of information will lead to errors in decisions.

3.3.1 Cost of thinking and type of reasoning

Categorical reasoning, ordinal or cardinal, represents a loss of information with respect to normative reasoning. An agent should only use categorical reasoning if the opportunity cost of making some bad decisions (or, more precisely, the expected opportunity cost of the errors in decision-making) is offset by something. We propose that the something that counterbalances the opportunity cost of bad decisions is a cost of thinking. We assume that cost of thinking $c(\cdot)$ increases with the degree of precision in representation, both in type

$$\theta = \begin{cases} ORD & \text{Categorical Ordinal} \\ CARD & \text{Categorical Cardinal} \\ PP & \text{Perfect Precision (Normative)} \end{cases},$$

and, for $\theta \in \{ORD, CARD\}$, in the number of categories N. We make an assumption that cardinal reasoning is always harder than ordinal reasoning by having

$$\forall N : c(ORD, N) < c(CARD, N) < c(PP, -)$$

and make $c(\cdot)$ increasing in N for $\theta \in \{ORD, CARD\}$

$$N_1 > N_2 \quad \Rightarrow c(\theta, N_1) > c(\theta, N_2)$$

to capture the cost of precision.

For decisions where the opportunity cost is very high, those when the expected value of an error is high, opportunity costs will dominate $c(PP, -)$ and the agent will use normative reasoning. On the other hand, for cases when the decision implies lower opportunity costs, the agent will save effort and use the categorical reasoning, either ordinal — for cases where the error is less important — or cardinal.

A different view of the same logic is that an agent that always uses the categorical reasoning will do better on average than one that will incur the setup cost for each decision. This view may be more palatable especially since deciding which type of reasoning to use may be more complex than making the decision (in the limit, the ideal way to choose which type of reasoning to use is to compute the solution for the decision problem using each of the three types and then choose the one that creates the best decision — which will always be PP, since all cost will be sunk by the time the decision solutions are computed).

The number of categories will be assumed fixed at some N (which is not computed and is left open for most of the paper), but under some assumptions on the structure of the cost, and after developing an endogenous categorization scheme, the optimal number could be found. The psychological literature (Miller, 1956) identifies a small number of discrimination levels (between five and seven) as being a fair description of many human decisions. Under some reasonable assumption on the cost of thinking, and including the results from the next section of the paper, we can derive results for the comparative statics of N that are in accordance with intuition: when the value of a decision increases, N increases, when the cost of thinking increases, N decreases.

We will now look into decision rules for categorical reasoning. We first look at ordinal decision rules, since these are different from the normative model (they are the subcases where dominance exists); then we look into cardinal decision rules in order to derive criteria for the choice of r_i in categorization.

3.3.2 Ordinal decision rules

In this section we will look at decisions that use only ordinal information. If the ordinal agent prefers A to B, we denote that by $A \succ_{ORD} B$.

We will impose the following constraint on category bins:

Assumption 6. *For all x, y in the same domain, $x < y$ implies $e(x) \leq e(y)$.*

This assumption rules out intermingling of categories (see figure 3.1) and makes the deriva-

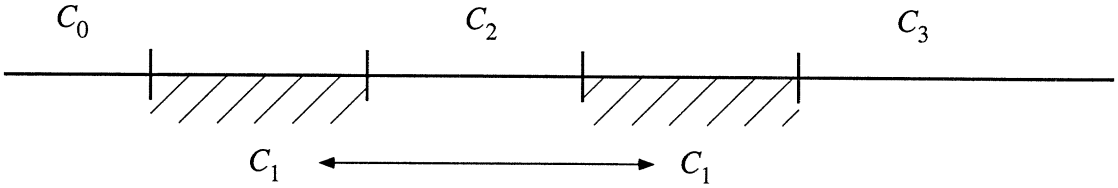

Figure 3.1: *Assumption 6 rules out intermingling of categories such as the shown in this figure. Elements of C_2 would at the same time be higher and lower than those in C_1, leading to a violation of the condition in one of the cases.*

tion of ordinal decision rules from dominance principles immediate. These rules are summarized next:

Assumption 7 (Ordinal decision rules). *Let x, y be payoffs, t_1, t_2 be times, and p_1, p_2 be probabilities. Then:*

$$e(x) > e(y) \textit{ and } e(p_1) \geq e(p_2) \textit{ implies } (x, p_1) \succ_{ORD} (y, p_2); \tag{3.4}$$

$$e(x) \geq e(y) \textit{ and } e(p_1) > e(p_2) \textit{ implies } (x, p_1) \succ_{ORD} (y, p_2); \tag{3.5}$$

$$e(x) > e(y) \textit{ and } e(t_1) \leq e(t_2) \textit{ implies } (x, t_1) \succ_{ORD} (y, t_2); \tag{3.6}$$

$$e(x) \geq e(y) \textit{ and } e(t_1) < e(t_2) \textit{ implies } (x, t_1) \succ_{ORD} (y, t_2). \tag{3.7}$$

These rules are equivalent to (and an extension to time of) condition * in Rubinstein (1988). To see that consider $x \sim y$ equivalent to $e(x) = e(y)$ and $p_1 \sim_p p_2$ equivalent to $e(p_1) = e(p_2)$. The rules in assumption 7 imply dominance:

Result 2. *Under assumption 6 the rules in assumption 7 are equivalent to dominance under the constraint that when $e(x) = e(y)$ the agent assumes $x = y$, be it for payoffs, times or probabilities.*

Proof: Immediate if we assume $e(a) = e(b)$ implies $a = b$. (End of proof.)

As pointed out above, an agent using categorical reasoning will lose some information. There are cases when that loss of information leads to reversal of preferences between normative and categorical reasoning:

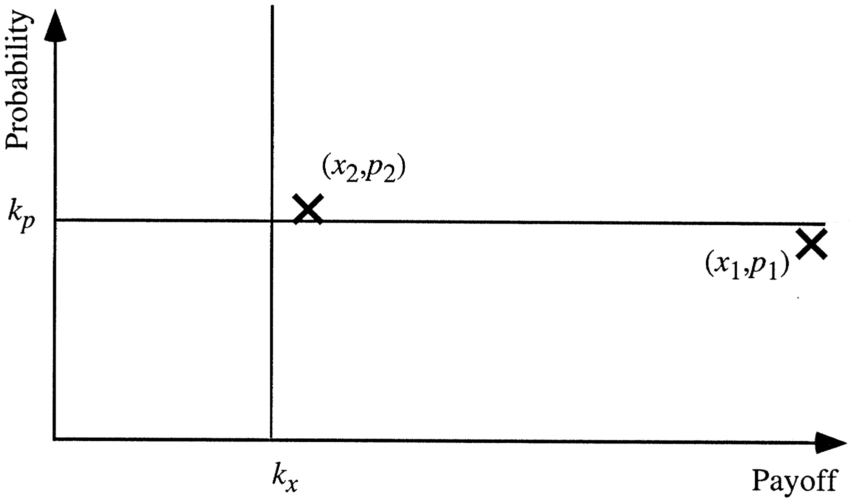

Figure 3.2: *For certain realizations of probabilities and payoffs, the categorical reasoning agent will make opposite choices from the normative model.*

Result 3. *There exist cases where $u(x_1)p_1 > u(x_2)p_2$ where $(x_2, p_2) \succ_{ORD} (x_1, p_1)$.*

Proof: Let $e(x_1) = e(x_2)$. Also, let $p_1 = k_p - \epsilon$ and $p_1 = k_p + \epsilon$, where k_p is a category boundary in probability space; this implies $e(p_1) < e(p_2)$. For a small ϵ, $u(x_1)p_1 \approx u(x_1)k_p$ and $u(x_2)p_2 \approx u(x_2)k_p$. Then, with $X_1 > x_2$, we can have $u(x_1)k_p > u(x_2)k_p)$, at the same time as $(x_2, p_2) \succ_{ORD} (x_1, p_1)$. Figure 3.2 has an illustration of this construction. (End of proof.)

For a fixed categorization, the ordinal agent will not have intransitive choices:

Result 4. *If $(x_1, p_1) \succ_{ORD} (x_2, p_2)$ and $(x_2, p_2) \succ_{ORD} (x_3, p_3)$, then $(x_1, p_1) \succ_{ORD} (x_3, p_3)$.*

Proof: Simple aplication of the rules in assumption 7. (End of proof.)

Result 5 (Special case rules). *Assume that $t = 0$ is a separate category, and $p = 0$ and $p = 1$ are also separate categories. Then, the following simple rules apply:*

$$e(w) < e(x) = e(y) \textit{ implies } x \succ_{ORD} (y, p; w, (1-p)) \quad \forall_{p \in (0.1)}; \tag{3.8}$$

$$e(w) = e(x) < e(y) \textit{ implies } (y, p; w, (1-p)) \succ_{ORD} x \quad \forall_{p \in (0.1)}; \tag{3.9}$$

$$e(x) \geq e(y) \textit{ implies } x \succ_{ORD} (y, t) \quad \forall_{t>0}. \tag{3.10}$$

Proof: Specialization of rules in assumption 7 to cases where the categories include separate singletons. (End of proof.)

These rules will be of special use in the illustration of prospect theory and hyperbolic discounting phenomena. Note that the categorization is endogenous, so these rules can only apply if the final result of the categorization satisfies the condition.

3.3.3 Cardinal categorical decisions

We will now focus on the cardinal decision rules. These use more information, since they are based on the r_i. The r_i are in the scale of the original variables (payoffs, time, and probability) and are unique only up to a linear transformation.

Definition 5 (Cardinal Preference). *An alternative (x_1, p_1, t_1) is preferred by a cardinal agent to a second alternative (x_2, p_2, t_2), denoted $(x_1, p_1, t_1) \succ_{CARD} (x_2, p_2, t_2)$ if and only if*

$$u(r_{e(x_1)}) \exp(-\rho r_{e(t_1)}) r_{e(p_1)} > u(r_{e(x_2)}) \exp(-\rho r_{e(t_2)}) r_{e(p_2)}.$$

If we compare the cardinal agent's equivalent of expected discounted utility

$$u(r_{e(x_1)}) \exp(-\rho r_{e(t_1)}) r_{e(p_1)} \tag{3.11}$$

with the normative expected discounted utility

$$u(x_1) \exp(-\rho t_1) p_1, \tag{3.12}$$

and assume that instances of x, t and p are drawn independently, we can derive criteria for the r_i in the following proposition:

Result 6 (Criteria for r_i). *Suppose x, t, p are drawn independently. Then, for cardinal preference $\succ_{CARD}$ to model normative preferences $\succ$, the r_i have to satisfy:*

$$\begin{aligned} u(r_{e(x)}) &= E[u(z)|e(z) = e(x)] & (3.13)\\ \exp(-\rho r_{e(t)}) &= E[exp(-\rho\tau)|e(\tau) = e(t)] & (3.14)\\ r_p &= E[\pi|e(\pi) = e(p)]. & (3.15) \end{aligned}$$

Proof: if x, t, p are independent random variables, then $r_{e(x)}$, $r_{e(t)}$ and $r_{e(p)}$ are also independent random variables, as are $u(r_{e(x)}$ and $\exp(-\rho r_{e(t)})$. Equations (3.13 - 3.15) imply that $u(r_{e(x)}$ is an unbiased estimator for $u(x)$, $\exp(-\rho r_{e(t)})$ is an unbiased estimator for

$\exp(-\rho t)$, and $r_{e(p)}$ is an unbiased estimator for p. Since the product of unbiased estimators of independent variables is an unbiased estimator of the product, the criteria in Proposition 6 make the cardinal estimator in equation (3.11) an unbiased estimator of the normative utility in equation (3.12). (End of proof.)

Result 7. $(x_1, t_1, p_1) \succ_{ORD} (x_1, t_1, p_1)$ *implies* $(x_1, t_1, p_1) \succ_{CARD} (x_1, t_1, p_1)$.

So, we can understand the behavior of a categorical cardinal reasoning by extending that of an ordinal categorical reasoning agent. We will use this to simplify the examples in section 3.5.

3.3.4 Summary of categorical decision

In this section we have presented rules for categorical reasoning. These rules were derived from optimization under the following constraints:

- When two numbers (be they payoffs, times, or probabilities) fall into the same category bin, the decision-maker assumes that they are the same number. This happens because she cannot distinguish between the original numbers. If the categories are to be used in a cardinal way, there is a representative element r_ℓ and all elements in the category are processed as if they were r_ℓ.

- Given this equality within categories, if an option O_1 dominates another O_2, the decision-maker should choose O_1 over O_2.

With these two assumptions and the normative ordering of payoffs (more is better), probabilities given sign of payoff (more probability is better for positive payoffs, and so on), and time given sign of payoff (less is better if the payoffs are positive, etc.), we derived ordinal decision rules. Cardinal decision rules were derived from applying expected discounted utility to the representative values.

3.4 Endogenous categorization

Categorical decision-making requires a categorization scheme. In this paper the agent is supposed to be as rational as possible under the constraints of the type of reasoning she is using. Therefore, the derivation of the categorization should be consistent with the utility formulation that supposedly would underlie the decision-making process of the agent were she to use the payoffs, probabilities and times directly.

Definition 6 (Categorization problem). *The problem of finding a categorization for payoffs x with N different bins, given a p.d.f for x, a norm $||\cdot||$ and a utility function $u(\cdot)$ is*

$$\min_{\substack{\{C_\ell\} \in \mathcal{P}(\mathbb{R}) \\ [r_\ell] \in \mathbb{R}^N}} \left\{ \int_{\mathbb{R}} \left|\left| u(x) - \sum_{\ell=1}^{N} u(r_\ell)\ 1(x \in C_\ell) \right|\right| \ dF_X(x) \right\}. \tag{3.16}$$

This general problem can be simplified if we choose the norm to be the squared error, $||x-y|| = (x-y)^2$; in that case, the representative category values are such that

$$u(r_\ell) = E[u(x)|x \in C_\ell],$$

as required in section 3.3.3 to make the elements in equation (3.11) unbiased estimators of those in equation (3.12)[3]; the C_ℓ are simple intervals defined by $N-1$ cut-off points k_ℓ,

$$C_1 = (-\infty, k_1]\,, C_2 = (k_1, k_2]\,, \ldots\,, C_N = (k_{N-1}, +\infty);$$

and the minimization problem in equation (3.16) reduces to choosing the cutoff points (with some abuse of notation we consider $k_0 = -\infty$ and $k_N = +\infty$) to solve

$$\min_{[k_\ell] \in \mathbb{R}^{N-1}} \left\{ \sum_{\ell=1}^{N} \mathrm{Var}\big[u(x) \,|\, k_{\ell-1} < x < k_\ell\big] \cdot \big(F(k_\ell) - F(k_{\ell-1})\big) \right\}. \tag{3.17}$$

The solution for this problem with linear utility and a Uniform distribution is a well-known result in information theory[4]: choose bins such that

$$\Pr(x \in C_i) = \Pr(x \in C_j)\,, \quad \forall C_i, C_j \in \mathcal{C}. \tag{3.18}$$

[3] Nota bene: different norms lead to different representative values. The choice of a absolute error norm, for instance, would lead to $u(r_\ell) = \mathrm{Median}[u(x)|x \in C_\ell]$, which deviates from the requirement in section 3.3.3. However, for many norms of the form $||x-y|| = (|x-y|)^m, m \in \mathbb{N}$, the results of categorization would be similar in the most important respects.

[4] The result can be found in many reference texts of information theory, for instance (Cover and Thomas,

For non-uniform distribution, this result approximates the optimal solution, being called a compandor-expandor solution. (See appendix C for details of how one relates to another.)

