How Categorization Shapes the Probability Weighting Function

Dan Schley [*], Hang-Yee Chan, [†] Manissa Gunadi, [‡] Alina Ferecatu, [§]

June 17, 2020

Abstract

The shape of the probability weighting function is one of the most well-known empirical observations in judgment and decision-making research. The tendency to overweight low probability events and underweight high probability events stems from the categorical distinction between “not happening,” “a chance," and “happening." We demonstrate that individuals categorize uncertainty differently across contexts (e.g., numeric risks vs. verbal labels of uncertainty). These categorical perceptions produce additional reference points in the probability space. The resulting behavioral patterns necessitate revising the traditional inverse-S shaped probability weighting function. Using experimental and model-based approaches, we demonstrate that probability sensitivity, and thus the shape of the probability weighting function, depends crucially on how decision-makers represent categories of uncertainty.

[*] Rotterdam School of Management, Erasmus University, schley@rsm.nl

[†] Amsterdam School of Communication Research, University of Amsterdam, h.y.chan@uva.nl

[‡] ESADE, Universitat Ramon Llull, manissa.gunadi@esade.edu

[§] Rotterdam School of Management, Erasmus University, ferecatu@rsm.nl

1 Introduction

The empirical observation that individuals predominantly overweight low probability events and underweight high probability events has become a staple of decision theories (e.g., Kahneman and Tversky 1979, Tversky and Kahneman 1992). As such, the inverse S-shaped probability weighting function is ubiquitous across a myriad of decision-making fields. In the current manuscript, we aim to reconsider the nature of probability weighting. In particular, whereas the traditional probability weighting function is assumed to have two reference points (i.e., 0 and 1) and thus exhibit an inverse S-shape, we will argue and demonstrate that probability weighting can exhibit multiple reference points, and the number of reference points is contextually dependent.

Predominant theories and models of probability weighting have long assumed that the cyclical nature of the probability weighting function is due to nonlinearities in the probability weights near 0 and 1 (Kahneman and Tversky 1979, Tversky and Kahneman 1992, Camerer and Ho 1994, Prelec 1998, Wu and Gonzalez 1996, Gonzalez and Wu 1999, Brandstätter et al. 2002, Mukherjee 2010). This notion of nonlinear perceptions of the probabilities originates in psychophysics. In particular, Kahneman and Tversky (1979) and Tversky and Kahneman (1992) argued that individuals have two reference points, at 0 and 1, and they exhibit diminishing sensitivity from those reference points. This high sensitivity from 0% to 1%, and from 100% to 99%, and lower sensitivity from 1% to 99% has its origins in how the mind affectively interprets risk (Loewenstein et al. 2001, Slovic et al. 2000, Rottenstreich and Hsee 2001). When evaluating a 0% chance of a car crash the mind generates a mental image of a car in a safe condition. However, when evaluating any chance greater than 0% the mind conjures an image of a car crash. These two sets of mental images have quite distinct affect attached to them, and this stark difference in affect underlies the high probability sensitivity moving away from the reference points. On the other hand, the mental imagery conjured by a 10% chance is quite similar to that of a 20% chance.

This affect-based account builds on the idea that decision makers interpret information categorically (e.g., mentally imagining a safe car versus a car crash). This categorical perspective offers a straightforward intuition about probability weighting. The mind interprets a 0% chance as "not happening." Thus, a 1% chance is categorically distinct from "not happening." Similarly, a 100% chance is interpreted as"certainly happening", but a 99% chance is perceived as categorically distinct from "certainly happening." Said differently, because 0% and 100% are categorically distinct from the rest of the probability spectrum, they act as reference points. The ubiquitous inverse S-shape is the consequence of diminishing marginal sensitivity relative to these reference points. In the current manuscript, we entirely agree with the proposition that distinctive categories of uncertainty act as reference points resulting in curvature of the probability weighting function. However, we argue that there are more than three distinct categories and that these categories are an emergent property of the particular decision context. In the following section we explore the psychology of categorization and its consequences on the probability weighting function.

2 Categorization and Probability Weighting

First, we review the origins of psychophysics, and in particular how discriminability amongst stimuli leads to sensitivity. Next, we review how distinct categorical boundaries influence this discrimination process and therefore sensitivity. We then provide evidence from the literature for our proposition that there exists multiple categorical boundaries along the probability continuum. Throughout this paper we use the term "categorical boundaries" as approximately equivalent to how "reference points" are used in the decision-making literature. We do this because we believe it helps facilitate understanding of the psychological determinants of where and how reference points are formed, particularly because we aim to integrate research from categorical perception in the current investigation.

2.1 Sensitivity and Categorization

Research on categorical perception and psychophysics have been entwined since the earliest days of modern psychology. Indeed, Weber (1834) and Fechner (1860) identified that the sensitivity to a change in a stimulus required a minimum just-noticeable difference between two stimulus values. That is, if the intensity of stimulus A was insufficiently different from that of stimulus B, the *perceived* intensity of A would be equal to that of B. While this may appear tautological, it was an important proposition for its time because it necessitates stimuli to be assigned to different categories for differences in the stimuli to be perceived.

This idea of stimulus discrimination is the foundation of signal-detection theory, whereby a decision-maker's sensitivity between two stimuli is governed by the degree of overlap between the two stimuli (for a review see Macmillan and Creelman 2005). For example, individuals are more sensitive (i.e., more accurate in their ability to identify differences between) the colors red and blue than they are between light blue and baby blue. In the former case, the visual inputs for processing red and blue do not overlap at all, whereas in the latter case the visual inputs overlap considerably.

Beyond differentiating nominal stimuli, the same principles apply to sensitivity to quantitative or cardinal stimuli. For instance, human perception of quantity, whether physical dots or symbolic numbers, has its roots in the approximate number system (for a review see Dehaene 2011). This system encodes quantitative inputs with error, and this error increases as a function of total quantity. For example, it is very easy to differentiate whether a room has 5 chairs versus 4 chairs, but it is incredibly difficult to differentiate whether a room has 105 chairs versus 104 chairs. The consequence of this increasing error is that the ability to discriminate between, and therefore sensitivity to quantities is marginally diminishing (Stevens 1957).

2.2 How Categorical Boundaries Facilitate Sensitivity

Given the above discussion, it is apparent that the ability to categorize quantities is necessarily related to individuals' sensitivity to those quantities. As such, any feature of a stimulus that facilitates its discrimination therefore produces sensitivity. For example, Pelham et al. (1994) demonstrated that taking a whole visual stimulus and chopping it up into pieces led participants to perceive the physical area as larger than in the case when the visual stimulus was presented as a whole. Splitting the whole into segments produces clear categorical boundaries facilitating discrimination. And, because the ability to discriminate stimuli underlies sensitivity, the visual area of the stimulus was perceived to be greater.

Similarly, Lembregts and Van Den Bergh (2019) presented participants scenarios about quantities of sugar, describing them as 10 versus 65 grams or as 2 versus 13 cubes. The authors found that participants perceived the difference between 2 and 13 cubes to be greater than that between the equivalent 10 and 65 grams. Furthermore, the authors demonstrated that this increased sensitivity is related to the fact that units of sugar cubes have clear discrete visual and conceptual boundaries, whereas units of mass (i.e., grams in this case) of sugar do not.

2.3 Probability Sensitivity

As discussed above, diminishing marginal sensitivity is the product of increased difficulty in the discrimination between equivalent pairs of stimuli as the intensity/quantity increases (Stevens 1957, Dehaene 2011). Traditional demonstrations of psychophysical curves are typically concave, indicating diminishing sensitivity from a single categorical reference point (i.e., 0). This reflects the principle of absolute threshold in psychophysics (Fechner 1860), whereby there is some minimum stimulus intensity required for the percept to cross from not perceiving to perceiving a stimulus. Kahneman and Tversky (1979) assumed that sensitivity to probabilities has two distinct reference points at 0 and 1. The product, therefore, is a probability weighting function that exhibits diminishing marginal sensitivity as probabilities move away from 0 and 1, leading to the ubiquitous inverse S-shaped function. This functional form explicitly assumes that decision makers have categorical boundaries for risk at 0 and 1, but also tacitly assumes no other categorical boundaries exist between 0 and 1. In the current paper, we explore this tacit assumption and uncover additional systematic categorical boundaries along the probability continuum.

2.3.1 Demarcations in the Probability Space.

Where might we expect to observe distinct categorical boundaries that produce increased probability sensitivity, beyond those found at 0 and 1? Hollands and Dyre (2000) explored an analogous problem in perceptions of ratios. In particular, past work had assumed a one-cycle (i.e., inverse S-shaped) function describing the relationship between the true proportion of a set of stimuli and the judged proportion (e.g., Erlick 1964, Varey et al. 1990, Nakajima 1987). The common inverse S-shape came from the same proposition of diminishing marginal sensitivity from proportions of 0 and 1. Hollands and Dyre (2000) demonstrated that demarcating a physical stimulus facilitates the discrimination of stimuli between categories. For example, consider a cylindrical container with no discernible demarcations. The glass has two clear reference points, empty and full. Decision-makers evaluating different quantities of liquid in the glass will therefore exhibit an inverse S-shape relationship between objective volume and perceived volume. Demarcating the cylinder at the halfway point will create three perceptible categorical boundaries (i.e., 0%, 50%, and 100%), resulting in a psychophysical function with concavity starting at 0%, flipping to convexity as the proportion moves closer to 50%, returning to concavity after 50% and again to convexity as it approach 100% (i.e., a cyclical function with 2 cycles). Further demarcating the cylinder at 25% and 75% adds additional cycles to the psychophysical function as it becomes easier to discriminate quantities below and above 25% and 75%. If psychophysical principles produce the ubiquitous inverse S-shaped probability weighting function, then it seems reasonable that additional categorical boundaries in the probability space would act as reference points, producing a probability weighting function with multiple cycles.

What then might demarcations look like across the entire probability space? In the domain of decisions under risk, risks are predominantly expressed numerically. The base-10 system of Arabic numerals exhibit a simple nested property. Digits from 0 to 9 are nested under sets of 10, and sets of 10 are nested under sets of 100, and so on. With this property in mind, past researchers have identified left-digit effects in riskless value-based judgment and choice domains (e.g., Thomas and Morwitz 2005, Manning and Sprott 2009, Allen et al. 2017). Since we use numbers to express risk, and those numbers serve for efficient categorization of the probability space, then we might expect probability weighting to exhibit greater sensitivity when risks cross these left digits.

One might ask, if decision-making researchers have been studying the probability weighting function for over four decades, why were these additional reference points in the probability weighting function not observed in previous data? The reason is that, because risky choice involves both probabilities and outcomes, researchers must design experimental paradigms specifically to capture nonlinearities in the probability space. For example, to demonstrate the overweighting of low probability events and underweighting of high probability events Kahneman and Tversky (1979) selected pairs of gambles that were specifically designed to test probability weighting near 0 and 1 (i.e., the certainty effect paradigm). In most studies, we use variations of the Wu and Gonzalez (1996) laddering paradigm and demonstrate that decision makers represent probabilities just below and just above the left digit as categorically distinct. If categorical perceptions drive probability sensitivity, then we would expect higher sensitivity for probabilities that cross these boundaries and lower sensitivity for probabilities within categories. For example, if Gamble A is a 1% chance of winning $20 and a 12% chance of winning $5 and Gamble B is a 16% chance of winning $5, the probability of the common outcome (winning $5) does not cross a categorical boundary (i.e., 12% and 16% are in the same [10-20)% category). Now, if Gamble C is a 1% chance of winning $20 and an 18% chance of winning $5 and Gamble D is a 22% chance of winning $5, then the probability of the common outcome for Gamble C resides in a different categorical space (i.e., the [10-20)% category) than Gamble D (i.e., the [20-30)% category). We would therefore expect probability sensitivity for the common $5 outcome to be greater between Gambles D and C than between Gambles B and A. This greater probability sensitivity results in the probability weighting function exhibiting additional reference points (i.e., additional points of nonlinearity). We demonstrate this empirical observation with our preregistered experiments. In addition, using a model-based approach we demonstrate the existence of these additional reference points in our data, and in those from the Wu and Gonzalez (1996) studies.

2.3.2 The Subjectivity of Numeric Categories.

The above discussion highlights that numeric expressions of risk can exhibit probability weighting with additional reference points beyond those traditionally observed at 0 and 1. But does this mean that for any uncertain outcome there is a stable probability weighting function with reference points spaced at every tenths digit from 0 to 1? This is unlikely. While each left digit provides an easily accessible demarcation of a categorical boundary, they need not always act as categorical boundaries. For example, if Gamble A is a 1% chance of winning $20 and a 12% chance of winning $5 and Gamble B is an 28% chance of winning $5, the categorical interval for the common outcome will lie between 10% and 30%. But, because none of the gambles cross less than two left digits, there is not sufficient data to observe nonlinearities in the probability weighting function at every left digit. So while there may theoretically exist categorical boundaries at every left-digit, because the probability-weighting function is descriptive, it's observable shape is bound to the context of the data; and in this case, the gamble pairs do not afford the observation of a function with nonlinearities at each left digit.

Now consider if Gamble C is a 1% chance of winning $20 and an 18% chance of winning

$5 and Gamble D is a 36% chance of winning $5. In this context the categorical interval for the common outcome will lie between 10% and 40%. Compared to the context with Gambles A and B, probability sensitivity to the common outcome will be greater in Gambles C and D. This means that the descriptive probability weighting function inferred from Gambles C and D would have a different shape than that inferred from A and B. In sum, this line of reasoning supposes that the shape of the probability weighting function, and in particular, the number of categorical boundaries will be endogenous to the stimulus set.

2.3.3 Verbal Expressions of Uncertainty.

We can take this line of reasoning a step further and ask about other expressions of uncertainty. Potentially more ubiquitous than numeric probabilities of risk, verbal expressions of uncertainty allow individuals to efficiently communicate risk information. Considerable research has mapped verbal expressions of uncertainty to numeric risks (Wallsten et al. 1993a,b, Budescu et al. 1988) often demonstrating distinct patterns of preferences between the two formats (Olson and Budescu 1997, Windschitl and Wells 1996). Because the categorical boundaries we identified for numeric risks are an inherent feature of the 10-based system of numbers, we might wonder what categorical boundaries exist for verbal expressions of uncertainty?

Rather than assuming a linguistic structure, we aim to take a data-generative approach to the exploration of verbal expressions. Using the verbal expressions from Wallsten et al. (1993a), we use a simple model-free approach to clustering the verbal expressions. We propose that sensitivity to decisions under uncertainty will be greater between gambles with verbal expressions that cross categorical boundaries than to those within a category.

3 Empirical Studies

With the aim of identifying additional systematic features of the probability weighting function, we build our empirical framework by examining the entire probability space. In particular, because we aim to identify intermediate reference points between 0 and 1, we explored the probability space in greater detail than past studies. Additionally, we identified other stimulus-driven categorical boundaries and the resulting non-linearities in the probability weighting function. To this end, we conducted a series of preregistered experiments.

