

Alós-Ferrer, Carlos; Garagnani, Michele

**Working Paper**

## Strength of preference and decision making under risk

Working Paper, No. 330

**Provided in Cooperation with:**

Department of Economics, University of Zurich

*Suggested Citation:* Alós-Ferrer, Carlos; Garagnani, Michele (2020) : Strength of preference and decision making under risk, Working Paper, No. 330, University of Zurich, Department of Economics, Zurich,  
<https://doi.org/10.5167/uzh-172509>

This Version is available at:

<https://hdl.handle.net/10419/222356>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*



**University of  
Zurich**<sup>UZH</sup>

University of Zurich  
Department of Economics

Working Paper Series  
ISSN 1664-7041 (print)  
ISSN 1664-705X (online)

---

Working Paper No. 330

# **Strength of Preference and Decision Making Under Risk**

Carlos Alós-Ferrer and Michele Garagnani

Revised version, July 2020

---

# Strength of Preference and Decision Making Under Risk\*

Carlos Alós-Ferrer<sup>†1</sup> and Michele Garagnani<sup>1</sup>

<sup>1</sup>Department of Economics, University of Zurich, Blümlisalpstrasse 10, CH-8006 Zurich, Switzerland

This Version: July 2020

## Abstract

Influential economic approaches as random utility models or quantal-response equilibria assume a monotonic relation between error rates and choice difficulty or “strength of preference,” in line with widespread evidence from discrimination tasks in psychology and neuroscience. However, while the latter define difficulty through objective dimensions (weight, length, etc), for economic decisions under risk the appropriate dimension remains unclear, with candidates including payoff-irrelevant numerical magnitudes. The very existence of the effects remains largely untested, because fitting models which assume such relations to data simply produces spurious findings. We provide a systematic empirical validation relying on two parsimonious experimental designs. Strength-of-preference effects are explained by expected values if objectively-correct answers exist, and by cardinal differences in independently-estimated utilities for preferential choices. Numerical magnitudes produce additional but minor effects. Finally, response times are inversely related to strength of preference, confirming that the observed relations are not “as if” explanations.

**JEL Classification:** D01 · D81 · D91

**Keywords:** Stochastic choice · Strength of preference · Decision errors · Risk attitude · Choice difficulty

## 1 Introduction

Errors are everywhere. Economics has by now embraced the view that economic choices are subject to noise (e.g., McFadden, 2001). Research in stochastic choice has provided

---

\*The authors thank José Apesteguía, Larbi Alaoui, Miguel Ballester, Michael Birnbaum, Andrew Caplin, Giorgio Coricelli, Ernst Fehr, Fabio Maccheroni, Paola Manzini, Paulo Natenzon, Nick Netzer, Antonio Penta, Antonio Rangel, Aldo Rustichini, Michael Woodford, and seminar audiences at the TIBER Symposium (Tilburg, 2018), the Conference on Decision Sciences (Konstanz, 2018), the Workshop on Stochastic Choice (Barcelona, 2019), the Sloan-Nomis Workshop on the Cognitive Foundations of Economic Behavior (Vitznau, 2019), as well as the SAET (Ischia, 2019), AISC (Lucca, 2019), and IAREP (Dublin, 2019) conferences.

<sup>†</sup>Corresponding author: carlos.alos-ferrer@econ.uzh.ch

extensive evidence that human beings often make different choices even when repeatedly confronted with the same set of options<sup>1</sup> (e.g., Fechner, 1860; Cattell, 1893; Tversky, 1969; Camerer, 1989; Hey and Orme, 1994; Ballinger and Wilcox, 1997; Hey, 2001; Moffatt, 2005; Conte, Hey, and Moffatt, 2011; Agranov and Ortoleva, 2017). There is, however, no universally-accepted view on the origins and determinants of noise or errors in economic decision making. How often do economic agents make mistakes, and what does the number of mistakes depend on? These questions are important both for positive and normative reasons. On the one hand, forecasting economic choices requires accurate models of decision errors, beyond the simple assertion that people indeed do make mistakes. On the other hand, predicting the effects of economic policies and evaluating their consequences is only possible if the consequences of human errors in response to them are understood. Indeed, large individual error rates are reflected in significant behavioral heterogeneity and can cause potentially large welfare losses at the aggregate level (e.g., Choi et al., 2014; Harrison and Ng, 2016; Alekseev et al., 2019).<sup>2</sup>

The key question is whether error rates are associated with directly or indirectly measurable *economic* variables. To understand the sources of mistakes (or stochastic choice) in economic decisions, however, it is useful to briefly step back and examine evidence from the cognitive sciences (chiefly cognitive psychology and neuroscience) on tasks which are significantly simpler than the ones proper of economics. In the domain of *psychophysics*, decades of research have concentrated on *perceptual discrimination tasks*, where two stimuli are presented and human participants are asked to estimate which one scores higher along an objective scale, for instance which of two sounds is louder, which of two lights is brighter, or which of two lines is longer. In such simple tasks, there is an objective, direct measure of choice difficulty: choices become gradually harder as the difference between the stimuli becomes smaller (along the objective scale). There are two firmly-established stylized facts in this literature. The first is that the percentage of correct choices is strictly decreasing with choice difficulty, that is, error rates are larger when stimuli are more similar (Cattell, 1893; Moyer and Landauer, 1967; Laming, 1985; Klein, 2001; Wichmann and Hill, 2001). The second is that choices are slower as choice difficulty increases, that is, response times are larger when the stimuli are more similar (Cattell, 1902; Dashiell, 1937; Moyer and Landauer, 1967). The latter fact is commonly taken as evidence that the effect on error rates derives from basic (gradual) neural mechanisms in the human brain. That is, decisions might derive from gradual, noisy processes of internal evidence accumulation, which are more error-prone

---

<sup>1</sup>“Common experience suggests, and experiment confirms, that a person does not always make the same choice when faced with the same options, even when the circumstances of choice seem in all relevant aspects to be the same.” (Davidson and Marschak, 1959).

<sup>2</sup>If a normative view is adopted where (except for knife-edge indifference cases) only one choice is considered correct (or consistent with underlying preferences), the statement that choice is stochastic is equivalent to the empirically-ubiquitous observation of positive error rates. It is in this sense that we speak of “errors” in this work. This is also in line with a positive-economics view, where one aims to understand the extent to which economic decision makers will deviate from choices deemed “rational.”

and time-consuming if the quantities that need to be teased apart are closer (Shadlen and Kiani, 2013; Shadlen and Shohamy, 2016).

In this work, we ask the question of whether these gradual effects are relevant for economic decisions under risk and, if so, which economic variables do determine them. This question is obviously also important for conceptual reasons, as the phenomena we discuss imply a *cardinal* effect of economic variables on choices, as opposed to the classical, purely-ordinal view of preferences. However, studying the dependence of error rates on underlying economic variables is far from straightforward, for two reasons which we will elaborate on below. The first is that it is by no means clear what the gradual effects predicted by psychology and neuroscience should depend on for *economic* choices, where a natural scale as weight, brightness, or length is usually not part of the problem’s formulation, and utilities are neither directly observable nor objective. The second is that economic models incorporating stochastic choice cannot be used to test for the presence of a monotonic relation between error rates and choice difficulty. This is because such models, e.g. random utility models, quantal response equilibria, or evidence accumulation models, *assume* this relation from the onset. Hence, fitting them to data will necessarily produce patterns in agreement with the assumed relation just because the underlying dimension (e.g., imputed utilities in random utility models, or fitted drift rates in evidence accumulation models) is estimated in order to produce a good fit. In other words, the estimation method (very especially within-sample fitting procedures) might often create apparent regularities where none exists, hence obscuring the actual origin of the investigated phenomena (we provide an example illustrating this point below).

The first problem is an obvious one. In a sense, and with apologies to those fields, psychophysicists and perceptual psychologists face easier problems than economists. It is *a priori* not clear whether objectively-given scales might play the role of weight or length for economic decisions, or even for some of them. For instance, on the basis of the available evidence, a good case could be made for numerical magnitudes, independently of whether they are payoff-relevant or not. Results by Moyer and Landauer (1967) and Dehaene, Dupoux, and Mehler (1990) (see also Dehaene, 1992; Dehaene et al., 2008) show that the gradual effects on error rates and response times exist even when humans are asked to discriminate among single-digit numbers. That is, (astonishingly) people make more mistakes (and take longer to decide) when asked whether 6 is larger than 5 than when asked whether 9 is larger than 2. This is compatible with evidence from electroencephalography (EEG), which suggests that the neural representations of numbers vary in a continuous, gradual way with numerical distance (Spitzer, Waschke, and Summerfield, 2017). Recently, Khaw, Li, and Woodford (2018) and Frydman and Jin (2018) have suggested that the mere imprecise representation of numerical magnitudes along these lines may explain the large estimates of risk aversion which are typically observed in laboratory experiments in economics. Given that many economic tasks

come with a numerical framing, it is necessary to tackle the question of which is the (most) relevant dimension underlying possible gradual effects on economic choices.

Although error rates might indeed be affected by the imprecise representation of observational variables as e.g. numerical magnitudes, this is unlikely to be the only determinant of errors. More natural candidates, painting a less bleak picture of human rationality, are related to economic gains, as is the case of expected value or (estimated) expected utility. Inspired by models from psychology (Thurstone, 1927), random utility models, as pioneered by Marschak (1960) and McFadden (1974, 2001), assume that errors depend on underlying (unobservable) utility differences. However, very few studies have actually empirically demonstrated a monotonic relation between error rates and differences in underlying utilities. A notable exception is the early study of Mosteller and Noguee (1951), which used utilities estimated through an interpolation procedure. An added difficulty is that in the empirical work of Mosteller and Noguee (1951) (and other experiments), as well as in theoretical random utility models an error is *defined* as a choice which does not maximize utility, that is, there is no *ex ante* definition of error independent of the (estimated or assumed) utility.

This leads us straight to the second problem. Most applied work in discrete choice microeconomics and beyond *assumes* a gradual relation between underlying utility differences and choice probabilities, often with a specific logit or probit form, in order to parametrically estimate the utilities themselves. This is true for random utility models (Anderson, Thisse, and De Palma, 1992; McFadden, 2001; Moffatt, 2015), but is also an assumption for the choice rules underlying quantal response equilibria (McKelvey and Palfrey, 1995; Goeree, Holt, and Palfrey, 2005), which are extensively used in game theory. A similar implication follows from evidence accumulation models from cognitive psychology and neuroscience (e.g., the Drift-Diffusion Model of Ratcliff, 1978), which are receiving increasing attention in economics (Fudenberg, Strack, and Strzalecki, 2018; Baldassi et al., 2019).<sup>3</sup> Those imply logit choice probabilities, and hence again incorporate a gradual relation between a form of choice difficulty and error rates. While all these approaches are invaluable to compare the fit of different utility- or payoff-based models of choice and have delivered many important insights, they are not appropriate to test the basic hypothesis that gradual effects exist, or to pin down measurable determinants thereof. A structural property cannot be tested by fitting a structural model to a dataset if the model incorporates the property among its assumptions. In particular, we specifically contend that there is, to date, no systematic analysis testing for an actual relation between utility differences derived from choices and subsequent choice errors *out of sample*. Studies fitting models which assume properties in this vein can be misconstrued to “show” such relations, but such an interpretation would be both incorrect and misleading. This is important, because, as pointed out above, in the absence of clear tests, there is no justification to actually assume a particular variable (say,

---

<sup>3</sup>As pointed out by Webb (2019), there is a relation between drift-diffusion models and random utility models, which in particular explains why the structural relation mentioned above is present in both.

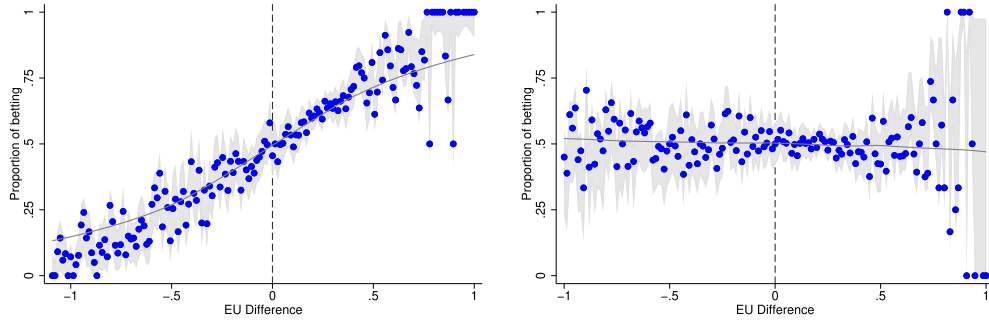


Figure 1: Analysis of a dataset of random, simulated choices. Left-hand panel: Choices as function of expected utility difference using a standard, within procedure. An appearance of order and gradual effects of expected utility differences on error rates emerges, even though no regularity is present in the data. Right-hand panel: The same choices as function of expected utility difference using an out-of-sample procedure. No regularity can be identified. Gray areas indicate 95% binomial proportion confidence intervals.

expected utility, expected value, or simply payoff-irrelevant numerical magnitudes) as the determinant of error rates.