We now deviate from the information theoretic result to include the characteristics of the utility and discount functions in the categorization. When the utility function is nonlinear, the equiprobable partitions that satisfy equation (3.18) are no longer the solution to equation (3.17). The partition will depend on the curvature of $u(\cdot)$ and on the distribution. Using a Uniform distribution for illustration, we have the following results: if $\partial u/\partial x > 0$ and $\partial^2 u/\partial x^2 < 0$, the solution to equation (3.17) will be a partition that satisfies

$$\forall C_i, C_j \in \mathcal{C}, \quad i > j \quad \Rightarrow \quad \Pr(x \in C_i) > \Pr(x \in C_j), \tag{3.19}$$

(recall that the order of labels reflects the underlying order of the payoffs in the support of their categories, in the sense that if $i > j$, for any $x \in C_i$ and any $y \in C_j$, $x > y$), and, conversely, if $\partial u/\partial x > 0$ and $\partial^2 u/\partial x^2 > 0$, the partition that solves equation (3.17) will have

$$\forall C_i, C_j \in \mathcal{C}, \quad i > j \quad \Rightarrow \quad \Pr(x \in C_i) < \Pr(x \in C_j).$$

We now apply these generic results to the cases of payoffs, time and probability.

3.4.1 Categorization of payoffs

As shown in the previous section, categorization of payoffs requires a utility function and a distribution. In this section we will make reasonable assumptions on the decision-making process in order to derive the distribution.

Solving the optimal categorization problem (equation 3.17) for each decision-making instance requires a substantial effort. An assumption that will underlie most of the remainder of this paper is that the categorization scheme is shared among different decision occasions, at least to some point. This means that the agent may have several generic categorization schemes, say $\mathcal{C}_1, \mathcal{C}_2, \ldots$ and select among these based on the characteristics of the problem, but will not create a specific $\mathcal{C}$ for each problem. For instance, if the agent has two different degrees of risk aversion over problems with large versus small stakes, ρ_L and ρ_S, this may be modeled by having two different solutions for equation (3.17), say $\mathcal{C}_L$ and $\mathcal{C}_S$. When

1991; Gersho and Gray, 1992); the non-Uniform case is solved in Max (1960) and Lloyd (1982), with more general conditions derived in Trushkin (1982) and Trushkin (1993); in the behavioral literature Parducci (1965) and Krumhansl (1978) provide some empirical evidence that categorization of stimuli follows the general principles of encoding for best information usage.

faced with a decision problem, the agent chooses between $\mathcal{C}_L$ and $\mathcal{C}_S$ based on the stakes of the problem at hand and then uses the appropriate categorization for decision-making. (In most of the paper we will consider only one categorization for simplicity.)

Given that the categorization scheme is shared across different decision problems, the payoffs should be standardized for maximum efficiency in the use of bins. Suppose the payoffs are categorized into three categories, with intuitive meanings of "significant loss", "insignificant change", "significant gains". Consider that a categorization $\mathcal{C}$ is optimal for the changes in the price of a high variance stock. Then, $\mathcal{C}$ is probably too coarse for changes in a stable stock (most payoffs would fall into the "insignificant change" category bin). Standardizing the payoffs solves this problem, as the scale becomes a function of variance (e.g.: the payoffs fall into a category "large gain" if they are at least three times the standard deviation above the mean). Standardization is a trivial transformation: x is mapped into $(x - \mu_x)/\sigma_x$, where $\mu_x = E[x]$ and $\sigma_x = \sqrt{\text{Var}[x]}$.

The source of the information required for the standardization may be either memory or inference. When information about the distribution of outcomes for a decision problem like the one under consideration exists, the decision-maker may recall the necessary parameters and use these to normalize the inputs. Norm theory (Kahneman and Miller, 1986) implies precisely this: when there is a norm for a stimulus, the stimulus is perceived in a scale that takes into account the expected value and the variance of the distribution in the norm. When there is no norm to guide standardization, the required parameters may be inferred from the problem data. Inference from problem data in the absence of a strong adequate norm is very common, as seen in Gourville (1998), Kahneman and Miller (1986), Prelec et al. (1997), Simonson and Tversky (1992), Tversky and Kahneman (1991), Tversky and Simonson (1993), and Wernerfelt (1995).

For the rest of this section we will ignore standardization in the notation, and assume that the payoffs are already standardized to have mean 0 and variance 1.

We now need a criterion to choose a distribution. In keeping with the idea of creating the most generic partition possible, we choose to select a distribution based on how informative the realizations of random variables drawn from that distribution are. To illustrate the concept of how informative a distribution is consider the case where a random variable x is set to a constant $x = x_0$. In this case, any realization of x is completely predictable, hence uninformative. The functional form of the distribution contains all the information in the case of this constant random variable. What we want in our case is for the assumed distribution to be as generic as possible in the sense that any realization conveys as much information as possible (the exact opposite of having a constant random variable). This is the sense in which the functional form for the distribution contains the weakest assumptions. To do this we need a measure of information; we will use entropy (Shannon, 1948).

To determine the distribution with the weakest assumptions, and therefore the one that has the most informative realized values, we maximize the the entropy of the distribution $f(x)$, denoted $\mathcal{H}(f)$, and defined as

$$\mathcal{H}(f) = -\int_{\mathbb{R}} f(x)\log(f(x))dx, \tag{3.20}$$

subject to

$$f(x) > 0 \quad \forall_x, \tag{3.21}$$

$$\int_{\mathbb{R}} f(x)\,dx = 1 \quad \forall_x, \tag{3.22}$$

$$\int_{\mathbb{R}} x\,f(x)\,dx = 0 \quad \forall_x, \tag{3.23}$$

$$\int_{\mathbb{R}} x^2\,f(x)\,dx = 1 \quad \forall_x. \tag{3.24}$$

(Equations (3.21) and (3.22) are required for $f(\cdot)$ to be a probability density function; equations (3.23) and (3.24) are a result of the standardization. They also establish necessary conditions for the calculus of variations problem of minimizing equation (3.20) to be well-defined.) The solution to equation (3.20) under these constraints is known: $f(x)$ is the p.d.f. of a standard Normal.

Assumption 8. $f(x) = (2\pi)^{-1/2}\exp(-x^2/2)$.

A Normal distribution has important implications regarding the categorization bins. First let us consider a linear utility function, $u(x) = a + b(x), b \neq 0$. In this case, equation (3.18) can be used as an approximation and the C_i are equiprobable[5]. Given the form of the Normal p.d.f. we can derive an important result regarding the support of the categories: denote the bin for the category that contains the number zero (for the case with an odd number of bins; for the case with an even number of bins consider the zero point) by C_0, with

[5] As is shown in the appendix, these results are conservative; Max (1960) shows that the optimal categorization of a Normal has bins that obey equations (3.25) and (3.26); these optimal category bins are narrower near zero and wider away from zero than the compandor-expandor category bins. This is what is required for the paper, so the approximation is used for illustration.

label 0, so that for categories below C_0 the labels are negative numbers and for categories above C_0 the labels are positive numbers; then equations (3.18) and (8) imply:

$$\ell_i > \ell_j \geq 0 \quad \Rightarrow \quad \mathrm{supp}(C_{\ell_i}) > \mathrm{supp}(C_{\ell_j}) \tag{3.25}$$

$$\ell_i < \ell_j \leq 0 \quad \Rightarrow \quad \mathrm{supp}(C_{\ell_i}) > \mathrm{supp}(C_{\ell_j}) \tag{3.26}$$

This result leads to decreasing sensitivity over payoffs: small differences will be less likely to be discriminated when they are further from zero, as the support of the categories is larger. If we solve the optimal problem, the results in equations (3.25) and (3.26) are the same, as shown in Max (1960).

For small values of the concavity of $u(\cdot)$, defined in equation (C.3), the changes in k_ℓ resulting from equation (3.19) are small and the property that the support of category bins increases with distance to the category containing zero is maintained. (This can be seen by making a change of variable into the space of utility; see appendix C.) Note that with three categories this result is trivial for any cutoffs: the bins are $(-\infty, k_1]$, $(k_1, k_2]$, and $(k_2, +\infty)$. With $N > 3$, and for very high concavity of $u(x)$, some categories may have supports that are not increasing in the distance to the category containing zero (the extreme categories have trivially larger support, since it is infinite). We will assume for most of the paper that only a small amount of concavity (in the sense that the increasing support of categories is maintained) is present. Note that with a concave $u(\cdot)$, the result in equation (3.25) is preserved; however the result in equation (3.26) may apply only for part of the category bins in on the negative side of $\mathbb{R}$.

At this point we should note two things: first, although the Normal distribution has characteristics that are important for the final result, it is not the only distribution that exhibits those characteristics; second, there are other supporting arguments for distributions that exhibit the desired characteristics other than that it achieves maximum entropy. Elaboration on these points follows.

Any distribution where the density decreases away from the mean is adequate for the derivation of the main result in this section (equations 3.25 and 3.26). Depending on the number of bins and location of the cut-off points, the decreasing density may have small areas where it is increasing, as long as the probability mass inside a bin is not changed. The distribution should thus be unimodal when viewed from the output (density of r_i) but that does not necessitate a unimodal density for x. Hence, the Normal distribution is but one for which the main result holds. It is however one that is supported by several arguments, from entropy maximization and the central limit theorem on the theoretical side to empirical data.[6]

A general form of the result in this section is

Result 8 (Categorization of Payoffs). *If $f(x)$ and $u(x)$ are such that*

$$f'_X(u^{-1}(u))((u^{-1}(u))')^2 + f_X(u^{-1}(u))(u^{-1}(u))'' \begin{cases} < 0 & u > u(0) \\ > 0 & u < u(0) \end{cases}$$

and the constraint in equation (C.3) holds, then, for a fixed Δ,

$$\Pr(e(x) = e(x+\Delta)) \leq \Pr(e(y) = e(y+\Delta)) \quad 0 < x < y; \tag{3.27}$$

$$\Pr(e(x) = e(x+\Delta)) \leq \Pr(e(y) = e(y+\Delta)) \quad 0 > x > y; \tag{3.28}$$

$$\Pr(e(x) = e(x+\Delta)) \geq \Pr(e(-x) = e(-x-\Delta)) \quad x > 0. \tag{3.29}$$

This result shows how the decreasing sensitivity properties of the prospect theory value function result from endogenous categorization. Equation (3.27) is a model of decreasing sensitivity over gains; equation (3.28) is a model of decreasing sensitivity over losses; and equation (3.29) is a model of steeper sloper (larger effects) over losses than over gains.

3.4.2 Categorization of time

In order to categorize time we have to go through the same steps as in section 3.4.1. We will require standardization for the same reason as with payoffs: if a categorization scheme with bins $\mathcal{C}$ is optimal to describe cooking times for assorted food products, it will prove too coarse in the low end for timing the 100 yard dash (all results would be in the "very

[6] A different supporting argument for distributions where the probability density is concentrated near the mean is an incentive-compatibility story. Suppose there is a large set of problems with all possible distributions of payoffs $\mathcal{L} \in \mathbb{L}$. An agent with a concave utility $u(\cdot)$ should choose to invest effort to reduce the likelihood of being given a problem $\mathcal{L}_1$ that has high probability of low payoffs and trade it for one that has higher payoffs, say $\mathcal{L}_0$. Similarly, if a problem $\mathcal{L}_2$ has low probability of extreme events and another problem $\mathcal{L}_3$ has high probability of a high payoff, the agent will expend less effort to trade $\mathcal{L}_2$ for $\mathcal{L}_3$ than she would to trade $\mathcal{L}_1$ for $\mathcal{L}_2$, given the concavity of $u(\cdot)$. Therefore, by selecting where to invest effort to change the set of problems she encounters, the agent selects a distribution where most payoffs are near the mean (the final mean) and few are away from the mean, that is a distribution like the one that is required for the main result of this section.

quick" category bin) and too coarse in the high end to represent vesting periods for stock options (all periods would be in the "very slow" category bin).

A second point is that of usage. Whereas with payoffs we assumed that these were operated upon by a concave utility function, in fact locking the payoffs into the sole role of decision inputs, the use of time in decision-making is more diversified: discounting, when time means the epoch of realization of a payoff; cost, e.g. when creating the quote for a consulting service; and gain, e.g. when stalling for time while winning a Super Bowl game. For the time being, we will consider only discounting[7].

Given a discount rate ρ, the categorization problem for time is to find C_i and r_i to

$$\min_{\substack{\{C_\ell\} \in \mathcal{P}(\mathbb{R}_0^+) \\ [r_\ell] \in (\mathbb{R}_0^+)^N}} \left\{ \int_{\mathbb{R}_0^+} \left\| \exp(-\rho\, t) - \sum_{\ell=1}^{N} \exp(-\rho\, r_\ell)\ 1(t \in C_\ell) \right\|\ dF_T(t) \right\}. \tag{3.30}$$

Again under the assumption that $\exp(-\rho\, r_\ell) = E[\exp(-\rho\, t) | t \in C_\ell]$, the problem is to minimize weighted variance as in equation (3.17). In order to do that we need a probability distribution. Assuming the domain of time, for decision making purposes, to be the future (as indicated by the choice of $\mathbb{R}_0^+$ for the domain in equation 3.30), the same criterion as used for payoffs (minimizing the information content of the distribution functional form as given by $\mathcal{H}(f)$ in equation (3.20)) leads to

Assumption 9. $f(t) = \exp(-t)$.

Since the discount function is convex and decreasing, we solve the optimal category bin problem using the variance formulation, equation (3.17), variant of problem (3.30) — because, as implied in section 3.3.3, we want $\exp(-\rho r_i) = E[\exp(-\rho t) | t \in C_i]$. This leads to bins where

$$i > j \quad \Rightarrow \quad \Pr(t \in C_i) > \Pr(t \in C_j). \tag{3.31}$$

When we put together the result in equation (9) with the implication in equation (3.31), we have, for any $\rho > 0$, a decreasing sensitivity result over time:

[7] These different applications are not innocuous: the categorization problem depends on the shape of the function applied to the variable under categorization. With discounting, the function is decreasing convex, with cost it is reasonable to assume increasing and convex, whereas with gain it is plausible to choose increasing concave; these differences would make for different behaviors of $\Pr(x \in C_i)$ vs $\Pr(x \in C_j)$ for fixed $i > j$.

Result 9. *For any two categories for time C_i, C_j, if $i > j$ then* $\operatorname{supp}(C_i) > \operatorname{supp}(C_j)$.

Since the choice of probability distribution is again very important for the result, it is interesting to note that the supporting arguments for the use of a Normal distribution over payoffs apply *mutatis mutandis* to the use of an exponential discounting over time. The first observation is that the result in equation (9), which is the main point of time categorization, does not require a distribution that is exponential; all it really requires is a distribution that has density decreasing with increasing distance from the mean. Second, empirically, most decisions involve mainly payoffs that are close to the decision reference time.

3.4.3 Categorization of probabilities

Categorization of probabilities should be straightforward: since, for decision-making purposes, probabilities are used unchanged (in the normative model), the loss function is linear, hence the optimization problem can be written as

$$\min_{\substack{\{C_\ell\} \in \mathcal{P}([0,1]) \\ [r_\ell] \in [0,1]^N}} \left\{ \int_0^1 \left\| p - \sum_{\ell=1}^{N} r_\ell \, 1(p \in C_\ell) \right\| \, dF_P(p) \right\}. \tag{3.32}$$

and we can use the result in equation (3.18) to create a categorization with equiprobable category bins given a distribution. The distribution, however, is not as simple as for the previous cases.

We assume that decisions without uncertainty are more common than those with uncertainty. Also, in this case it is very important to separate cases that are certain from cases that are impossible. This leads to the following assumption (separating out the zeros and the ones):

Assumption 10. *There are two singleton categories $C_0 \doteq \{0\}$ and $C_1 \doteq \{1\}$.*

This provides the second part of the justification for the antecedent in proposition 5. Note that this assumption embodies a very strong discontinuity at zero and at one.

To derive a distribution for probabilities we start from a log-odds representation of uncertainty. The measure of uncertainty is thus

$$o \doteq \log\left(\frac{p}{1-p}\right).$$

Using the same maximum entropy argument as for the previous two dimensions, we will assume $o \sim N(0, 1)$. This implies that probabilties will have a distribution where the density decreases away from the extremes. This, in turn, leads to a pattern of decreasing sensitivity that matches that of the probability weighting function.