3.1 Overview

Study 1 applies a fine-grained approach by examining the entire probability range from 0% to 100%. In particular, we investigated whether probability sensitivity changes as a function of whether or not probabilities among gamble pairs both sit within a category or between categories (i.e., in the same decile group). Importantly, this study presents evidence for this categorical-based account in an incentive-compatible context. Study 2 further explores context dependency by manipulating the width between gamble pairs. To ensure that our conclusions are sufficiently generalizable, we tested our account using a different paradigm in Study 3. Whereas all other studies involved choices between risky gambles, Study 3 uses certainty equivalents. Furthermore, Study 4 tests our hypothesis using verbal expressions of uncertainty. As a supplement to the experiments, we also introduce a novel two-parameter model of the probability weighting function that parsimoniously accounts for our empirical findings and existing data from (Wu and Gonzalez 1996).

3.2 Laddering Paradigm

In order to examine the properties of the probability weighting function, in the majority of the studies reported in this paper, we modified the Wu and Gonzalez (1996) laddering technique to specifically investigate our categorical-based account. Because we use this paradigm in Studies 1, 2 and 4, we first provide a detailed overview of the procedures. Specific information pertaining to a particular study is discussed in the respective study sections.

Participants' task was to choose between pairs of gambles. Each gamble pair consisted of a riskier (R) and safer (S) option. The R option always presented an $p\%$ chance of winning a smaller payout ($Payout_{common}$) and 1% chance of winning a larger payout ($Payout_{extra}$). The S option did *not* have the 1% chance of the larger extra payout ($Payout_{extra}$), but always presented a higher $p+q\%$ chance of winning the smaller common payout ($Payout_{common}$). The expected values of both options were kept the same.

The focal dependent variable of the laddering experiments (Studies 1, 2, and 4) was whether participants choose the safer (S) option (0 or 1).[1] As noted by Wu and Gonzalez (1996), because within each gamble pair the two options had the same expected value, non-monotonicity in preference for one option as p traverses the probability space indicates non-linearity in the probability weighting function. To illustrate, consider the case where a gamble pair straddles over a categorical boundary. That is, the $p\%$ (riskier) and $p+q\%$ (safer) chance for $Payout_{common}$ differ by one left digit (e.g., 18% versus 22%). Given the same $q\%$, *higher* preference for the S option for that specific gamble pair would imply that this $q\%$ difference has created *higher* sensitivity when the resultant $p+q\%$ differs by one left digit (compared to when it does not, e.g., 12% versus 16%). This observed heightened sensitivity would be, we reason, the consequence of the crossing of a distinct category boundary in decision-makers' minds.

In sum, *stronger* preference for the S option signals *heightened* probability sensitivity. We contend that the proportion of choices for the S option across the probability space approximates the probability weighting function.

3.2.1 Further Details on Laddering Experiments.

In Study 1, the gambles are consequential for participants, as one of their decisions in the gamble pairs is randomly selected and carried out with actual payment. In the other studies, the gambles are hypothetical. Following the Wu and Gonzalez (1996) procedure, these gamble pairs are structured into a ladder, with each rung of the ladder representing a single gamble pair with varying values of p. Within a ladder for a specific study, q is held equal. The total number of rungs in the ladder differs by study. Study 1 has three distinct ladders, Study 2 has two ladders, and Study 4 has one ladder with random subset of rungs.

All studies were conducted online using either Prolific Academic or Amazon Mechanical Turk. Participants completed every rung (gamble pair) presented in randomized order within one

[1] Note that in all our preregistrations, we mentioned the focal dependent variable in our studies was the proportion of participants choosing the *riskier* option, which is mathematically equivalent.

ladder. Note that in Study 1 where participants did multiple ladders, the order of the ladders was also randomized. At each rung, participants viewed on the computer screen the R and S options positioned horizontally (either R-S or S-R, decided randomly for each rung), and indicated which of the two options they would prefer.

3.3 Study 1 — Incentive Compatible Ladders

We began with a comprehensive examination of the entire probability space from 0% to 100%. Our aim was to precisely identify probability sensitivities along the probability spectrum. As discussed in the introduction, we expected decision makers to rely on easily accessible cues to facilitate categorical distinctions between stimuli. With numeric probabilities, it is likely that the left-digit of the number provides a simple diagnostic categorical boundary. If such categorical boundaries act as reference points, then we would expect increased sensitivity when probabilities straddle across categorical boundaries compared to when they remain within the boundaries. For example, when the left digit of $Payout_{common}$ changes from 18% to 22%, the probability crosses from the “10s" to the “20s". Conversely, when the left digit of the $Payout_{common}$ probability changes from 12% to 16%, it stays within the “10s" category.

Our goal in this study was to establish initial evidence for our predictions, and to demonstrate whether probability sensitivity increases at these left-digit categorical boundaries. Importantly, we utilized an incentive-compatible paradigm, where participants' decisions would be carried out for actual payment following the completion of the study. The hypothesis, procedures, and analysis for this study were preregistered on AsPredicted: `http://aspredicted.org/blind.php?x=9vh8i9`.

3.3.1 Procedure.

Participants (n = 100, 39.0% female; M_{age} = 27.66, SD = 8.85, Prolific Academic) evaluated three different gamble ladders, with each ladder consisting of 15 gamble rungs. Each gamble rung represented a choice between riskier (R) and safer (S) options. The R option involved a 1% chance of getting $Payout_{extra}$, and a *p*% chance of getting a $Payout_{common}$. The S option involved a *p*+5% chance of getting a $Payout_{common}$. Thus, the difference between the focal probabilities for the R and S options (*q*) was set to be the same for all ladders (5%). The gamble rungs were evenly spaced within the ladder, with a difference of 7% between each rung, apart from R7-R8 which differed by 6% and R14-R15 which differed by 5%. The instructions and the list of the gamble pairs used in this study can be found in Web Appendix WA7.

All ladders have the same set of probabilities for the riskier and safer options, but differ in financial outcomes (i.e., $Payout_{common}$ and $Payout_{extra}$). This ladder design is particularly suited to infer risk preferences and probability weighting using a behavioral model a la Prospect Theory, which we attempt in the modeling section (§4) of this paper. This is because we allow for variation in financial outcomes while keeping the objective probabilities of the gambles constant across the three ladders. As our sample consisted of participants from the United Kingdom, we used the British currency (pence) for this study. For Ladder 1, $Payout_{extra}$ was 500 pence and $Payout_{common}$ was 100 pence. For Ladder 2, it was 250 pence and 50 pence; and for Ladder 3, 100 pence and 20 pence, respectively. Importantly, these payouts ensured that the expected values were the same between the riskier and safer options for each rung.

At the beginning of this study, participants were informed that one of the gamble pairs they evaluated would be randomly selected and subsequently carried out for real payment at the end of the study. Thus, participants' payment included a completion fee following the hourly rate imposed by Prolific and bonus amount from the incentivized gamble, in line with their submitted choice.

3.3.2 Results.

Descriptive results are presented for each rung of each ladder in Figure 1. Using the Wu and Gonzalez (1996) laddering paradigm, if decision makers weighted probabilities linearly, there should be no change in the preference for the safer option as a function of the $Payout_{common}$ probability. Observing a non-monotonic (particularly U-shaped) relationship serves as evidence of an inverse S-shape in the probability weighting function, whereby decision makers are more sensitive to equivalent changes in probability near 0 and 1 than in the middle of the probability spectrum.[2] Consistent with Wu and Gonzalez (1996), we do observe this non-monotonicity in preference for the safer option across the probability space, such that sensitivity to the difference in probability of the common outcome is higher near the probabilities of 0 and 1.

In Figure 1, gamble pairs with probabilities that straddle over the left-digit boundary are indicated by dark red triangles, while probabilities within categorical boundaries are indicated by light gray circles. Visually, it appears that preference for the safer option was substantially higher for these gamble pairs. Our explanation is that when the $Payout_{common}$ probability in the safer option ($p+q$%) differs by the left-digit, the subjective probability of the safer option feels substantially higher than that of the riskier option since the probability of the safer option falls into a distinct category in decision makers' minds.

In line with our preregistered analysis plan, we conducted a mixed effects logistic regression predicting the choice for the safer option, with participant-specific intercepts. We included three sets of categorical variables in the regression: (a) a predictor indicating whether a rung crosses a categorical boundary; (b) dummy variables for R01 and R15 (to demonstrate that the effect of the categorical boundary is not solely driven by probabilities of 0 and 1), and dummy variables binning R2-R5, R6-R10, and R11-R14 (to account for any overall trends across the probability space), with the middle bin (Bin 3) as default; (c) dummy variables for the three ladders, with ladder 1 as default; and (d) a covariate accounting for the order with which participants' saw the gamble pairs. Table 1 reports the results of the model.

Of key interest, we observed a significant positive coefficient on the left-digit boundary, β = 0.240, p = .0045. Consistent with the descriptive results, this means that the preference for the safer option was significantly higher when the $Payout_{common}$ probability ($p+q$%) had different left digits. This result is evidence of increased probability sensitivity when crossing the left-digit boundaries. Importantly, because of Bins 1 and 5 in the model (i.e., the rungs containing the gambles with 0% and 100%), the effect of the left-digit boundary reflects the influence of categorization, after controlling for any effect produced by probabilities of 0 and 1.

[2] Wu and Gonzalez (1996) plot the proportion choosing the *riskier* option while we plot the proportion choosing the *safer* option. The result is simply a flip of the vertical axis of the plot. We find plotting the proportion of choice for the safer option more intuitive because higher proportions indicate greater probability sensitivity.

Table 1: Mixed effects logistic regression predicting safer choice (Study 1)

	Dependent variable:
	Safer choice
Order	−0.044
	(0.039)
Ladder 2	0.027
	(0.095)
Ladder 3	0.075
	(0.095)
Riskier probability = 0 (Bin 1)^a	1.141***
	(0.164)
Bin 2	0.529***
	(0.100)
Bin 4	−0.337***
	(0.104)
Safer probability = 1 (Bin 5)	0.202
	(0.165)
Left-digit boundary	0.240***
	(0.084)
Constant	−1.163***
	(0.224)
Observations	4,500
Log Likelihood	−2,155.594
Akaike Inf. Crit.	4,331.189
Bayesian Inf. Crit.	4,395.307
Note:	^{a}Bin 3 as reference; *p<0.1; **p<0.05; ***p<0.01

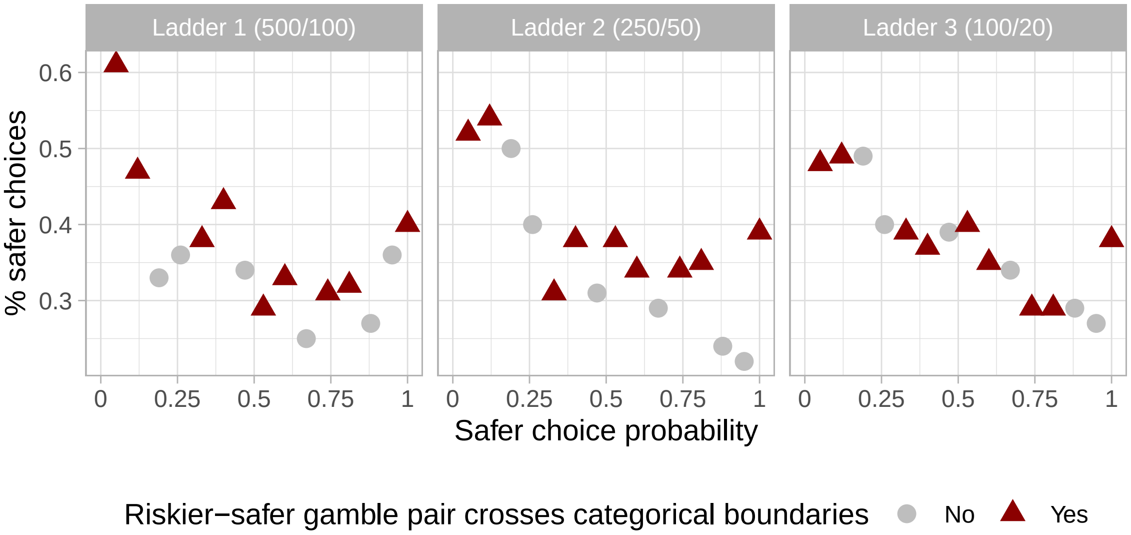

Figure 1: Proportion of safer choices of the three incentive compatible ladders (Study 1)

3.4 Study 2 — Context-Dependent Ladders

Study 1 presented initial evidence that participants were more sensitive to normative-equivalent differences in probability when those probabilities crossed a categorical (i.e., left-digit) boundary. Our interpretation is that these categorical boundaries act as additional reference points at each decile of the probability space.We interpret these findings as evidence that participants are sensitive to a change in probability from 0% to 1%, diminished sensitivity between 1% and 9%, but again quite sensitive from 9% to 10%, followed by diminished sensitivity from 10% to 19%, and so on.

Rather than the traditional probability weighting function with 2 reference points, do these results suggest that the probability weighting function has 11 equally spaced reference points from 0 to 1? In other words, are these reference points fixed within the probability space for numeric risks? Because categorization is an emergent property of the stimulus context (for a review see Rosch 1999), we surmise that this is not the case. Instead, we expect the relevant categorical boundaries, and thus reference points, to be contingent on the focal stimuli and the particular spacing of probabilities.

We aim to test this prediction by means of two gamble ladders with different numeric contexts. We contend to show that, given the specification of the gamble ladders, it is possible that there is, for instance, a discontinuity in preference for the safer option at 30% in the first ladder, but there may not necessarily be the same discontinuity at 30% in another ladder. An important design factor for the two ladders used in this study is that the difference between the focal riskier and safer options was set to be the same for both ladders (p% and $p + 14$%), as well as the difference or step-size between each gamble rung (8%). However, the two ladders had different starting points, which consequently allowed for the comparison between the boundaries that are crossed in Ladders 1 and 2. Referring to Table WA2 in the Web Appendix, in rungs 4, 5, 9 and 10, only Ladder 1 had a difference of 2 in the left digit of the $Payout_{common}$ probability, while Ladder 2 had this feature in rungs 2, 3, 7 and 8. If categories were an emergent property of the stimuli, then we would not expect additional reference points at each change in the left digit along the probability space. This means that in a context like Ladder 1, we should expect that the probability weighting function would have reference points at different places compared to Ladder 2. This study was preregistered on AsPredicted: `https://aspredicted.org/blind.php?x=f68va3`.

3.4.1 Procedure.

Participants (n = 1003, 55.8% female; M_{age} = 36.91, SD = 11.93, MTurk) were randomly assigned to one of the two ladders. Similar to the previous study, both ladders roughly covered the entire probability space. More specifically, the ladders did not include the extremes at 0% and 100%, but the probabilities range from 4% to 96%. For both ladders, the $Payout_{extra}$ was \$70 and $Payout_{common}$ was \$5. The instructions and the list of the gamble pairs used in this study can be found in Web Appendix WA8.