To drive this point home, we constructed a dataset by simulating fictitious subjects who made completely random decisions among alternative risky choices. We then treated the dataset as if it would come from actual decision makers and used a standard fitting approach estimating an alleged risk propensity, assuming that errors depend on utility differences. Specifically, we estimated a CARA utility function using an additive random utility model and heteroskedastic, normally-distributed errors, as commonly done in the literature (e.g., McFadden, 2001; Moffatt, 2015); for more details on the estimation procedures, see Section 4.2 below. Plotting choice frequencies against the estimated utility differences yields a regular sigmoidal curve (as in any logit or probit model), which creates the appearance of order (and gradual effects arising from utility differences) for the nonsensical dataset. This is shown on the left-hand panel of Figure 1. Actually, this appearance is a mere artifice of the method, as can be shown by estimating utility *out of sample*, i.e., using part of the choices for estimation purposes and plotting the rest of the choices against the resulting estimated utility differences. Specifically, we estimated a CARA utility function from even-numbered choices and used it to plot data from odd-numbered choices, and vice versa. This approach shows that there is no actual regularity in the dataset, as depicted on the right-hand panel of Figure 1. We conclude that structural models where utility is estimated can mistakenly create an appearance of gradual effects, and hence direct tests are needed.<sup>4</sup>

<sup>4</sup>The point that fitting a structural model can produce spurious findings is widely acknowledged in the literature. Thus, various criteria have been proposed to examine the validity of structural models, as the prominent Akaike's Information Criterion (e.g. Wilcox, 2008). However, measures of goodness of fit are intrinsically relative and do not provide an actual test of the effects we target here.

In this work, we aim to test and clarify the dependence of error rates (the variable of interest) in decisions under risk on economic variables. We attack the problem on three fronts. First, we conduct an experiment (Experiment 1) with normatively-correct answers (but high error rates) where the explanatory variable can be determined in advance. This is made possible by employing a binary-choice gambling task where the winning probability or, equivalently, expected value, is an unequivocal, objectively measurable indicator of choice difficulty. In this study, the definition of error can be made *ex ante*, independently of any estimation of utility, simply because correct responses are independent of attitudes toward risk. In this sense, this is representative of tasks in the judgment and decision making domain, as those employed to study many heuristics and biases (e.g. Bar-Hillel, 1980; Grether, 1980; Charness and Levin, 2005). The key is that we commit to the explanatory variable *before collecting the data*, and utility estimation plays no role. The task is simple in the sense that rational decision makers could “figure it out” with relative ease, as is the case of many problems in judgment and decision making, where an optimal decision under risk has to be made on the basis of individual (updated) beliefs. We find a significant fraction of errors (above 25%), and we demonstrate that far from being pure noise, error rates stand in a clear monotonic relation with differences in expected value. That is, we demonstrate the existence of gradual effects of an *objective economic distance* among alternatives on error rates. Importantly, the design also allows us to test for dependence on payoff-irrelevant numerical effects as in Moyer and Landauer (1967), and we find that there is indeed some relation, but it is a second-order phenomenon compared with the dependence on expected-value differences. This delivers a first, objective confirmation of psychophysical, gradual effects in decision making under risk arising from economic variables.

Second, we conduct a different experiment (Experiment 2) using a betting paradigm where whether a decision is correct or not depends on individual attitudes toward risk, as is the case for most lottery-choice tasks. Rather than fitting the data to an estimated utility, we employ an out-of-sample estimation procedure excluding any artifices arising from the estimation method.<sup>5</sup> Again, we find a monotonic relation, with larger error rates arising when the differences in the expected utility of the options are smaller. This relation is robust to different structural assumptions on the estimated utility. In contrast, the dependence on expected value differences is considerably weaker. That is, we demonstrate that the gradual effects on choice observed in psychophysics can be readily found in standard economic tasks, but they will in general arise from a *subjective economic distance* which arises from integrated, unobservable variables (“utility”). Further, economic distance (subjective or objective) can then be considered a *cardinal* measure of “strength of preference,” because its cardinal magnitude determines a measurable, continuous variable (error rates).

---

<sup>5</sup>Precisely due to the problems pointed out above, out-of-sample testing is gaining ground both in economics and in psychology. Examples include studies on test-retest reliability (e.g. Weber, Blais, and Betz, 2002), or the leave-one-out approach (e.g. Lopez-Guzman et al., 2018).

Third, we conduct a further test to confirm the cardinal nature of the uncovered effects. In both experiments, we collect response time data as an independent variable, which in particular plays no role for the estimation of underlying utilities. Psychophysics predicts a robust relation, with decisions where stimuli are closer being slower. We find this relation in both experiments. In Experiment 1, response times increase as differences in expected value decrease, but they are relatively unaffected by payoff-irrelevant numerical magnitudes. In Experiment 2, the analysis of response times confirms that differences in underlying utilities are a better candidate for economic distance (which replaces the choice difficulty of psychophysics) than differences in expected value (or numerical magnitudes). In both cases, the relation with response times (again a measurable, continuous variable) confirms the cardinal content of economic distance.

Taken together, our evidence provides a systematic demonstration showing that the psychophysical effects found in the cognitive sciences are indeed very relevant for economic decisions under risk, but they depend more on economic variables than on perceptual or numerical ones. Decision-irrelevant factors (numerical magnitudes) influence error rates, but they play a secondary role in comparison with purely-economic variables. In settings where objectively-optimal answers can be derived from (correct) beliefs, it is possible to give an exogenous definition of errors, which in turn allows for a straightforward observation of the link between economic distance and error rates (in particular, one which is free of estimation problems). When risk attitudes play a role, the explanatory variable is a subjective, integrated one capturing “strength of preference,” which needs to be estimated. This result validates the ideas and assumptions behind random utility models and other approaches. Further, the relation to response times shows that the effects are more than “as if” accounts of decision making and have their origin in brain processes of a gradual nature, as assumed e.g. by evidence accumulation models (Ratcliff, 1978; Fudenberg, Strack, and Strzalecki, 2018).

The paper is structured as follows. Section 2 briefly reviews the related literature. Sections 3 and 4 discuss Experiments 1 and 2, respectively. The analysis of response times is conducted in the last subsections within those sections. Section 5 concludes.

## 2 Related Literature

Our work is related to long-standing problems in economics and to several strands of the recent literature. Classical studies of stochastic choice endorsed the view that utilities should be understood as reflecting choice probabilities (e.g., Debreu, 1958; Luce, 1959), in direct opposition with the neoclassical view that they reflect preferences of an exclusively ordinal nature (Hicks and Allen, 1934). The proliferation of experimental data showing the stochastic nature of economic choice has led to increased attention on theoretical models of stochastic choice in the recent years (e.g. Manzini and Mariotti, 2014; Gul, Natenzon, and Pesendorfer, 2014; Fudenberg and Strzalecki, 2015; Apesteguía, Ballester, and Lu, 2017; Apesteguía and Ballester, 2018; Natenzon, 2019). In game theory, models

of stochastic (logit) choice based on observable payoffs and unobservable idiosyncratic shocks have given rise to new equilibrium concepts as quantal response equilibria (McKelvey and Palfrey, 1995; Goeree, Holt, and Palfrey, 2005). In microeconometrics, random utility models (Anderson, Thisse, and De Palma, 1992) have become standard for fitting experimental data and recovering underlying utility functions, assuming specific functional forms (e.g., logit or probit) for error terms. Those models assume exact functional forms mapping differences in utilities to error terms, which are highly valuable as structural assumptions but are in general not directly tested. In stark contrast, Alós-Ferrer, Fehr, and Netzer (2018) have recently shown that certain properties of the empirical distribution of response times allow to recover the underlying preferences in random utility models without imposing any substantive assumptions on the distribution of random terms.

The relation between choice difficulty and choice frequencies has also repeatedly been illustrated in a domain which bridges the gap between psychophysics and decisions under risk. In experimental studies on consumer choice (typically snack food items), participants make choices based on their personal preferences. Although they generally have a very different focus, several works in this domain have illustrated a relation between differences in subjective, self-reported ratings and choice frequencies (Krajbich, Armel, and Rangel, 2010; Fisher, 2017; Clithero, 2018; Polania, Woodford, and Ruff, 2019, e.g.). However, the approach in those works is typically to fit an evidence accumulation model, which assumes a relation of this sort (as discussed above) and is hence inadequate to test for the postulated effects, and in any case does not extend to decisions under risk.

To the best of our knowledge, the first study to point at a connection between utility differences and choice frequencies in decisions under risk was the inspiring experiment of Mosteller and Nogee (1951) on poker dice gaming, which aimed to “test the validity of the construct” represented by (expected) utility. Their analysis included illustrations which suggested a sigmoidal relation between utility differences and choice frequencies, although, as the authors admitted, those were at the individual level and cherry-picked among all experimental participants. While suggestive, their illustrations were not a test for the presence of gradual effects (and were actually not meant to be), because their utility functions were constructed exclusively out of observed indifferences. For instance, although their illustrations map zero utility difference to 50 percent choice frequency, “this finding was built into the expected utilities by the construction leading to the utility curves” (Mosteller and Nogee, 1951, p. 202).

Conceptually, our work is also related to the study of Khaw, Li, and Woodford (2018), who carried out an experiment on risky choice where participants chose between a sure amount and lotteries with a single non-zero outcome and a fixed probability of winning varying amounts (that is, the winning probability was identical for all choices). By varying the sure amount and the lottery outcome, Khaw, Li, and Woodford (2018) explored the reaction of choice frequencies to changes in payoffs and argued that the data could be explained assuming an imprecise internal representation of numerical

magnitudes, in line with Moyer and Landauer (1967) and Dehaene (1992). Hence, their work speaks in favor of a direct effect of numerical magnitudes in error rates. However, by design, their numerical magnitudes stand in a monotonic relation to payoffs, and hence in their data it is not possible to disentangle the effects of numerical magnitudes and the effects of expected values (or utilities). In our experiments, different numerical magnitudes are associated with the same expected payoffs and vice versa, allowing us to study the effects separately.

Although our main variable of interest are decision errors, we also examine response times for two reasons. First, well-established effects in psychophysics encompass both error rates and response times, hence the analogy would not be complete without the latter. Second, while an explanation of error rates alone might be challenged as a pure “as if” story, response times allow reasonable inferences on the actual decision processes generating the errors. In this sense, our work is related to the small but growing literature examining response times in economics (see Spiliopoulos and Ortmann, 2018, for a recent review). The most-closely related contribution to this aspect of our studies is the work of Chabris et al. (2009), who studied intertemporal decisions and found a monotonic relationship between response times and estimated utility differences (discount factors). More recently, Konovalov and Krajbich (2019) found a negative relation between response times and a distance measure between estimated, subject-level utility parameters and the parameter values which would have yielded indifference in trial-specific choices. Their data included an intertemporal choice task, a binary mini-dictator game, and a binary choice between a sure outcome and a gamble including losses; in the latter case, they abstract from risk aversion and the utility parameter they estimate captures loss aversion only. Also related is Alós-Ferrer et al. (2016), which postulated a model of lottery choice and evaluation including a relation between choice difficulty and response times to investigate the determinants of the preference reversal phenomenon (Lichtenstein and Slovic, 1971; Grether and Plott, 1979; Tversky, Slovic, and Kahneman, 1990). Other response-time studies in economics include Wilcox (1993, 1994) and the web-based studies of Rubinstein (2007, 2013). On a different front, Achtziger and Alós-Ferrer (2014) relied on response times to differentiate different decision processes in a framework where intuitive reinforcement might conflict with optimal decisions based on Bayesian updating of beliefs (see also Alós-Ferrer and Ritschel, 2018).

### 3 Experiment 1: Objective Domain

We first aim to demonstrate the gradual effects of “strength of preference” on error rates (for decisions under risk) in a domain where the variable influencing those effects is objectively given, and, as a consequence, utility estimation plays no role. The task we employ is representative of studies in the judgment and decision-making domain, where economic agents make decisions under risk or uncertainty but there is an objectively-correct answer, for example due to stochastic dominance. A prominent example is given

by tasks involving updating of previously-held beliefs (e.g. Bar-Hillel, 1980; Grether, 1980, 1992; Charness and Levin, 2005; Achtziger and Alós-Ferrer, 2014). We will rely on a simple gambling task with given probabilities, which is designed with two objectives in mind. The first is that objectively-correct decisions exist, independently of attitudes toward risk, and thus an exogenously-given measure of the strength of preference is available.<sup>6</sup> The second is that numerical differences (in a perceptual sense) can be disentangled from economic incentives, allowing us to investigate both possible dimensions of choice difficulty. Additionally, the design aims to avoid extraneous dimensions of choice difficulty (e.g., complexity of the options, competing heuristics, etc.) to concentrate on the role of strength of preference.

### 3.1 Design and Procedures

The experiment was computerized and programmed in Psychopy (Peirce, 2007), a software which ensures high precision in the measurement of response times. We recruited  $N = 96$  participants (54 females, age range 19 – 47, mean 24) using ORSEE (Greiner, 2015) at the Cologne Laboratory for Economic Research. Participants were university students enrolled in fields other than psychology and economics. They were provided with written instructions and answered five control questions before starting the task, to ensure correct comprehension of the procedures and payment mechanism. Three participants failed at least one comprehension question and were excluded from the analysis. Subjects were paid according to their performance in the experiment. Total earnings were the sum of the earnings in the 160 trials plus a show-up fee of EUR 4. Sessions lasted around 60 minutes and the average payoff was EUR 16.45 (around USD 17.60 at the time of the experiment). All trials in the experiment were paid. In our context, this payment mechanism is incentive compatible under mild assumptions on individual preferences, as shown by Azrieli, Chambers, and Healy (2018, 2020).<sup>7</sup>

The experimental task is as follows. Participants are confronted with three decks of cards, a red one (Diamonds) and two black ones (Clubs and Spades), containing ten cards each (numbered 1 to 10, see Figure 2) left-hand panel. The participants’ task is to choose *twice* from which of the two black decks a card should be randomly extracted from, and the game’s objective is to beat a card extracted from the red deck with the black one. Each trial starts with a participant choosing between the two (complete) black

---

<sup>6</sup>Of course, it is conceivable that even in this setting some subjects rely on alternative objectives, e.g. if they distort probabilities. The point is merely that normatively-correct answers can be identified in this paradigm, independently of any individual-level estimation, and thus we are able to commit to an objective scale before conducting the experiment.