Result 10. *For a fixed* Δ*,* $\Pr\left(e(p) = e(p+\Delta)\right)$ *decreases away from* $p = 1/2$*.*

Proof: By making the change of variable $o \rightarrow p$ and then applying the linear $u(\cdot)$ version of the proof in appendix C, the result is immediate. (End of proof.)

Note that a maximum entropy distribution for probabilities would lead to a uniform distribution over $[0, 1]$. This does not match empirical distributions of probability values. A better-fitting distribution will have higher density near the extremes and lower density at the center. This is the type of distribution that would emerge from risk avoidance, and the one that is derived from maximum entropy in log-odds formulation.

3.4.4 Summary of assumptions and conclusions of endogenous categorization

In this section we derived the categories[8] for the decision-maker. This derivation is made under some assumptions:

- Quantities to encode are standardized. The standardization parameters, μ and σ can be recalled from memory or inferred from problem characteristics.

- Categorization minimizes the mean squared error in utility for payoffs, discounting for time, and probabilities.

- Payoffs are distributed Normal, and are standardized to have mean 0 and variance 1.

- Time is distributed Exponential with parameter $\lambda = 1$.

[8] In this section we have defined the category bins as sets, so that an element either is in a category bin or not. There are reasons to believe that this model of categorization is not adequate for the way human decision-makers categorize (Zadeh, 1965).

Models of categories exhibiting graded membership can capture non-abrupt transitions between categories. In some of these models, an x may belong to different bins with different membership values. That allows the model to capture effects like intransitive indifference and fat indifference curves. We choose to stay with a model that uses classical sets as bins for two main reasons: one, it is an approximation, so we choose a model that is elegant in that it has fewer free parameters; two, the semantics of graded membership are not as well developed as those of classical sets, so that using them would add an extra layer of assumptions to the theory.

- Probability has two mass points at 0 and 1 and has a p.d.f. over $(0, 1)$ which is increasing with distance to $1/2$.

Using these assumptions, we conclude that for some not too concave utility functions the category bins have support increasing with distance to zero; for time, the bins have increasing support away from zero; and, for probabilities, 0 and 1 are separate categories and the bins for the interval $(0, 1)$ increase in size towards the center.

3.5 Ordinal reasoning and behavioral decision phenomena

In this section we show that the behavioral phenomena introduced in section 3.2 can be explained by applying the decision rules of section 3.3 to the categorization derived in section 3.4; this brings the paper together. To motivate the categorization-based explanation for behavioral phenomena, we will use the following examples:

1. *Risk-aversion over gains with moderate probabilities.* In the gains domain, a decision-maker prefers a sure outcome over a lottery with the same expected value, when the probabilities in the lottery are not close to one and zero. In particular, for $s_i > 0$ and a moderate p_i, $s_i \succ (s_i/p_i \text{ w.p. } p_i; 0 \text{ w.p. } 1-p_i)$. The illustration will show that a decision maker prefers a gain of \$1 over a .5 probability of a \$2 gain.

2. *Risk-seeking over losses with moderate probabilities.* In the losses domain, a decision-maker prefers a lottery over a sure payoff with the same expected value, when the probabilities in the lottery are not close to one and zero. In particular, for $s_i < 0$ and a moderate p_i, $(s_i/p_i \text{ w.p. } p_i; 0 \text{ w.p. } 1-p_i) \succ s_i$. The illustration will show that a decision maker prefers to risk a .5 probability of a \$2 loss over the sure loss of \$1.

3. *Risk-seeking over gains with extremely low probabilities.* In the gains domain, a decision-maker prefers a lottery over a sure payoff with the same expected value, when the probabilities in the lottery are close to one and zero. In particular, for $s_i > 0$ and an extreme p_i, $(s_i/p_i \text{ w.p. } p_i; 0 \text{ w.p. } 1-p_i) \succ s_i$. The illustration will show that the decision maker prefers a one in a million probability of gaining \$1,000,000 over the sure gain of \$1.

4. *Risk-aversion over losses with extremely low probabilities.* In the losses domain, a decision-maker prefers a sure outcome over a lottery with the same expected value, when the probabilities in the lottery are close to one and zero. In particular, for $s_i < 0$ and an extreme p_i, $s_i \succ (s_i/p_i \text{ w.p. } p_i; 0 \text{ w.p. } 1-p_i)$. The illustration will show that the decision maker prefers a sure loss of \$1 over a one in a million probability of a \$1,000,000 loss.

5. *Loss aversion.* A decision-maker prefers the *status quo* to a lottery with expected value zero. In particular, for any $s_i > 0$, $0 \succ (s_1 \text{ w.p. } 1/2; -s_i \text{ w.p. } 1/2)$. The illustration will show that a decision-maker prefers zero to a 50-50 chance of loss or gain of \$1.

6. *Hyperbolic discounting.* Agents exhibit inter-temporal choice behavior that is inconsistent with changes in the moment of first event. It is possible for a decision maker to prefer one payoff s_1 now over s_2 at time t, and prefer s_2 at time $t+\tau$, $\tau > 0$, over s_1 at time τ.

For illustration we will consider a three category partition of the payoff space,

$$C_{-1} = (-\infty, -2/3], \quad C_0 = (-2/3, 1], \quad C_1 = (1, +\infty)$$

(these numbers are stylized for illustration, but they could be derived from a Normal distribution with a concave utility function, since they comply with equations (3.19), (3.25) and (3.26)) and a two category partition of the time space,

$$C_0 = [0, 1], \quad C_1 = (1, +\infty).$$

These numbers are selected for illustration, but it is important to notice that they are consistent with the distributions for payoffs and time and the concavity of utility and convexity of exponential discounting. We will not need a partition for probabilities for these examples, but one could also be considered, for instance

$$C_{never} = \{0\}, \quad C_0 = (0, 1/10], \quad C_1 = (1/10, 9/10), \quad C_2 = [9/10, 1), \quad C_{sure} = \{1\}.$$

We will use these categories to demonstrate how the conjunction of categorization (including standardization) and ordinal reasoning rules provides an explanation for the examples 1 – 6 above.

3.5.1 Decision under uncertainty — risky choice examples

For most part in section 3.4 we assumed that the problem data were already standardized in order to simplify notation. In this section we will look at decision-making, so standardization needs to be explicit. We assume that there is no strong norm for the value of σ; it is to be inferred from problem data. Since the problems are choices over payoffs, we will assume that there is a norm implying $\mu = 0$, so that zero is always treated as zero.

Decision-making requires a three step process: first, infer σ from the choice payoffs and probabilities; second, encode x/σ and find a label $e(x/\sigma)$ and a representative value $r_{e(x/\sigma)}$; third, apply decision rules (ordinal rules are sub-cases of a cardinal categorical comparison, so if we can reach a conclusion using ordinal rules, this conclusion is the same as that reached using r_i) using $e(x/\sigma)$ and/or $r_{e(x/\sigma)}$.

Beginning with example 1, we note that

$$\begin{aligned} 0 < \sigma < 1 \Rightarrow \quad & e(0) < e(1/\sigma) = e(2/\sigma) \Rightarrow \quad \text{Choose 1} && (3.33)\\ 1 < \sigma < 2 \Rightarrow \quad & e(0) = e(1/\sigma) < e(2/\sigma) \Rightarrow \quad \text{Choose lottery} && (3.34)\\ \sigma > 2 \Rightarrow \quad & e(0) = e(1/\sigma) = e(2/\sigma) \Rightarrow \quad \text{No decision.} \end{aligned}$$

The choices are driven by the ordinal decision rules; equation (3.33) is an application of rule (3.8) and equation (3.34) is an application of rule (3.9).

The plausibility of the different values for σ is not the main question here, although one could imagine that the lower values are more likely, given the payoff ranges and the probabilities, than the higher ones. The question is how likely each of these partitions for σ is, conditional on the problem data, and especially when compared across different cases[9]. The key question is, does the standardization variance change in the way that is consistent with the data in the decision (namely the distribution of the payoffs), for different problems? To answer that we will look at all four risky choice examples 1 – 4.

Moving on to example 2, we note that

$$\begin{aligned} 0 < \sigma < 3/2 \Rightarrow \quad & e(0) > e(-1/\sigma) = e(-2/\sigma) \Rightarrow \quad \text{Choose lottery} && (3.35)\\ 3/2 < \sigma < 3 \Rightarrow \quad & e(0) = e(-1/\sigma) > e(-2/\sigma) \Rightarrow \quad \text{Choose } -1 && (3.36)\\ \sigma > 3 \Rightarrow \quad & e(0) = e(-1/\sigma) = e(-2/\sigma) \Rightarrow \quad \text{No decision.} \end{aligned}$$

(rule (3.9) implies result (3.35); rule (3.8) implies result (3.36).) So, again for the smaller values of the σ[10], the desired result, risk-seeking over losses, is obtained.

Example 3 is similar to example 1, except for the fact that the numbers involved are much higher. Notice that in this case a high value for σ, especially compared to the numbers

[9] In this paper we have no model of how the values for the standardization are inferred from problem data, except for the most likely comparative statics. However, consider for illustration that the agent computes a naive approximation of σ by giving the sure payoff a over-representation in the payoff sample to account for its higher probability. Suppose that the over-representation makes the sample vector be $[0, 1, 1, 2]$ (note that the distribution having continuous support and the payoffs being discrete numbers complicates the problem of creating a maximum likelihood estimate without a separate theory of realization; we assume that the agent will steer clear of these complications and act naive). With this sample vector, the empirical unbiased estimate for the standard deviation is $1/\sqrt{3}$, hence the agent would act as predicted by prospect theory. Since the numbers in the categorization are arbitrary, this is not a strong argument, it is just meant as an illustration.

[10] Including the same naive estimate as for illustration of example 1, $\hat{\sigma} = 1/\sqrt{3}$.

used in example 1 is much more likely[11]; the cases are

$$0 < \sigma < 1 \Rightarrow \quad e(0) < e(1/\sigma) = e(1000000/\sigma) \Rightarrow \quad \text{Choose 1} \tag{3.37}$$

$$1 < \sigma < 1000000 \Rightarrow \quad e(0) = e(1/\sigma) < e(1000000/\sigma) \Rightarrow \quad \text{Choose lottery} \tag{3.38}$$

$$\sigma > 1000000 \Rightarrow \quad e(0) = e(1/\sigma) = e(1000000/\sigma) \Rightarrow \quad \text{No decision.}$$

Example 4 leads to similar (yet symmetric) conclusions:

$$0 < \sigma < 3/2 \Rightarrow \quad e(0) > e(-1/\sigma) = e(-1000000/\sigma) \Rightarrow \quad \text{Choose lottery} \tag{3.39}$$

$$3/2 < \sigma < 1500000 \Rightarrow \quad e(0) = e(-1/\sigma) > e(-1000000/\sigma) \Rightarrow \quad \text{Choose } -1 \tag{3.40}$$

$$\sigma > 1500000 \Rightarrow \quad e(0) = e(-1/\sigma) = e(-1000000/\sigma) \Rightarrow \quad \text{No decision.}$$

When we compare examples 1 through 4, we note that a consistent standardization scheme would select small values of σ for narrow ranges of payoffs and larger values of σ for broader ranges of payoffs. So, a consistent standardization scheme, coupled with the categorization scheme proposed, when operated upon by the ordinal choice rules, will create the same effects as the conjunction of an s-shaped value function $v(x)$ and a probability weighting function $w(\pi)$.

The main point of these four examples can be clarified when the numbers are put together. The standardization has been assumed endogenous, therefore it is a question of plausibility of the σ in equations (3.33) – (3.40). First, note that the ranges of σ for which the choices with a median probability value (3.33) and (3.35) behave as predicted by the prospect theory value function contain much smaller values than those for (3.38) and (3.40), where the choices also conform with prospect theory, via the interaction of the value function and the probability weighting function. Therefore, if the increases in problem range lead to increasing σ, that fact plus the categorization create same effect as the shape of the value function and the probability weighting function (except for the loss aversion, which is illustrated in the next section). Since the standardization may be inferred from the problem data, if we have two sets of payoffs, one with a narrow range $\mathcal{N} = \{0, 1, 2\}$ and one with a broad range, $\mathcal{B} = \{0, 1, 1000000\}$, and we have two distributions, one with low variance σ_L^2 and one with higher variance σ_H, if we have to match payoff sets with distributions, it is more plausible that $\mathcal{N}$ matches σ_L^2 and $\mathcal{B}$ matches σ_H^2 than $\mathcal{B}$ matches σ_L^2 and $\mathcal{N}$ matches σ_H^2.

[11] Using the same inference rationale as in footnote 9, the value for a naive estimate of σ is $\hat{\sigma} \approx 1000$.

In summary, if increasing range of payoffs leads to higher σ, the predictions of prospect theory regarding choices between sure payoffs and lotteries can be explained by ordinal reasoning.

3.5.2 Loss aversion

In order to show that the conjunction of standardization and categorical reasoning also creates the loss aversion effect, we turn to example 5, and note that for the cases where it is possible to make an ordinal decision, we have loss aversion:

$$0 < \sigma < 1 \Rightarrow \quad e(-1/\sigma) < e(0) < e(1/\sigma) \quad \text{No decision;} \tag{3.41}$$
$$1 < \sigma < 3/2 \Rightarrow \quad e(-1/\sigma) < e(0) = e(1/\sigma) \quad \text{Choose 0;} \tag{3.42}$$
$$\sigma > 3/2 \Rightarrow \quad e(-1/\sigma) = e(0) = e(1/\sigma) \quad \text{No decision.} \tag{3.43}$$

so that any preferences derived from the ordinal decision process are $0 \succ (-1, 1/2; 1, 1/2)$. The origin of this effect is the asymmetric nature of the category bins, which are derived from the concavity of the utility function. This is fortunate, since the utility function itself exhibits — by virtue of its concavity — loss aversion.

A further point is that for the case in equation (3.41), the cardinal categorical decision will be a choice of the sure 0 over the lottery, given that the r_i are computed to create an unbiased representation of the utility, and the utility exhibits decreasing marginal returns.

3.5.3 A more general view of how categorization of payoffs and probabilities leads to the same effects as prospect theory

As the heading suggests, this section goes beyond examples 1 – 5 above and illustrates how categorization that is built from the normative models leads to effects that match those created by the prospect theory value function and by the probability weighting function.

We start by noting that a description of the value function over gains is that it has decreasing sensitivity: that the effect of a change $\Delta > 0$ in payoffs, from $x > 0$ to $x + \Delta$ is decreasing in x. Now, note that with the categorization in the illustration,

$$C_{-1} = (-\infty, -2/3], \quad C_0 = (-2/3, 1], \quad C_1 = (1, +\infty)$$

for a given Δ, whenever $x \in C_1$, $x+\Delta \in C_1$, but there are cases (when $0 < x < 1$, $0 < \Delta < 1$ and $x+\Delta > 1$) when $x \in C_0$ and $x+\Delta \in C_1$; that is, as x increases, the discrimination ability (in the sense of ability to determine whether two payoffs are different) of the categorization decreases. This is a case of decreasing sensitivity: the same effect (say $\Delta = 2/3$ that can be detected for small numbers (say $x = 1/2$) can no longer be detected for large numbers (say $x = 3$).

With a higher number of categories this effect holds, because of the result in equation (3.25). With a large number of categories the agent is capable of discriminating ever smaller Δs for lower values of x, but for any Δ there is a limit $\bar{x}$ beyond which it is very likely that $x > \bar{x}$ implies $e(x) = e(x + \Delta)$. Given equation (3.25) we can in general state that

$$0 < x < y \quad \Rightarrow \quad \Pr\left(e(x) = e(x + \Delta)\right) < \Pr\left(e(y) = e(y + \Delta)\right) \tag{3.44}$$

When put together with the ordinal choice rule (3.8), this leads to decreasing sensitivity in choice for the same difference Δ, as the base payoff x increases.