Again referring to Table WA2 in the Web Appendix, the left digit could differ by either 1 or 2 (e.g. 4% vs. 18%, by 1; 28% vs. 42%, by 2) in both ladders. Furthermore, this design provided 4 boundary crossings that are unique to each specific ladder. This allows us to test whether categorical boundaries are inherent or contextually determined. If our account is correct, we should expect the probability weighting function inferred from Ladder 1 to have reference points at different places compared to the function inferred from Ladder 2. To illustrate, in Ladder 2 we should expect to see an increase in preference for the safer option when the $Payout_{common}$ probability differs by 2 in the left digit in Rung 2 (i.e. 18% vs. 32%). In other words, there could potentially be increased probability sensitivity near 30% in this ladder. However, in Ladder 1, we need not necessarily expect to observe an increase in preference to the same extent when the probability of the safer option changes by 1 digit only and crosses 30% in Rung 1 (i.e. 20% vs. 34%).

Now consider probability weighting near 40%. At this point in the probability space, we could expect increased probability sensitivity in Rung 4 (i.e. 28% vs. 42%) of Ladder 1 but not in Rung 4 of Ladder 2 (i.e. 34% vs. 48%). Thus, this design allows us to test whether the findings of Study 1 reflect a stable left-digit effect or categorical boundaries that are an emergent property of the stimulus context.

3.4.2 Results.

Descriptive results are presented for each rung of the Ladder 1 (solid line) and Ladder 2 (dashed line) in Figure 2. In both ladders, we expected to find greater probability sensitivity (as evidenced by an increase in preference for the safer option) for gamble pairs where the left digit changed by 2, but at different points of the probability space. Indeed, we observed this pattern: for example, for Rung 3 of Ladder 1 (R [.01, \$70; .20, \$5]; S3 [.34, \$5]) 55.1% of participants chose the safer option, while Rung 2 of Ladder 2 (R [.01, \$70; .18, \$5]; S [.32, \$5]) the safer choice proportion jumped to 69.7%, even though these two rungs were separated by 2% in the probability space. That is, participants appeared sensitive to categorical boundaries within the context of a particular ladder.

In line with our preregistered analysis plan, we ran two mixed effects logistic regressions on choice for the safer option, with participant-specific intercepts (Table 2). In the first model (partial), we excluded rungs 1 and 6 where both ladders have the same left-digit difference of 1. We then created a dummy that denotes rungs 2, 3, 7, and 8, where Ladder 2 has left-digit difference of 2 (i.e., categorical boundaries specific to Ladder 2), while categorical boundaries occur at the remaining

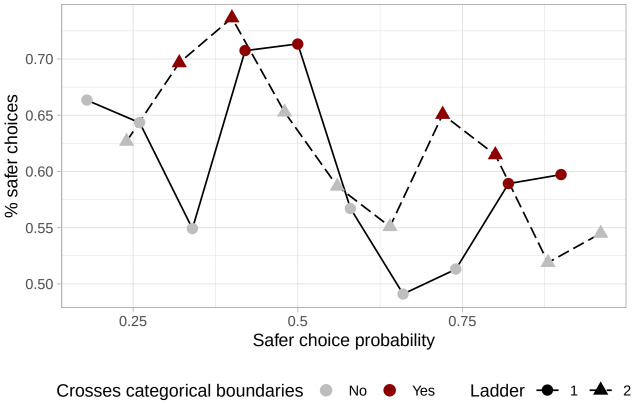

Figure 2: Proportion of safer choices across the two ladders showing different contexts (Study 2)

rungs for Ladder 1. Consistent with the preregistered prediction, we observed an overall significant positive effect on the interaction term (Ladder 2 × Ladder 2-only boundary), $\beta = 1.353$, $p < .0001$.

We then implemented a second model with the full data set. We created a dummy variable denoting rungs with a left-digit difference of 2 in Ladder 1 (i.e., rungs 4, 5, 9, and 10), and a dummy variable denoting rungs with a left-digit difference of 2 in Ladder 2 (i.e., rungs 2, 3, 7, and 8); the intercept term therefore reflects rungs 1 and 6 where both ladders have the same left-digit difference of 1. Again, we observed a significant positive coefficient on the interaction term Ladder 2 × Ladder 2-only boundary, $\beta = 1.028$, $p < .0001$, and a negative coefficient on the interaction term Ladder 2 × Ladder 1-only boundary, $\beta = -0.331$, $p = .0164$. Notably, the main effect coefficient of the ladder dummy was not significant, showing that there was no difference in overall risk preference between the two groups who saw different ladders.

3.5 Study 3 — Certainty Equivalence

In contrast to the previous studies where participants were asked to compare probabilities between riskier and safer options in a gamble, in this study we examine whether we could document similar probability sensitivity using certainty equivalents. Participants read four hypothetical gambles and indicated the desired sure payout for which they were indifferent between the prospects. We focus on probabilities clustered around the 50% level, and predict that 50% would be treated as a distinct categorical boundary. As a result, probabilities below the 50% level should be subjectively perceived to be substantially less likely to occur than those above the 50% level. This study was preregistered on AsPredicted: `https://aspredicted.org/blind.php?x=rh8n9e`.

3.5.1 Procedure.

Participants (n = 1008, 57.7% female; $M_{age} = 37.42$, SD = 12.03, MTurk) were presented with four separate gambles in random order. Each gamble involved a specific probability (expressed in percentage form) of winning \$200, otherwise \$0. The probabilities were 44%, 48%, 52%, and 56%. Of interest, we expected the difference in certainty-equivalence judgments between 48% and 52% (which crosses the 50% level) to be significantly greater than that between 44% and 48%, and between 52% and 56%.

Table 2: Mixed effects logistic regression predicting safer choice (Study 2)

	Dependent variable:	
	Safer choice	
	(1)	(2)
	Partial	Full
Order	−0.011	−0.014
	(0.029)	(0.025)
Ladder 2	-0.502^{***}	−0.160
	(0.148)	(0.170)
Left-digit boundary (Ladder 1 only)		0.242^{**}
		(0.097)
Left-digit boundary (Ladder 2 only)	-0.665^{***}	-0.422^{***}
	(0.079)	(0.096)
Ladder 2 × Left-digit boundary (Ladder 1 only)		-0.331^{**}
		(0.138)
Ladder 2 × Left-digit boundary (Ladder 2 only)	1.353^{***}	1.028^{***}
	(0.114)	(0.139)
Constant	1.000^{***}	0.762^{***}
	(0.106)	(0.120)
Observations	8,024	10,030
Log Likelihood	−4,409.019	−5,399.392
Akaike Inf. Crit.	8,830.038	10,814.780
Bayesian Inf. Crit.	8,871.980	10,872.490
Note:	*p<0.1; **p<0.05; ***p<0.01	

On each separate page, participants were told to imagine that they are presented with a gamble that offers a [44/48/52/56]% chance of winning \$200, otherwise \$0. They had a chance to receive a sure payout for this gamble, before knowing the eventual outcome. Participants were required to indicate the lowest amount they would accept for this sure payout (i.e., certainty equivalents, or CE), which served as our focal dependent variable. (The instructions can be found in Web Appendix WA9.) Based on pretests, we expected large outliers for this open-ended response. As such, participants' responses were restricted to be between \$0 and \$200, as the maximum payout of the gambles was \$200.

3.5.2 Results.

Descriptive results are shown in Figure 3. To analyze the data we calculated the arithmetic differences between participants' four judgments, corresponding to the differences in certainty equivalents (ΔCE) between probabilities of 44% and 48%, 48% and 52%, and 52% and 56%. As stipulated in our preregistration, we then used Wilcoxon paired signed rank test to determine whether the difference between 48% and 52%, which crossed the 50% boundary, was greater than between 44% and 48%, and between 52% and 56%.

The increases in certainty equivalents between successive levels were ΔCE_{44-48} = 4.19 (SD = 19.23); ΔCE_{48-52} = 7.55 (SD = 21.13); and ΔCE_{52-56} = 3.81 (SD = 20.46) respectively. Wilcoxon paired signed rank tests showed that crossing the 50% boundary (ΔCE_{48-52}) resulted in the biggest increase (V = 93871, $p < .001$ compared to ΔCE_{44-48}; V = 82942, $p < .001$, and compared to ΔCE_{52-56}). The difference between ΔCE_{44-48} and ΔCE_{52-56} was also significant (V = 99169, p = 0.048).

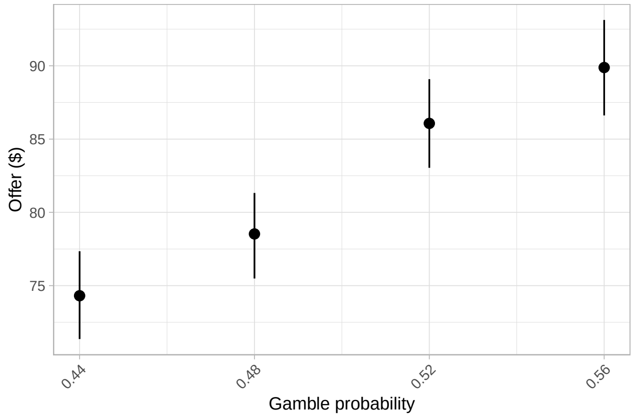

Figure 3: Means and 95% confidence intervals of the offer distributions across different gamble probabilities (Study 3)

3.6 Study 4 — Verbal Labels

In the previous studies, we explored probability sensitivity in contexts where risks were expressed numerically. In Study 4, we aim to build on the insights from the previous studies and explore whether similar categorical boundaries are formed when uncertainty is expressed verbally. Because there is no normative benchmark by which to objectively map verbal expressions of uncertainty, this study does not directly assess probability weighting. But, if the unique probability weighting patterns observed in the previous studies are due to a broader feature of how the mind categorizes uncertainty, then we should be able to observe analogous patters with verbal labels. We expect these verbal labels to sit in semantic clusters, and sensitivity to the difference between verbal labels to be greater across such clusters than within a cluster.

3.6.1 Pretest Procedure and Results.

To minimize the potential of cherry-picking verbal labels, we selected a set of verbal labels from Wallsten et al. (1993a). The authors investigated a set of 11 phrases expressing uncertainty: *impossible*, *doubtful*, *slight chance*, *unlikely*, *improbable*, *tossup*, *probable*, *likely*, *good chance*, *sure*, and *certain*. Importantly, Wallsten et al. (1993a) selected these verbal labels independent of the current paper's hypothesis.

We first conducted a pretest in order to establish the ordinal rank of these labels and identify potential categorical boundaries between them. Participants (n = 400, 57.3% female; M_{age} = 37.85, SD = 12.17, MTurk) were tasked to rank the 11 labels, which were listed in a randomized order, in a way they deemed most appropriate. Participants established the rank of the 11 labels using a drag-and-drop interface. We then calculated the mean rank of each label, with smaller rank number being less likely. We enforced this rank directionality by coding their responses such that that the rank of *impossible* was always smaller than the rank of *certain*. The mean rank of the labels is shown in Figure 4.

Using verbal labels to express uncertainty could arguably result in less well-defined categorical boundaries, compared to when risks are expressed numerically. Expressing risks numerically should lend itself to clearer clustering and formation of categorical boundaries. For example, a gamble pair involving a 34% chance of A occurring and a 42% chance of B occurring involves a jump in the left digit—from '3' to '4'. This potentially signifies a change in categorical boundary, as opposed to when the pair involves a 30% and a 38% chance, where the left digit remains the same at '3', despite the same objective difference (i.e., 8% chance). In the case of verbal labels, the exact distribution of the categories is less clear. In this pretest, we attempted to identify categorical boundaries using the following procedure. After establishing the rank order of the labels, we calculated the mean signed differences between consecutive labels (Figure 5, left panel). Categorical boundaries likely happen between labels with larger mean signed rank differences (i.e., more consistently ranked by the participants). We thus reordered the label-pairs (i.e., '*slight chance—probable*' as the most likely categorical boundary, '*improbable—tossup*' as the next likely, and so on) according to this measure.

3.6.2 Main Study Procedure.

Participants (n = 800, 50.2% female; M_{age} = 37.13, SD = 11.77, MTurk) evaluated 20 randomly selected gamble pairs from a total set of 80 (see Web Appendix WA10 for study details). The procedure for forming the gamble ladder was similar to the laddering paradigm described in previous studies, barring a few crucial differences. Similar to previous studies, the gamble pairs consisted of riskier (R) and safer (S) options. The R option involved a 1% chance of winning \$50 ($Payout_{extra}$) and a chance of winning \$5 ($Payout_{common}$) at varying levels of uncertainty,

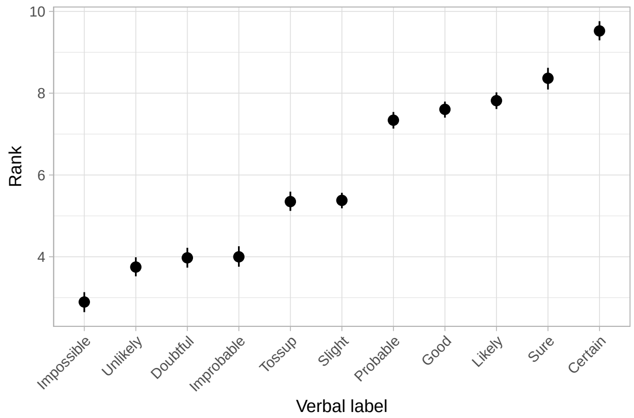

Figure 4: Mean rank of verbal labels (Study 4 pretest)

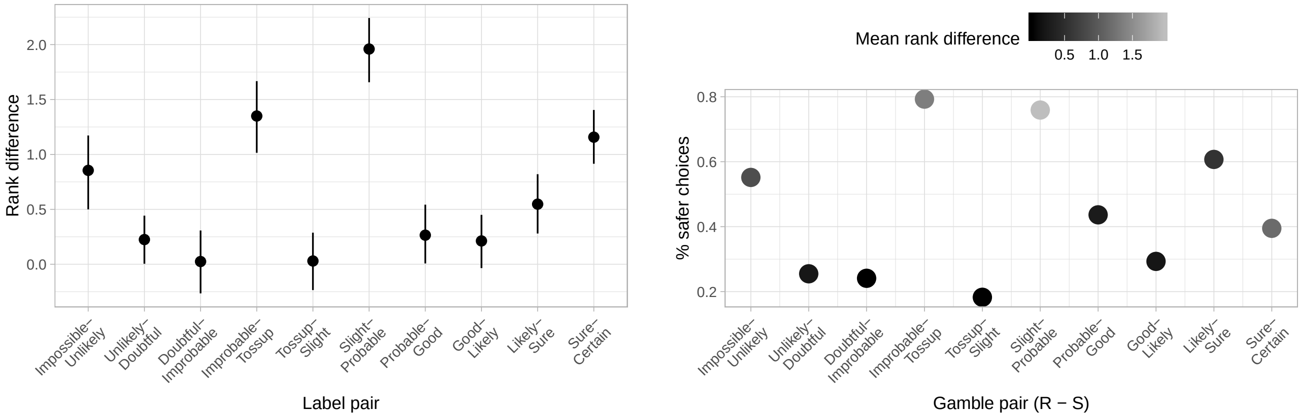

Figure 5: Mean signed rank difference between consecutive label pairs (Study 4 pretest—left) and proportion of safer choices across corresponding gamble pairs (Study 4 main—right). The left plot shows mean signed differences in rank (1-11) between consecutive label pairs in the pretest study; the right plot shows the proportion of safer choice of the gambles using the same consecutive label pairs in the main study. The darker gradient on the left plot signifies less category overlap (i.e., more likely to have a categorical boundary), while the lighter gradient signifies overlapping categories.

which were expressed verbally. The S options presented chances of the same $Payout_{common}$ ($5) at varying levels of uncertainty, which were also expressed verbally. Thus, the R option read: “Chance to win $50: 1% and Chance to win $5: [$impossible/doubtful/...$]” whereas the S option read: “Chance to win $5: [$impossible/doubtful/...$]”. This study was preregistered on AsPredicted: `https://aspredicted.org/blind.php?x=9k4vs6`.