<sup>7</sup>The pay-all mechanism is incentive compatible if subjects’ preferences fulfill a condition called “no complementarities at the top” when evaluating bundles of outcomes (Azrieli, Chambers, and Healy, 2018). In the context of the current experiment, this assumption is immediately fulfilled as long as participants prefer more money over less (but the assumption might impose stronger restrictions in more complex environments). In the binary choice setting of the current experiment, in every trial one option is dominated by the other. Choosing the correct option in all rounds dominates any other alternative bundle of choices, hence making this payment mechanism incentive-compatible according to Azrieli, Chambers, and Healy (2018).

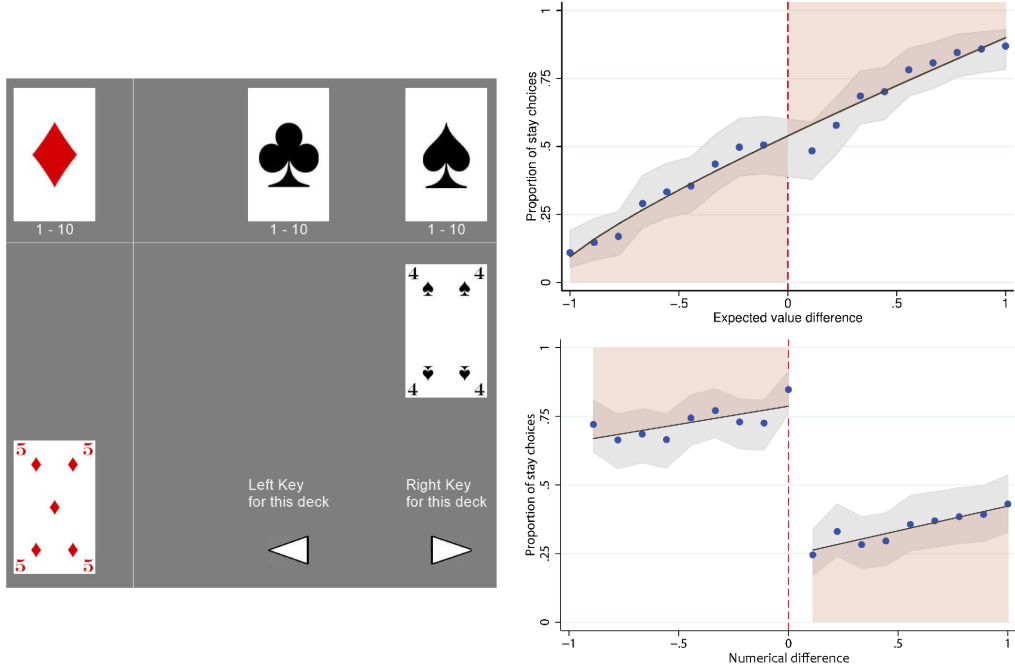


Figure 2: Experiment 1. A trial starts with participants choosing between the black decks (left). A black card is extracted from the chosen deck and a red card is displayed. No replacement happens after the first choice. Participants then choose a black deck again, and a card is extracted from that deck. The participant wins if and only if the second extracted black card is strictly larger than the red one. The right-hand panels depict stay choices as function of expected value difference (stay minus switch; upper panel), and of the numerical distance (black minus red; lower panel). Gray areas indicate 95% binomial proportion confidence intervals.

decks, but this first choice is irrelevant for our purposes since at this point both decks are identical. It is also unpaid (to avoid possible reinforcement or valence effects). The choice, however, creates an asymmetry which is the essence of the task. After the first choice is made and the first black card is extracted, that card remains on the table (there is no replacement). A card is extracted from the red deck, and the participant is asked to choose between the black decks a second time. This is the choice we are interested in. A (black) card is extracted from the chosen deck, and the participant received EUR 0.15 if and only if that black card has a number strictly bigger than the red card, otherwise she receives nothing. Subsequently, the trial ends, all cards are placed back in their decks, and decks are reshuffled before a new repetition starts. Participants knew that trials were independent, so the outcome as well as the cards displayed in one trial were unrelated to those of subsequent trials. Each participant completed 160 of such trials.

After the first choice, one black deck has either one winning card less or one losing card less. Hence, by design, there is an optimal decision pattern for the second choice, which is to bet on the deck with a higher proportion of winning cards. That is, if the

first black card was smaller or equal than the red card, the participant should choose the same deck, and if the first black card was strictly larger than the red card, the participant should choose the other deck. In the example depicted in Figure 2 left-hand panel, the red card is a 5 and the first black card is a 4 (of spades), so the spades deck contains only 9 cards, 4 losing and 5 winning ones. The clubs deck still contains 10 cards, 5 losing and 5 winning cards. Therefore the deck of spades contains 1 losing card less than the untouched deck of clubs and the optimal choice is to choose it again. On the contrary, if the first extracted black card had been strictly larger than the red card, the chosen deck contains 1 winning card less than the other one, and the optimal decision would be to choose the untouched deck. Hence, independently of risk aversion, there is always a normatively-correct decision for the second choice.

In spite of the fact that all choices are either objectively correct or objectively wrong, some choices are “more correct” than others, because opportunity costs are different. Let  $r, b_1 \in \{1, \dots, 10\}$  be the red card and the first extracted black card, respectively. Let  $\pi^0(r, b_1)$  and  $\pi^1(r, b_1)$  be the probability of winning by choosing the same deck or by shifting to the other deck, respectively. Then  $V(r, b_1) = |\pi^0(r, b_1) - \pi^1(r, b_1)|$  is the cardinal difference (distance) between the probability of winning by making the correct choice and the probability of winning while making an error. Since participants are paid only in case they win, up to a rescaling of monetary units this is also the difference in expected values between a correct decision and an error. If  $r \geq b_1$ , one obtains  $\pi^0(r, b_1) = (10 - r)/9$  and  $\pi^1(r, b_1) = (10 - r)/10$ . If  $r < b_1$ , one obtains  $\pi^0(r, b_1) = (10 - r - 1)/9$  and  $\pi^1(r, b_1) = (10 - r)/10$ . Hence,

$$V(r, b_1) = \begin{cases} (10 - r)/90 & \text{if } r \geq b_1, \\ r/90 & \text{if } r < b_1. \end{cases}$$

By design,  $V(r, b_1)$  assumes values in the set  $\{1/90, 2/90, \dots, 9/90\}$ . These differences in expected value indicate the opportunity cost of (not) choosing the right answer and reflect how far away from “indifference” the participants were in every decision, and are hence a natural measure for the “strength of preference.” Thus, we take  $V(r, b_1)$  as the potential driver for stochastic choice and refer to this magnitude as *(objective) economic distance*.

The probabilities of winning by staying or switching, and the economic distance, are monotonic functions of the numerical value of the red card. However, for computing the optimal choice the only necessary information is the sign of the relation between the first black card and the red card. That is, the actual magnitude of the difference between the values of these two cards is economically inconsequential. However, Moyer and Landauer (1967), Dehaene, Dupoux, and Mehler (1990), and others have shown that, in simple comparisons, errors do depend on the numerical differences between stimuli. Therefore, we also contemplate the possibility that the distance between the numerical values of the first black card and the red one influences choice frequencies (and response times).

There are ten possible distances between the two cards, ranging from 0 to 9. We refer to this magnitude as the *numerical distance*.

To ensure enough variability in the stimuli, the set of initial stimuli (first black card and red card) was predetermined and pseudorandomized. Furthermore, the red card was never a 10, since in this case winning would be impossible, hence the choice would be inconsequential (the instructions did not claim that the red card was randomly selected, since the procedure by which it was selected was payoff-irrelevant once the actual choice was faced). The key second black card was randomly selected among the remaining cards.

### 3.2 Choices and Errors

In spite of the simplicity of the task, the mean error rate across participants was 28.93%, with a median of 31.25% (SD = 18.21, min 0.63%, max 60.00%). We start by examining the dependence of error rates on both economic and numerical distance. Figure 2 (right-hand panels) plots the frequency of “stay” decisions (choosing the same deck as in the first decision) for each possible value of each variable. The upper-right panel plots the dependence on economic distance, i.e. differences in expected values. The red-shaded areas correspond to errors, as a rational decision maker should stay for a positive expected value difference and switch for a negative one. To facilitate the comparison, in all figures and regressions the economic and numerical distances are normalized to be between 0 and 1.<sup>8</sup> The probability to stay with the same deck stands in a clear positive relation with the difference in expected values. The frequency of errors becomes smaller as the difference becomes larger (no matter the sign), and it is largest (essentially 1/2) when the difference approaches zero. This pattern is radically different from that predicted by neoclassical economic theory. Even accounting for noise, neoclassical predictions would prescribe choice frequencies with a flat slope somewhere above zero for negative expected value differences (where stay is an error), and a flat slope somewhere below one for positive ones (for which stay is the correct option). This is clearly not the case. Subjects gradually make less errors as the objective economic distance between the options becomes larger.

In contrast, the lower-right panel of Figure 2 depicts the relation between choice frequency and the numerical difference between the values of the (first) black card and the red card. Again, shaded areas correspond to errors, since the correct decision is to stay when the first black card was strictly smaller than the red one, and switch otherwise. The gradual relation is essentially absent in this case (with slopes being relatively flat), and there is a clear discontinuity at zero (the normative switching point). That is, subjects on average understood the task, but the numerical distance between card values

---

<sup>8</sup>For all figures, unless otherwise specified, each point represents each distinct value of the variable in the  $x$ -axis and the corresponding average value of the variable in the  $y$ -axis (choice frequencies or average response times). Therefore each point is an average across potentially different subjects and trials. The depicted curves are estimated using a fractional regression with a polynomial of second degree.

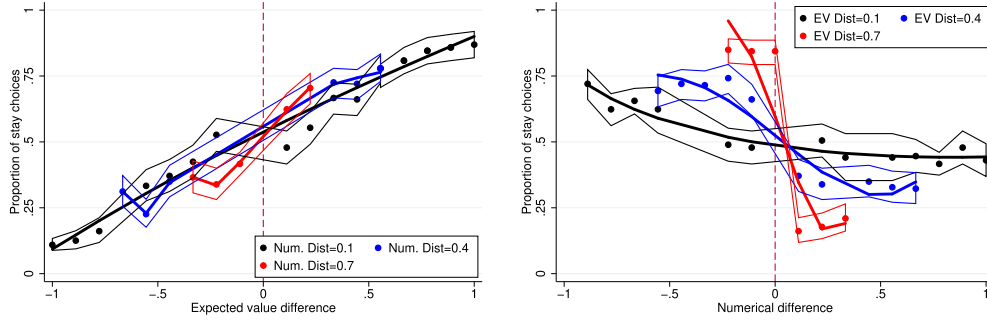


Figure 3: Experiment 1. Left-hand panel: Stay choices as a function of expected value differences for various, fixed numerical distances. Right-hand panel: Stay choices as a function of numerical distance for various, fixed expected value differences. The 95% binomial proportion confidence intervals are plotted.

does not play a large role. The comparison between panels suggests that the variability in responses arises mainly from differences in expected values, and not from numerical differences.

Figure 3 further investigates the relative contribution of the two dimensions of choice difficulty by letting one variable vary while keeping the other fixed (which is made possible by our design). In the left-hand panel, we plot choice frequencies as a function of expected value differences, separately for trials where the numerical distances correspond to three particular, fixed values (1, 2, or 3, corresponding to 0.1, 0.4, and 0.7 after normalization). The positive relation between the proportion of stay choices and expected value difference is unchanged, with the depicted curves overlapping. In contrast, the right-hand panel plots choice frequencies as a function of numerical distances, separately for trials where the expected value differences correspond to three particular, fixed values ( $1/90$ ,  $2/90$ , or  $3/90$ , again corresponding to 0.1, 0.4, and 0.7 after normalization). The relation changes drastically for different expected value differences, uncovering a negative, monotonic, and gradual relation between choices and the numerical difference between the stimuli, which becomes stronger for larger expected value differences. The figure suggests again that expected value differences are the determinant factor (gradually) influencing error rates, but also that, when keeping the economic dimension of choice difficulty fixed, second-order effects appear which are compatible with common findings from the perceptual literature.

We now turn to a regression analysis. The data form a strongly balanced panel with 160 trials for each of the 93 participants. Table 1 shows random-effects Probit regressions where the dependent variable is 1 in case of a correct answer.<sup>9</sup> Model 1 establishes the basic effect, namely that larger (objective) economic distances lead to less errors. Model 2 introduces the numerical distance and shows that this variable also

<sup>9</sup>For this and all other regressions below, relying on a fixed effects regression instead does not affect the main results, showing that they are not determined by heterogeneity among subjects.

Table 1: Experiment 1. Random-effects Probit regressions on correct answers.