Conversely, if $x < 0$, a small change from x to $x - \Delta$, with $\Delta > 0$, is more likely to be detected for small $|x|$: when $|x|$ is small (say $x = -1/2$), $x \in C_0$ and for some small values of Δ (say $\Delta = -2/3$), $x \in C_0$ and $x - \Delta \in C_{-1}$; for larger values of $|x|$ (say $x = -2$), $x \in C_{-1}$ and $x - \Delta \in C_{-1}$, hence $e(x) = e(x - \Delta)$. For a higher number of categories, and given the result in equation (3.26), we know that the symmetric result of equation (3.44) holds (note the change in order in the premise of the implication):

$$y < x < 0 \quad \Rightarrow \quad \Pr\left(e(x) = e(x - \Delta)\right) < \Pr\left(e(y) = e(y - \Delta)\right), \tag{3.45}$$

and this, together with the ordinal choice rule (3.9), leads to decreasing sensitivity in choice for the same difference Δ, as the magnitude of the base payoff $|x|$ increases.

The reversal of risk-attitudes in the cases with extreme probabilities come through standardization: the decreasing sensitivity of the categorization is offset by the larger σ inferred from very broad ranges (note that if $p_i \approx 0$, then a comparison of s_i with s_i/p_i has a much broader range than if $p_i \approx 1/2$, since the first s_i/p_i is a very large number).

Regarding the loss aversion, there are essentially three cases, as in the illustration. Either all payoffs fall into the same category, in which case there is no possible decision using the categorization (as in equation 3.43); the payoffs fall into three different categories as in equation (3.41), in which case the ordinal decision-maker will not be able to determine dominance (and will possibly resort to cardinal decision making, as explained below) or two payoffs will fall into the same category whereas another will not, like in equation (3.42). In this last case, the asymmetry of the categorization with respect to the zero point guarantees that it is the lower payoff that falls into the separate category: since

$|k_{-1}| < k_0$, for any $x > 0$, if $-x > k_{-1}$ then necessarily $x > k_0$, but the converse is not true $(e(-3/4) < e(0) = e(3/4))$.

Regarding the case when $e(-x) < e(0) < e(x)$, for a positive x, we note that the representative values for the categories are derived from a concave utility function; for the three category case that means that $r_0 - r_{-1} > r_1 - r_0$, therefore whenever an agent has to make a cardinal categorical decision, she will choose 0 since

$$1/2\, r_{-1} + 1/2\, r_1 < r_0.$$

At this point it has been shown that other than when $e(x) = e(0) = e(-x)$ the agent will act driven by loss aversion. As the number of categories increases the probability of $e(x) = e(0) = e(-x)$ decreases, but the ordinal choice and the cardinal choice (which becomes the most likely one as N increases) continue to create the effect of loss aversion.[12]

Note also that the ordinal decisions in equations (1) – (5) are consistent with cardinal categorical decisions (as ordinal decisions are based on dominance, hence they imply dominance in categorical values also). Therefore, even if the agent is more sophisticated than a simple ordinal decision-maker, she will still behave as predicted by these equations.

Regarding the probability weighting function, we note that the categorization of probabilities as derived in section 3.4.3 leads to decreasing sensitivity near the center of the $[0, 1]$ interval. Therefore, effects as those created by the probability weighting function may be explained by this decreasing sensitivity. Another explanation, as derived from the illustration, is based on the inference of σ from payoff range and standardization prior to categorization.

3.5.4 Decisions over time — hyperbolic discounting-type results

Regarding example 6, inter-temporal inconsistency in choice, there are several differences. The first is that we will require categorization of time as well as of payoffs for the example. A second difference is that the trade-offs between the standardization of time and the standardization of payoffs make the results less clear. We will denote the standard deviation for time by σ_t and the one for payoffs by σ (the same as with the other cases). We will also denote the encoding of time by $e_T(t)$ to separate it from the encoding of payoffs $e(x)$. For clarity we will ignore the standardization of time altogether for now, and discuss its implications later.

[12] It is worth repeating here that the loss aversion in this model is derived from the concavity of the utility function, which exhibits loss aversion itself.

Consider the following choice alternatives: $(1 \otimes t)$ versus $(2 \otimes t + 1)$. Assume that σ is such that $e(2) > e(1)$. Given the categories for illustration of time above,

$$C_0 \doteq [0, 1], \quad C_1 \doteq (1, +\infty),$$

we note that for small t, $e_T(t) < e_T(t + 1)$. In these cases, the decision-maker will resort to cardinal decision-making (categorical); in many cases the discount rate will be enough to make

$$r_{e(1)}\, r_{e_T(t)} > r_{e(2)}\, r_{e_T(t+1)}, \tag{3.46}$$

which implies

$$(1 \otimes t) \succ (2 \otimes t + 1) \quad \text{for small } t. \tag{3.47}$$

When t increases, the likelihood that $e_T(t) = e_T(t + 1)$ increases, since the support of C_0 is much smaller than that of C_1; hence as t increases, and without changing any of the quantities that lead to the results in equation (3.46) — that is, without change in the underlying discount rate and utility function — the decision maker will be faced with a situation where

$$e(2) > e(1) \quad \text{and} \quad e_T(t) = e_T(t + 1)$$

at which point rule (3.6) applies and implies

$$(2 \otimes t + 1) \succ (1 \otimes t) \quad \text{for large } t. \tag{3.48}$$

Note that if the decision-maker were to make the decision using cardinal categorical reasoning, to maintain consistency with equation (3.46), the result would be the same: $e(2) > e(1)$ implies

$$r_{e(1)}\, r_{e_T(t+1)} < r_{e(2)}\, r_{e_T(t)}, \tag{3.49}$$

since the $r_{e_T(\cdot)}$ are equal in this case. Therefore the result is not an artifact of using two different kinds of reasoning for (3.47) and (3.48)), but a result of categorization itself.

So, we have shown that categorization can explain the reversal of inter-temporal preferences (from equation (3.47) to equation (3.48) or from equation (3.46) to equation (3.49) without any change in the underlying normative decision-making elements, utility (which makes $e(2) > e(1)$ and provides unbiased $r_{e(1)}$ and $r_{e(2)}$) and discount rate (which, coupled with the distribution of time, is responsible for C_0 and C_1 and unbiased estimators for $r_{e_T(t)}$), or in the distribution assumptions for optimal categorization.

At this point we want to get back to the question of standardization of time. The basis for the hyperbolic discounting is essentially that in the cases when t is small $e_T(t) < e_T(t{+}1)$, whereas when t is large $e_T(t) = e_T(t+1)$. Standardization changes this somewhat, as it makes the difference $((t+1)-t)/\sigma_T)$ smaller with increasing σ_T. So, if we assume that σ_T will be increasing in t, the result becomes weaker. It need not disappear, however. As long as the rate increase in σ_T with t is smaller than 1 (this critical rate is adequate for the numbers in the example), or more precisely, as long as the σ_T increases slowly with t — something that is consistent with an inference story behind the standardization[13], hence consistent with the theory behind categorization — the result still holds.

Another point pertaining to standardization is that it may be done by categories itself. So, the choices where the range is $t \in [0, 10]$ may all elicit the same $\sigma_t = 1$; in that case, the result in equations (3.47) and (3.48) holds.

3.5.5 How the model works

There are elements to the model, and these account for specific parts of the results. In this section we trace the effects back to the origin.

- The decreasing sensitivity of the value function is a result of the Normal distribution of payoffs. This creates risk-aversion over gains and risk-seeking over losses.

- Loss aversion is created by the concavity of the utility function.

- Reversal of attitudes towards risk with low probabilities has two separate explanations: the probability weighting function — which is derived in the model — and standardization. Therefore the phenomenon is over-determined and empirical tests are necessary to separate the two possible causes.

- The probability weighting function is a direct result of the distribution assumed for probability values.

[13] For any fixed Δ, the unbiased estimator for the standard deviation given two realizations from a distribution, t and $t+\Delta$, is $\hat{\sigma}_T = \Delta/2$, independent of t; this simplifies the result with standardization to the cases in equations (3.47) and (3.48). Since the maximum likelihood estimator for the standard deviation, using the fact that the distribution is exponential, is $\sigma_T^* = t + \Delta/2$, a combination of these two estimates would lead to a choice of σ_T that increases slowly with t, hence making the results still hold for cases with standardization.

- Decreasing sensitivity over time (and the consequent hyperbolic discounting) is a result of the Exponential distribution for payoffs.

Given the causes above, why categorical reasoning at all? It is not cited anywhere. But, without the need for categorization there is no need for generic distributional assumptions (why would they matter?). At most, the distribution of each problem should be considered; and in this case, categorization would not exhibit the decreasing sensitivity created by a Normal (or Exponential) distribution.

3.6 Other implications of the model

In this section we look at model predictions beyond those made by prospect theory and hyperbolic discounting.

3.6.1 Inference versus recall of standardization parameters

The inference of σ and μ from problem data leads to different results than the recall of those parameters from a norm in memory. For illustration suppose the decision-maker is asked to categorize the set of payoffs

$$-2000, -1500, -1000, -500, -100, -50, -20, -5.$$

If these are lottery payoffs, she may use the generic categorization, assume $\mu = 0$ as the reference point for lottery payoffs and categorize them in the following sets:

$$\{-2000, -1500, -1000, -500, -100\}, \{-50, -20\}, \{-5\}.$$

Suppose that the payoffs are prices for computers, and the decision-maker has a strong norm on these with $\mu = -1000$ and $\sigma = 100$; then, the categorization will look like

$$\{-2000, -1500\}, \{-1000\}, \{-500, -100, -50, -20, -5\}.$$

There is a problem with this thought experiment: given the set of observed payoffs, the decision-maker may doubt the recalled norm, or infer that the problem is too "make-believe" to be credible.

3.6.2 Intransitive versus transitive indifference

All other things being equal, the categorization shown above should lead to transitive indifference: if $x \sim y$ then $e(x) = e(y)$ and therefore, if $y \sim z$ we can conclude $x \sim z$ since $e(x) = e(y) = e(z)$. There are two caveats to this logic. The first, partial membership in a category, was assumed away in the choice of sets as category bins, but is one possible counter force for this transitivity. The second is a part of the model, standardization driven by inference.

If we model the standardization (and for illustration assume it is just inference of μ, not σ, which will be fixed at $\hat{\sigma} = 1$), we have

$$x \sim y \quad \Rightarrow \quad e(x - \mu(\{x, y\})) = e(y - \mu(\{x, y\}))$$

and

$$z \sim y \quad \Rightarrow \quad e(z - \mu(\{z, y\})) = e(y - \mu(\{z, y\})).$$

To illustrate how this creates intransitive indifference, let the estimator of μ be the empirical mean, $x = 1$, $y = 2$, $z = 3$, and the categories be $C_{-1} = (-\infty, -2/3), C_0 = [-2/3, 1)$ and $C_1 = [1, +\infty)$. Now it is clear that $x \sim y$, since $e(x - \mu(\{x, y\})) = e(-1/2) = 0 = e(1/2) = e(y - \mu(\{x, y\}))$, and $y \sim z$, since $e(y - \mu(\{y, z\})) = e(-1/2) = 0 = e(1/2) = e(z - \mu(\{z, y\}))$, but not $x \sim z$ since $\mu(\{x, z\}) = 2$ and

$$e(x - 2) = e(-1) = -1 < e(z - 2) = e(1) = 1.$$

The other possible source of intransitive indifference in this categorization is the use of partial membership in category bins, but this requires going into the formalism of decision-making with fuzzy sets; this is beyond the scope of this paper.

Inference of μ can lead to intransitive choice (how?) Note that this is not inconsistent with prospect theory, as μ is used in this model as the reference point, and inference of the reference point may lead to inconsistent choices even in standard models of prospect theory.

3.6.3 Category bins follows specific patterns

The categorization derived from the generic assumptions has distinctive properties (assuming a not too concave utility function):

- *Asymmetric categorization of payoffs.* If the respondent exhibits loss aversion, then her category bins have to be narrower on the loss side (for cut-off points with labels centered at zero, $|k_{-i}| < k_i$).

- *Decreasing sensitivity for payoffs.* Categories for payoffs have narrow bins near zero and wide bins away from zero

- *Decreasing sensitivity for time.* Categories for time have narrow bins for small t and wide bins for large t.

- *Non-monotonic categorization of probabilities.* Categories for probabilities have narrow bins near the extremes and wide bins near the middle.

Note that these categorizations are non-optimal for each small problem by itself. For instance, if an agent were given the following payoffs

$$-180 - 110, -100, -90, -5, -2, 0, 2, 5, 90, 100, 110, 180$$

to categorize into five bins (and assuming for a moment no risk-aversion, so that the problem is very simple), the optimal choice would be

$$\{-180\}, \{-110, -100, -90\}, \{-5, -2, 0, 2, 5\}, \{90, 100, 110\}, \{180\}.$$

However, categorization using a generic distribution will lead to something like

$$\{-180, -110, -100, -90\}, \{-5, -2\}, \{0\}, \{2, 5\}, \{90, 100, 110, 180\}$$

or

$$\{-180, -110, -100, -90\}, \{-5\}, \{-2, 0, 2\}, \{5\}, \{90, 100, 110, 180\}.$$

The same type of results applies for time and probability.

3.6.4 Cardinal values and skewness

The derivation of category bins in section 3.4 creates cardinal representative values that lead to unbiased representations of utility, discounting, and probability. However, since the distributions, within each bin, are skewed, this creates interesting likelihood effects. For instance, consider the categories for linear utility (or no utility, just encoding of numbers). For positive x, the probability density in each category is higher towards the lower end of the bin; therefore, the median of each bin is lower than the mean. The opposite happens for negative x. This creates a likelihood distortion: for $x > 0$, $\Pr\left(x < r_{e(x)}\right) > 1/2$; for $x < 0$, $\Pr\left(x > r_{e(x)}\right) > 1/2$. So, given that a payoff $x > 0$ falls into a category C_ℓ, the estimate used by the decision-maker is more likely to under-estimate x than to over-estimate x. Note again that this is not a systematic bias; this is conditional on each value of x, and the effect disappears when integrated over the support of C_i.

This is a result that depends on the chosen norm not being the absolute value, otherwise this would not hold. If the norm had been the absolute value, the results would be biased towards the center. So there is a trade-off in the model, but there is no way around it in reality either: if the norm is chosen to avoid likelihood effects, it will lead to worse decisions, because the cardinal categorical utility, equation (3.11) in section 3.3.3, will be a biased estimator of the correct normative value, equation (3.12).

With concave utility and Normal distribution, we have that the categories are skewed towards the center. For negative values of x we can use this and the Schwartz inequality to conclude

$$\Pr\left(u(x) > u(r_{e(x)}) \middle| x < 0\right) > 1/2, \quad x < 0.$$

This result does not have a positive side counterpart; we know that $\Pr(r_{e(x)} < x|x > 0) > 1/2$ but that does not allow conclusions derived from Schwartz's inequality, since $u(E[x]) > E[u(x)])$.

When estimating probabilities, the following happens (for $p \in (0,1)$, since 1 and 0 are categories on their own):

$$\Pr\left(p < r_{e(p)}|p < 1/2\right) > 1/2$$

and

$$\Pr\left(p > r_{e(p)}|p > 1/2\right) > 1/2.$$

So, although there is no bias in estimation ($r_i = E[p|p \in C_i]$), there is a higher likelihood of over-estimating probabilities than of under-estimating them for low values of p, and the reverse for high values of p.

3.6.5 Effect of different concavities of the utility function

Suppose there are two different categorizations for payoffs, $\mathcal{C}_1$ and $\mathcal{C}_2$, each being the result of solving problem (3.16) for a different utility function, respectively u_1 and u_2. If the concavity of u_2 is higher than that of u_1, $\mathcal{C}_2$ will be more asymmetric than $\mathcal{C}_1$, in the sense that (using superscripts for the categorization)

$$k_i^1 - |k_{-i}^1| < k_i^2 - |k_{-i}^2|.$$

It also means that the cut-off points in $\mathcal{C}_2$ will be to the right of equivalent points in $\mathcal{C}_1$:

$$k_i^1 < k_i^2. \tag{3.50}$$

This is independent of k_i^j's sign. Therefore, we can conclude that an increase in the concavity of the utility function leads to more loss aversion, more risk-aversion over gains and *less* risk-seeking over losses. In fact, this can be made more general: any change in the attitude towards risk on the gains side (say an increase) due to changes in the utility function, leads to a change *in the same direction* on the losses side (so risk-aversion would increase).