3.6.3 Main Study Results.

For the purpose of this manuscript, we present only participants’ responses to the 10 focal gamble pairs where (a) the verbal labels for the $Payout_{common}$ of the two options are always consecutively ranked, and (b) the riskier option label always has the lower rank than the safer option. (Note that the rank order of the labels was independently established by the pretest study; see Figure 4.) A supplemental analysis using all non-dominant gamble pairs provides similar conclusions and can be found in the Web Appendix (Table WA4).

The proportions of safer choices across the ten focal gamble pairs are presented in the right panel of Figure 5. We observe that preference for the safer choice in gamble pairs with consecutively-ranked labels in the main study were largely consistent with the average rank difference between these labels in the pretest study. For example, in the pretest study, the mean signed rank difference between *slight chance* and *probable* was largest among all label pairs (Δrank = 1.96), meaning that participants most consistently ranked the two labels in the same order. As such, it is likely that *probable* is perceived to be categorically distinct from *slight chance*. In the main study, we can see a corresponding increase in safer choice preference, indicating that the safer option is more attractive with *probable* common payout than the safer choice of *slight chance* common payout and a 1% chance of extra payout.

Conversely, the mean signed rank difference between *tossup* and *slight chance* in the pretest study was second smallest (Δrank = 0.03), meaning that participants did not consistently rank the two labels in the same order. Thus, it is likely that *tossup* and *slight chance* belong to the same category as expressions of uncertainty. In the main study, we can see a lower proportion of safer choice, indicating that the safer option of *slight chance* of the common payout and the riskier choice of *tossup* of the common payout were comparable to participants, rendering the additional 1% chance of the extra payout in the riskier option more attractive.

We conducted a mixed effects logistic regression with participant-specific intercepts, and a dummy variable indicating whether a particular gamble pair crossed a categorical boundary. Results are presented in Table 3. Since no objectively-defined criterion for a categorical boundary with verbal labels exists (unlike, for example, left-digit change), we included dummy variables indicating label pairs. Each successive model applied a finer criterion for categorical boundaries determined by the mean signed rank difference in the left panel of Figure 5. For example, column 2 is a model that assumes only one categorical boundary (i.e., between *slight chance* and *probable*). The model in column 3 assumes two categorical boundaries (i.e., between *slight chance* and *probable* and between *improbable* and *tossup*), and so on. For models in Table 3, we observe a significant positive coefficient for most of the categorical boundaries. That is, whether we assume one categorical boundary or nine, the model indicates that for those gamble pairs that cross the assumed categorical boundary/boundaries, preference for the safer option increases. This result is consistent with the categorical-boundary effects observed in the previous studies. Additionally, these results demonstrate that the effect of categorization reflects a psychological process that extends beyond the domain of numbers.

It is important to acknowledge that the observed findings could be argued as tautological. That is, if verbal expressions of uncertainty are unequally spaced in decision-makers' minds, then that unequal spacing would produce the non-monotonicities in Figure 5. At the same time, these verbal expressions represent how individuals attempt to communicate the uncertainty they face in their environment. It is possible that individuals experience outcomes that are not uniformly spaced along the probability spectrum (e.g., Stewart et al. 2006). Or, if probabilities are experienced evenly across the probability space, why has language adapted to cluster verbal expressions of uncertainty in the manner observed in Figure 5? While these points are sobering for interpreting the current results, they do not negate the focal message of the findings documented in this study – that categorical boundaries produce increased probability sensitivity.

Table 3: Mixed effects logistic regression predicting safer choice (Study 4 main)

	Dependent variable:									
	Safer choice									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Order	−0.107**	−0.079*	−0.093*	−0.093*	−0.087*	−0.080	−0.086	−0.086	−0.090*	−0.086***
	(0.047)	(0.048)	(0.051)	(0.051)	(0.051)	(0.053)	(0.054)	(0.054)	(0.054)	(0.002)
Slight- Probable		1.521***	1.788***	1.809***	1.966***	2.276***	2.510***	2.536***	2.723***	2.535***
		(0.183)	(0.192)	(0.194)	(0.197)	(0.205)	(0.213)	(0.219)	(0.238)	(0.002)
Improbable- Tossup			2.014***	2.033***	2.197***	2.527***	2.744***	2.771***	2.948***	2.771***
			(0.201)	(0.203)	(0.207)	(0.217)	(0.223)	(0.229)	(0.245)	(0.002)
Sure- Certain				0.140	0.283*	0.554***	0.781***	0.807***	0.983***	0.790***
				(0.164)	(0.168)	(0.175)	(0.182)	(0.189)	(0.207)	(0.002)
Impossible- Unlikely					0.964***	1.256***	1.490***	1.520***	1.685***	1.494***
					(0.167)	(0.175)	(0.184)	(0.193)	(0.208)	(0.002)
Likely- Sure						1.548***	1.757***	1.787***	1.952***	1.774***
						(0.185)	(0.191)	(0.200)	(0.215)	(0.162)
Probable- Good							0.975***	1.003***	1.174***	0.975***
							(0.179)	(0.187)	(0.204)	(0.002)
Unlikely- Doubtful								0.109	0.271	0.077***
								(0.204)	(0.218)	(0.002)
Good- Likely									0.492**	0.298***
									(0.216)	(0.002)
Tossup- Slight										−0.393***
										(0.002)
Constant	−0.203***	−0.339***	−0.559***	−0.578***	−0.735***	−1.045***	−1.263***	−1.291***	−1.461***	−1.276***
	(0.048)	(0.052)	(0.059)	(0.064)	(0.072)	(0.087)	(0.099)	(0.113)	(0.139)	(0.002)
Observations	1,963	1,963	1,963	1,963	1,963	1,963	1,963	1,963	1,963	1,963
Log Likelihood	−1,347.082	−1,305.880	−1,239.768	−1,239.406	−1,222.086	−1,183.093	−1,168.328	−1,168.187	−1,165.617	−1,164.540
Akaike Inf. Crit.	2,700.163	2,619.760	2,489.536	2,490.812	2,458.172	2,382.187	2,354.657	2,356.374	2,353.235	2,353.079
Bayesian Inf. Crit.	2,716.910	2,642.089	2,517.447	2,524.305	2,497.248	2,426.844	2,404.897	2,412.196	2,414.639	2,420.066
Note:										*p<0.1; **p<0.05; ***p<0.01

4 Modelling the Impact of Crossing Categorical Boundaries on Probability Weighting

In the previous section of this paper we provided experiment-based evidence of these categorical-boundary effects in probability weighting. In this section, we develop a behavioral model to assess the extent to which categorization shapes the probability weighting function. Our theoretical development, and the studies above, demonstrated that probability weighting is contextual. Crossing categorical boundaries changes how individuals weight probabilities. Budescu et al. (2011) argue that different contexts (type of events) impact the probability weighting functions used by participants. In turn, this could lead to preference reversals, and impact the elicitation and encoding of subjective probabilities. In the same vein, we encode subjective probability as context-specific, and model different probability weighting functions depending on whether outcome probabilities cross a categorical boundary. Here, we build a behavioral model that benchmarks against the previous literature by using the extensively validated Prospect Theory framework (Tversky and Kahneman 1992). We then integrate our categorization-based perspective on probability weighting. We test our behavioral model on the data sets collected for Study 1. Additionally, we test our behavioral model using the original data from Wu and Gonzalez (1996), which inspired the experimental design of our studies.[3] Note, all choices across our experiments were in the gains domain.

Before presenting the development of the behavioral model, we explored the Wu and Gonzalez (1996) data in Figure 6. The data set consisted of five ladders, with variation across rungs in the values of p, or the $Payout_{common}$ probabilities of the safer options, as well as in the q values, or the increase in $Payout_{common}$ probabilities for the safer options. Wu and Gonzalez (1996) evidence the inverse S-shape of the probability weighting function by demonstrating that the percentage of choice for the safer option was convex as a function of the safer choice probability; this is clearly evident in Figure 6. While Wu and Gonzalez (1996) did not intend to study probability sensitivity across categorical boundaries, their data appear consistent with our studies. In particular, we indicate gamble pairs whose $Payout_{common}$ probabilities cross left digit boundaries with dark red triangles. Compared to those gambles whose $Payout_{common}$ probabilities do not cross categorical boundaries indicated by light grey circles, the red triangles appear notably higher, suggesting increased probability sensitivity. (In Web Appendix WA12, we present a similar mixed effects logistic regression analysis as those reported in Studies 1 and 2. Consistent with our studies, we find a significant categorical-boundary effect in the Wu and Gonzalez (1996) data.)

4.1 Model Development

We model participants' choices between two prospects. Prospect U_j is defined as:

$$U_j(p_1,\, x_1;\, p_2,\, x_2) = w(p_1)v(x_1) + [w(p_1 + p_2) - w(p_1)]v(x_2) \tag{1}$$

Our behavioral model embeds a value function $v(x)$ and a probability weighting function $w(p)$, in line with Prospect Theory. Following the ladders structure presented in §3.2, Equation 1 translates into utility $U(R) = w(1\%)v(Payout_{extra}) + [w(1\% + p\%) - w(1\%)]v(Payout_{common})$ for Option R, and utility $U(S) = [w(p + q\%)]v(Payout_{common})$ for Option S.

[3] We are grateful to the authors for sharing the data with us.

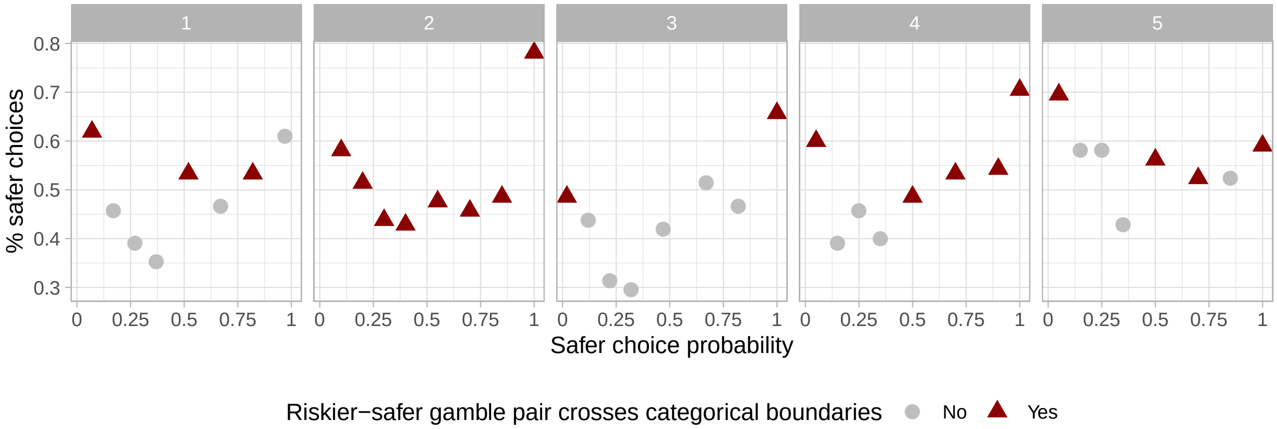

Figure 6: Proportion of safer choices across gambles, in the Wu and Gonzalez (1996) data, across the five ladders

We used a power specification for the value function:

$$v(x) = x^{\alpha} \tag{2}$$

where parameter α indicates the curvature of the value function. We assumed that probabilities are weighted following the classic Tversky and Kahneman (1992) model, to allow for a straightforward comparison of our results with those reported in Wu and Gonzalez (1996).[4]

The one-parameter probability weighting function is:

$$w(p) = \frac{p^{\gamma}}{(p^{\gamma} + (1-p)^{\gamma})^{1/\gamma}} \tag{3}$$

We specified the probability of choosing prospect U_j using a categorical logit model[5]:

$$Pr(U_j) = \frac{exp(\phi U_j)}{\Sigma_J exp(\phi U_j)} \tag{4}$$

We accounted for the sensitivity of participants' choices to the prospects' utility via the parameter $\phi \geq 0$. When ϕ is 0, choices between prospects are random. As ϕ increases, participants' choices are more sensitive to the values of the prospects.

4.2 A Categorization-Based Probability Weighting Specification

Prospect Theory assumes that individual probabilities are imputed in an absolute manner. That is, the $Payout_{common}$ probability for the R option doesn't directly impact the subjective representation of the probability for the S option. Recent advances in models of multi-alternative-multiattribute choice explicitly model the value for one option as the consequence of a relative comparison to

[4] In appendix WA15, we test the robustness of our results to the Prelec (1998) probability weighing specification.

[5] The model is equivalent to the Luce choice rule (Luce 1959) commonly employed and shown to lead to best model fit when estimating parameters (Wu and Gonzalez 1996, Stott 2006, Toubia et al. 2013)

other options (for a review see Turner et al. 2018). Similarly, social utility models, such as the Equity-Reciprocity-Competition (ERC) model (Bolton and Ockenfels 2000) and the Fehr-Schmidt model (Fehr and Schmidt 1999) document decision makers' both absolute (pecuniary) and relative (fairness) payoff considerations. In an ultimatum game, bargainers' utility increases in monetary gain, but decreases as their payoff deviates below a fair settlement. More directly, de Langhe and Puntoni (2015) argued that decision makers rely not on the absolute outcomes in gambles but on the ratio between gains and losses when making decisions under risk.

In line with the above discussion, we contend that probabilities are interpreted in both an absolute and a comparative manner. We conceptualized the absolute component as is traditionally done in Prospect Theory. The comparative component is a weight dependent on the difference between the $Payout_{common}$ probabilities of the R versus the S option. Specifically, we suggest that when this safer choice probability crosses a categorical boundary, probability sensitivity increases.

We argue that nonlinearities in probability weighting are due to changes in probability sensitivity within and across categorical boundaries. Despite the fact that the classic probability weighting function only inputs probabilities in an absolute manner, it nevertheless captures this comparison of within and cross categorical boundaries at probabilities 0 and 1. Because absolute and comparative evaluations at 0 and 1 provide the same information, 0 and 1 will *always* act as reference points. But, as seen in our previous studies, for probabilities between 0 and 1, absolute and comparative evaluations do not necessarily provide equivalent information. In particular, categorical boundaries depend on the comparison between the $Payout_{common}$.

To model both the absolute and comparative nature of our account, we augmented the probability weighing function to allow for changes in probability weights as gamble pairs cross categorical boundaries. As most of the variation in probability weights in the classic model is driven by their comparison with the reference points at 0 and 1, the function is always crossing categorical boundaries.