Correct	Model 1	Model 2	Model 3	Model 4
Econ. Dist.	1.229*** (0.099)	1.447*** (0.107)	1.622*** (0.140)	1.620*** (0.141)
Num. Dist.		0.362*** (0.087)	0.584*** (0.094)	0.582*** (0.093)
Econ. Dist. $\times$ Num. Dist.			-0.010*** (0.004)	-0.010*** (0.004)
Round	-0.001*** (0.000)	-0.001** (0.000)	-0.001** (0.000)	-0.007** (0.003)
Constant	0.233** (0.095)	0.003 (0.096)	-0.061 (0.103)	0.886*** (0.284)
Controls	No	No	No	Yes
Log L.	-7358.011	-7337.550	-7329.264	-7265.955
Wald test	163.304***	191.995***	188.653***	251.391***
Obs.	14880	14880	14880	14880

Robust standard errors in brackets, \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

leads to lower error rates, revealing perceptual effects on top of value-induced ones.<sup>10</sup> Since both distances are normalized, we can compare the magnitude of the two effects. The regression coefficients for economic distance range from 1.2 to 1.6, and the regression coefficients for numerical distance range from 0.36 to 0.58. We can also calculate the relative elasticity of the two variables. A percentage variation in the economic distance predicts an average increase of 21.84% in the probability of a correct answer, while a percentage variation in the numerical distance predicts an average increase of only 5.40% in the probability of a correct answer. This provides further evidence for the predominant role of the economic dimension of choice difficulty over the perceptual.

As Figure 3 illustrates, our design allows to examine trials with identical expected value differences but different numerical distances, and vice versa. However, a purely mechanical effect prevents both variables from being fully orthogonal, as a larger numerical distance between the cards allows a larger number of feasible values of economic distance (the Spearman correlation between numerical and economic distances across the set of decisions is  $\rho = -0.6491$ ,  $N = 160$ ,  $p < 0.001$ ). Hence, in Model 3, we introduce the interaction between the two dimensions of choice difficulty as a control. The coefficient is significant and negative, reflecting the mechanical relation in the dimensions across the entire dataset. However, the main effects are unaffected by this control. Last, Model 4 adds a number of other controls: gender, native language, left-handedness, and accumulated earnings. The regression shows that the main effects are robust (see Appendix B for details on the controls).

<sup>10</sup>The results are unchanged if we define the numerical distance through the log of the numerical values of the cards, as suggested by Moyer and Landauer (1967).

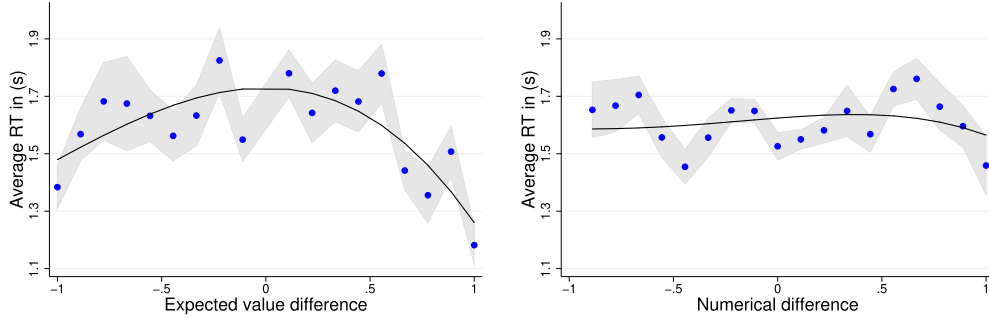


Figure 4: Average response times as a function of expected value difference (left-hand panel) and as a function of numerical distance (right-hand panel).

### 3.3 Response Times and the Underlying Processes

As an additional, independent test of the cardinal effects of economic distance, we measured response times for all decisions. For our purposes, the key decision is the second one within each trial. For this decision, we computed individual average response times. The average of those across individuals was 1.612 seconds (SD = 0.648, median 1.560, min = 0.272, max = 3.797).

The left-hand panel of Figure 4 plots average response times as a function of the expected value difference between stay and switch. There is clear evidence of gradual effects as postulated in psychophysics. An inverted U-shape is apparent, indicating a negative relation between response times and the distance in expected values between stay and switch. Choices closer to indifference (zero expected value difference) are associated with the longest response times. In contrast, the right-hand panel of Figure 4, which depicts the relation between response times and numerical differences, shows an essentially flat trend. That is, unlike in the case of expected value differences, there is no discernible pattern. In summary, response times suggest a gradual effect of (objective) economic distance (but not of numerical distance), confirming that the postulated relationship goes beyond a simple *as if* story and reflects actual decision processes.

We now turn to a regression analysis. Response times are a noisy variable, usually presenting a skewed, non-normal distribution and rare extreme observations. To account for these features it is common practice to take the logarithm of response times as the variable of interest in regression analyses (Fischbacher, Hertwig, and Bruhin, 2013; Achtziger and Alós-Ferrer, 2014). Table 2 reports random-effects regressions of log-transformed response times, taking advantage of the panel structure of the data. To control for individual differences in mechanical swiftness, we use the log-transformed response time for the non-rewarded, first black card and the log-transformed response time for pressing a space bar, which was required before the start of each trial. Model 1 establishes the basic effect, namely that responses are faster for larger (objective) economic distances, confirming that the phenomena we study reflect basic properties of actual

Table 2: Experiment 1. Random-effects regressions on log response times.

Log RT	Model 1	Model 2	Model 3	Model 4
Econ. Dist.	−0.174*** (0.020)	−0.237*** (0.031)	−0.321*** (0.035)	−0.316*** (0.035)
Num. Dist.		−0.106*** (0.029)	−0.224*** (0.038)	−0.209*** (0.039)
Econ. Dist. × Num. Dist.			0.005*** (0.001)	0.005*** (0.001)
Round	−0.001*** (0.000)	−0.001*** (0.000)	−0.001*** (0.000)	−0.005*** (0.003)
Constant	0.644*** (0.048)	0.712*** (0.055)	0.746*** (0.057)	0.853*** (0.253)
Controls	No	No	No	Yes
$R^2$ overall	0.108	0.109	0.110	0.130
Wald test	219.477***	220.200***	257.260***	259.572***
Obs.	14880	14880	14880	14880

Robust standard errors in brackets, \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

decision processes. Model 2 adds numerical distance. The coefficient is also significantly negative, although of a smaller magnitude (recall that both variables are normalized to have the same range). This shows that, in spite of the relatively flat shape of the aggregate relation as depicted in Figure 4, response times are also influenced by numerical differences, at least as a second-order determinant. As in the case of choice frequencies, Model 3 shows that the effects are robust to controlling for the interaction between the two distances. Finally, Model 4 shows that the results are robust to additional controls (see Appendix B for details).

### 3.4 Discussion of Experiment 1

The first experiment is probably as close as one can get to pure psychophysics in the economics domain. By using a gambling task with objectively correct answers, we can commit to the exact values of the explanatory variable *before* running the experiment; that is, we can rely on expected value differences and no utility estimation is needed. Still, the task is representative of the judgment and decision-making domain and remains intrinsically interesting for decision making under risk. We find a robust gradual relation between cardinal, objective economic distance, as captured by expected value differences, and error rates. Error rates gradually decrease as the distance between alternatives becomes larger (decisions become easier). Purely numerical effects (by how much a number is larger than another one, even if the comparison is payoff-irrelevant) do influence error rates as predicted by psychophysical studies, but this is a second-order effect and the main explanatory variable remains economic distance.

The fact that high error rates are observed in decisions under risk even when objectively-correct answers exist is of course not new. For instance, this is reflected in any experiment where subjects choose dominated options, in many tasks from the heuristics and biases literature, and in the probability matching literature (which shows that humans often choose intermediate probabilities even when they should go for corner solutions; Vulkan, 2000). What our analysis adds is that such errors are not simply white noise, and that an underlying, objective variable can be fruitfully used to better understand their structure.

Response times confirm the gradual relationship. As predicted by psychophysics, easier decisions (in the sense reflected by objective economic distance) are faster. This is important, because response times are a direct reflection of the underlying decision processes. Hence the relationship further confirms that the gradual effects of choice difficulty do reflect actual decision processes and not just a characteristic of how the statistical model of errors fits the data.

## 4 Experiment 2: Subjective Domain

Experiment 1 can be seen as a streamlined proof of concept which does away with the problems inherent in utility estimation. In Experiment 2, we parsimoniously go one step further by reproducing the analysis for more complex decisions under risk where what is “correct” depends on the individual risk attitude, and hence utility estimation is unavoidable. In this sense, Experiment 2 studies choices in the subjective domain, while Experiment 1 belonged to the objective domain. Crucially, to avoid the problems pointed out in the Introduction, we will strictly adhere to an out-of-sample approach where the utility used to test the gradual dependencies in the data is always estimated from a different part of the dataset. This ensures that the estimation allows us to test for the presence of gradual effects, instead of artificially creating them. Since the estimation procedure requires to focus on a parametric family of utility functions, we conducted several robustness tests to ensure that the effects do not hinge on the specific parametric form. We report on an estimation using CARA utilities below, but Appendix C repeats the analysis for CRRA functions instead, and Appendix D reproduces it relying on a random parameter model.

As in Experiment 1, we focus on error rates. We will have three (explanatory) variables of interest in sight. Of course, we will focus on expected utility as just described. Additionally, in this experiment we can examine the differences between expected value and expected utility differences as determinants of gradual effects on error rates. For completeness, we will also examine the potential effects of (payoff-irrelevant) numerical distance. Last, and again as a confirmatory exercise, we will examine the effects of those variables on response times.

In order to concentrate on the effects of strength of preference, we again strive to streamline the choice environment to reduce possible additional sources of choice diffi-

culty. In particular (and again as in the previous experiment), all decisions are embedded within a fixed, naturalistic environment.

#### 4.1 Design and Procedures

Implementation, procedures, and data collection were as in Experiment 1. Participants were  $N = 96$  (different) university students (66 females, age range 18 – 36, mean 24). Sessions lasted around 60 minutes and the average payoff was EUR 13.45 (around USD 14.40 at the time of the experiment). Three participants were unable to understand the task and were excluded from the analysis.

The experimental task is as follows. Participants are confronted with two decks of cards, a red one (Diamonds) and a black one (Clubs), containing ten cards each (numbered 1 to 10). At the beginning of each of the 170 trials, *two* cards are extracted from the black deck, one red card is extracted from the red deck, and a monetary prize is displayed (see Figure 5). The participants’ task is to decide whether to bet or to pass. After this decision, a further black card will be extracted from the remaining eight cards in the black deck, and the objective is to beat the red card with that new card. Betting is costly: placing a bet costs EUR 0.10 (fixed for all trials), independently of the outcome of the trial. If the participant bets and if the newly-extracted black card is strictly larger than the displayed red card, the participant receives the displayed monetary prize (minus the cost). Otherwise, the payment is zero (resulting in a net loss equal to the cost of betting). If the participant does not bet, there is neither a payment nor a cost. Before a new trial starts, all cards are returned to their respective decks and those are reshuffled. Hence, each trial reflects an independent decision situation. All trials were paid. In our particular context, this payment mechanism is incentive-compatible under mild assumptions on the participants’ preferences, as shown by Azrieli, Chambers, and Healy (2018, 2020).<sup>11</sup>

The set of initial stimuli (red card, first two black cards, and prize) was predetermined and pseudorandomized across trials to achieve adequate stimuli variance. The crucial third black card was randomly selected among the cards remaining in the deck. Red cards were extracted in such a way that there was always some probability of winning, so as to avoid trivial decisions. Hence, there were eight possible distinct probabilities of winning, ranging from 12.5% to a sure win. Prizes ranged from 10 to 120 cents, and were determined trial-by-trial as deviations from the actuarially-fair prize, the amount that leaves a risk-neutral agent indifferent between betting and passing. Eleven different distortions from the fair prize were implemented, ranging from 50% below to 50% above, in 10% steps.

---

<sup>11</sup>Specifically, the requirements are monotonicity and a weak condition called *show-up fee invariance*, which Azrieli, Chambers, and Healy (2018, footnote 26) argue to be a reasonable assumption in contexts as ours where choices are independent and feedback is given independently of choices. This is because each decision problem can be considered independent, since participants receive feedback on what would have happened for every choice they could have made before, and hence incentives to experiment (hedge) are eliminated (Azrieli, Chambers, and Healy, 2018).

In each trial, at the moment of the decision, the black deck contains eight cards, and the two already-extracted cards are displayed. The probability to win when betting depends on the magnitude of the red card and on whether the displayed black cards are winning or losing cards. In the example depicted in the left-hand panel of Figure 5, the red card is an 8 and the two extracted black cards are a 2 and a 4, hence both are losing cards. That is, the black deck contains two winning cards and six losing ones, yielding a probability of winning of  $1/4$ . Since the cost of betting is 10 cents, the actuarially-fair prize is 40 cents, but the offered prize is 24 cents. Hence, a risk-averse or risk-neutral agent should decline to bet, while a risk-loving one might rationally decide to bet. That is, there are no objectively-correct decisions in this task; rather, what is “correct” depends on the individual risk attitude. Therefore, we hypothesized that the natural measure of choice difficulty or *subjective economic distance* would be the difference between the expected utilities of betting and passing, referred to as EU distance for clarity, which requires us to estimate the underlying individual utilities of money.

By design, however, the expected value of betting depends on the distortion of the fair prize. For risk neutral individuals, the difference in expected value between passing and betting reflects how far away from “indifference” the participants were, and are hence a natural, alternative measure for “strength of preference.” Therefore, another candidate determinant of gradual effects is simply the absolute value of the expected value differences between betting and passing, which we refer to as EV distance. In contrast to Experiment 1, the comparison between these two measures of economic distance is informative of which is the relevant measure of strength of preference in this context.