In order to see why this is a relevant prediction, let us contrast it with an equivalent prediction for a different model of $v(x)$. A simple decreasing sensitivity model of $v(x)$ can be built by choosing a increasing concave function $g(x)$, with $g(0) = 0$ and a penalty factor

$\lambda > 1$. The model is

$$\hat{v}(x) = \begin{cases} g(x) & \text{If } x \geq 0 \\ -\lambda\, g(-x) & \text{If } x < 0 \end{cases}. \tag{3.51}$$

This model has the shape of $v(\cdot)$: concave over gains, convex over losses, and the losses half has steeper slopes (by λ) than the gains part. It also makes predictions regarding the effects of curvature changes: if $\hat{v}(x)$ becomes more concave over gains (i.e., risk-aversion over gains increases), then $\hat{v}(x)$ becomes more *convex* over losses (i.e., risk-aversion over losses *decreases*). The functional form in (3.51) makes contrasting predictions to those of the categorical reasoning model.

There is also an interaction between the effect of the concavity of $u(\cdot)$ the choice being in the losses or gains domain. We will use θ to denote the concavity of $u(\cdot)$; since this is a ill-defined concept in general, we assume a CARA form for $u(\cdot)$; this form is not used at all, but it makes the θ have a meaning: the risk-aversion coefficient. Pick two payoffs x, y with $y > x >> 0$. Define $L(y) = (y, x/y; 0, 1 - x/y)$ and $L(-y) = (-y, x/y; 0, 1 - x/y)$. Denote the probability of a choice of x from the set $\mathcal{O}$ given a concavity θ by $N_x(\mathcal{O}; \theta)$. Then the model implies

$$N_x(\{x, L(y)\}; \theta) > N_{-x}(\{-x, L(-y)\}; \theta).$$

But the model also implies an interaction with the concavity:

$$\frac{\partial}{\partial\, \theta} \left(N_x(\{x, L(y)\}; \theta) - N_{-x}(\{-x, L(-y)\}; \theta)\right) > 0.$$

This happens since increasing θ leads to the k_i moving to the right (except possibly near zero, hence the $x >> 0$ above), proportionally decreasing the sensitivity on the losses side away from zero more than on the gains side.

3.6.6 Probability weighting function or standardization?

Both this model and prospect theory predict that it is possible to have

$$1 \succ (2, 1/2; 0, 1/2)$$

and

$$(10^6, 10^{-6}; 0, 1 - 10^{-6}) \succ 1.$$

This model predicts this through one of two mechanisms: standardization or the probability weighting function. Standardization states that 1 in the presence of $\{0,2\}$ is a different number than in the presence of $\{0,10^6\}$, since the encoded number is $1/\sigma(\{0,2\})$ or $1/\sigma(\{0,10^6\})$ and monotonicity of σ implies $\sigma(\{0,2\}) << \sigma(\{0,10^6\})$.

Prospect theory states that $v(1)$ is independent of the payoffs in the other prospect, so the effect comes from the probability weighting function. So, in cases where the standardization dominates the effect of the probability weighting function, this model and prospect theory make different predictions

3.6.7 Interaction between ρ and time in myopic decisions

As the discount rate, ρ, increases, the optimal categories for time become narrower (towards zero). Denoting the cutoffs as a for a given ρ by k_i^ρ,

$$\rho_1 > \rho_2 \quad \Rightarrow \quad k_i^{\rho_1} < k_i^{\rho_2}. \tag{3.52}$$

This creates an interaction of ρ and time differences in decisions. Consider the following preferences, where $e(y) > e(x)$ is assumed:

$$(x \otimes 0) \succ (y \otimes 1).$$

Consider two time shifts, t_1 and t_2, with $0 << t_1 << t_2$. Given the optimal categorization of time, with an Exponential distribution, for any ρ,

$$\Pr(e(t_2+1) = e(t_2)) > \Pr(e(t_1+1) = e(t_1)).$$

Plus, equation (3.52) implies that this difference is increasing in ρ. Now, if we denote the probability of choosing payoff y off a set of choices $\mathcal{O}$ given a discount rate ρ by $N_y(\mathcal{O};\rho)$, we have

$$\frac{\partial}{\partial \rho}\left(N_y(\{(x\otimes t_2),(y\otimes t_2+1)\};\rho) - \right.$$
$$\left. N_y(\{(x\otimes t_1),(y\otimes t_1+1)\};\rho)\right) > 0.$$

So with increasing ρ we should see not only an increase in reversal of preferences, but also an interaction with t: the further away t is, the highest the increase in preference reversals due to a given increase in ρ will be, or

$$\frac{\partial^2}{\partial \rho\, \partial t} N_y(\{(x\otimes t),(y\otimes t+1)\};\rho) > 0.$$

3.6.8 Time reflection

Decision making has no place for the past. But, were the decision-maker to evaluate past investment events, she should have increasing sensitivity towards the past (since events in the past have more and more value accrues to them the further they are). Assuming that the domain of time is now the real line, for our model, the distribution becomes a Normal, and similar arguments as for the utility apply (categorization with a decreasing convex function leads to the same type of loss functions as categorizing with a increasing concave function - losses per unit decrease with increasing x or t). Therefore, this model predicts decreasing sensitivity over the past.

Exponential discounting predicts

$$(100 \otimes -100) \succ (110 \otimes -110) \Rightarrow (10 \otimes 0) \succ (20 \otimes -10).$$

Hyperbolic discounting leads to negative discount numbers for sore regions of the negative numbers, so it is inadequate. (If we assume that the times are symmetric, then even the exponential will lead to strange results.)

The above model predicts combinations such as

$$(110 \otimes -110) \succ (100 \otimes -100) \text{ and } (10 \otimes 0) \succ (20 \otimes -10) \tag{3.53}$$

are possible.

A change to hyperbolic discounting, where the effects of time are reflected, under the constraint that $D(0) = 1$, is

$$D(t) = \frac{1-2t}{1-t} \quad t < 0. \tag{3.54}$$

This form would explain reversal of preferences, but there is no "law of effect" support for it. The derivation of the categories is based on the exponential discounting and is consistent with normative theory. Also, if we assume that we can reverse time, there is an exponential equivalent to equation (3.54),

$$\delta(t) = 2 - \exp(-t)$$

that also explains the result in equation (3.53). So, reversal of time, per se, not hyperbolic discounting, is sufficient for the effect in (3.53).

If instead of evaluation of past investment, we consider the evaluation of past utility from the standpoint of now, the model changes considerably. Although there is no standard way to evaluate utility in the past, one simple model is that of decay: the consumption creates a stock of utility s which decays with time. Assuming simple dynamics $\dot{s} = -d\,s$, where the parameter d is a decay rate, the current stock of utility of a payoff x at time $-t$ is $u(x)\ \exp(-d\,t)$. This explains phenomena like stated preference for recent events as opposed to past events. But it also makes specific predictions regarding preferences. In particular, it $(x \otimes t_1) \succ (y \otimes t_2)$, then $(x \otimes t_1 + \tau) \succ (y \otimes t_2 + \tau)$. With categorization of time, this no longer happens. The results towards the past are similar to those of hyperbolic discounting of the future, even if the categories are derived taking into account the exponential depreciation of utility stocks.

3.7 Conclusions and contributions to the literature

This paper extends our knowledge of decision-making in two different directions. First, it provides one model that explains prospect theory and hyperbolic discounting starting with rationality assumptions. Second, using that model, the paper makes predictions that are beyond the scope of prospect theory and of hyperbolic discounting.

This paper can be seen as expanding Rubinstein (1988) model of similarity. In that model, a symmetric similarity relation is assumed as a primitive axiom and, from it, the Allais Paradox is explained. The categorization developed in section 3.4 makes the similarity relationship endogenous to decision-making. The categorization does not allow for irreflexive similarity (Tversky, 1977), however, unless stimuli are used to infer both μ and σ for standardization (the China and Korea example in Tversky's paper could be explained by having similarity over size take the second element of the relation as the source of information for both μ and σ). Since the theory of standardization is not developed in this paper, effects as irreflexive similarity are left out of the model. (The only rule for standardization in the paper is that increasing empirical variance should lead to increased inferred variance.)

Appendix A

Proofs and computations for Distortions from Encoding

(To save space, I have omitted unenlightening intermediate steps from these proofs and computations.)

A.1 Illustration of the complexity of optimal encoding and decoding

I will illustrate the effects of a change in mean of the underlying distribution on the design of a encoder/decoder pair. I will also show the effects of changing concavity of u in encoder/decoder pair design.

To keep the example simple, I will assume a Uniform distribution with unit support for the payoffs, and a two-label encoder. I use θ to represent the parameter varying; for the first illustration it will be the lower bound of the support and for the second illustration it will be the concavity parameter[1]. The encoder is completely defined by a cutoff point $k(\theta)$. The decoder is defined by two levels, $\tilde{u}(\ell = 0; \theta)$ for numbers smaller than $k(\theta)$ and

[1] Note that standardization of x leads to different concavities around the number to be encoded $(x - \mu)/\sigma$, hence the importance of varying concavity. A effect of concavity worth noting is that optimal encoding will create bins that are smaller on regions of $\mathbb{R}$ where slopes are steeper and larger in regions of $\mathbb{R}$ where slopes are flatter; this leads to optimal encoders where bins are not equiprobable. The variation in concavity will be one of the main reasons the generic encoding (as opposed to optimal) will have equiprobable bins: with varying concavity, for which should the bins be designed, and how reusable would the encoding be for other domains?

$\tilde{u}(\ell = 1; \theta)$ for numbers bigger than $k(\theta)$.

Table A.1 shows the effect of varying the lower bound $\underline{x}$ of the distribution (so $x \sim \mathcal{U}([\underline{x}, \underline{x} + 1])$). The utility function for this case is $u(x) = \sqrt{x}$, conveniently concave over all $\mathbb{R}^+$. In this case, the problem is to find k, $\tilde{u}_0$ and $\tilde{u}_1$ defined by

$$\min_{k \in (\underline{x}, \underline{x}+1)} \int_{\underline{x}}^{k} (\sqrt{x} - \tilde{u}_0)^2 dx + \int_{k}^{\underline{x}+1} (\sqrt{x} - \tilde{u}_1)^2 dx \tag{A.1}$$

where

$$\begin{aligned} \tilde{u}_0 &= \frac{1}{k - \underline{x}} \int_{\underline{x}}^{k} \sqrt{x} dx \\ \tilde{u}_1 &= \frac{1}{\underline{x} + 1 - k} \int_{k}^{\underline{x}+1} \sqrt{x} dx \end{aligned} \tag{A.2}$$

which, using simple but tedious calculus (substituting (A.2) in (A.1), expanding and simplifying, differentiating the result, and solving for zeros in the $[\underline{x}, \underline{x} + 1]$ interval), yields the results in table A.1.

Table A.2 shows the effect of varying the concavity of the utility function. In this case, the support of x is $[0, 1]$, and the utility function is $u(x) = x - \eta x^2/2$. Since $d^2u/dx^2 = -\eta$, η is a concavity parameter. Here, the problem is to find k, $\tilde{u}_0$ and $\tilde{u}_1$ defined by

$$\min_{k \in (0,1)} \int_{0}^{k} (x - \eta x^2/2 - \tilde{u}_0)^2 dx + \int_{k}^{1} (x - \eta x^2/2 - \tilde{u}_1)^2 dx \tag{A.3}$$

where

$$\begin{aligned} \tilde{u}_0 &= \frac{1}{k} \int_{0}^{k} x - \eta x^2/2 dx \\ \tilde{u}_1 &= \frac{1}{1 - k} \int_{k}^{1} x - \eta x^2/2 dx \end{aligned} \tag{A.4}$$

which, again using simple but tedious calculus on equations (A.3) and (A.4), yields the results in table A.2.

From table A.1 and table A.2, two observations indicate the need for a specific setup for each combination $(F_X(x), u(x))$:

- The change in cutoff point is informationally complex: the position of $k(\theta)$ relative to $\underline{x}$ changes with $\underline{x}$ and η in a way that depends on the marginal utility, and cannot be solved with a simple translation or projection parameterized on $\underline{x}$ or η.

Table A.1: *Demonstration of the effect of varying support of $F_X(x)$ on encoder and decoder design.*

Origin ($\underline{x}$)	Cutoff ($k^*(\underline{x})$)	$\tilde{u}(\ell = 0\|\underline{x})$	$\tilde{u}(\ell = 1\|\underline{x})$
0	0.3820	0.4120	0.8248
1/2	0.9561	0.8497	1.1060
1	1.4716	1.1100	1.3162
3/2	1.9789	1.3178	1.4963
2	2.4832	1.4965	1.6552
5/2	2.9860	1.6557	1.8003
3	3.4881	1.8007	1.9346
7/2	3.9896	1.9348	2.0600
4	4.4907	2.0601	2.1781

- The values of $\tilde{u}(\ell|\theta), \ell = 0, 1$, also change in a non-translational or projective way.

In short, what this means is that change of just the support (not even the shape) of the distribution requires recomputing cutoff points and valuations of labels from first principles every time, by solving equations (1.4) and (1.5). Same with changes to the utility function.

A.2 Proof of main result and supporting observations

Part 1

Claim: For a distribution $\mathcal{D}$, the optimal encoder has $\Pr(x \in C|x \sim \mathcal{D}) = 1/N$ for all $C \in \mathcal{C}$. *Proof*: Since the encoding losses depend only on of $|x - e(x)|$, but not on x, given that $|x - e(x)|$ is accounted for (in other words, that $|x - e(x)|$ is a sufficient statistic for x as far as encoding is concerned), there is a result in information theory, the Shannon-McMillan-Breiman theorem (Shannon, 1948; MacMillan, 1953; Breiman, 1957), which states that in this case optimal maximum entropy encoding for $\mathcal{D}$ will have $\Pr(x \in C_i) = \Pr(x \in C_j)$ for all $C_i, C_j \in \mathcal{C}$. This implies $\Pr(x \in C) = 1/N$. (*End of proof.*)

Claim: Categories are intervals. *Proof*: Losses from encoding are

$$\int_{-\infty}^{+\infty} ||x - \tilde{x}(e(x))||dF_X(x)$$

Table A.2: *Demonstration of the effect of varying concavity of $u(x)$ on encoder and decoder design.*

Concavity (η)	Cutoff ($k^*(\eta)$)	$\tilde{u}(\ell = 0\|\eta)$	$\tilde{u}(\ell = 1\|\eta)$
0	0.5000	0.2500	0.7500
1/8	0.4889	0.2395	0.7085
1/4	0.4763	0.2287	0.6672
3/8	0.4620	0.2177	0.6263
1/2	0.4458	0.2063	0.5858
5/8	0.4274	0.1947	0.5460
3/4	0.4069	0.1828	0.5069
7/8	0.3843	0.1706	0.4687
1	0.3596	0.1583	0.4317

where $\tilde{x}(\ell) \doteq E[x|x \in C_\ell]$. For one category, and since $F_X(x)$ is continuous almost everywhere, these losses can be written as

$$\sum_{c=1}^{M} \int_{x \in A_c} ||x - \tilde{x}|| dF_X(x) \tag{A.5}$$

where M can be infinite and the A_c are intervals. Since $|| \cdot ||$ is increasing in its argument, and $F_X(x)$ is smooth except possibly on a set of measure zero, minimizing (A.5) under the constraint that

$$\Pr(x \in C) \doteq \sum_{c=1}^{M} \int_{x \in A_c} dF_X(x) = 1/N$$

implies minimizing the distances between the supports of the sets A_c. The minimum distance is attained when all A_c collapse into one contiguous interval C_ℓ. (*End of proof.*)

Claim: The optimal encoder for a distribution $\mathcal{D} \in \mathbb{D}$, has $k_\ell - k_{\ell-1}$ decreasing in ℓ for $\ell < 0$, and increasing in ℓ for $\ell > 0$. *Proof*: From the Shannon constraint and the previous point we know $\Pr(x \in C) \doteq \int_{k_{\ell-1}}^{k_\ell} dF_X(x) = 1/N$. Also, from assumption 3, point 1, $f_X(x)$ is symmetric and decreasing in $|x|$. Therefore, taking any $k_\ell, k_{\ell+1}, k_{\ell+2} > 0$, we have

$$\int_{k_\ell}^{k_{\ell+1}} f_X(x)dx = \int_{k_{\ell+2}}^{k_{\ell+2}} f_X(x)dx$$

but since all $f_X(x)$ in the second integral are strictly smaller than those in the first integral except possibly for a set of measure zero, the only way to achieve equality is to integrate over a larger support. Hence $k_{\ell+2} - k_{\ell+1} > k_{\ell+1} - k_\ell$.