The studies reported in Section 3 documented a lower preference for the safer option in gamble pairs where probabilities do not cross categorical boundaries. This suggests that our model should allow for a decrease in the probability weight of the safer option when a probability pair lies within a categorical interval. Probabilities are similarly weighted when the gamble pair crosses categorical intervals, as previously implemented by probability weighting models. We accomplished this by specifying a parsimonious two-parameter model embedded in the Tversky and Kahneman (1992) specification, as follows:

$$w(p) = \frac{p^{\gamma'}}{(p^{\gamma'} + (1-p)^{\gamma'})^{1/\gamma'}} \text{, where:}$$

$$\gamma' = \begin{cases} \gamma_1 : \text{Weight parameter of the probability for riskier option,} \\ \text{and for the safer option, when the latter crosses a categorical boundary;} \\ \\ \gamma_2 : \text{Weight parameter of the probability for the safer option,} \\ \text{ when gamble probabilities lie within the same categorical boundary.} \end{cases} \tag{5}$$

When probability pairs cross categorical boundaries, the model reduces to the classic Tversky and Kahneman (1992) specification.[6] Within the same categorical interval, the probability of the

[6] We chose to set γ_1 as the weighting function when probabilities cross categorical boundaries, as opposed to the

safer option is weighted differently than the riskier option probability, by the amount identified through γ_2. We expect and test that $\gamma_2 \leq \gamma_1$, to account for the decrease in the weight of the safer option probability.[7]

Figure 7 displays a conceptual representation of the probability weighting function we specify. We plot the probability weighting function for the $Payout_{common}$ of the riskier option overlaid with the probability weighting function for the $Payout_{common}$ of the safer option when q varies between 1% and 10% (i.e., the probability of the $Payout_{common}$ for the safer option crosses a categorical boundary when it is 5% higher than that for the riskier option). Because of the comparative nature of our model, the probability weighting function of the safer option depends on whether the the probabilities of the $Payout_{common}$ are within or across categorical boundaries. Thus, when considering the influence of probability weighting on preference, differences in probability weighting between the riskier and safer options will lead to increased preference differentiation between the options.

Conceptually, our account presumes a cyclical probability weighting function with multiple cycles. The equations above do not directly model a multi-cycle function. Our model does, however, offer an efficient analogy to our conceptual framework while still maintaining a parsimonious two-parameter structure. We approximated our conceptual model by overlaying two traditional inverse S-shape functions; one for γ_1 and another for γ_2. Because γ' depends on a comparison between the common probabilities for options R and S, the model folds multiple inverse S-shaped functions over itself across the probability space, resulting in the wave-shaped function highlighted in Figure 7. Importantly, this folding happens across the probability space because of the comparisons identifying γ_1 and γ_2.

4.3 Behavioral Model Estimation

We then applied our behavioral model to the experimental data presented in Study 1, and to the Wu and Gonzalez (1996) data. We use a Bayesian framework (Nilsson et al. 2011, Baillon et al. 2020) to analyse the data. For the experimental data in Study 1, we implemented a hierarchical Bayesian model where parameters are estimated at the individual level, but assume that individual parameters are generated from a common population level distribution. The relatively large number of observation per individual (n = 45), combined with variation in the outcome values, allows us to estimate reliable individual level parameters of our behavioral model, as well as the correlations between the parameters.

For the Wu and Gonzalez (1996) data set, we considered the data as being generated by a representative agent, and assume that participants have similar preferences and behaviors. The limited number of observations per individual in each ladder (n = 4) does not permit a reliable

reverse. This is because the shape of the classic probability weighting function is driven by the categorical boundaries at 0 and 1. If indeed there were no additional categorical boundaries between 0 and 1, our model would reduce to the classical model. But, if we specified γ_1 as the weighting function when probabilities do *not* cross categorical boundaries, that implies that probability sensitivity is increased compared to the classic model. This would not be correct since the high probability sensitivity in the classic model is driven by the categorical boundaries at 0 and 1.

[7] Gonzalez and Wu (1999) propose a two-parameter probability weighting function where one parameter governs the elevation of the function, and the second parameter governs its curvature. The specification assumes that probability weights are inputted in an absolute manner. Our two-parameter model holds a different underlying assumption. We allow for probability weights to be context-dependent, and implement categorical jumps at different reference points.

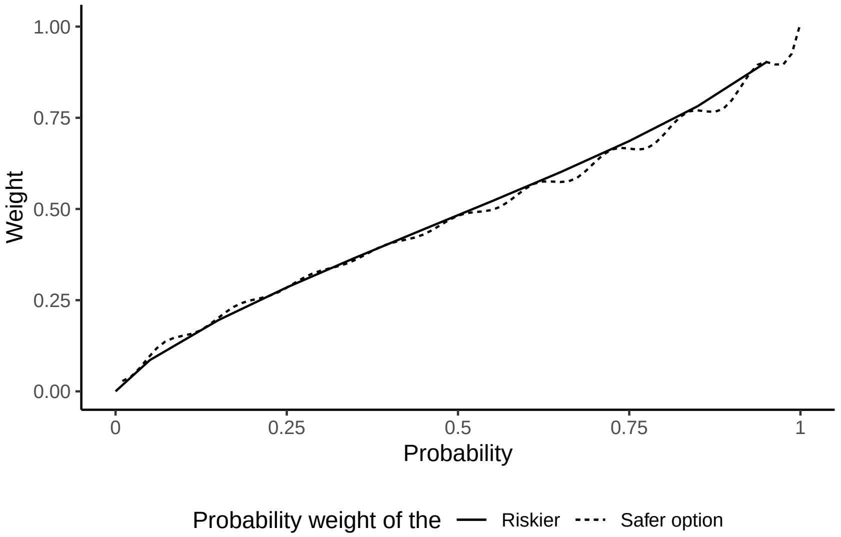

Figure 7: Conceptual representation of the categorization-based probability weighting specification. The objective riskier choice probabilities are set at 0%, 5%, and increasing by 10 percentage points afterwards. The objective safer choice probabilities are between 1% and 10% higher than their riskier choice counterparts, crossing categorical boundaries when 5% higher. The riskier choice probability weight follows the classic probability weighting function (the solid line). The (smoothed) safer choice probability weight is context-dependent, and follows the dotted line. When the objective safer choice probability crosses a categorical boundary, it is estimated following the classic model. When objective probabilities of the gamble pairs remain within two categorical boundaries, the safer choice probability weight is estimated following a probability weighting function with a γ parameter below the one used for riskier choice probabilities ($\gamma_2 \leq \gamma_1$). This effectively allows for a decrease in the safer choice probabilities when gamble pairs remain within the same categorical interval, and proxies the limited sensitivity shown by participants in our studies.

estimation of individual-level parameters, therefore we use an aggregate-level counterpart of the hierarchical Bayesian model. Nonetheless, finding evidence of category-driven probability weighting would solidify our theory, especially since the data set was gathered 24 years prior to this study, and with a different research objective in mind.

Using Bayesian inference, we combined the likelihood of the data with (diffuse) prior information to obtain draws from the posterior distributions of the model parameters. We described parameter estimates in terms of central tendencies (means and medians), and quantiles of their posterior distributions distributions, used to assess Bayesian credible intervals. Appendix WA13 presents the details of our estimation method.

4.4 Model Selection

To test the fit and predictive accuracy of our proposed behavioral model, we benchmarked the categorization-based-two-parameter probability weighting specification (hereafter the "categorization-based model") against the Prospect Theory model embedding a one-parameter probability weighting function (hereafter the "classic model").

4.4.1 In-Sample Fit.

To assess in-sample fit, we discuss likelihood-based and likelihood-free model selection criteria.

We first report the likelihood-based Watanabe-Akaike Information Criterion (WAIC). The measure is computed based on the log-predictive density evaluated at the posterior parameter draws, and corrected for model complexity using the effective number of parameters. Lower values of the WAIC denote better fit.

We also report a set of likelihood-free criteria, and compute the sum of squared error between the observed and predicted choices made by participants. At each level of the safer choice probability, we computed the squared error as the square of the difference between the observed proportion of safer choices and the average choice probabilities predicted by the model. We then summed the squared errors. We use this measure because it allows for direct comparison with the results reported in Wu and Gonzalez (1996).

4.4.2 Out-of-Sample Predictive Accuracy.

We used cross-validation to assess our model's predictive accuracy out of sample. We partitioned the data set into a calibration sub-sample D_{-k} and a validation sub-sample D_k. We do this by removing one observation k per individual from the calibration data set. For the Study 1 data set, our calibration sub-sample D_{-k} thus consists of 44 observations per individual, for the 100 participants. We predicted choices out-of-sample for the observations in our validation sub-sample D_k, which consists of 100 observations, one for each individual. We removed the observation per individual from the calibration data in the randomized order in which participants evaluated the gamble pairs, thus ensuring that the full probability space is covered when estimating the probability weighing function across all individuals. We repeated the validation exercise 45 times, thus ensuring that we predict each observation in our original data set out of sample.

We used the same procedure for the Wu and Gonzalez (1996) data set. Each ladder-specific data set has four observations per individual. We removed one observation per individual from the calibration data set, and predict it out-of-sample based on the parameters estimated using the calibration data.

For each model, the out-of-sample predictive performance was measured by its mean squared error. For both data sets, we computed the squared error as the difference between observed and predicted choice probabilities at the rung level, and average across rungs. Table 4 shows the measures of in-sample fit, and the out-of-sample predictive accuracy.

There is substantial evidence that categorical boundaries explain differences in probability weighing, when estimated using the experimental data in Study 1. Our proposed categorization-based (CB) model fits the data substantially better than the model which embeds the classic probability weighting specification, both in terms of model fit ($WAIC_{Classic}$ = 4,254 vs. $WAIC_{CB}$ = 4,050, $SSE_{Classic}$ = .088 vs. SSE_{CB} = .056), as well as out-of-sample predictive accuracy ($MSE_{Classic}$ = .034 vs. MSE_{CB} = .032).

In the Wu and Gonzalez (1996) data set, our proposed model fits the data better in Ladders 1 and 4 across all measures. The classic model fits similarly to the categorization-based model in Ladders 2 and 3, which is unsurprising since there isn't sufficient variability in left-digit crossing to estimate

Table 4: Model selection

	WAIC		In-sample (SSE)		Out-of-sample (MSE)	
Data sets	Classic[a]	CB[b]	Classic	CB	Classic	CB
Study 1	4,254.7	4,050.0	0.088	0.056	0.034	0.032
Wu and Gonzalez (1996) - Ladder 1	1,155.5	1,154.4	0.034	0.026	0.015	0.015
Wu and Gonzalez (1996) - Ladder 2	1,129.1	1,129.1	0.007	0.007	0.009	0.009
Wu and Gonzalez (1996) - Ladder 3	1,131.1	1,132.4	0.022	0.020	0.020	0.021
Wu and Gonzalez (1996) - Ladder 4	1,145.1	1,143.8	0.024	0.014	0.012	0.012
Wu and Gonzalez (1996) - Ladder 5	1,150.4	1,152.1	0.028	0.028	0.014	0.014

[a]Classic: The Tversky and Kahneman (1992) probability weighting function. [b]CB: Categorization-based, i.e., the probability weighting function specified in Equation 4.2.

the effect of categorical boundaries because Ladder 2 has categorical-boundary crossings for *all* gamble pairs and Ladder 3 has crossings *only* for the gambles that include probabilities of 0 and 1 (see Figure 6). The categorization-based model performs worse than the classic model when fitted to the Ladder 5 data set when looking at likelihood-based measures ($WAIC_{Classic}$ = 1,150.4 vs. $WAIC_{CB}$ = 1,152.1), while the two models performs similarly well judging by the likelihood-free measures ($SSE_{Classic} = SSE_{CB}$ = .028, $MSE_{Classic}$ = .014 = MSE_{CB} = .014).[8]

Overall, the model comparison results support our conceptual idea that probability sensitivity is driven by categorical boundaries. In the following section, we use our model-based approach as a tool for hypothesis testing.

4.5 Behavioral Parameter Estimates

Table 5 summarises the posterior distributions of our model parameters. The estimates are in line with the results reported in the literature. For the Study 1 data, we report the population-level means of all parameter estimates, along with the 95% Bayesian credible intervals. For the Wu and Gonzalez (1996) data set, we report the aggregate-level parameter estimates, along with the 95% Bayesian credible intervals (BCI). The parameters of the classic model are remarkably in line to those reported in Wu and Gonzalez (1996), despite differences in both the specification of the model (the choice sensitivity parameter ϕ was not estimated in Wu and Gonzalez 1996), and the estimation method (Wu and Gonzalez 1996 use a nonparametric technique.).

The parameter γ_2 quantifying the impact of crossing categorical boundaries on choice is lower than the parameter γ_1 in five of the six data sets used for estimation, confirming the effect of crossing categorical boundaries, or the left-digit effect documented by our studies.

[8] To test whether our proposed model best explains the data when the underlying data generating process is aligned with the categorization-based model, we conducted the following simulation study. Using the Study 1 data set and parameter values close to the estimates reported in Table 5, we simulated data following the categorization-based model. We estimated both the classic and the categorization-based models on the simulated data, and compared model fit. The categorization-based model fits the data better than the classic specification. This study also allowed us to test the parametric identification of our model. Further details of this simulation study can be found in Web Appendix WA14.

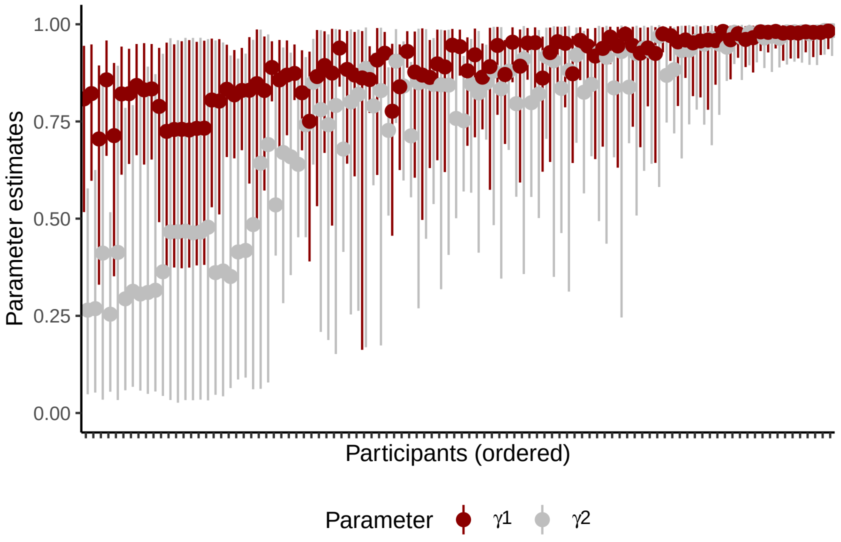

Figure 8: Posterior distributions (means and 95% BCIs) of all individual-specific probability weighting parameters for participants in Study 1. Most individual-level γ_{2i} parameters are on average lower than the γ_{1i} parameters.

In Study 1, participants appear on average slightly risk averse ($\bar{\alpha}$=.85, 95% BCI=[.75, .93]). Parameter γ_2 ($\bar{\gamma}_2$=.82, 95% BCI=[.72, .90]) is lower than γ_1 ($\bar{\gamma}_1$=.91, 95% BCI=[.87, .95]).