We remark also that the probability of winning does *not* depend on the numerical distances between the black cards and the red one, but only on whether the former are larger or smaller than the latter. Hence, numerical distances in themselves are payoff-irrelevant (but the sign of the numerical differences is not). Analogously to Experiment 1, this allows us to disentangle the numerical closeness of stimuli as a further possible dimension of choice difficulty, which is the closest one to standard measures of perceptual similarity used in psychophysics. Since there are two black cards, we have different possible candidates for numerical distance. We present here the analysis using the distance between the red card and the second, most recent black card, since a large literature has advocated the prominence of the recency effect (Deese and Kaufman, 1957; Murdock Jr., 1962). There are ten possible perceived distances between the red card and the second black card, ranging from 0 to 9. We refer to this magnitude as *numerical distance*. We also carried out analyses with other definitions of numerical distance; the main results described below are unaffected.<sup>12</sup>

---

<sup>12</sup>We also considered the distance between the highest black card and the red one, the distance between the average of the two black cards and the red card, and the distance between the highest or lowest black card and the red one, and the distances between log-transformed numbers.

## 4.2 Utility Estimation

We estimate out-of-sample risk attitudes for each subject. Specifically, we use the choices made in odd trials to estimate risk attitudes and use this estimation to predict the expected utility in the even trials, and vice versa.<sup>13</sup>

Following a standard approach, we estimate an additive random utility model (RUM) which considers a given utility function plus an additive noise component (e.g., Thurstone, 1927; Luce, 1959; McFadden, 2001). The estimation procedure employs well-established techniques as used in many recent contributions (Von Gaudecker, Van Soest, and Wengström, 2011; Conte, Hey, and Moffatt, 2011; Moffatt, 2015). We provide a short description in Appendix A.

For the functional form of the utility function, we adopt a normalized constant absolute risk aversion (CARA) function as in Conte, Hey, and Moffatt (2011), which is given by

$$u(x) = \begin{cases} \frac{1 - \exp(-rx)}{1 - \exp(-rx_{\max})}, & \text{if } r \neq 0 \\ \frac{x}{x_{\max}}, & \text{if } r = 0, \end{cases}$$

where  $x_{\max}$  is the upper bound of the outcome variable  $x$ . For the noise term, we consider normally-distributed errors. All our results remain qualitatively unchanged if we assume a CRRA utility function instead; we provide the corresponding analysis in Appendix C. However, random utility models have been recently criticized (see Wilcox, 2008, 2011; Bruner, 2017; Apesteguía and Ballester, 2018; Vieider, 2018). In Appendix D we repeat the analysis using a random preference model (RPM; Loomes and Sugden, 1998; Apesteguía and Ballester, 2018), which follows a different approach for the specification of noise. Again, the results remain qualitatively unchanged. That is, our results do not hinge on the particular functional form of the utility function or error specification.

The estimated risk propensities in our dataset have an average  $r = 0.026$  (SD = 0.016, median 0.025, min = -0.007, max = 0.086). The risk propensity estimated on odd trials ( $r = 0.027$ ) is not significantly different from the one estimated on even trials ( $r = 0.025$ ; Wilcoxon Signed-Rank test,  $N = 93$ ,  $z = 0.738$ ,  $p = 0.463$ ). Given the estimated risk attitudes 3 subjects are classified as risk-seeking, while some subjects display values of  $r$  close to 0, indicating risk neutrality. However, the majority of subjects are estimated to be risk averse. As an illustration, an agent with a risk propensity equal to the average in our sample,  $r = 0.025$ , would have a certainty equivalent of about \$4.69 when facing a lottery paying \$10 with 50% probability and zero otherwise.

For the simulation reported in Figure 1, we generated a dataset where each of 93 fictitious subjects randomly chose 170 times between accepting certain bets or not (the dataset mimics the basic features of Experiment 2). Bets involved a certain probability of a positive prize, and led to the loss of a small amount of money with the remaining probability. The outside option always yielded zero payoffs. Prizes and probabilities

---

<sup>13</sup>Our results do not change if we use different out-of-sample approaches, as e.g. using an initial block of observations for the estimation and predicting the expected utility for the remaining trials.

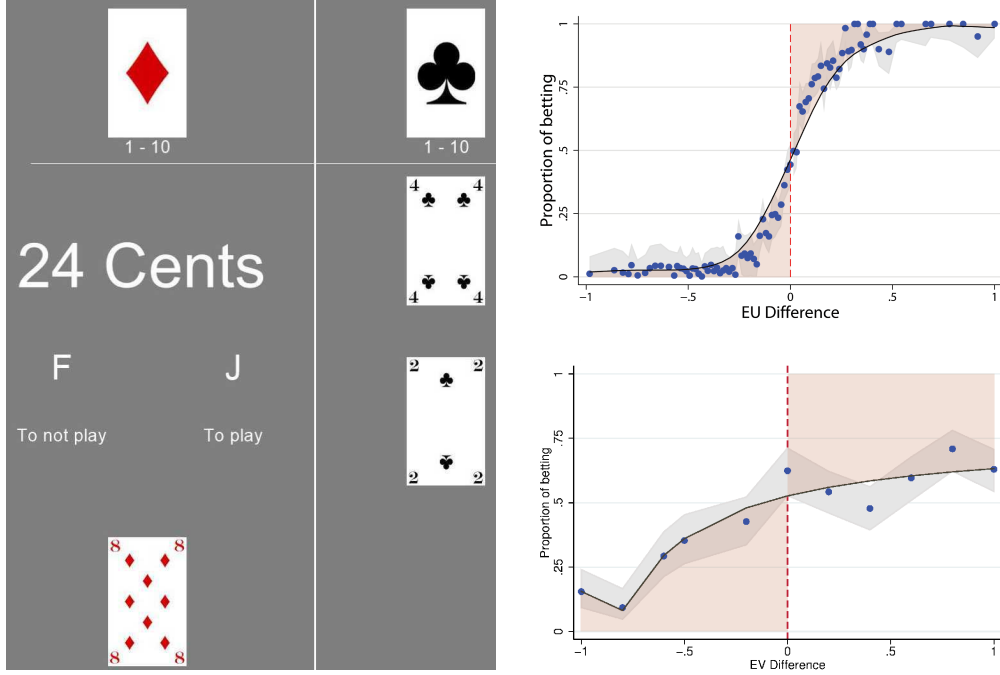


Figure 5: Experiment 2. Each trial starts by extracting two black cards, a red card, and displaying a prize (left). Participants then decide whether to bet or not (betting is costly). After a bet, a black card is extracted, and the participant wins if and only if it is strictly larger than the red card. Proportion of betting decisions as a function of expected utility differences (upper right) and expected value differences (lower right). Gray areas indicate 95% binomial proportion confidence intervals. Shaded areas indicate the proportion of errors.

changed across trials, but the amount potentially lost was fixed. The set of bets was such that a risk-neutral subject would accept half of the times. “Decisions” were fully random and unrelated to the options. The right-hand panel of Figure 1 corresponds to an estimation performed exactly as described above for Experiment 2. The left-hand panel depicts the results of an estimation using the same CARA functional form, but with a standard within-sample approach as common in the literature. Specifically, we implemented a RUM assuming heteroskedastic Fechner errors (Fechner, 1860; Hey and Orme, 1994). We used the estimated risk attitudes to compute, *within sample*, the expected utility difference between the two options (betting minus passing), and then plotted this difference against the proportion of times one option was chosen over the other. As argued in the introduction, the difference between both approaches shows that the estimation procedure might create apparent gradual effects simply because they are assumed in the underlying random utility model. Our out-of-sample procedure ensures that the regularities we uncover correspond to actual features of the data.

### 4.3 Choices and Errors

We define an error as a choice which gives a negative expected utility, e.g. deciding to bet when the expected utility (as estimated out of sample) of betting is strictly smaller than the expected utility of passing. The mean error rate across participants was 27.31%, with a median of 28.24% (SD = 8.44, max 44.12%, min 10.00%). Figure 5 (right-hand panels) plots the frequency of betting decisions for each possible value of each variable. As in previous pictures, to facilitate the comparison, in all figures and regressions the various distances are normalized to be between 0 and 1. The upper panel plots the dependence on expected utility differences. The shaded areas correspond to errors with the definition above. We observe that the relation between betting frequency and expected utility differences has a sigmoidal shape resembling a cumulative normal distribution or a logistic curve. This shape indicates that error rates decrease gradually as the difference in expected utilities between the options becomes larger. For very large differences, error rates are close to zero. For differences close to zero, error rates are close to 50%. This stands in sharp contrast to deterministic, neoclassical models, which would predict that subjects always bet when expected utility differences are positive and always pass when they are negative.

The lower right-hand panel of Figure 5 plots the proportion of betting choices as a function of the differences in expected value (betting minus passing). We observe a positive but non-monotonic trend with greater expected values corresponding roughly to a higher frequency of betting.<sup>14</sup> This is not surprising, since as long as utility is increasing on monetary amounts, there will be some positive correlation between expected utility and expected values in a dataset. However, the figure strongly suggests that expected utility differences better explain gradual effects on error rates than differences in expected values.

Similarly to Experiment 1, one can depict the proportion of betting decisions as a function of numerical distances as defined above (errors, however, cannot be derived from numerical distance alone in this experiment). For the sake of brevity, we omit this figure (see Figure B1 in Appendix B) and simply comment that a graphical representation suggests a weak, noisy monotonic relation which might hint to second-order effects but offers no strong evidence of an impact of purely numerical, payoff-irrelevant perceptions on choice frequencies.

In summary, our data shows that, as in Experiment 1, there is a gradual relation between economic distance and error rates, but the former now corresponds to differences in expected utilities. We remark that this and subsequent results do not hinge on the specific functional form or error structure that we assume. Equivalent results are obtained assuming CRRA instead of CARA (Appendix C) or considering random parameter models instead of RUMs (Appendix D).

<sup>14</sup>Errors in this panel are defined as decisions which contradict expected value differences. According to this risk-neutral definition, the mean error rate across participants was 36.29%, with a median of 36.47% (SD = 6.60, min 19.41%, max 54.12%).

We now turn to a regression analysis. The data form a strongly balanced panel with 170 trials for each of the 93 participants. We ran random-effects panel Probit regressions where the dependent variable is 1 in case of a correct answer. For completeness, we provide separate analyses for expected utility (Table 3) and expected value differences (Table 4), while controlling for numerical distance in both. Recall that Expected Utility distance (EU distance), Expected Value distance (EV distance), and numerical distance are all normalized to range from 0 to 1. The various regression models are built in a completely analogous way, and hence we discuss them simultaneously. Note that the definitions of errors is the natural one in each table, i.e. choices against expected utility differences in Table 3 and choices against expected value differences in Table 4.

In Model 1 of both tables we see that larger economic distances lead to less errors, confirming the basic prediction. However, there is a considerable difference in the magnitude of the estimated coefficients, with EU distance having a coefficient almost 30 times bigger than EV distance. To conduct a proper comparison, we calculated the relative elasticities. A percentage variation in EU distance increases the probability of a correct answer by an average of 18.31%, while the analogous percentage for EV distance is only 11.98%. This confirms the message from Figure 5 that differences in expected utility, and not in expected value, are the relevant dimension of strength of preference in this context.<sup>15</sup>

Model 2 in both tables introduces numerical distance as an additional control. In the presence of EU distance, numerical similarity does not seem to play a role. The effect is also insignificant when controlling for the interaction between numerical distance and EU distance (Model 3), and when adding further controls (Model 4; Appendix B contains the details on the controls). In the presence of EV distance, numerical effects are not statistically significant. They only become significant when we further control for the interaction between numerical distance and EV distance (Model 3) as well as other controls (Model 4). The results for numerical distance should be attributed to the fact that there is a (mechanical) correlation between the expected value and numerical distance across all decisions in the dataset (Spearman’s  $\rho = 0.1453$ ;  $N = 170$ ,  $p = 0.0587$ ), but there is no correlation between numerical distance and expected utility (Spearman’s  $\rho = 0.0931$ ,  $N = 170$ ,  $p = 0.2271$ ). In all models we control for learning effects. Participants appear to improve as the experiment advances (Round, 1–170).

#### 4.4 Response Times and the Underlying Processes

The previous section shows that differences in expected utilities are the best candidate as an explanatory determinant of gradual effects on errors. Expected value differences and numerical differences also display significant effects, but those are of a smaller magnitude

---

<sup>15</sup>The same conclusion is obtained in regressions using utilities estimated with a CRRA functional form (Appendix C) or considering random parameter models instead of RUM (Appendix D).

Table 3: Experiment 2. Random-effects Probit regressions on correct answers for EU distance, defined as passing when  $EU \leq 0$  and betting when  $EU \geq 0$ .

Correct	Model 1	Model 2	Model 3	Model 4
EU_Dist.	13.854*** (0.443)	13.881*** (0.463)	16.345*** (0.625)	16.277*** (0.639)
Num._Dist.		0.008 (0.044)	0.006 (0.043)	0.003 (0.044)
Num._Dist. $\times$ EU_Dist.			-5.544*** (0.924)	-5.332*** (0.907)
Round	0.003*** (0.000)	0.003*** (0.000)	0.003*** (0.000)	0.003*** (0.000)
Constant	-0.130*** (0.040)	-0.131*** (0.038)	-0.139*** (0.038)	0.536** (0.257)
Controls	No	No	No	Yes
Log L.	-8392	-8392	-8379	-8320
Wald test	984.525***	985.811***	1024.929***	1041.316***
Obs.	15810	15810	15810	15810

Robust standard errors in brackets, \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Table 4: Experiment 2. Random-effects Probit regressions on correct answers for EV distance, defined as passing when  $EV \leq 0$  and betting when  $EV \geq 0$ .