Using the same expression, but with $k_\ell, k_{\ell+1}, k_{\ell+2} < 0$, we have each of the $f_X(x)$ in the second integral strictly larger than any of the $f_X(x)$ on the first integral (again, almost everywhere); hence in this case $k_{\ell+2} - k_{\ell+1} < k_{\ell+1} - k_\ell$. (*End of proof.*)

At this point it is proved that for each $\mathcal{D} \in \mathbb{D}$ the optimal encoder, parameterized on $\mathcal{C}^*(\mathcal{D})$ (note the dependency on the distribution) exhibits support($\mathcal{C}^*(\mathcal{D})$) increasing with distance of C to zero.

Claim: a generic encoder can be found in the set of optimal encoders for distributions in $\mathbb{D}$. *Proof*: For each pair of distributions define a loss function

$$L(\mathcal{D}_i, \mathcal{D}_j) \doteq \int_{x \sim \mathcal{D}_i} [||u(x) - \tilde{u}(e_{\mathcal{C}^*(\mathcal{D}_j)}(x))||] dF_{X_i}(x).$$

where $\tilde{u}$ is the optimal decoder for u and $\mathcal{D}_j$, as defined in equation 1.5.

For each distribution in $\mathbb{D}$ we can compute

$$EL \doteq \int_{\mathcal{D}_i \in \mathbb{D}} L(\mathcal{D}_i, \mathcal{D}_j) dF_{\mathcal{D}}(\mathcal{D}_i) \tag{A.6}$$

and from assumption 1, we know EL is bounded[2]. Therefore, select $\mathcal{D}^* \in \mathbb{D}$ which minimizes equation (A.6) as the generic encoder and use EL as a lower bound for S. If the minimum is not attained, then select any $\mathcal{D}$ for which EL is in a neighborhood η of the minimum, use $EL + \eta$ as the bound for S, and $e_{\mathcal{C}^*(\mathcal{D})}$ as the generic encoder. (*End of proof.*)

Hence, the generic encoder has support($\mathcal{C}^*(\mathcal{D})$) increasing with distance of $\mathcal{C}$ to zero. That proves part 1 of the result.

Part 2

Define $\tilde{v}(x)$ as the following piecewise linear function:

$$\tilde{v}(x) = \begin{cases} \tilde{u}(-L) & \text{If } x \le 2k_{-L+1} - k_{-L+2} \\ (x - 2k_{-L+1} + k_{-L+2}) \frac{\tilde{u}(-L+2) - \tilde{u}(-L+1)}{2(k_{-L+2} - k_{-L+1})} + \tilde{u}(-L) & \text{If } 2k_{-L+1} - k_{-L+2} \ge x \ge k_{-L+1} \\ (x - k_\ell) \frac{\tilde{u}(\ell+1) - \tilde{u}(\ell-1)}{2(k_{\ell+1} - k_\ell)} + \frac{(\tilde{u}(\ell+1) - \tilde{u}(\ell))}{2} & \text{For } \ell \ne -L, L \text{ and } x \in C_\ell \\ (x - k_{L-1}) \frac{\tilde{u}(L) - \tilde{u}(L-1)}{2(k_{L-1} - k_{L-2})} + \frac{\tilde{u}(L-1) + \tilde{u}(L)}{2} & \text{If } k_{L-1} \ge x \ge 2k_{L-1} - k_{L-2} \\ \tilde{u}(L) & \text{If } x \ge 2k_L - k_{L-1} \end{cases}$$

Note: this is a linear interpolation of the discrete categories, as illustrated in figure 1.1.

[2] Note that this uses a bound on the kurtosis of distributions in $\mathbb{D}$ implied in assumption 1.

Claim: $|\tilde{u}(e(x)) - \tilde{v}(x)| <$ some ϵ. *Proof*: From the definition of $\tilde{v}$, it is clear that $\tilde{u}(x) - \tilde{v}(x) = 0$ for $x \leq 2k_{-L+1} - k_{-L+2}$ or $x \geq 2k_L - k_{L-1}$. For each category C_ℓ, the quantity $|\tilde{u}(e(x)) - \tilde{v}(x)|$ is maximized at an extreme point of C, since both $\tilde{u}$ and $\tilde{v}$ are linear functions. Hence, either $|\tilde{u}(e(x)) - \tilde{v}(x)| \leq (\tilde{u}(\ell+1) - \tilde{u}(\ell))/2$ or $|\tilde{u}(e(x)) - \tilde{v}(x)| \leq (\tilde{u}(\ell) - \tilde{u}(\ell-1))/2$; since $\tilde{u}$ has decreasing marginals, $(\tilde{u}(\ell) - \tilde{u}(\ell-1))/2$ is the binding value. (*End of proof.*)

Claim: $\lim_{N\to\infty} |\tilde{u}(e(x)) - \tilde{v}(x)| = 0$. *Proof*: We know that

$$|\tilde{u}(e(x)) - \tilde{v}(x)| \leq (\tilde{u}(\ell) - \tilde{u}(\ell-1))/2.$$

Since the range of $\tilde{u}$ is fixed, when $N \to \infty$, we know that $(\tilde{u}(\ell)-\tilde{u}(\ell-1)) \to 0$, except maybe for a finite set $l_1, l_2, \ldots$; otherwise the range would become infinity. However, condition (1.7) and the decreasing marginals constraint on $\tilde{u}$ preclude the existence of large differences in $\tilde{u}(\ell) - \tilde{u}(\ell-1)$ when $k_\ell - k_{\ell-1} \to 0$; therefore, if there existed $l_1, l_2 \ldots$, that condition would not hold. So,

$$\lim_{N\to\infty} |\tilde{u}(e(x)) - \tilde{v}(x)| = \lim_{N\to\infty} \max_{\ell}\{\tilde{u}(\ell) - \tilde{u}(\ell-1)\} = 0.$$

(*End of proof.*)

At this point, it is proved that $\tilde{u}(e(\cdot))$ can be made to approximate $\tilde{v}$ with any accuracy $\epsilon > 0$ desired. It remains to be proved that $\tilde{v}(\cdot)$ is s-shaped for all N, with losses looming larger than gains of equivalent magnitude.

Claim $\tilde{v}(z)$ is concave for $z > 0$. *Proof*: the slope of $\tilde{v}$ for category C_ℓ is $(\tilde{u}(\ell+1) - \tilde{u}(\ell-1))/2(k_{\ell+1} - k_\ell)$; the slope for category $C_{\ell+1}$ is $(\tilde{u}(\ell+2) - \tilde{u}(\ell))/2(k_{\ell+2} - k_{\ell+1})$. The difference of these slopes is

$$\frac{\tilde{u}(\ell+2) - \tilde{u}(\ell)}{2(k_{\ell+2} - k_{\ell+1})} - \frac{\tilde{u}(\ell+1) - \tilde{u}(\ell-1)}{2(k_{\ell+1} - k_\ell)} \tag{A.7}$$

which is negative since $\tilde{u}(\ell+2) - \tilde{u}(\ell) < \tilde{u}(\ell+1) - \tilde{u}(\ell-1)$ and $k_{\ell+2} - k_{\ell+1} > k_{\ell+1} - k_\ell$ for $\ell \geq e(0)$. This means slopes are decreasing with increasing category label ℓ and ℓ is weakly increasing in z. Therefore, slopes are decreasing in z, which means $\tilde{v}$ is weakly concave.(*End of proof.*)

Claim $\tilde{v}(z)$ is convex for $z < 0$. *Proof*: Looking again at the difference in slopes in consecutive categories, equation (A.7), we know that $k_{\ell+2} - k_{\ell+1} < k_{\ell+1} - k_\ell$ for $\ell \leq e(0)$; therefore, if marginals were constant, $\tilde{v}$ would be convex (would have increasing slopes) for $z < 0$. Since marginals are decreasing, it is necessary to limit their rate of decrease in order to obtain the convexity. In order for the quantity in (A.7) to be positive, it is sufficient that

$$\frac{k_{\ell+2}-k_{\ell+1}}{k_{\ell+1}-k_{\ell}} \leq \frac{\tilde{u}(\ell+2)-\tilde{u}(\ell)}{\tilde{u}(\ell+1)-\tilde{u}(\ell-1)}$$

which is equivalent to condition (1.7). (*End of proof.*)

Claim For $z > 0$, $\tilde{v}(z) - \tilde{v}(0) \leq |\tilde{v}(-z) - \tilde{v}(0)|$. *Proof*: Define $\Delta_{C_\ell} = \max_{z \in C_\ell}\{\tilde{v}(z)\} - \min_{z \in C_\ell}\{\tilde{v}(z)\}$. Note that since $\tilde{u}$ has decreasing marginals and $\tilde{v}$ is a linear interpolation thereof, $\Delta_{C_\ell} \leq \Delta_{C_{-\ell}}$ for all $\ell > 0$. I will prove this result separately for N odd and N even. (For N even.) If $e(z) = l$ then $e(-z) = -l$, from symmetry of $e(\cdot)$, and

$$\begin{aligned} \tilde{v}(z) - \tilde{v}(0) &= \sum_{i=1}^{l-1} \Delta_{C_i} + b(z)(z - k_l) \\ |\tilde{v}(-z) - \tilde{v}(0)| &= \sum_{i=1}^{l-1} \Delta_{C_{-i}} + b(z)|z - k_{-l}| \end{aligned}$$

where the last slope is defined by

$$b(z) = \begin{cases} 0 & \text{If } z \leq 2k_{-L+1} - k_{-L+2} \\ \frac{\tilde{u}(-L+2)-\tilde{u}(-L+1)}{2(k_{-L+2}-k_{-L+1})} & \text{If } 2k_{-L+1} - k_{-L+2} \geq z \geq k_{-L+1} \\ \frac{\tilde{u}(\ell+1)-\tilde{u}(\ell-1)}{2(k_{\ell+1}-k_{\ell})} & \text{For } \ell \neq -L, L \text{ and } z \in C_\ell \\ \frac{\tilde{u}(L)-\tilde{u}(L-1)}{2(k_{L-1}-k_{L-2})} & \text{If } k_{L-1} \geq z \geq 2k_{L-1} - k_{L-2} \\ 0 & \text{If } z \geq 2k_L - k_{L-1} \end{cases}$$

Each quantity in the top equation has an equivalent one in the bottom equation, but the magnitudes are higher in the bottom equation for b, and all Δs.

(For N odd.) Define $\Delta_0 = \tilde{v}(0) - min_{z \in C_0}\{\tilde{v}(z)\}$ and note that Δ_0 is also equal to $|\tilde{v}(0) - max_{z \in C_0}\{\tilde{v}(z)\}|$, since the symmetry of $e(\cdot)$ requires 0 to be the center of C_0. Then, if $e(z) \neq 0$,

$$\begin{aligned} \tilde{v}(z) - \tilde{v}(0) &= \Delta_0 + \sum_{i=1}^{l-1} \Delta_{C_i} + b(z)(z - k_l) \\ |\tilde{v}(-z) - \tilde{v}(0)| &= \Delta_0 + \sum_{i=1}^{l-1} \Delta_{C_{-i}} + b(z)|z - k_{-l}| \end{aligned}$$

Each quantity in the top equation has an equivalent one in the bottom equation, but the magnitudes are higher in the bottom equation for b, and all Δs, except for Δ_0. If $e(z) = 0$, then $\tilde{v}(z) - \tilde{v}(0) = |\tilde{v}(z) - \tilde{v}(0)|$. (*End of proof.*)

At this point it is proved that $\tilde{v}$ is weakly concave for $z > 0$, weakly convex for $x < 0$ and losses loom larger than gains in $\tilde{v}$. This completes the proof of the main result. QED.

A.3 Importance of *not too concave* condition

Condition (1.7) is a necessary condition for the s-shape; however, it is mostly a technical condition, not one that I would expect to have significant effects. To illustrate this, consider the following evaluation function:

$$\tilde{u}(\ell) \doteq \ell - \frac{\eta}{2}(\ell - 1)\ell$$

where the labels are rescaled to $0 \ldots N - 1$. This evaluation has constant second-order difference $-\eta$. For this functional form, the *not too concave* condition is equivalent to

$$\frac{2 - \eta(\ell^2 + 3\ell + 3)}{2 - \eta(\ell^2 + \ell + 1)} \geq \frac{k_{\ell+3} - k_{\ell+1}}{k_{\ell+2} - k_\ell}.$$

Note that $\tilde{u}$ has to be increasing, so that implies $\eta < 1/(N - 1)$, from the definition above. Putting these constraints together (and checking all ℓ for each N) yields the following results for a Normal distribution:

N	Increasing $\tilde{u}$	Not too concave $\tilde{u}$
5	$\eta < 1/4$	$\eta < 0.132$
6	$\eta < 1/5$	$\eta < 0.152$
7	$\eta < 1/6$	$\eta < 0.212$
8	$\eta < 1/7$	$\eta < 0.228$

So, at least for constant second-order difference $\tilde{u}$, the not too concave condition is mostly a technical condition required for the reconstruction with small number of labels but not important for large N.

A.4 Comparative statics of N

Up to now, I have been using N as exogenous. In fact, since expected losses are a function of N, one would expect it to be a function of the decision's importance. The other important fact is the cost of number of labels (which is the second part of the cost of thinking in this paper — the first part being the setup cost).

To determine the optimal number of labels for a decision, it is necessary to have a measure of the cost of a number of labels and a measure of the value of a number of labels.

The value of a number of labels, $m(N, \bar{u})$ depends on the losses given N and the overall value of the decision, $\bar{u}$. For any given $\bar{u}$, losses are decreasing in N, but the decreases are ever smaller (a consequence of optimal use of encoding capacity). Also, increases in the overall value of the decision $\bar{u}$ lead to increasing value of precision, or increasing value of N. These conditions are summarized by

$$\frac{\partial m}{\partial N} > 0 \quad , \quad \frac{\partial^2 m}{\partial N^2} < 0 \quad , \quad \frac{\partial m}{\partial \bar{u}} > 0 \quad , \quad \frac{\partial^2 m}{\partial N \partial \bar{u}} > 0. \tag{A.8}$$

Cost of a number of labels depends on the number of labels and the cost of thinking. This cost of thinking is the cost of using a information processing unit. It may change, for instance, with the time available, or with the physical conditions of the decision-maker. (For more on cost of thinking see Shugan (1980).)

So, what is meant by an increase in the cost of thinking? I assume that the cost of using each information processing unit goes up, such that each additional capacity requirement goes up even faster: since recruiting of units to process information should be done selecting the cheapest first, increases in marginal cost should themselves be non-decreasing.