Table 5: Parameter estimates: Models with the classic (left) and categorization-based (right) probability weighting functions

	Classic model[a]			Categorization-based (CB) model			
Data sets	γ	α	ϕ	γ_1	γ_2	α	ϕ
Study 1	0.95	0.93	3.12	0.91	0.82	0.85	1.77
	(0.91, 0.98)	(0.86, 0.98)	(1.52, 6.19)	(0.87, 0.95)	(0.72, 0.9)	(0.75, 0.93)	(1.1, 2.59)
Wu and Gonzalez[b]	0.59	0.69	0.63	0.56	0.54	0.53	1.54
Ladder 1	(0.33, 0.79)	(0.31, 1.1)	(0.11, 2.09)	(0.37, 0.72)	(0.35, 0.72)	(0.22, 0.9)	(0.18, 5.23)
Wu and Gonzalez	0.62	0.49	1.46	0.62	0.5	0.49	1.46
Ladder 2	(0.24, 0.86)	(0.22, 0.74)	(0.27, 2.65)	(0.24, 0.86)	(0.13, 0.88)	(0.22, 0.74)	(0.27, 2.65)
Wu and Gonzalez	0.63	0.74	0.44	0.61	0.62	0.76	0.36
Ladder 3	(0.51, 0.71)	(0.57, 0.9)	(0.13, 0.93)	(0.48, 0.7)	(0.48, 0.71)	(0.55, 0.94)	(0.09, 0.84)
Wu and Gonzalez	0.67	0.53	0.9	0.6	0.57	0.4	1.08
Ladder 4	(0.37, 0.88)	(0.21, 0.82)	(0.2, 2.01)	(0.34, 0.85)	(0.15, 0.93)	(0.13, 0.7)	(0.19, 2.78)
Wu and Gonzalez	0.84	0.46	1.45	0.81	0.77	0.37	1.59
Ladder 5	(0.69, 0.94)	(0.16, 0.8)	(0.58, 3.17)	(0.66, 0.91)	(0.6, 0.9)	(0.12, 0.66)	(0.52, 3.49)

[a]Means of posterior distributions, with 95% Bayesian credible intervals in brackets. [b]The Wu and Gonzalez (1996) data set.

Figure 8 shows the posterior distributions of the individual-specific probability weight parameters for participants in Study 1. Results indicate that most individual-level γ_{2i} parameters are on average lower than the γ_{1i} parameters.

To test whether the γ_2 parameters are significantly lower than the γ_1 parameters on average, we computed point estimates of the individual-specific probability weight parameters. An ANOVA test revealed that the γ_2 parameters are significantly lower than the γ_1 parameters ($F(1,198) = 37.77$, $p < .001$).

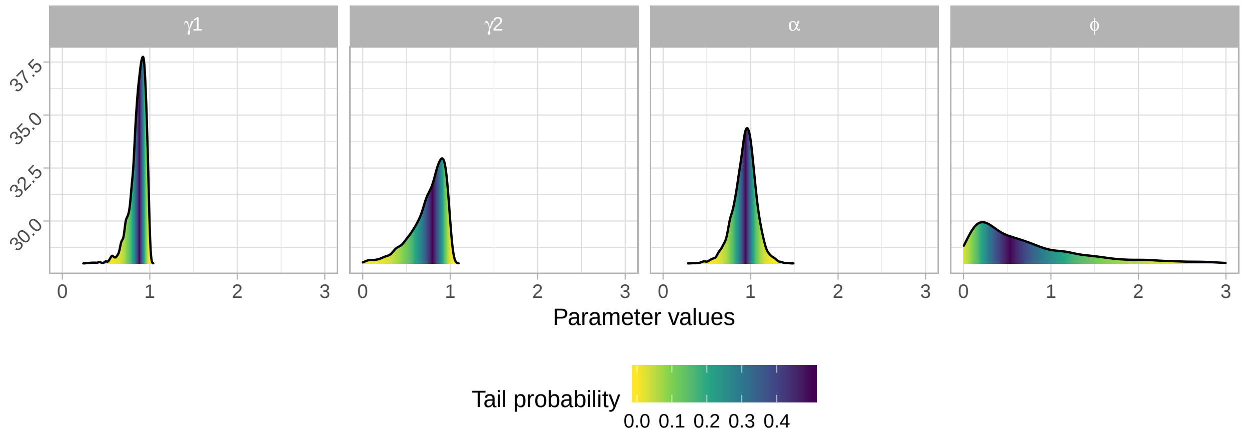

Figure 9: Posterior probability density functions of the model parameters for participant 28

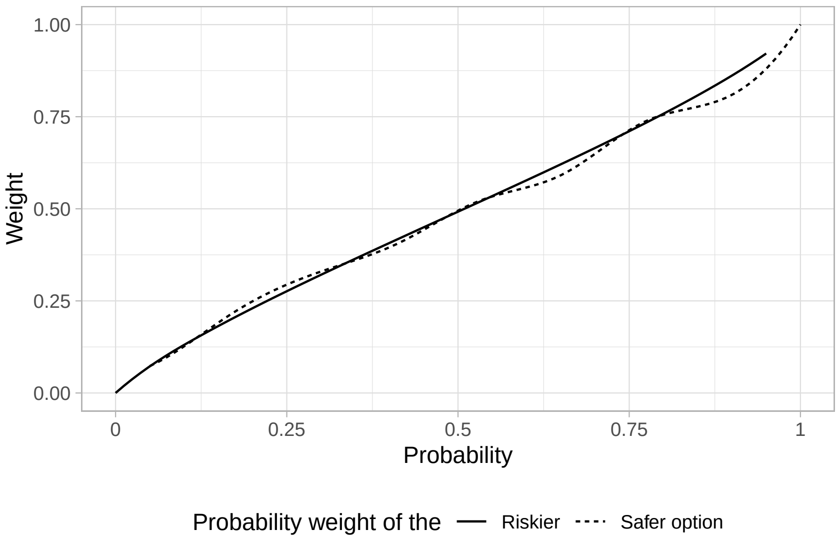

Figure 10: Categorization-based probability weighting specification for participant 28

Figures 9 and 10 display the probability densities for all model parameters, and illustrate the categorization-based probability weighting function for participant 28, with preferences similar to the population averages.

5 Discussion

The current manuscript offers a new perspective of the probability weighting function. Original conceptualizations of probability weighting assumed the ubiquitous inverse S-shape was the result of categorical distinctions between 0, 1, and the probabilities in between (Kahneman and Tversky 1979, Tversky and Kahneman 1992, Wu and Gonzalez 1996, Camerer and Ho 1994, Gonzalez and Wu 1999, Brandstätter et al. 2002, Prelec 1998, Mukherjee 2010). The literature has long presumed that the mind interprets a 0% chance as “not happening" and a 1% chance as categorically distinct from “not happening." Similarly, a 100% chance is “certainly happening" but a 99% chance is categorically distinct from “certainly happening." Said another way, because 0% and 100% are categorically distinct from the rest of the probability spectrum, they act as reference points, with diminishing marginal sensitivity away from those reference point, resulting in the traditional inverse S-shaped function. We fully agree with the proposition that categorical distinctions produce reference points which result in nonlinear probability sensitivity. However, the crucial contribution of the current manuscript is in testing the tacit assumption that decision makers only have category-based reference points at 0 and 1. We presented experimental evidence in four empirical studies, demonstrating that probability sensitivity increased substantially when probabilities crossed categorical boundaries. Additionally, our model-based approach offered convergent evidence consistent with this proposition. Below, we outline a few key implications of our account on models of preference.

5.1 Contextually Determined Reference Points in the Probability Space

In Study 1, we found higher probability sensitivity when the probabilities crossed a categorical/left-digit boundary. Importantly, evidence from Study 2 suggests that the categorical boundaries a decision maker relies on for interpreting uncertainty depend critically on the context of the choice environment. We consider this notion highly consistent with the constructed preference perspective that has become pervasive within theories of judgment and decision making (Payne et al. 1993, Slovic 1995, Lichtenstein and Slovic 2006). While the notion of contextually-driven preferences is ubiquitous in the decision-making literature it is not always sufficiently descriptive. In the following sections we highlight a few particular contexts whereby categorical boundaries influence decisions.

5.1.1 Numeric Risks Whose Categories Are Derived from the Stimuli.

As seen from Study 1, when making choices between gambles whose risks are expressed numerically, the left digit of the probability can act as a clear categorical boundary. But as evidenced from Study 2, there need not always be a categorical boundary at every 10% along the probability spectrum. Instead, a particular probability for the common outcome (e.g., 30%) can act as a bigger categorical boundary in one context (Ladder 1) than in another (Ladder 2). That is, in a context whereby probabilities differ by, for instance, 14%, gamble pairs whose probabilities differ by 2 left digits will feel more different than those that differ by 1. Conversely, if decision makers are evaluating risks that differ by very small amounts (e.g., 3.8% vs. 4.0%) then we could expect to see categorical boundaries at every 1% interval in the probability space. Thus, the precision of the particular context will govern the relevant width of the category.

5.1.2 Numeric Risks Whose Categories Are Inferred from Memory.

While the current studies explore contexts where decision makers can make comparisons between gamble pairs, not all decisions are made in joint evaluation (see Hsee et al. 1999, Hsee and Zhang 2010). Often when decisions are made in isolation, decision makers will pull relevant contextual information from memory. For example, Birnbaum (1999) showed that when evaluating “how large is the number 9?” on a 10-point scale ranging from *very very small* to *very very large*, participants offered a mean judgment of 5.13. But, another group of participants evaluating “how large is the number 221?” offered a substantially lower mean judgment of 3.10. The reason is that when evaluating “9” in isolation, decision makers will often infer a context of 1-10. Within this categorical context, 9 is relatively large. But, when evaluating “221” in isolation, decision makers will often infer a context of 1-1000. Within this categorical context, 221 is relatively small. The reason is that we have more experience interacting with 9s in the 1-10 context than in the 1-1000 context and more experience interacting with the 221s in the 1-1000 context than in the 1-10 context. Because of this asymmetry in experience, when seeing the “9” in isolation, decision makers are far more likely to sample the 1-10 context from memory (Stewart et al. 2006). Therefore, in any given context, the numbers previously experienced likely will define the contextual category even when choices are made in isolation. For example, if a decision maker has a lot of experience with gambles that involve roughly-equal chance binary events they may naturally form a reference point near 50% such that a novel risky choice involving a chance of 48% would feel substantially lower than a 52% chance.

5.1.3 Non-numeric Risks and Uncertainty.

Our account is not limited to decisions under risk with numeric probabilities. Study 4 demonstrated that verbal phrases commonly used to express uncertainty also exhibit categorical boundaries and these categorical boundaries result in greater sensitivity when uncertain options cross these boundaries. These results suggest that the current research has implications for decisions under uncertainty as well.

In addition, risk information can be expressed numerically or non-numerically. For example, past work has demonstrated very different risk preferences when risks are expressed numerically versus graphically (Stone et al. 1997, 2003). Whereas numeric risks offer efficient categorization along the left digit of the number, graphical presentations (e.g., pie charts) rely on perceptual cues. In such cases, we would expect perceptually-driven categorical boundaries like those reviewed by Hollands and Dyre (2000) to greatly influence the shape of the probability weighting function. For example, differentiating whether a pie chart is slightly more or less than 25%, 50%, or 75% is much easier than differentiating whether it is more or less than 21%, 46%, or 71%. As such, we would expect greater probability sensitivity across the former thresholds than across the latter.

Another important domain is how probabilities are categorized when they are experienced. Considerable research over the past two decades has looked at difference in probability weighting when probabilities are expressed numerically versus learned through experiential sampling (Hertwig et al. 2004, Hertwig and Erev 2009, Wulff et al. 2018). One potential route for exploring these differences in probability weighting is to investigate how decision makers categorize risks when sampling them experientially. Do decision makers categorize experienced risks as “not happening,” “maybe happening”, and “certainly happening,” or are there other intermediate categories as well?

5.2 Generalizing Beyond Risk

It should be clear at this point that the insights in the current paper apply to contexts beyond probability weighting. In principle, all quantities can be interpreted relative to a particular context (for a review see Parducci 1965), and these contexts are efficiently organized within distinct categorical boundaries (Pelham et al. 1994, Lembregts and Van Den Bergh 2019). For example, Allen et al. (2017) explored marathon running times and found evidence of multiple reference points in runners’ preferences. In particular, marathon running times clustered just before distinct time horizons (e.g., 3 hours, 3.5 hours, 4 hours, etc...) and dropped immediately after those horizons. Runners often set goals at these concrete time horizons, and these goals can act as reference points (Heath et al. 1999). Similar patterns have been documented for baseball and SAT performance as well (Pope and Simonsohn 2011). Importantly, whether considering the value function, goal progress, or any other relevant domain, categorical boundaries need not occur at each left digit. Instead, categorical boundaries are endogenous markers within any particular stimulus set. The left digit can be an efficient categorical boundary for numeric quantities but other quantitative expressions (e.g., verbal or physical) will have unique categorical structures. That is, the take-home-message of this research extends beyond the left digits of numbers and focuses more broadly on how the mind categorizes quantities more broadly.

5.3 On the “Shape" of the Probability Weighting Function

We propose a novel functional form for the probability weighting function. But at the same time, we suppose that there is, in fact, no stable functional form. We suggest that categorical boundaries produce reference points in the probability space, resulting in greater sensitivity for probabilities that cross these boundaries. Conceptually this produces a cyclical power function — like the traditional inverse S-shape function — but with more cycles. The observation that the number of reference points is contextually determined therefore means that the number of cycles in the function is similarly contextually determined.In sum, there is no set functional form for probability weighting, but its form is an emergent property of the particular stimulus context.

6 Conclusions

The current paper follows a long line of judgment and decision-making research that attempts to understand the nature of human preference via theories of human perception. The perceptual system not only offers an efficient approximation of decision processes, but likely acts as a direct input into decision. We put forth a simple but powerful point: the probability weighting function is shaped by categorical perceptions and the result is a function that needs not be inverse S-shaped.

References

Allen EJ, Dechow PM, Pope DG, Wu G (2017) Reference-Dependent Preferences: Evidence from Marathon Runners. *Management Science* 63(6):1657–1672.

Baillon A, Bleichrodt H, Spinu V (2020) Searching for the Reference Point. *Management Science* 66(1):93–112.

Betancourt MJ, Girolami M (2015) Hamiltonian Monte Carlo for Hierarchical Models. *Current Trends in Bayesian Methodology with Applications* 79(30), arXiv: 1312.0906.

Birnbaum MH (1999) How to show that 9> 221: Collect judgments in a between-subjects design. *Psychological Methods* 4(3):243.

Bolton GE, Ockenfels A (2000) ERC: A Theory of Equity, Reciprocity, and Competition. *The American Economic Review* 90(1):166–193.

Brandstätter E, Kühberger A, Schneider F (2002) A cognitive-emotional account of the shape of the probability weighting function. *Journal of Behavioral Decision Making* 15(2):79–100.

Budescu D, Abbas A, Wu L (2011) Does probability weighting matter in probability elicitation? *Journal of Mathematical Psychology* 55(4):320–327.