Correct	Model 1	Model 2	Model 3	Model 4
EV_Dist.	0.463*** (0.053)	0.464*** (0.053)	1.100*** (0.069)	1.113*** (0.070)
Num._Dist.		-0.013 (0.030)	0.628*** (0.061)	0.639*** (0.062)
Num._Dist. $\times$ EV_Dist.			-1.552*** (0.116)	-1.600*** (0.124)
Round	0.002*** (0.000)	0.002*** (0.000)	0.002*** (0.000)	0.002*** (0.000)
Constant	-0.011 (0.021)	-0.006 (0.025)	-0.285*** (0.035)	-0.159 (0.258)
Controls	No	No	No	Yes
Log L.	-10170	-10170	-10102	-10037
Wald test	223.538***	223.868***	453.437***	466.95***
Obs.	15810	15810	15810	15810

Robust standard errors in brackets, \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

and appear less robust. In this section, we further compare the gradual effects of all three variables by focusing on response times. The main objective is to show that, while there appears to be a strong, clear correspondence between expected utility differences and actual human decision processes as reflected by response times, that relation is far from clear when it comes to other alternative variables.

The variable of interest is the time participants took to decide whether to bet or to pass. The average across individual average response times for this decision was 2.051

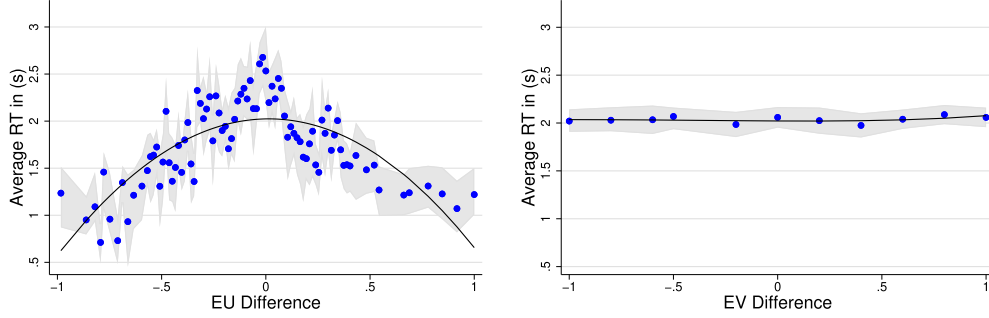


Figure 6: Experiment 2. Average response times on EU Difference (left) and EV difference (right). Gray areas indicate 95% confidence intervals.

seconds ( $SD = 0.150$ , median = 2.090, min = 1.354, max = 2.200). Figure 6 plots average response times as a function of expected utility differences (left) and of expected value differences (right). Response times and EU distance clearly show an inverted U-shaped relation. Harder decisions, resulting in longer response times, are those corresponding to smaller expected utility differences. However, the figure shows no systematic relation with EV differences (the corresponding coefficient is not significantly different from zero,  $\text{coef.} = 0.012, z = 0.85, p = 0.393$ ). A similar result to EV differences is obtained for the numerical distance ( $\text{coef.} = 0.017, z = 0.25, p = 0.803$ ; see Figure B1 in Appendix B). This provides an independent confirmation that a larger strength of preference, in the sense of larger subjective economic distance, can be linked to easier decisions.

As for Experiment 1, we conducted a panel regression analysis for log-transformed response times. Tables 5 and 6 report the corresponding regressions using expected utility distances and expected value distances as a measure of strength of preference, respectively. In all models with further control for individual differences in mechanical swiftness using the log of the response time for pressing the space bar to move to the next trial.

Response times are significantly shorter for larger EU distances across all models in Table 5. This effect is robust to controlling for numerical distance and additional controls. Additionally, numerical distance does have an effect on response times, validating the view from psychophysics (Moyer and Landauer, 1967; Dehaene, Dupoux, and Mehler, 1990) that even payoff-irrelevant perceptual differences might influence actual choice difficulty. That is, in addition to the effects of subjective economic distance, response times are shorter for more perceptually distinguishable stimuli (larger numerical distance).

In contrast, the effect of expected value differences is less clear. In Model 1 of Table 6, we observe larger response times for *higher* values of EV distance, contrary to expectations if EV distance was taken to explain the gradual effects of strength of preference. However, the effect becomes non-significant when we control for the relation

Table 5: Experiment 2. Random-effects regressions on log response times, EU distance.

Log RT	Model 1	Model 2	Model 3	Model 4
EU_Dist.	-1.286*** (0.223)	-1.482*** (0.214)	-1.149*** (0.206)	-1.154*** (0.206)
Num._Dist.		-0.055*** (0.013)	-0.136*** (0.019)	-0.136*** (0.019)
Num._Dist. $\times$ EU_Dist.			1.823*** (0.286)	1.824*** (0.285)
Round	-0.002*** (0.000)	-0.002*** (0.000)	-0.002*** (0.000)	-0.003*** (0.000)
Constant	1.313*** (0.040)	1.320*** (0.040)	1.317*** (0.040)	1.268*** (0.309)
Controls	No	No	No	Yes
$R^2$ overall	0.135	0.136	0.140	0.159
Wald test	458.704***	523.572 ***	559.366***	579.216 ***
Obs.	15810	15810	15810	15810

Robust standard errors in brackets, \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Table 6: Experiment 2. Random-effects regressions on log response times, EV distance.

Log RT	Model 1	Model 2	Model 3	Model 4
EV_Dist.	0.054*** (0.015)	0.068*** (0.015)	0.009 (0.023)	0.009 (0.023)
Num._Dist		-0.149*** (0.018)	-0.211*** (0.024)	-0.211*** (0.024)
Num._Dist. $\times$ EV_Dist.			0.145*** (0.044)	0.145 (0.043)
Round	-0.003*** (0.000)	-0.003*** (0.000)	-0.003*** (0.000)	-0.003*** (0.000)
Constant	1.231*** (0.037)	1.288*** (0.039)	1.315*** (0.040)	1.290*** (0.310)
Controls	No	No	No	Yes
$R^2$ overall	0.132	0.136	0.136	0.154
Wald test	472.772***	607.972***	665.532***	687.710***
Obs.	15810	15810	15810	15810

Robust standard errors in brackets, \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

between EV distance and numerical distance as well as other controls (Models 3 and 4; Appendix B adds details on the controls). Again, the analysis is consistent with the view that expected utility differences are the key variable explaining the gradual effects that we investigate. In all models we control for time trends, reproducing the standard observation that subjects become slightly faster over time. Other controls deliver no additional insights.

## 4.5 Discussion of Experiment 2

The second experiment considers standard economic decisions under risk (betting), which are an example of preferential choice where there is no objectively correct alternative. Contrary to the first experiment, the appropriate dimension explaining gradual effects on error rates needs to be estimated from the data. Our evidence shows that expected utility, and not expected value, is the appropriate integrated variable capturing strength of preference. Choices with a larger expected utility difference between the alternatives result in lower error rates and shorter response times. The effects are robust and obtain even though we use a strictly out-of-sample approach, that is, they are not an artifice of the estimation method. Further, the link to response times shows that the relationship between expected utility differences and choice difficulty reflects the characteristics of actual decision processes, rather than being just “as if” modeling.

Numerical distance, seen as a more perceptual dimension of choice difficulty, plays a minor role. The effects on error rates are small and not robust to the addition of controls. Response times suggest that a second-order effect is present, but expected utility differences are the major determinant of the effects we study.

## 5 Discussion

*Homo oeconomicus* does not play dice (but *homo sapiens* might). A fully rational economic agent would be consistent, choosing an option 100% of the time if it delivered a slightly larger payoff than the alternative, and 0% if a minute payoff reduction left it worse than the alternative. However, considerable evidence suggests that the implementation of decision processes in the human brain follows processes of a more gradual nature (e.g., Shadlen and Kiani, 2013). We have demonstrated the existence of a stable, gradual relation between error rates in decisions under risk and an underlying, cardinal “strength of preference,” and shown that the latter is best represented by integrated variables of an exclusively economic nature. That is, decisions become more error-prone as the economic distance between the alternatives becomes smaller.

Our research strategy has followed three complementary approaches. First, we have shown that, in decisions in the domain of judgment and decision making where objectively-correct options can be identified, expected value differences are enough to explain error rates. This is important, because such an explanatory variable is independent of any estimation of subjective values and hence constitutes the direct parallel to psychophysical studies which have identified gradual effects as a function of objective differences in weight, length, brightness, etc. The typical candidate explanatory variable derived from purely psychophysical approaches for economic tasks, (payoff-irrelevant) numerical differences (Moyer and Landauer, 1967; Dehaene, Dupoux, and Mehler, 1990), does play a role but can be safely considered a second-order variable.

Second, we have shown that, in decisions under risk in the subjective domain, where what is correct depends on individual risk attitudes, strength of preference can be characterized by an integrated variable reflecting differences in expected utility, while expected value differences do a considerably worse job. Again, numerical differences do play a role, but appear to be relatively less important than pure economic distance. Crucially, our approach has followed a strictly out-of-sample procedure where utility functions are estimated on one part of the dataset and the test of gradual effects between utility differences and error rates is conducted using the choices in a different part of the dataset. This is important, because fitting a dataset with, say, a random utility model merely *assumes* that errors follow a smooth distribution; that is, gradual effects are *assumed* and would appear to be present after the fact even if they did not exist at all.

Third, we have shown that the relation between strength of preference, as captured by notions of economic distance, and error rates reflects more than an *ex post* and *as if* model. The same gradual effects are obtained when examining response times, with easier decisions (where economic distance is large) being made faster than harder ones (where economic distance is small). Response times are an easily-measurable reflection of the actual functioning of human decision processes. Most importantly, they are unrelated to estimation and fitting procedures and hence serve as an independent confirmation of the postulated effects.

Our results provide empirical support and explicit foundation for the literature on stochastic choice, which has been long advocated as a realistic building block for theories of microeconomic decision making (Debreu, 1958; Davidson and Marschak, 1959; Luce, 1959; Machina, 1985). It is in line with modern empirical contributions pointing out the ubiquitousness of stochastic choice and decision inconsistencies (Camerer, 1989; Hey and Orme, 1994; Hey, 2001; Agranov and Ortoleva, 2017), but goes beyond those by precisely examining the content of elusive concepts as “strength of preference” and isolating them from possibly-artificial phenomena derived, e.g., from the underlying assumptions of models used to estimate noisy utility.

The results also speak to the role of choice difficulty in economic decision making. Of course, our experiments are (on purpose) stylized and, by design, shut down a number of additional dimensions of choice difficulty that are bound to also play a role in economic decisions. Those range from the complexity of the options’ description to the presence of transparent relationships (e.g., dominance) and whether the decision environment cues in cognitive shortcuts or not. Our claim is merely that strength of preference is one of the relevant dimensions underlying choice difficulty, and in particular it is a dimension that can be characterized by measurable variables with an explicitly economic content.

The analysis is broadly in line with the psychophysics and neuroscience literature, where the presence of gradual effects on decision making is regarded as an elementary, firmly-established fact (e.g., Weber’s Law), but goes beyond it by showing that economic distance is not as simple as objectively-measurable weight or length. Economic decisions are decisions, and hence it is unsurprising that they might share (neural) mechanisms

with perception-based judgments. However, economic decisions are *complex* decisions, and it is equally unsurprising that simple applications of psychophysics (as, say, focusing on numerical magnitudes only) fall short of the task of accounting for economic errors.

Conceptually, our results agree with earlier studies as Mosteller and Nogee (1951) and with recent contributions as Khaw, Li, and Woodford (2018). Both report gradual increases in the proportion of risky choices in lottery experiments as the reward increases. Khaw, Li, and Woodford (2018) argue in terms of an imprecise perception of stimuli. Those are payoff-relevant numerical magnitudes and hence aligned with economic distance as we consider it.

It is also important to remark that the sigmoidal relation between economic distance and choice frequencies arises spontaneously from the data, and either require no utility estimation (Experiment 1) or are robust to alternative utility and error specifications (Experiment 2; Appendices C and D). Hence, our results provide empirical support for random utility models as typically used in applied microeconomics, which often employ logit or probit error distributions. By taking a step back from fitting approaches, our analysis highlights the presence of a systematic structure of noise terms reflecting the gradual effects of choice difficulty. This observation builds upon earlier arguments by Hey and Orme (1994) and Harless and Camerer (1994), which attempted to shift the focus in microeconomics away from deterministic choice models as alternatives to expected utility theory.<sup>16</sup>

The implications of our results are of broad significance for economic modeling. First, the systematic demonstration of the gradual relation between economic integrated variables and errors provides a foundation for theories of stochastic choice and empirical approaches to preference revelation alike. Second, the fact that these effects are a natural extension of those observed in psychophysics provides a tangible bridge to other disciplines, most notably neuroscience, through which new techniques and ideas can travel (in both directions). Third, accounting for strength of preference might be important to gauge the actual extent of received anomalies in economic decision making, ranging from the endowment effect (Knetsch, 1989; Kahneman, Knetsch, and Thaler, 1990) to decoy effects (Huber, Payne, and Puto, 1982; Soltani, De Martino, and Camerer, 2012) or preference reversals (Grether and Plott, 1979; Alós-Ferrer et al., 2016). The reason is that these anomalies are often demonstrated for choices where the options are of similar value along some economic dimension (e.g., the monetary value of the options in Knetsch, 1989), and thus larger error rates are to be expected. Fourth, the results pose a significant conceptual challenge to traditional, neoclassic *as if* modeling, because the latter is based on deterministic and, more importantly, *purely ordinal* preferences. Our results on gradual mappings from economic variables to error rates and response times go beyond any *as if* interpretation, and, in our opinion, are best viewed in the context of an inherently cardinal view of preferences.