So, $c = c(N, \underline{c})$ is in fact a function of N and a parameter $\underline{c}$, the unit cost of processing to which I just referred. The preceding paragraph can be summarized by

$$\frac{\partial c}{\partial N} > 0 \quad , \quad \frac{\partial^2 c}{\partial N^2} \geq 0 \quad , \quad \frac{\partial c}{\partial \underline{c}} > 0 \quad , \quad \frac{\partial^2 c}{\partial N \partial \underline{c}} > 0 \tag{A.9}$$

These lead to the following comparative statics:

Observation 7 (Comparative statics of N^*). *For $c(N, \underline{c})$ and $m(N, \bar{u})$ as defined by equations (A.8) and(A.9), the comparative statics of the optimal number of labels N^* are*

$$\frac{\partial N^*}{\partial \underline{c}} < 0 \quad , \quad \frac{\partial N^*}{\partial \bar{u}} > 0$$

So, increases in the unit cost of thinking lead to fewer labels, increases in decision utility lead to more labels. These are fairly obvious, but it is reassuring to see that they result from the formal model.

Proof: (I will use N instead of N^*, to reduce clutter.) The required comparative statics for cost are:

$$\begin{aligned}
\frac{\partial N}{\partial \underline{c}} &= -\left[\frac{\partial}{\partial \underline{c}}\left(\frac{\partial c}{\partial N} - \frac{\partial m}{\partial N}\right)\right] / \left[\frac{\partial}{\partial N}\left(\frac{\partial c}{\partial N} - \frac{\partial m}{\partial N}\right)\right] = \\
&= -\left[\frac{\partial^2 c}{\partial \underline{c} \partial N} - \frac{\partial^2 m}{\partial \underline{c} \partial N}\right] / \left[\frac{\partial^2 c}{\partial N^2} - \frac{\partial^2 m}{\partial N^2}\right] = \\
&= -\frac{(> 0) - (0)}{(\geq 0) - (< 0)} < 0
\end{aligned}$$

using the envelope and implicit function theorems on the solution of the optimality condition, $c'(N) - m'(N) = 0$, and the independence of $m(\cdot)$ and $\underline{c}$ (there is no reason why the value of a set of labels changes with the unit cost of representing these labels).

Comparative statics with respect to overall utility of a decision are

$$\frac{\partial N}{\partial \bar{u}} = -\left[\frac{\partial^2 c}{\partial \bar{u} \partial N} - \frac{\partial^2 m}{\partial \bar{u} \partial N}\right] / \left[\frac{\partial^2 c}{\partial N^2} - \frac{\partial^2 m}{\partial N^2}\right] = -\frac{(0) - (> 0)}{(> 0) - (< 0)} > 0$$

QED.

Appendix B

Model used for empirical testing

The information processing processing model has three components: a standardization part, an encoder, and a decoder. The standardization is necessary to make the encoding generic: several different payoffs, in different decisions, may use the same encoder, thus limiting the amount of effort in set-up. The encoder translates a standardized payoff into an internal representation. It essentially creates a partition of the payoff space into some categories; inside these categories there is loss of precision, in that the decision-maker does not separate between any two elements in the same category. Finally, the decoder attaches meaning to the internal representation. In our case the meaning will be a concave utility function, defined over the categories.

The first element is standardization. Mathematically this is a trivial process, where a payoff x is mapped into a standardized payoff $z = (x - \mu_x)/\sigma_x$ where μ_x is the mean of the distribution from which x is drawn, and σ_x^2 its variance. The interpretation of this process in psychological terms is akin to norm theory (Kahneman and Miller, 1986), the idea that the type of stimulus brings with it norms on its dimensionality. For instance, when evaluating a stock fluctuation, a \$1 change may be important, but when evaluating the Dow-Jones Industrial Average, a \$1 fluctuation is not relevant. Another parallel can be established with mental accounting (Thaler, 1985): dimensions of payoffs are relative to a reference point (in this case the normalization is based on inference of the μ_x as the reference point) and are measured in terms of the shadow cost of the account (in income per util), which is roughly equivalent to σ_x in this formulation. Note that the shadow cost of the account in Thaler (1985) is a budgeting device but in this parallel corresponds to a normalization of the scale; this is just an instance of the dual nature of constraints.

In the context of this paper, standardization is not an interesting part of the model; therefore, for the remainder of this appendix, we will simplify notation and assume $\mathrm{E}[x] = 0$ and $\mathrm{Var}[x] = 1$.

Given the standardized payoffs, the second task is to encode them into some labels. An encoder takes a payoff, x, and maps it into a label $\ell = e(x)$; the label corresponds to a category C_ℓ, and the encoding function is simple: if $x \in C_\ell$, then $e(x) = \ell$. Different sets of C_ℓ create different encoding schemes. We therefore need a criterion for evaluating encoding schemes $e(\cdot)$. We choose to minimize the loss of information incurred by the encoder, measured by the Kullback-Leibler distance between the input variables x and the best possible reconstruction of the inputs given the encoding, $\tilde{x}(e(x))$. This means that we want a decoder that destroys the minimal amount of information — as measured by how much could be recovered with the optimal recovery mechanism[1]. The formulation of the criterion requires two things. Firstly, a distribution for the input variables, that we will leave open for the moment and denote its p.d.f. by $f(x)$. Secondly, the best possible reconstruction of a payoff, $\tilde{x}$, given that the corresponding encoding label is $\ell = e(x)$, will be given by $\tilde{x} = \mathrm{E}[x|e(x) = \ell]$, and the distribution of $\tilde{x}$ will be denoted (again in p.d.f. form) by $\tilde{f}(\tilde{x})$. With these two elements, we can now state the problem of finding $e(\cdot)$ as the minimization of the following Kullback-Leibler distance:

$$-\int_{\mathbb{R}} f(x) \log\left(\frac{\tilde{f}(\tilde{x})}{f(x)}\right) dx. \tag{B.1}$$

With the above definition for $\tilde{x}$, the solution to the problem in equation (B.1) is well-known in the information theory literature (Lloyd, 1982; Gersho and Gray, 1992). It called an equiprobable quantizer, and has the characteristic that all categories are equally likely. This can be understood by noting that if errors are equally costly, independently of their position (so that just the magnitude is relevant), each error should be ex-ante equally surprising. The output of a quantizer can either be a quantized value (the r_i above, which are also the optimal $\tilde{x}$), or the category labels, as we have been assuming. The category labels create a loss of information with respect to the r_i, because they convey only ordinal information, not cardinal. On the other hand, categorical reasoning is the main assumption behind this model, in that the decision maker has some form of discrete, category-based,

[1] We are not assuming that the information is reconstructed for purposes of evaluation; we are designing an encoder that would allow for the best reconstruction of the inputs *if that were the task at hand.* This is a measure of information loss for a general-purpose encoding, so this criterion is adequate, even though we are not going to reconstruct the information when the task is evaluation.

ordinal reasoning, and all the cardinal information is at the evaluation level (and that is just a convenient way to refer to preferences, not necessarily the mechanism that leads to the decisions[2]). We denote the label of category C_i by ℓ_i and will use interchangeably $e(x)$ and ℓ as arguments for the utility function.

To develop a more workable model, first we note that the labels themselves are arbitrary objects; therefore, we can choose (for reasons that will be clear in the next paragraph) labels to be intervals:

$$\begin{aligned} \ell_1 &= [0, 1/N); \\ \ell_2 &= [1/N, 2/N); \\ &\vdots \\ \ell_N &= [(N-1)/N, 1]. \end{aligned}$$

We now note that with equiprobable partitioning, each of these category labels, ℓ_i, corresponds to the values of the c.d.f for the x in the category C_i. As the number of labels tends to infinity, $e(\cdot)$ approximates $F(x)$ ever more closely. In other words, it can be shown that, for any $\epsilon > 0$ there exists a N such that a partition C_i, such that an encoder using C_i, $e(x)$, with N labels is within ϵ of the c.d.f, or $|e(x) - F(x)| < \epsilon$. We will use this fact to make a convenient continuous approximation $e(x) = F(x)$; such approximation allows the manipulation of derivatives instead of differences and simplifies the algebra, without any significant effect in the results themselves.

So, at this point we need a probability distribution. A reasonable candidate in the Normal distribution: since the payoffs are standardized to have mean zero and variance one, it is more likely that, over a large set of distributions for the payoffs, realizations come near the mean than away from the mean. (The average payoff will be distributed normally, as a result of the central limit theorem; if we were to design the partition for that average payoff, the normal would be a required choice. However, the argument based on the information minimization is more compelling given the rest of the assumptions in the model.) A stronger formal argument can be made by noting what criteria the choice of distribution for a generic encoder should obey. One such criterion is that the distribution should contain the minimum number of assumptions; to measure the number of assumptions

[2] As in the microeconomics literature, where only preferences are assumed to be knowable, and these are purely ordinal, but a cardinal utility function simplifies the argument; the parallel is actually more precise than that, since both $u(\cdot)$ and $\tilde{u}(\cdot)$ are only relevant up to a affine transformation, and share other characteristics (cannot be compared numerically across subjects, for instance).

in a distribution $f(x)$, one uses the entropy $\mathcal{H}(f)$, defined as

$$\mathcal{H}(f) = -\int_{\mathbb{R}} f(x)\log(f(x))dx.$$

This functional measures the information gain from observing a realization of the variable described by the density $f(x)$[3]. Therefore, maximizing $\mathcal{H}(f)$ yields the distribution with the least information in the assumptions generating its functional form. We need to impose the conditions that the function is a probability distribution ($f(x) > 0, \forall_x$, and $\int f(x)dx = 1$), and the standardization ($\int xf(x)dx = 0$ and $\int x^2 f(x)dx = 1$). Under these constraints, the solution is a standard normal, with $f(x) = (2\pi)^{-1/2}\exp(-x^2/2)$.

Putting the Normal distribution together with the continuous approximation $e(x) = F(x)$, we have the first part of the main result: $e(x)$ is an increasing, s-shaped function with inflection point at zero. We now focus on the decoding, $\tilde{u}(e)$.

Using a continuous set of labels implies a continuous $\tilde{u}(e)$. Since we are assuming as a primitive that this is a concave function, the composition $\tilde{u}(e(x))$ will have loss aversion. To illustrate that, note that if $\tilde{u}$ were linear, $\tilde{u}(e(x))$ would be symmetric; with a small degree of concavity, each $\tilde{u}(e(x))$ is shifted down, but the shift increases in the direction of negative numbers; therefore, for a $x > 0$, $e(x) - \tilde{u}(e(x))$ is a smaller difference than $e(-x) - \tilde{u}(e(-x))$. Since $e(x) - e(0) = |e(-x) - e(0)|$, having $e(x) - \tilde{u}(e(x)) < e(-x) - \tilde{u}(e(-x))$ implies $\tilde{u}(e(x)) - \tilde{u}(e(0)) < |\tilde{u}(e(-x)) - \tilde{u}(e(0))|$, loss aversion. The question then is, does this function $\tilde{u}(e(x))$ have a s-shape? The answer depends on the degree of concavity of $\tilde{u}$. For very concave $\tilde{u}$, the overall result will not be s-shaped. A condition that guarantees the s-shape is

$$-\frac{\tilde{u}''}{\tilde{u}'}(e(x)) < |x|, \quad \forall_{x<0}. \tag{B.2}$$

(*Proof*: The composition of increasing concave functions is increasing and concave; therefore it is only necessary to assure that $\tilde{u}(e(\cdot))$ is convex for $x < 0$. Since $\frac{\partial^2}{\partial x^2}\tilde{u}(e(x)) = \tilde{u}''e' + \tilde{u}'e''$, with $\tilde{u}''e' < 0$ and $\tilde{u}'e'' > 0$ for $x < 0$, and we want $\frac{\partial^2}{\partial x^2}\tilde{u}(e(x)) > 0$, the condition is that $-\tilde{u}''e' < \tilde{u}'e''$ or, substituting the Normal c.d.f. for $e(x)$, that $-\tilde{u}''/\tilde{u}' < |x|$. *End of proof.*)

[3] To understand this concept, note that if $X = k$, a constant, there would be no information to be gained from any realization of X: before the observation there would be a perfect prediction of its value. If the support of X is $\mathbb{R}$, but we know $X = 1$ or $X = 0$ with equal probability, there will be more information than in the case of a constant, but the limitation to the two values make for very simple predictions, hence very little information (we know, for instance, that the result will never be -3). For any bounded support, we can maximize the surprise (hence informativeness) of any observation by making them all equally likely. With unbounded support, it is necessary to solve the entropy functional subject to boundary constraints.

Condition (B.2) is a bound on the "internal" risk aversion, or risk aversion measured over the internal representation. This condition implies that for some cases of very high risk aversion the behavior will resemble more that of a normative decision-maker than in cases of low risk-aversion.

Appendix C

Encoding of payoffs with concave utility

Note: this proof draws heavily from well-known results in information theory (Max, 1960; Lloyd, 1982; Trushkin, 1982, 1993). The decreasing sensitivity result for the Normal distribution was derived originally by Max (1960) and also by Lloyd (1982). Some intermediate algebra has been omitted in the interest of space.

To solve the categorization problem in equation (3.16) we note that the problem of minimizing

$$\int \left(u(x) - u\left(r_{e(x)}\right)\right)^2 f_X(x)\, dx$$

can be transformed into a classic information theory problem (the linear case) with a change of variable to $u \doteq u(x)$, because by definition of r_i, $u(r_i)$ is a conditional mean for u. (We used the explicit $f_X(x)$ to denote the p.d.f. of x since we will be making changes of variables.) This implies a probability density function for u, $f_U(u) = f_X\left(u^{-1}(u)\right)\left(u'(u)\right)^{-1}$, which is well-defined since $u(x)$ is one-to-one and continuous. We will denote the cut-off points in u space by θ_i and the representative elements by $\bar{u}_i$. In this space, we want to solve the problem of minimizing

$$E_u\left[\left(u - \bar{u}_{e(u)}\right)^2\right]. \tag{C.1}$$

We denote the cut-off points for this problem by θ_i. It can be shown with simple derivation that the minimization of the squared error in equation (C.1) implies

$$\theta_i = \left(\bar{u}_{i-1} + \bar{u}_i\right)/2$$

called the *nearest neighbor condition.* The solution is then to find sets of θ_i and $\bar{u}_i$ that fit both the nearest neighbor condition and the expected utility condition, $u_i = E[u|\theta_i \leq u \leq \theta_{i+1}]$.

When the distribution of a given random variable Z is uniform, minimizing $E[(z - \bar{z})^2]$ is very simple: each representative value $\bar{z}_i$ is the middle point of the category bin, and the bins are all the same size (i.e. they are equiprobable). This is called *uniform quantizing.* The categories in uniform quantizing are equiprobable. We can use this problem if we find a way to make the problem distribution uniform.