Budescu DV, Weinberg S, Wallsten TS (1988) Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance* 14(2):281–294.

Camerer CF, Ho TH (1994) Violations of the betweenness axiom and nonlinearity in probability. *Journal of Risk and Uncertainty* 8(2):167–196.

Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A (2017) Stan: A probabilistic programming language. *Journal of statistical software* 76(1).

de Langhe B, Puntoni S (2015) Bang for the Buck: Gain-Loss Ratio as a Driver of Judgment and Choice. *Management Science* 61(5):1137–1163.

Dehaene S (2011) *The number sense: How the mind creates mathematics, Rev. and updated ed* (New York, NY, US: Oxford University Press), ISBN 978-0-19-975387-1.

Erlick DE (1964) Absolute judgments of discrete quantities randomly distributed over time. *Journal of Experimental Psychology* 67(5):475–482.

Fechner GT (1860) *Elemente der Psychophysik*, volume 2 (Leipzig: Breitkopf und Härtel).

Fehr E, Schmidt KM (1999) A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics* 114(3):817–868.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) *Bayesian Data Analysis* (CRC Press), ISBN 978-1-4398-9820-8.

Gelman A, Rubin DB (1992) Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* 7(4):457–472.

Gonzalez R, Wu G (1999) On the Shape of the Probability Weighting Function. *Cognitive Psychology* 38(1):129–166.

Heath C, Larrick RP, Wu G (1999) Goals as reference points. *Cognitive psychology* 38(1):79–109.

Hertwig R, Barron G, Weber EU, Erev I (2004) Decisions from experience and the effect of rare events in risky choice. *Psychological science* 15(8):534–539.

Hertwig R, Erev I (2009) The description–experience gap in risky choice. *Trends in cognitive sciences* 13(12):517–523.

Hoffman MD, Gelman A (2014) The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research* 15(1):1593–1623.

Hollands JG, Dyre BP (2000) Bias in proportion judgments: The cyclical power model. *Psychological Review* 107(3):500–524.

Hsee CK, Loewenstein GF, Blount S, Bazerman MH (1999) Preference reversals between joint and separate evaluations of options: a review and theoretical analysis. *Psychological bulletin* 125(5):576.

Hsee CK, Zhang J (2010) General evaluability theory. *Perspectives on Psychological Science* 5(4):343–355.

Kahneman D, Tversky A (1979) Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47(2):263–291.

Lembregts C, Van Den Bergh B (2019) Making Each Unit Count: The Role of Discretizing Units in Quantity Expressions. *Journal of Consumer Research* 45(5):1051–1067.

Lewandowski D, Kurowicka D, Joe H (2009) Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis* 100(9):1989–2001.

Lichtenstein S, Slovic P (2006) *The construction of preference* (Cambridge University Press).

Loewenstein GF, Weber EU, Hsee CK, Welch N (2001) Risk as feelings. *Psychological Bulletin* 127(2):267–286.

Luce RD (1959) On the possible psychophysical laws. *Psychological Review* 66(2):81–95.

Macmillan NA, Creelman CD (2005) *Detection theory: A user's guide, 2nd ed.* Detection theory: A user's guide, 2nd ed (Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers), ISBN 978-0-8058-4230-2 978-0-8058-4231-9.

Manning KC, Sprott DE (2009) Price Endings, Left-Digit Effects, and Choice. *Journal of Consumer Research* 36(2):328–335.

Mukherjee K (2010) A dual system model of preferences under risk. *Psychological Review* 117(1):243–255.

Nakajima Y (1987) A model of empty duration perception. *Perception* 16(4):485–520.

Nilsson H, Rieskamp J, Wagenmakers EJ (2011) Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology* 55(1):84–93.

Olson MJ, Budescu DV (1997) Patterns of preference for numerical and verbal probabilities. *Journal of Behavioral Decision Making* 10(2):117–131.

Parducci A (1965) Category judgment: a range-frequency model. *Psychological review* 72(6):407.

Payne JW, Payne JW, Bettman JR, Johnson EJ (1993) *The adaptive decision maker* (Cambridge university press).

Pelham BW, Sumarta TT, Myaskovsky L (1994) The easy path from many to much: The numerosity heuristic. *Cognitive Psychology* 26(2):103–133.

Pope D, Simonsohn U (2011) Round numbers as goals: Evidence from baseball, sat takers, and the lab. *Psychological science* 22(1):71–79.

Prelec D (1998) The Probability Weighting Function. *Econometrica* 66(3):497–527.

Rosch E (1999) Principles of categorization. Margolis E, Laurence S, eds., *Concepts: Core Readings*, 189–206 (MIT Press).

Rottenstreich Y, Hsee CK (2001) Money, kisses, and electric shocks: On the affective psychology of risk. *Psychological Science* 12(3):185–190.

Slovic P (1995) The construction of preference. *American psychologist* 50(5):364.

Slovic P, Monahan J, MacGregor DG (2000) Violence risk assessment and risk communication: The effects of using actual cases, providing instruction, and employing probability versus frequency formats. *Law and Human Behavior* 24(3):271–296.

Stevens SS (1957) On the psychophysical law. *Psychological Review* 64(3):153–181.

Stewart N, Chater N, Brown GD (2006) Decision by sampling. *Cognitive psychology* 53(1):1–26.

Stone ER, Sieck WR, Bull BE, Yates JF, Parks SC, Rush CJ (2003) Foreground: Background salience: Explaining the effects of graphical displays on risk avoidance. *Organizational behavior and human decision processes* 90(1):19–36.

Stone ER, Yates JF, Parker AM (1997) Effects of numerical and graphical displays on professed risk-taking behavior. *Journal of Experimental Psychology: Applied* 3(4):243.

Stott HP (2006) Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty* Vol.32(No.2):101–130.

Thomas M, Morwitz V (2005) Penny Wise and Pound Foolish: The Left-Digit Effect in Price Cognition. *Journal of Consumer Research* 32(1):54–64.

Toubia O, Johnson E, Evgeniou T, Delquié P (2013) Dynamic Experiments for Estimating Preferences: An Adaptive Method of Eliciting Time and Risk Parameters. *Management Science* 59(3):613–640.

Turner BM, Schley DR, Muller C, Tsetsos K (2018) Competing theories of multialternative, multiattribute preferential choice. *Psychological Review* 125(3):329–362.

Tversky A, Kahneman D (1992) Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty* 5(4):297–323.

Varey CA, Mellers BA, Birnbaum MH (1990) Judgments of proportions. *Journal of Experimental Psychology: Human Perception and Performance* 16(3):613–625.

Wallsten TS, Budescu DV, Zwick R (1993a) Comparing the Calibration and Coherence of Numerical and Verbal Probability Judgments. *Management Science* 39(2):176–190.

Wallsten TS, Budescu DV, Zwick R, Kemp SM (1993b) Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society* 31(2):135–138.

Weber EH (1834) *De pulsu, resorptione, auditu et tactu. Anatationes anatomicae et physiologicae* (Leipzig: Koehler).

Windschitl PD, Wells GL (1996) Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied* 2(4):343–364.

Wu G, Gonzalez R (1996) Curvature of the Probability Weighting Function. *Management Science* 42(12):1676–1690.

Wulff DU, Mergenthaler-Canseco M, Hertwig R (2018) A meta-analytic review of two modes of learning and the description-experience gap. *Psychological bulletin* 144(2):140.

WA7 Experiment Details of Study 1

Participants first read the following at the beginning of the task:

> In this task, you will be presented with a series of short binary-choice questions. Each question involves a choice between two gambles. The task will be split into three parts, each part consists of 15 binary choice questions. Therefore, you will make 45 choices in total. You will be informed when a specific part begins and ends, and you can move to the next part. We expect the study to take roughly 10 minutes of your time.
>
> BONUS PAY: In addition to the advertised compensation, you may earn a bonus amount. The bonus you may earn depends on the choices you make throughout the task. At the end of the task, of the 45 choices you will make, we will choose one and actually play out the gamble that you choose. If you win the gamble, we will pay out the bonus to your Prolific account.

They then completed three ladders in random order, each of which had 15 rungs, presented in random order on separate pages. The difference between R and S options (i.e., q) is 5%.

Table WA1: Gamble ladder for Study 1 (incentive compatible)

Rung	Bin	Riskier Option	Safer Option	Δ_{LD}
1	1	[.01, 500/250/100p]	[.05, 100/50/20p]	1
2	2	[.01, 500/250/100p; .07, 100/50/20p]	[.12, 100/50/20p]	1
3	2	[.01, 500/250/100p; .14, 100/50/20p]	[.19, 100/50/20p]	0
4	2	[.01, 500/250/100p; .21, 100/50/20p]	[.26, 100/50/20p]	0
5	2	[.01, 500/250/100p; .28, 100/50/20p]	[.33, 100/50/20p]	1
6	3	[.01, 500/250/100p; .35, 100/50/20p]	[.40, 100/50/20p]	1
7	3	[.01, 500/250/100p; .42, 100/50/20p]	[.47, 100/50/20p]	0
8	3	[.01, 500/250/100p; .48, 100/50/20p]	[.53, 100/50/20p]	1
9	3	[.01, 500/250/100p; .55, 100/50/20p]	[.60, 100/50/20p]	1
10	3	[.01, 500/250/100p; .62, 100/50/20p]	[.67, 100/50/20p]	0
11	4	[.01, 500/250/100p; .69, 100/50/20p]	[.74, 100/50/20p]	1
12	4	[.01, 500/250/100p; .76, 100/50/20p]	[.81, 100/50/20p]	1
13	4	[.01, 500/250/100p; .83, 100/50/20p]	[.88, 100/50/20p]	0
14	4	[.01, 500/250/100p; .90, 100/50/20p]	[.95, 100/50/20p]	0
15	5	[.01, 500/250/100p; .95, 100/50/20p]	[1.00, 100/50/20p]	1

Δ_{LD}: difference in left digit

WA8 Experiment Details of Study 2

Participants first read the following at the beginning of the task:

> In this task, you will be presented with 10 questions. On each of the following pages, there will be a scenario with two gamble options. You will have to indicate which option you prefer. Please make sure you take the time to review the scenarios and choose carefully.

They then completed one of the two ladders, each of which had 10 rungs, presented in random order on separate pages. In either ladder, the difference between R and S option (i.e., q) was 14%. The step-size between gamble pair rungs was 8%.

Table WA2: Gamble ladders for Study 2 (context dependent)

Ladder 1				Ladder 2				
Rung	Riskier Option	Safer Option	Δ_{LD}	Rung	Riskier Option	Safer Option	Δ_{LD}	Diff. in Δ_{LD}
1	[.01, $70; .04, $5]	[.18, $5]	1	1	[.01, $70; .10, $5]	[.24, $5]	1	0
2	[.01, $70; .12, $5]	[.26, $5]	1	2	[.01, $70; .18, $5]	[.32, $5]	2	+1
3	[.01, $70; .20, $5]	[.34, $5]	1	3	[.01, $70; .26, $5]	[.40, $5]	2	+1
4	[.01, $70; .28, $5]	[.42, $5]	2	4	[.01, $70; .34, $5]	[.48, $5]	1	−1
5	[.01, $70; .36, $5]	[.50, $5]	2	5	[.01, $70; .42, $5]	[.56, $5]	1	−1
6	[.01, $70; .44, $5]	[.58, $5]	1	6	[.01, $70; .50, $5]	[.64, $5]	1	0
7	[.01, $70; .52, $5]	[.66, $5]	1	7	[.01, $70; .58, $5]	[.72, $5]	2	+1
8	[.01, $70; .60, $5]	[.74, $5]	1	8	[.01, $70; .66, $5]	[.80, $5]	2	+1
9	[.01, $70; .68, $5]	[.82, $5]	2	9	[.01, $70; .74, $5]	[.88, $5]	1	−1
10	[.01, $70; .76, $5]	[.90, $5]	2	10	[.01, $70; .82, $5]	[.96, $5]	1	−1

Δ_{LD}: difference in left digit

WA9 Experiment Details of Study 3

Participants answered four questions on separate pages, presented in random order:

> Imagine you have a gamble that offers a [44/48/52/56]% chance of winning $200, otherwise $0. Someone offers you cash (a sure payout) for this gamble before you know the outcome of the gamble. What is the lowest amount you would accept? Please enter the amount below (numbers only).

WA10 Experiment Details of Study 4 (Main)

The 11 labels expressing uncertainty from Wallsten et al. (1993a) form a total of 55 possible label pairings (each producing two gamble pairs, R-S and S-R). To narrow down the stimulus set, we conducted another pilot test (n = 102), in which each participant completed 22 randomly chosen gamble pairs. Based on their responses, we chose a subset of 40 label pairings that were least lopsided. (An example of lopsided pairing would be *impossible–sure*, where in the pilot test no participant chose the safer option when presented with [R:*sure*, S:*impossible*], while all but one participant chose the safer option for [R:*impossible*, S:*sure*].) These 40 label pairings are shown in Table WA3, resulting in 80 gamble pairs in total for the main study (R-S and S-R). Both the pilot and main studies have instructions similar to Study 2.

Table WA3: Verbal label pairings for Study 4 (main)

Doubtful-Impossible	Probable-Improbable
Slight Chance-Impossible	Likely-Improbable
Unlikely-Impossible	Good Chance-Improbable
Improbable-Impossible	Sure-Improbable
Tossup-Impossible	Certain-Improbable
Probable-Impossible	Probable-Tossup
Slight Chance-Doubtful	Likely-Tossup
Unlikely-Doubtful	Good Chance-Tossup
Improbable-Doubtful	Sure-Tossup
Tossup-Doubtful	Certain-Tossup
Probable-Doubtful	Likely-Probable
Unlikely-Slight Chance	Good Chance-Probable
Improbable-Slight Chance	Sure-Probable
Tossup-Slight Chance	Certain-Probable
Probable-Slight Chance	Good Chance-Likely
Likely-Slight Chance	Sure-Likely
Improbable-Unlikely	Certain-Likely
Tossup-Unlikely	Sure-Good Chance
Probable-Unlikely	Certain-Good Chance
Tossup-Improbable	Certain-Sure

WA11 Regression Analysis on Non-dominant Gamble Pairs in Study 4 (Main)

We include here (Table WA4) an analysis using all non-dominant gamble pairs (i.e., the riskier option's $Payout_{common}$ label is always lower ranked than the safer option's).