---

<sup>16</sup>“Perhaps we should now spend some time on thinking about the noise, rather than about even more alternatives to expected utility?” (Hey and Orme, 1994).

## References

- Achtziger, Anja and Carlos Alós-Ferrer. 2014. “Fast or Rational? A Response-Times Study of Bayesian Updating.” *Management Science* 60 (4):923–938.
- Agranov, Marina and Pietro Ortoleva. 2017. “Stochastic Choice and Preferences for Randomization.” *Journal of Political Economy* 125 (1):40–68.
- Alekseev, Aleksandr, Glenn W. Harrison, Morten Lau, and Don Ross. 2019. “Deciphering the Noise: The Welfare Costs of Noisy Behavior.” Working Paper.
- Alós-Ferrer, Carlos, Ernst Fehr, and Nick Netzer. 2018. “Time Will Tell: Recovering Preferences when Choices are Noisy.” Working Paper, University of Zurich.
- Alós-Ferrer, Carlos, Dura-Georg Granić, Johannes Kern, and Alexander K. Wagner. 2016. “Preference Reversals: Time and Again.” *Journal of Risk and Uncertainty* 52 (1):65–97.
- Alós-Ferrer, Carlos and Alexander Ritschel. 2018. “The Reinforcement Heuristic in Normal Form Games.” *Journal of Economic Behavior and Organization* 152:224–234.
- Anderson, Simon P., Jacques-François Thisse, and André De Palma. 1992. *Discrete Choice Theory of Product Differentiation*. Cambridge, MA: MIT Press.
- Apesteguía, José and Miguel A. Ballester. 2018. “Monotone Stochastic Choice Models: The Case of Risk and Time Preferences.” *Journal of Political Economy* 126 (1):74–106.
- Apesteguía, José, Miguel A. Ballester, and Jay Lu. 2017. “Single-Crossing Random Utility Models.” *Econometrica* 85 (2):661–674.
- Azrieli, Yaron, Christopher P. Chambers, and Paul J. Healy. 2018. “Incentives in Experiments: A Theoretical Analysis.” *Journal of Political Economy* 126 (4):1472–1503.
- . 2020. “Incentives in Experiments with Objective Lotteries.” *Experimental Economics* 23 (1):1–29.
- Baldassi, Carlo, Simone Cerreia-Vioglio, Fabio Maccheroni, and Massimo Marinacci. 2019. “A Behavioral Characterization of the Drift Diffusion Model and its Multi-Alternative Extension to Choice under Time Pressure.” *Management Science* forthcoming.
- Ballinger, T. Parker and Nathaniel T. Wilcox. 1997. “Decisions, Error and Heterogeneity.” *Economic Journal* 107 (443):1090–1105.
- Bar-Hillel, Maya. 1980. “The Base-Rate Fallacy in Probability Judgments.” *Acta Psychologica* 44:211–233.
- Bruner, David M. 2017. “Does Decision Error Decrease with Risk Aversion?” *Experimental Economics* 20 (1):259–273.
- Camerer, Colin F. 1989. “Does the Basketball Market Believe in the ‘Hot Hand’?” *American Economic Review* 79:1257–1261.
- Cattell, James McKeen. 1893. “On Errors of Observation.” *The American Journal of Psychology* 5 (3):285–293.

- . 1902. “The Time of Perception as a Measure of Differences in Intensity.” *Philosophische Studien* 19:63–68.
- Chabris, Christopher F., Carrie L. Morris, Dmitry Taubinsky, David Laibson, and Jonathon P. Schuldt. 2009. “The Allocation of Time in Decision-Making.” *Journal of the European Economic Association* 7 (2-3):628–637.
- Charness, Gary and Dan Levin. 2005. “When Optimal Choices Feel Wrong: A Laboratory Study of Bayesian Updating, Complexity, and Affect.” *American Economic Review* 95 (4):1300–1309.
- Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman. 2014. “Who Is (More) Rational?” *American Economic Review* 104 (6):1518–1550.
- Clithero, John A. 2018. “Improving Out-of-Sample Predictions Using Response Times and a Model of the Decision Process.” *Journal of Economic Behavior and Organization* 148:344–375.
- Conte, Anna, John D. Hey, and Peter G. Moffatt. 2011. “Mixture Models of Choice Under Risk.” *Journal of Econometrics* 162 (1):79–88.
- Dashiell, John F. 1937. “Affective Value-Distances as a Determinant of Aesthetic Judgment-Times.” *American Journal of Psychology* 50:57–67.
- Davidson, Donald and Jacob Marschak. 1959. “Experimental Tests of a Stochastic Decision Theory.” In *Measurement: Definitions and Theories*, vol. I, Part I, edited by West Churchman and Philburn Ratoosh. New York: Wiley, 233–269.
- Debreu, Gerard. 1958. “Stochastic Choice and Cardinal Utility.” *Econometrica* 26 (3):440–444.
- Deese, James and Roger A. Kaufman. 1957. “Serial Effects in Recall of Unorganized and Sequentially Organized Verbal Material.” *Journal of Experimental Psychology* 54 (3):180–187.
- Dehaene, Stanislas. 1992. “Varieties of Numerical Abilities.” *Cognition* 44 (1-2):1–42.
- Dehaene, Stanislas, Emmanuel Dupoux, and Jacques Mehler. 1990. “Is Numerical Comparison Digital? Analogical and Symbolic Effects in Two-Digit Number Comparison.” *Journal of Experimental Psychology: Human Perception and Performance* 16 (3):626–641.
- Dehaene, Stanislas, Véronique Izard, Elizabeth Spelke, and Pierre Pica. 2008. “Log or Linear? Distinct Intuitions of the Number Scale in Western and Amazonian Indigene Cultures.” *Science* 320 (5880):1217–1220.
- Fechner, T. Gustav. 1860. *Elemente der Psychophysik [Elements of Psychophysics]*. Leipzig: Breitkopf & Härtel.
- Fischbacher, Urs, Ralph Hertwig, and Adrian Bruhin. 2013. “How to Model Heterogeneity in Costly Punishment: Insights from Responders’ Response Times.” *Journal of Behavioral Decision Making* 26 (5):462–476.
- Fisher, Geoffrey. 2017. “An Attentional Drift-Diffusion Model Over Binary-Attribute Choice.” *Cognition* 168:34–45.

- Frydman, Cary and Lawrence J. Jin. 2018. "Efficient Coding and Risky Choice." Working Paper.
- Fudenberg, Drew, Philipp Strack, and Tomasz Strzalecki. 2018. "Speed, Accuracy, and the Optimal Timing of Choices." *American Economic Review* 108 (12):3651–3684.
- Fudenberg, Drew and Tomasz Strzalecki. 2015. "Dynamic Logit with Choice Aversion." *Econometrica* 83 (2):651–691.
- Goeree, Jacob K., Charles A. Holt, and Thomas R. Palfrey. 2005. "Regular Quantal Response Equilibrium." *Experimental Economics* 8:347–367.
- Greiner, Ben. 2015. "Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE." *Journal of the Economic Science Association* 1:114–125.
- Grether, David M. 1980. "Bayes Rule as a Descriptive Model: The Representativeness Heuristic." *Quarterly Journal of Economics* 95:537–557.
- . 1992. "Testing Bayes Rule and the Representativeness Heuristic: Some Experimental Evidence." *Journal of Economic Behavior and Organization* 17:31–57.
- Grether, David M. and Charles R. Plott. 1979. "Theory of Choice and the Preference Reversal Phenomenon." *American Economic Review* 69 (4):623–638.
- Gul, Faruk, Paulo Natenzon, and Wolfgang Pesendorfer. 2014. "Random Choice as Behavioral Optimization." *Econometrica* 82 (5):1873–1912.
- Harless, David W. and Colin F. Camerer. 1994. "The Predictive Utility of Generalized Expected Utility Theories." *Econometrica* 62 (6):1251–1289.
- Harrison, Glenn W. and Jia Min Ng. 2016. "Evaluating the Expected Welfare Gain from Insurance." *Journal of Risk and Insurance* 83 (1):91–120.
- Hey, John D. 2001. "Does Repetition Improve Consistency?" *Experimental Economics* 4 (1):5–54.
- Hey, John D. and Chris Orme. 1994. "Investigating Generalizations of Expected Utility Theory Using Experimental Data." *Econometrica* 62 (6):1291–1326.
- Hicks, John R. and Roy G.D. Allen. 1934. "A Reconsideration of the Theory of Value. Part I." *Economica* 1 (1):52–76.
- Huber, Joel, John W. Payne, and Christopher Puto. 1982. "Adding Symmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis." *Journal of Consumer Research* 9 (1):90–98.
- Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler. 1990. "Experimental Tests of the Endowment Effect and the Coase Theorem." *Journal of Political Economy* 98 (6):1325–1348.
- Khaw, Mel Win, Ziang Li, and Michael Woodford. 2018. "Cognitive Imprecision and Small-Stakes Risk Aversion." Working Paper, Columbia University.
- Klein, A. Stanley. 2001. "Measuring, Estimating, and Understanding the Psychometric Function: A Commentary." *Attention, Perception, & Psychophysics* 63 (8):1421–1455.

- Knetsch, Jack L. 1989. "The Endowment Effect and Evidence of Nonreversible Indifference Curves." *American Economic Review* 79 (5):1277–1284.
- Kononov, Arkady and Ian Krajbich. 2019. "Revealed Strength of Preference: Inference from Response Times." *Judgment and Decision Making* 14 (4):381–394.
- Krajbich, Ian, Carrie Armel, and Antonio Rangel. 2010. "Visual Fixations and the Computation and Comparison of Value in Simple Choice." *Nature Neuroscience* 13 (10):1292–1298.
- Laming, Donald. 1985. "Some Principles of Sensory Analysis." *Psychological Review* 92 (4):462–485.
- Lichtenstein, Sarah and Paul Slovic. 1971. "Reversals of Preference Between Bids and Choices in Gambling Decisions." *Journal of Experimental Psychology* 89 (1):46–55.
- Loomes, Graham and Robert Sugden. 1998. "Testing Different Stochastic Specifications of Risky Choice." *Economica* 65 (260):581–598.
- Lopez-Guzman, Silvia, Anna B. Konova, Kenway Louie, and Paul W. Glimcher. 2018. "Risk Preferences Impose a Hidden Distortion on Measures of Choice Impulsivity." *PloS One* 13 (1):e0191357.
- Luce, R. Duncan. 1959. *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.
- Machina, Mark J. 1985. "Stochastic Choice Functions Generated from Deterministic Preferences over Lotteries." *Economic Journal* 95 (379):575–594.
- Manzini, Paola and Marco Mariotti. 2014. "Stochastic Choice and Consideration Sets." *Econometrica* 82 (3):1153–1176.
- Marschak, Jacob. 1960. "Binary Choice Constraints on Random Utility Indicators." In *Stanford Symposium on Mathematical Methods in the Social Sciences*, edited by Kenneth J. Arrow. Stanford, CA: Stanford University Press, 312–329.
- McFadden, Daniel L. 1974. "Conditional Logit Analysis of Qualitative Choice Behavior." In *Frontiers in Econometrics*, edited by P. Zarembka. New York: Academic Press, 105–142.
- . 2001. "Economic Choices." *American Economic Review* 91 (3):351–378.
- McKelvey, Richard D. and Thomas R. Palfrey. 1995. "Quantal Response Equilibria for Normal Form Games." *Games and Economic Behavior* 10 (1):6–38.
- Moffatt, G. Peter. 2015. *Experimentetrics: Econometrics for Experimental Economics*. London: Palgrave Macmillan.
- Moffatt, Peter G. 2005. "Stochastic Choice and the Allocation of Cognitive Effort." *Experimental Economics* 8 (4):369–388.
- Mosteller, Frederick and Philip Nogee. 1951. "An Experimental Measurement of Utility." *Journal of Political Economy* 59:371–404.
- Moyer, Robert S. and Thomas K. Landauer. 1967. "Time Required for Judgements of Numerical Inequality." *Nature* 215 (5109):1519–1520.