If we make a new change of variable, to $F = F_U(u)$, have a problem where the variable is uniformly distributed, and the solution in F space is uniform quantizing: choose cutoff points η_i, in u space, such that the bins in F are equiprobable:

$$\int_{\eta_i}^{\eta_{i+1}} dF_U(u) = 1/N.$$

This implies that for regions of u where $F(u)$ is a convex function of u we have

$$\eta_{i+1} - \eta_i > \eta_{i+2} - \eta_{i+1},$$

and for regions of u where $F(u)$ is a concave function of u we have

$$\eta_{i+1} - \eta_i < \eta_{i+2} - \eta_{i+1}.$$

These two conditions are the solution for the problem of minimizing

$$E_F\left[(F - \bar{F})\right],$$

called the compandor-expandor problem of u, not the problem in equation (C.1). We now compare the first-order conditions for each problem. We know that

$$\frac{\partial}{\partial \eta_i} \int_{\eta_i}^{\eta_{i+1}} (F - \bar{F}_i)^2 dF(u) = \frac{\partial}{\partial \eta_{i+1}} \int_{\eta_i}^{\eta_{i+1}} (F - \bar{F}_i)^2 dF(u) = 0,$$

since the η_i are the solution to the minimization problem for F. Since $F(u)$ is a non-linear function, we also know that, for $\theta_i = \eta_i$ and $\theta_{i+1} = \eta_{i+1}$,

$$\frac{\partial}{\partial \theta_i} \int_{\theta_i}^{\theta_{i+1}} (u - \bar{u}_i)^2 dF(u) = -(\theta_i - \bar{u}_i)^2 f_U(\theta_i) < 0,$$

and

$$\frac{\partial}{\partial \theta_{i+1}} \int_{\theta_i}^{\theta_{i+1}} (u - \bar{u}_i)^2 dF(u) = (\theta_{i+1} - \bar{u}_i)^2 f_U(\theta_{i+1}) > 0.$$

We now consider the two separate cases of $u > u(0)$ and $u < u(0)$. For $u < u(0)$, $f_U(u)$ is increasing. If $f_U(u)$ were constant $\theta_{i+1} - \bar{u}_i = \theta_{i+1} - \bar{u}_i$; with $f_U(u)$ increasing, we have

$$(\theta_i - \bar{u}_i)^2 f_U(\theta_i) > (\theta_{i+1} - \bar{u}_i)^2 f_U(\theta_{i+1})$$

Therefore, an increase in θ_i lowers the minimization value more than an decrease in θ_{i+1}. This implies that for negative values, if $\eta_{i+1} - \eta_i > \eta_{i+2} - \eta_{i+1}$, this effect will be amplified:

$$\eta_{i+1} - \eta_i > \eta_{i+2} - \eta_{i+1} \Rightarrow \theta_{i+1} - \theta_i >> \theta_{i+2} - \theta_{i+1}.$$

For $u > u(0)$, we have

$$(\theta_i - \bar{u}_i)^2 f_U(\theta_i) < (\theta_{i+1} - \bar{u}_i)^2 f_U(\theta_{i+1}).$$

Therefore, an increase in θ_i lowers the minimization value less than an decrease in θ_{i+1}; hence, for positive values, if $\eta_{i+1} - \eta_i < \eta_{i+2} - \eta_{i+1}$, this effect will also be amplified, which in this case means

$$\eta_{i+1} - \eta_i < \eta_{i+2} - \eta_{i+1} \Rightarrow \theta_{i+1} - \theta_i << \theta_{i+2} - \theta_{i+1}.$$

Now we can derive sufficient conditions which create decreasing sensitivity over u: for $u < u(0)$ we need $F_U(u)$ convex and for $u > u(0)$ we need $F_U(u)$ concave. Solving for the fact that x is Normal, we have

$$\begin{cases} (u^{-1}(u))'' < u^{-1}(u)((u^{-1}(u))')^2 & u > u(0) \\ (u^{-1}(u))'' > u^{-1}(u)((u^{-1}(u))')^2 & u < u(0) \end{cases} \tag{C.2}$$

If we relax the assumption of the Normal, the conditions are

$$f'_X(u^{-1}(u))((u^{-1}(u))')^2 + f_X(u^{-1}(u))(u^{-1}(u))'' \quad \begin{cases} < 0 & u > u(0) \\ > 0 & u < u(0) \end{cases}$$

These conditions are very strong, as they would apply to any number of labels; but, by themselves are not sufficient. The result has to hold over x space, so it is necessary that u is not too concave in the sense that

$$\theta_{i+1} - \theta_i > \theta_{i+1} - \theta_{i+1} \Rightarrow u^{-1}(\theta_{i+1}) - u^{-1}(\theta_i) > u^{-1}(\theta_{i+1}) - u^{-1}(\theta_{i+1}). \tag{C.3}$$

Note that with a linear $u(x)$, the conditions on $f_X(x)$ are simply that it is unimodal and symmetric. (Again, this is a very strong condition, since it would have to apply to any number of labels.) This will create the required $f_U(u)$ convex over negative values and concave over positive values, hence the decreasing sensitivity.

Conditions (C.2) and (C.3) are sufficient for decreasing sensitivity. We now look into the asymmetry of categories. We have shown above two conditions must hold for the optimal partition. First, the nearest neighbor condition for the cut-off points in u

$$u(k_i) = (u(r_i) + u(r_{i+1}))/2$$

and, second, the centroid condition, also in u,

$$u(r_i) = E[u(x)|x \in C_i].$$

For linear $u(x) = ax$ and Normal x, the results are symmetric around 0, both for k_i and r_i. If u is concave, and for simplicity we choose $u(0) = 0$, the centroid condition will imply new $u(r_i)$ for the same cut-offs k_i. Since we are holding the k_i constant at their previous levels, the new $u(r_i)$ will be lower than the previous values, violating the nearest neighbor condition. In order to make that condition hold, we have to move the cut-off points either inward (reducing the centroids) or out (increasing the cut-offs).

For $x < 0$, $f_X(x)$ is increasing, and since $u(x)$ is concave (and $F(u^{-1}(x))$ is convex from condition (C.2) above), moving the cut-off points to the left will lead to a higher decrease in $E[x]$ than in $u(k_i)$. Therefore, if the cutoff points shift to the left, the only point that satisfies both the centroid and nearest neighbor conditions is $-\infty$. It is simple to show that this choice is not a minimum of (3.16). Hence, for $x < 0$ concavity of u leads to a shift of the cutoff points to the right of their symmetric locations. A corresponding effect takes place on the positive side of x, where the concavity of u coupled with the decreasing density of x lead to a shift of the cutoff points outward, also to the right.

References

Ainslie, G. (1975). "Specious reward: A behavioral theory of impulsiveness and impulse control." *Psychological Bulletin, 82*(4), 463-496.

Ainslie, G. (1991). "Derivation of 'rational' economic behavior from hyperbolic discount curves." *American Economic Review, 81*(2), 334-340.

Allais, M. (1953). "Le comportement de l'homme rationel devant le risque, critique des postulates et axiomes de l'ecole americaine." *Econometrica, 21*, 503-546.

Armstrong, S. L., Gleitman, L. R., and Gleitman, H. (1983). "What some concepts might not be." *Cognition, 13*, 263-308.

Azfar, O. (1999). "Rationalizing hyperbolic discounting." *Journal of Economic Behavior & Organization, 38*(2), 245-252.

Basu, K. (1984). "Fuzzy revealed preference theory." *Journal of Economic Theory, 32*, 212-227.

Beja, A., and Gilboa, I. (1992). "Numerical representations of imperfectly ordered preferences (a unified geometric exposition)." *Journal of Mathematical Psychology, 36*, 426-449.

Breiman, L. (1957). "The individual ergodic theorem of information theory." *Annals of Mathematical Statistics, 28*, 809-810.

Bridges, D. S. (1983). "A numerical representation of preferences with intransitive indifference." *Journal of Mathematical Economics, 11*, 25-42.

Camerer, C. (1995). "Individual decision making." In J. H. Kagel and A. Roth (Eds.), *Handbook of experimental economics.* Princeton, NJ: Princeton University Press.

Cover, T. A., and Thomas, J. A. (1991). *Elements of information theory.* New York: John Wiley and Sons.

Deacon, T. (1997). *The symbolic species.* New York: W. W. Norton.

Fennema, H., and Wakker, P. (1997). "Original and cumulative prospect theory: A discussion of empirical diferences." *Journal of Behavioral Decision Making*, *10*, 53-64.

Fischer, K., and Jungermann, H. (1996). "Rarely occurring headaches and rarely occurring blindness: is rarely = rarely? the meaning of verbal frequentistic labels in specific medical contexts." *Journal of Behavioral Decision Making*, *9*(3), 153-172.

Gersho, A., and Gray, R. M. (1992). *Vector quantization and signal compression.* Boston: Kluwer Academic Press.

Gilboa, I., and Lapson, R. (1995). "Aggregation of semiorders: intransitive indifference makes a difference." *Economic Theory*, *5*, 109-126.

Gingerenzer, G., and Goldstein, D. G. (1996). "Reasoning the fast and frugal way: models of bounded rationality." *Psychological Review*, *103*(4), 650-669.

González-Vallejo, C. C., Erev, I., and Wallsten, T. S. (1994). "Do decision quality and preference order depend on whether probabilities are verbal or numerical?" *American Journal of Psychology*, *107*(2), 157-172.

Gourville, J. T. (1998). "Pennies-a-day: The effect of temporal reframing on transaction evaluation." *Journal of Consumer Research*, *24*(4), 395-408.

Harless, D. W., and Camerer, C. F. (1994). "The predictive utility of generalized expected utility theories." *Econometrica*, *62*(6), 1251-1289.

Herrnstein, R., and Boring, E. (1965). *A source book in the history of psychology.* Harvard University Press.

Holyoak, K., and Alker, J. (1976). "Subjective magnitude information in semantic orderings." *Journal of Verbal Learning and Verbal Behavior*, *15*, 287-299.

Huttenlocher, J., and Hedges, L. V. (1994). "Combining graded categories: Membership and tipicality." *Psychological Review*, *101*(1), 157-165.

Kahneman, D., and Miller, D. (1986). "Norm theory: Comparing reality to its alternatives." *Psychological Review*, *93*(2), 136-153.

Kahneman, D., and Tversky, A. (1979). "Prospect theory: a theory of decision under uncertainty." *Econometrica*, *47*, 263 - 291.

Kahneman, D., and Tversky, A. (1996). "On the reality of conitive illusions." *Psychological Review*, *103*(3), 582-591.

Kaufman, B. E. (1999). "Emotional arousal as a source of bounded rationality." *Journal of Economic Behavior & Organization*, *38*(2), 135-144.

Kirby, K. N., and Marakovic, N. M. (1995). "Modeling myopic decisions: Evidence for hyperbolic delay-discounting within subjects and amounts." *Organizational Behavior and Human Decision Processes*, *64*(1), 22-30.

Krumhansl, C. L. (1978). "Concerning the applicability of geometric models to similarity data: the interrelationship between similarity and spatial density." *Psychological Review*, *85*(5), 455 - 463.

Laibson, D. (1997). "Golden eggs and hyperbolic discounting." *Quarterly Journal of Economics*, *112*(2), 443-477.

Lakoff, G. (1987). *Women, fire and dangerous things: What categories reveal about the mind.* Chicago, IL: University of Chicago Press.

Lloyd, S. P. (1982). "Least squares quantization in PCM." *IEEE Transactions on Information Theory*, *IT-28*, 127-135.

Luce, R. D. (1988). "Rank-dependent, subjective expected utility representations." *Journal of Risk and Uncertainty*, *1*, 305-332.

Luce, R. D., and Fishburn, P. (1991). "Rank- and sign- dependent linear utility models for finite first-order gambles." *Journal of Risk and Uncertainty*, *4*, 29-59.

Machina, M. J. (1987). "Choice under uncertainty: Problems solved and unsolved." *Journal of Economic Perspectives*, *1*, 121-154.

MacMillan, B. (1953). "The basic theorems of information theory." *Annals of Mathematical Statistics*, *24*, 196-219.

Macmillan, N. A., Kaplan, H. L., and Creelman, C. D. (1977). "The psychophysics of categorical perception." *Psychological Review*, *84*(5), 452-471.

Margolis, E., and Laurence, S. (1999). *Concepts: Core readings.* MIT Press.

Max, J. (1960). "Quantizing for minimum distortion." *IEEE Transactions on Information Theory*, *IT-6*, 7-12.

Miller, G. A. (1956). "The magical number seven, plus or minus two: some limits on our capacity for processing information." *Psychological review*, *63*, 81-97.

Olson, M. J., and Budescu, D. V. (1997). "Patterns of preference for numerical and verbal probabilities." *Journal of Behavioral Decision Making*, *10*, 117-131.

Parducci, A. (1965). "Category judgment: a range-frequency model." *Psychological review*, *72*, 406-418.

Pinker, S. (1994). *The language instinct.* New York: Harper Perennial.

Pinker, S., and Bloom, P. (1992). "Natural language and natural selection." In J. H. Barkow, L. Cosmides, and J. Tooby (Eds.), *The adapted mind.* New York: Oxford University Press.

Prelec, D. (1989). *Decreasing impatience: Definition and consequences.* (Harvard Business School Working Paper)

Prelec, D. (1998). "The probability weighting function." *Econometrica*, *66*, 497-527.

Prelec, D., and Lowenstein, G. (1991). "Decision making over time and under uncertainty: a common approach." *Management Science*, *37*(7), 770-786.

Prelec, D., Wernerfelt, B., and Zettelmeyer, F. (1997). "The role of inference in context effects: Inferring what you want from what is available." *Journal of Consumer Research*, *24*(1), 118-125.

Rabin, M. (1998). "Psychology and economics." *Journal of Economic Literature*, *36*, 11-46.

Rubinstein, A. (1988). "Similarity and decision-making under risk." *Journal of Economic Theory*, *46*, 145-153.

Rubinstein, A. (1998). *Models of bounded rationality.* Cambridge, MA: M.I.T. Press.

Shannon, C. E. (1948). "A mathematical theory of communication." *Bell Sys. Tech. Journal*, *27*, 379-423.

Shugàn, S. M. (1980). "The cost of thinking." *Journal of Consumer Research*, *7*, 99-111.

Simon, H. A. (1955). "A behavioral model of rational choice." *Quarterly Journal of Economics*, *69*, 99-118.

Simonson, I., and Tversky, A. (1992). "Choice in context: Tradeoff contrast and extremeness aversion." *Journal of Marketing Research*, *29*(3), 281-295.

Thaler, R. (1980). "Towards a positive theory of consumer choice." *Journal of economic behavior & organization.*, *1*, 39-60.

Thaler, R. (1985). "Mental accounting and consumer choice." *Marketing Science*, *4*(3), 199-214.

Trushkin, A. V. (1982). "Sufficient conditions for uniqueness of a local optimal quantizer for a class of convex error weighting functions." *IEEE Transactions on Information Theory*, *IT-28*(2), 187 -198.

Trushkin, A. V. (1993). "On the design of an optimal quantizer." *IEEE Transactions on Information Theory*, *IT-39*(4), 1180 -1194.

Tversky, A. (1977). "Features of similarity." *Psychological Review*, *84*(4), 327-352.

Tversky, A., and Kahneman, D. (1974). "Judgement under uncertainty: Heuristics and biases." *Science*, *185*, 1124-1130.

Tversky, A., and Kahneman, D. (1981). "The framing of decisions and the rationality of choice." *Science*, *211*, 453-458.

Tversky, A., and Kahneman, D. (1991). "Loss aversion in riskless choice: a reference-dependent model." *Quarterly Journal of Economics*, *106*, 1039-1061.

Tversky, A., and Kahneman, D. (1992). "Advances in prospect theory: cumulative representation of undertainty." *Journal of Risk and Uncertainty*, *5*, 297-323.

Tversky, A., and Simonson, I. (1993). "Context-dependent preferences." *Management Science*, *39*, 1179-1189.

Viswanathan, M., and Childers, T. (1996). "The encoding and usage of numerical and verbal product information by consumers." *Journal of Consumer Psychology, 5*(4), 359-385.

Viswanathan, M., and Childers, T. (1997). "'5' calories or 'low' calories? what do we know about using numbers or words to describe poroducts and where do we go from here." In M. Brucks and D. J. MacInnis (Eds.), *Advances in consumer research* (Vol. 24, p. 412-418). Provo, UT: Association for Consumer Research.

Von Neumann, J., and Morgenstern, O. (1944). *Theory of games and economic behavior.* Princeton: Princeton University Press.

Wallsten, T. S., Budescu, D. V., and Zwick, R. (1993). "Comparing the calibration and coherence of numerical and verbal probability judgements." *Management Science, 39*(2), 176-190.

Wernerfelt, B. (1995). "A rational reconstruction of the compromise effect: Using market data to infer utilities." *Journal of Consumer Research, 21*, 627-633.

Wu, G., and Gonzalez, R. (1996). "Curvature of the probability weighting function." *Management Science, 42*(12), 1676-1690.

Wu, G., and Gonzalez, R. (1999). "Nonlinear decision weights in choice under uncertainty." *Management Science, 45*(1), 74-85.

Zadeh, L. (1965). "Fuzzy sets." *Systems and Control, 8*, 338-353.

Zadeh, L. (1978). "Fuzzy sets as a theory of possibility." *Fuzzy Sets and Systems, 1*, 3-28.

Zadeh, L. (1982). "A note on prototype theory and fuzzy sets." *Cognition, 12*, 291-297.