Table WA4: Mixed effects logistic regression predicting safer choice (Study 4 main — all non-dominant gamble pairs)

	Dependent variable:									
	Safer choice									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Order	−0.078***	−0.068**	−0.060**	−0.059**	−0.058**	−0.060**	−0.061***	−0.059**	−0.061**	−0.057*
	(0.028)	(0.028)	(0.029)	(0.029)	(0.029)	(0.030)	(0.001)	(0.030)	(0.030)	(0.030)
Δrank^a	−0.935***	−0.707***	−0.448***	−0.395***	−0.260***	0.090	0.322***	0.547***	0.750***	0.192
	(0.034)	(0.039)	(0.043)	(0.046)	(0.050)	(0.060)	(0.001)	(0.094)	(0.110)	(0.163)
Slight-Probable		0.900***	1.069***	1.134***	1.384***	1.789***	1.742***	1.983***	2.163***	2.019***
		(0.074)	(0.078)	(0.081)	(0.089)	(0.097)	(0.001)	(0.115)	(0.126)	(0.130)
Improbable-Tossup			0.984***	1.069***	1.176***	1.693***	2.031***	2.250***	2.486***	2.156***
			(0.073)	(0.077)	(0.079)	(0.093)	(0.001)	(0.123)	(0.140)	(0.156)
Sure-Certain				0.295***	0.575***	0.201*	0.412***	0.538***	0.658***	0.347**
				(0.092)	(0.100)	(0.108)	(0.001)	(0.119)	(0.124)	(0.141)
Impossible-Unlikely					0.708***	1.118***	1.416***	1.488***	1.608***	1.290***
					(0.098)	(0.105)	(0.001)	(0.117)	(0.122)	(0.140)
Likely-Sure						1.211***	1.204***	1.421***	1.415***	1.091***
						(0.106)	(0.001)	(0.121)	(0.121)	(0.139)
Probable-Good							0.682***	0.879***	0.858***	0.584***
							(0.001)	(0.113)	(0.113)	(0.128)
Unlikely-Doubtful								0.317***	0.501***	0.018
								(0.114)	(0.125)	(0.164)
Good-Likely									0.397***	0.048
									(0.109)	(0.133)
Tossup-Slight										−0.603***
										(0.131)
Constant	0.992***	0.689***	0.273***	0.172**	−0.109	−0.752***	−1.159***	−1.560***	−1.922***	−0.962***
	(0.046)	(0.053)	(0.062)	(0.070)	(0.080)	(0.099)	(0.001)	(0.165)	(0.193)	(0.282)
Observations	7,920	7,920	7,920	7,920	7,920	7,920	7,920	7,920	7,920	7,920
Log Likelihood	−4,339.074	−4,263.780	−4,167.125	−4,162.002	−4,135.203	−4,067.219	−4,039.272	−4,035.026	−4,028.396	−4,017.654
Akaike Inf. Crit.	8,686.148	8,537.561	8,346.249	8,338.004	8,286.406	8,152.438	8,098.545	8,092.051	8,080.791	8,061.308
Bayesian Inf. Crit.	8,714.057	8,572.447	8,388.112	8,386.844	8,342.223	8,215.232	8,168.316	8,168.800	8,164.517	8,152.011

Note: ^{a} Difference in ordinal rank; *p<0.1; **p<0.05; ***p<0.01

WA12 Regression Analysis on Wu and Gonzalez (1996) Data

We analysed the Wu and Gonzalez (1996)'s data using a mixed effects logistic regression with participant-specific intercepts, similar to the analysis reported in the Studies 1 and 2 of the paper. We modeled the probability of choosing the safer option as a function of (1) a predictor indicating whether the rung probabilities cross categorical boundaries; (2) dummy variables for gamble pairs at the two extremes of the probability space, i.e., when the riskier choice's $Payout_{common}$ probability is 0 or the safer choice's $Payout_{common}$ probability is 1, again to demonstrate that the effect of the categorical boundary is not solely driven by these cases, and (3) dummy variables for the five ladders.[9] Consistent with our studies, this analysis of the Wu and Gonzalez (1996) data suggests that the safer choice probabilities significantly increase when crossing categorical boundaries ($\beta = 0.192$, $p = 0.029$) (see Table WA5).

WA13 Hierarchical Bayes Inference Details

Bayesian analysis integrates the likelihood function with a prior distribution placed on the model parameters to infer the joint posterior distribution of the model parameters. We used a Hamiltonian Monte Carlo (HMC) simulation (Gelman et al. 2013), with a No-U-Turn sampler (Hoffman and Gelman 2014) implemented using the software *stan* (Carpenter et al. 2017). This sampler ensured an efficient movement through the joint posterior distribution of our model parameters, particularly important given the nonlinearity of the model specification. A more common random walk Metropolis-Hastings algorithm can take a long time exploring the high dimensional target distribution, and can lead to autocorrelated Markov chains.

We set diffuse priors, to allow the data to drive the posterior distributions of our parameter estimates. The unconstrained priors were distributed multivariate normal with mean μ_0 and a full covariance matrix Σ_0.

To efficiently estimate the individual-level parameters, we implemented a non-centered parametrization (Betancourt and Girolami 2015). The non-centered parametrization breaks down the correlation between the individual and population levels of the hierarchy by deterministically transforming parameters, such that we can sample uncorrelated variables. In our model, instead of sampling directly from the MVN(μ_0,Σ_0), we broke down the dependency between the population-level parameters $\{\mu_0,\Sigma_0\}$ and the individual-level parameters $\{\gamma_{1,i}, \gamma_{2,i}, \alpha_i, \phi_i\}$ by sampling from:

$$vec(\gamma_{1i},\gamma_{2i},\alpha_i,\phi_i) \sim \text{vec}(\mu_0 + L\Sigma_0 \nu_i) \tag{WA1}$$

where $vec(\nu_i) \sim N(0,1)$ and $L\Sigma_0$ is the Cholesky factor of the covariance matrix Σ_0. This shifts the correlation between the data and the parameters to the hyperparameters. The elements in $vec(\nu_i)$ are uncorrelated, and are very easy to draw. We decompose the covariance matrix into a location and a scale prior:

$$\Sigma_0 = \text{diag}(\tau)\Omega\text{diag}(\tau) \tag{WA2}$$

[9] Introducing a binning variable similar to that specified in Study 1 is not advisable since the binning variable and the dummy variable encoding the categorical boundary crossings would be almost colinear due to the study design by Wu and Gonzalez (1996).

Table WA5: Mixed effects logistic regression predicting safer choice (data from Wu and Gonzalez 1996)

	Dependent variable:
	Safer choice
Ladder 2	−0.116
	(0.112)
Ladder 3	−0.270***
	(0.102)
Ladder 4	−0.068
	(0.102)
Ladder 5	0.152
	(0.101)
Riskier probability = 0	0.409***
	(0.109)
Safer probability = 1	0.813***
	(0.123)
Left-digit boundary	0.192**
	(0.088)
Constant	−0.142
	(0.090)
Observations	4,200
Log Likelihood	−2,834.651
Akaike Inf. Crit.	5,687.302
Bayesian Inf. Crit.	5,744.387
Note:	*p<0.1; **p<0.05; ***p<0.01

where $diag(\tau)$ is a diagonal matrix of standard deviations, and Ω is the correlation matrix. The hyperpriors of the parameters above are:

$$\mu_{0,l} \sim \mathrm{N}(0,1);\ \tau_l \sim \text{exponential}(1),\ for\ l = 1,...,4;\ \text{and}\ \ \Omega \sim \text{LKJcorr}(\xi) \qquad \text{(WA3)}$$

We set up exponential priors on the standard deviations, which allows for each parameter to be scaled differently. We work with an LKJ prior distribution (Lewandowski et al. 2009) on the correlation matrix. As the scaling factor ξ increases, the correlation prior concentrates around the unit matrix, thus favouring more independent individual-level parameters. When $\xi = 1$, all correlation matrices are equally possible, as the LKJ prior has an uniform density over the correlation matrices of order L, where L is the total number of parameters. We set $\xi = 2$, which put slighter higher prior density on correlation matrices that differ from the identity matrix. This is because the Prospect Theory parameters are likely to be correlated (Toubia et al. 2013).

These priors offer several advantages. It is not required for the priors to be conjugate, they are computationally efficient, and they provide standard deviations and a correlation matrix which make for a more natural interpretation. We constrain the model parameters to remain in acceptable ranges in the following way. We ensure that γ_1 and γ_2 parameters remain in the [0,1] interval using an inverse logit transformation. We apply an exponential transformation to constrain the parameter ϕ to be positive. We use a scaled inverse logit transformation to ensure $\alpha \in [0,2]$. $\alpha < 1$ indicates risk aversion, and $\alpha > 1$ implies risk seeking behavior.

The joint posterior distribution of our model parameters given the data is:

$$\begin{aligned} p(\{\gamma_{1i}\},\{\gamma_{2i}\},\{\alpha_i\},\{\phi_i\},\mu_0,\Sigma_0|\{U_j\}) \propto p(\{U_j\}|\{\gamma_{1i}\},\{\gamma_{2i}\},\{\alpha_i\},\{\phi_i\})\times \\ p(\{\gamma_{1i}\},\{\gamma_{2i}\},\{\alpha_i\},\{\phi_i\}|\mu_0,\Sigma_0)\times p(\mu_0,\Sigma_0|\bar{\mu},\Sigma_{LKJ})\times p(\tau_0) \quad \text{(WA4)} \end{aligned}$$

where the first term in the right-hand side of Equation WA4 represents the likelihood specified in Equation 4 and the remaining terms represent the prior distributions.

We generated 10,000 draws, from two chains with random starting values, and used the last 2,000 iterations of each chain for inference. The HMC sampler requires less posterior draws to converge than a typical MCMC sampler, because it leads to less autocorrelated draws. We assessed convergence using the Gelman-Rubin statistic (Gelman and Rubin 1992). The statistic measures the ratio of the average variance of samples within each chain to the variance of the pooled samples across chains. At convergence, the statistic is 1. The Gelman-Rubin statistic was below the acceptable threshold of 1.1 for all our model parameters, showing that the chains have converged.

To conclude, our results show that we obtained an accurate representation of the posterior distributions of the model parameters.

WA14 Model Comparison and Parameter Recovery Using Simulated Data

We conducted a simulation study to test whether the categorization-based model best explains the data when the underlying behavior follows the proposed model. The simulation also allows us to study the parametric identification of the model proposed in §4. Using the ladder and amounts designed for Study 1 and parameter values close to the estimates reported in Table 5, we simulate data following the categorization-based model, with parameters estimated at the individual level. We estimate both the classic and the categorization-based models on the simulated data, and compare model fit. As expected, both measures (WAICs and SSEs) reported in Table WA6 confirm that the categorization-based model fits the data better than the classic specification.

Table WA6: Model selection with simulated data

	Model	
	Classic	Categorization-based
WAIC	5,177.9	4814.6
SSE	0.096	0.013

Data generated following the categorization-based model.
Lower values of the WAIC denote better fit.

Table WA7: Parameter recovery with simulated data

		Group-level Parameter estimates	
	True	Classic model	Categorization-based model
$\bar{\gamma}$	—	0.9 (0.8, 0.96)	—
$\bar{\gamma}_1$	0.91	—	0.92 (0.87, 0.96)
$\bar{\gamma}_2$	0.82	—	0.75 (0.66, 0.83)
$\bar{\alpha}$	0.73	0.77 (0.61, 0.91)	0.7 (0.55, 0.83)
$\bar{\phi}$	1.65	0.94 (0.49, 1.5)	1.20 (0.71, 1.81)

Table WA9 reports the true and estimated population-level parameters for the classic and categorization-based models. We see that all 95% BCIs of the recovered parameters contain their true values, thus the model is efficient in recovering the underlying parameters. Of interest, the parameter γ_2 is well recovered and significantly lower than γ_1.

WA15 Robustness to Prelec Specification

In the main document, we proposed a categorization-based model where the probability weighting function is based on the Tversky and Kahneman (1992) specification. Here, we test the robustness of our results to the Prelec (1998) one-parameter probability weighing function. The specification of the model follows the individual-level model described in §4, with the exception that the probability weighting function is $w(p) = -exp(-ln(p)^{\gamma})$. In the categorization-based model, we estimated two parameters, γ_1 and γ_2, to account for the effect of crossing categorical boundaries. We estimated the model using the data set gathered for Study 1.

Table WA8 shows the measures of in-sample fit and out-of-sample predictive accuracy. The results in Table WA8 are qualitatively similar to those of our main analysis. Our proposed categorization-based (CB) model fits the data substantially better than the model which embeds the classic probability weighting specification, both in terms of model fit ($WAIC_{Classic}$ = 4,262 vs. $WAIC_{CB}$ = 4,105, $SSE_{Classic}$ = .107 vs. SSE_{CB} = .071), and when considering the out-of-sample predictive accuracy ($MSE_{Classic}$ = .035 vs. MSE_{CB} = .032).

Table WA8: Model selection

	Classic[a]	Categorization-based[b]
WAIC	4,262.9.7	4,105.9
In-sample (SSE)	0.107	0.071
Out-of-sample (MSE)	0.035	0.032

[a]Classic: The Prelec (1998) probability weighting function. [b]Categorization-based: The probability weighting function specified similarly to Equation 4.2.

Participants appear on average risk averse ($\bar{\alpha}$=.57, 95% BCI=[.48, .66]), more so than estimated in the main analysis. The robustness checks confirm that the population-level weight parameter γ_2 ($\bar{\gamma}_2$=.63, 95% BCI=[.45, .75]) is lower than the parameter γ_1 ($\bar{\gamma}_1$=.79, 95% BCI=[.71, .86]).

Table WA9: Robustness check: Prelec's probability weighing function

	Population-level parameter estimates	
	Classic model	Categorization-based model
$\bar{\gamma}$	0.78	—
	(0.66, 0.86)	
$\bar{\gamma}_1$	—	0.79
		(0.71, 0.86)
$\bar{\gamma}_2$	—	0.63
		(0.47, 0.75)
$\bar{\alpha}$	0.64	0.57
	(0.55, 0.71)	(0.48, 0.66)
$\bar{\phi}$	1.77	2.62
	(1.01, 2.7)	(1.68, 3.75)

Overall, results support our conjecture that the probability weighing function is shaped by categorical boundaries.

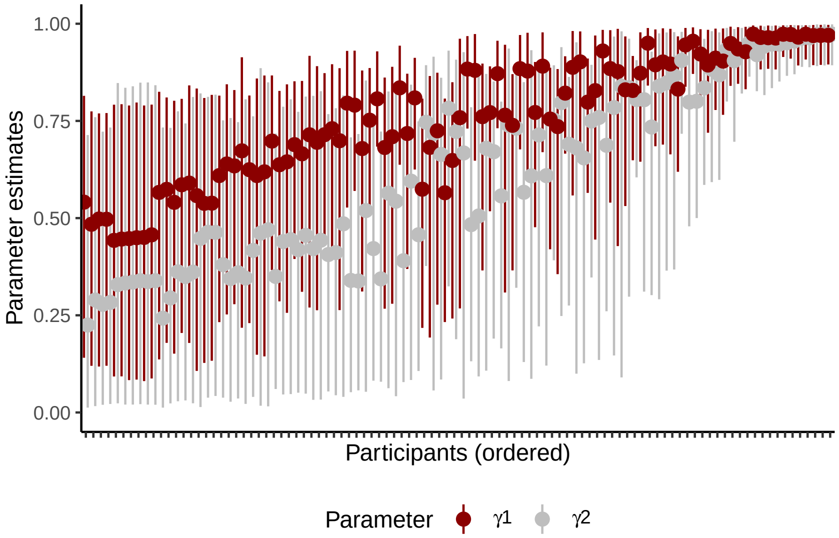

Figure WA1: Individual-specific probability weighting parameters estimated using the Prelec (1998) specification. Most individual-level γ_{2i} parameters are on average lower than the γ_{1i} parameters.