- Murdock Jr., Bennet B. 1962. “The Serial Position Effect of Free Recall.” *Journal of Experimental Psychology* 64 (5):482–488.
- Natenzon, Paulo. 2019. “Random Choice and Learning.” *Journal of Political Economy* 127 (1):419–457.
- Peirce, Jonathan W. 2007. “PsychoPy – Psychophysics Software in Python.” *Journal of Neuroscience Methods* 162 (1):8–13.
- Polania, Rafael, Michael Woodford, and Christian C. Ruff. 2019. “Efficient Coding of Subjective Value.” *Nature Neuroscience* 22 (1):134–142.
- Ratcliff, Roger. 1978. “A Theory of Memory Retrieval.” *Psychological Review* 85:59–108.
- Rubinstein, Ariel. 2007. “Instinctive and Cognitive Reasoning: A Study of Response Times.” *Economic Journal* 117 (523):1243–1259.
- . 2013. “Response Time and Decision Making: An Experimental Study.” *Judgment and Decision Making* 8 (5):540–551.
- Shadlen, Michael N. and Roozbeh Kiani. 2013. “Decision Making as a Window on Cognition.” *Neuron* 80:791–806.
- Shadlen, Michael N. and Daphna Shohamy. 2016. “Decision Making and Sequential Sampling from Memory.” *Neuron* 90:927–939.
- Soltani, Alireza, Benedetto De Martino, and Colin Camerer. 2012. “A Range-Normalization Model of Context-Dependent Choice: A New Model and Evidence.” *PLoS Computational Biology* 8 (7):1–15.
- Spiliopoulos, Leonidas and Andreas Ortmann. 2018. “The BCD of Response Time Analysis in Experimental Economics.” *Experimental Economics* 21 (2):383–433.
- Spitzer, Bernhard, Leonhard Waschke, and Christopher Summerfield. 2017. “Selective Overweighting of Larger Magnitudes During Noisy Numerical Comparison.” *Nature Human Behaviour* 1 (0145):1–8.
- Thurstone, Louis L. 1927. “A Law of Comparative Judgement.” *Psychological Review* 34:273–286.
- Tversky, Amos. 1969. “Intransitivity of Preferences.” *Psychological Review* 76:31–48.
- Tversky, Amos, Paul Slovic, and Daniel Kahneman. 1990. “The Causes of Preference Reversal.” *American Economic Review* 80 (1):204–217.
- Vieider, Ferdinand M. 2018. “Violence and Risk Preference: Experimental Evidence from Afghanistan, Comment.” *American Economic Review* 108 (8):2366–2382.
- Von Gaudecker, Hans-Martin, Arthur Van Soest, and Erik Wengström. 2011. “Heterogeneity in Risky Choice Behavior in a Broad Population.” *American Economic Review* 101 (2):664–694.
- Vulkan, Nir. 2000. “An Economist’s Perspective on Probability Matching.” *Journal of Economic Surveys* 14 (1):101–118.

- Webb, Ryan. 2019. “The (Neural) Dynamics of Stochastic Choice.” *Management Science* 64 (1):230–255.
- Weber, Elke U., Ann-Renée Blais, and Nancy E. Betz. 2002. “A Domain-Specific Risk-Attitude Scale: Measuring Risk Perceptions and Risk Behaviors.” *Journal of Behavioral Decision Making* 15 (4):263–290.
- Wichmann, A. Felix and N. Jeremy Hill. 2001. “The Psychometric Function: I. Fitting, Sampling, and Goodness of Fit.” *Attention, Perception, & Psychophysics* 63 (8):1293–1313.
- Wilcox, Nathaniel T. 1993. “Lottery Choice: Incentives, Complexity, and Decision Time.” *Economic Journal* 103 (421):1397–1417.
- . 1994. “On a Lottery Pricing Anomaly: Time Tells the Tale.” *Journal of Risk and Uncertainty* 8 (7):311–324.
- . 2008. “Stochastic Models for Binary Discrete Choice Under Risk: A Critical Primer and Econometric Comparison.” In *Risk Aversion in Experiments, Research in Experimental Economics*, vol. 12, edited by James C. Cox and Glenn W. Harrison. Bingley, UK: Emerald, 197–292.
- . 2011. “Stochastically More Risk Averse: A Contextual Theory of Stochastic Discrete Choice Under Risk.” *Journal of Econometrics* 162 (1):89–104.

# Strength of Preference and Decision Making Under Risk

Carlos Alós-Ferrer\* and Michele Garagnani

Department of Economics, University of Zurich,  
Blümlisalpstrasse 10, 8006 Zurich, Switzerland.

This version: July 2020

## APPENDICES – For Online Publication Only

### A Description of RUM Estimation

We now describe the details of the estimation procedure used in the main text, which follows the approach described in Moffatt (2015, Chapter 13). All trials  $t = 1, \dots, T$  used for the utility estimation involved binary choices between “bet” and “pass.” The option to bet at  $t$  corresponds to a lottery  $A_t = (p_t, x_t - y; 1 - p_t, -y)$  paying an amount  $x_t - y$  with probability  $p_t$ , and an amount  $-y$  with probability  $1 - p_t$ . The option to pass corresponds to the degenerate lottery  $B$  paying an amount 0 for sure. The amount  $y$  is the cost of betting, and is fixed for all subjects  $i = 1, \dots, N$  and trials. We use odd trials to estimate the utility of options in even-numbered trials out-of-sample, and vice versa. In the main text we assume a normalized constant absolute risk aversion (CARA) function as in Conte, Hey, and Moffatt (2011),

$$(1) \quad u(x \mid r) = \begin{cases} \frac{1 - e^{-rx}}{1 - e^{-rx_{\max}}}, & \text{if } r \neq 0 \\ \frac{x}{x_{\max}}, & \text{if } r = 0, \end{cases}$$

where  $x_{\max} = \max\{x_1, \dots, x_T\} - y$  is the maximum outcome across all  $T$  trials. However, the results are qualitatively unchanged when we assume a utility function with constant relative risk aversion (CRRA) instead (see Appendix C). Under the assumption of Expected Utility maximization, subject  $i$ , with utility function  $u(x \mid r_i)$ , chooses  $A_t$  over  $B$  if the expected utility of  $A$  is positive (since  $EU_i(B) = 0$ ), that is,

$$(2) \quad \nabla_t(r_i) := \frac{p_t(1 - e^{-r_i(x_t - y)}) + (1 - p_t)(1 - e^{r_i y})}{1 - e^{-r_i x_{\max}}} > 0$$

---

\*Corresponding Author: Email: carlos.alos-ferrer@econ.uzh.ch, phone: +41 44 634 10 16

(for  $r_i \neq 0$ ). To estimate individual utilities, we follow a standard approach based on a Random Utility Model (RUM). We add an error term  $\varepsilon_{it} \sim N(0, \sigma^2)$  with  $\sigma^2 > 0$  to (2). That is,  $A_t$  is chosen if and only if

$$(3) \quad \nabla_t(r_i) + \varepsilon_{it} > 0.$$

Define the binary choice indicator for trial  $t$  by

$$\gamma_{it} = \begin{cases} 1 & \text{if } A_t \text{ chosen by subject } i \\ -1 & \text{if } B \text{ chosen by subject } i. \end{cases}$$

The probability that subject  $i$  chooses  $A_t$  in trial  $t$  is

$$(4) \quad P(\gamma_{it} = 1) = P(\nabla_t(r_i) > -\varepsilon_{it}) = \Phi\left(\frac{\nabla_t(r_i)}{\sigma}\right)$$

where  $\Phi$  is the standard normal cumulative distribution function. Conversely, the probability of  $B$  is

$$(5) \quad P(\gamma_{it} = -1) = P(\varepsilon_{it} < -\nabla_t(r_i)) = \Phi\left(-\frac{\nabla_t(r_i)}{\sigma}\right).$$

The conditional probabilities described in Equations 4 and 5 above were derived conditional on a subject's risk parameter  $r_i$ . To account for individual heterogeneity, we assume that the risk parameter is distributed over the population and we estimate the parameters of this distribution (e.g., see Harless and Camerer, 1994; Moffatt, 2005; Harrison and Rutstrom, 2008; Bellemare, Kröger, and van Soest, 2008; Von Gaudecker, Van Soest, and Wengström, 2011; Conte, Hey, and Moffatt, 2011; Moffatt, 2015). This approach is standard by now, and is generally preferred to estimating risk parameters separately for each individual (e.g., see Hey and Orme, 1994). The main reason is that the former greatly reduces the degrees of freedom compared to individual-level estimates, avoiding possible overfitting problems (see Conte, Hey, and Moffatt, 2011, for a more detailed discussion). Further, the population estimation allows to obtain estimates even for extremely risk-averse or risk-seeking subjects, which is sometimes impossible at the individual level due to identification issues (Moffatt, 2015). Last, population-level estimates are generally preferable when predicting the behavior of the same or a similar group of individuals (Conte, Hey, and Moffatt, 2011).

Assume that the individual risk attitudes are distributed normally in the population (from which our subjects were drawn) according to

$$r \sim N(\mu, \eta^2).$$

Hence, the log-likelihood of a sample given by the matrix  $\Gamma = (\gamma_{it})$  consisting of  $T$  trials and  $N$  subjects is

$$(6) \quad \log L = \sum_{i=1}^N \ln \int_{-\infty}^{\infty} \prod_{t=1}^T \Phi \left( \gamma_{it} \frac{\nabla_t(r)}{\sigma} \right) f(r \mid \mu, \eta) dr$$

where  $f(r \mid \mu, \eta) = \frac{1}{\sqrt{2\pi\eta^2}} e^{-\frac{1}{2} \left( \frac{r-\mu}{\eta} \right)^2}$  is the density function of the risk parameter  $r$ .

The log likelihood can be maximized by standard methods to obtain the maximum likelihood estimates. Since the integral in the likelihood function does not have a closed-form solution, we approximate it using standard simulation techniques (see Train, 2003, for details). Specifically, we approximate this integral by the average

$$(7) \quad \frac{1}{H} \sum_{h=1}^H \left( \prod_{t=1}^T \Phi \left( \gamma_{it} \frac{\nabla_t(r_{ih}(\mu, \eta))}{\sigma} \right) \right)$$

using a sequence of  $H$  (transformed) Halton draws  $(r_{i1}, \dots, r_{iH})$  from  $N(\mu, \eta^2)$  for each subject  $i$  (fixed across trials  $t$ ). Halton draws, a by-now-standard procedure, are pseudo-random draws that ensure even coverage of the parameter space (e.g. avoiding clustering) using Halton sequences (Halton, 1960; Moffatt, 2015). For the estimation, we use the Stata implementation “mdraws” of this procedure (Cappellari and Jenkins, 2003) for a uniform distribution  $U(0, 1)$  and transform the resulting sequence  $(g_1, g_2, \dots, g_H)$  to obtain draws  $r_{ih}(\mu, \eta) = \mu + \eta \Phi^{-1}(g_h)$  from  $N(\mu, \eta^2)$ .

The maximum simulated likelihood (MSL) approach amounts to replacing the integral in (6) by (7) and then maximize the resulting function  $\log \hat{L}$ . Maximization of  $\log \hat{L}$  is carried out using standard MLE routines to obtain the estimates  $(\hat{\mu}, \hat{\eta}, \hat{\sigma})$ . Given those estimates we obtain the posterior expectation of each subject’s risk attitude  $\hat{r}_i$  conditional on their  $T$  choices approximating the conditional expectation by

$$(8) \quad \hat{r}_i = E(r_i \mid \gamma_{i1}, \dots, \gamma_{iT}) \approx \frac{\frac{1}{H} \sum_{h=1}^H r_{ih} \left( \prod_{t=1}^T \Phi \left( \gamma_{it} \frac{\nabla_t(r_{ih})}{\hat{\sigma}} \right) \right)}{\frac{1}{H} \sum_{h=1}^H \left( \prod_{t=1}^T \Phi \left( \gamma_{it} \frac{\nabla_t(r_{ih})}{\hat{\sigma}} \right) \right)}$$

where  $r_{ih} = r_{ih}(\hat{\mu}, \hat{\eta})$ . Given the estimated individual risk parameter  $\hat{r}_i$ , we obtain

$$\hat{u}_i(x) = \frac{1 - e^{-\hat{r}_i x}}{1 - e^{-\hat{r}_i x_{\max}}} \text{ for } \hat{r}_i \neq 0$$

as the estimated utility function of subject  $i$ .

## B Omitted Tables and Figures

This section contains the regressions with additional controls from Tables 1–6 in the main text (Model 4 in each table). In the main text, these regressions are just used to show that the main effects are robust to the inclusion of additional controls. Here

we present the actual regression coefficients and briefly discuss some interpretations and complementary tests.

The two regressions in Table 1 correspond to Model 4 in Tables 1 and 2, respectively, and refer to Experiment 1. The first is a random-effects probit model on correct answers. Females (54) make more errors than males, as can be confirmed by a direct, non-parametric test (females 33.08%, males 23.19%; Mann-Whitney-Wilcoxon test,  $N = 93$ ,  $z = -2.764$ ,  $p = 0.0068$ ). Native speakers (72) make less errors than other participants (natives 25.89%, others 39.35%; MWW test,  $N = 93$ ,  $z = 3.116$ ,  $p = 0.0018$ ). Also, participants who earned more in previous trials (Sum Won) are more likely to make a correct choice, which is merely an indication of heterogeneous skills among participants. The second regression is a random-effects model on log-transformed response times. To control for individual differences in mechanical swiftness, we use the log-transformed response time for the non-rewarded, first black card (RT1) and the log-transformed response time for pressing a space bar, which was required before the start of each trial (RT0). Gender and accumulated earnings did not affect response times, but native speakers took longer to respond than other participants.

The two regressions in Table 2 correspond to Model 4 in Tables 3 and 5, respectively, and refer to Experiment 2 using expected utility differences as a regressor (EU distance). Again, the first is a random-effects probit model on correct answers. There are no gender differences in errors defined according to EU distance, as confirmed by a non-parametric test (females 22.41%, males 23.46%; MWW test,  $N = 93$ ,  $z = 1.558$ ,  $p = 0.1201$ ). Likewise, native speakers (77) did not perform significantly differently from other participants (natives 22.29%, others 19.41%; MWW test,  $N = 93$ ,  $z = 1.298$ ,  $p = 0.1971$ ). The second regression is a random-effects model on log-transformed response times. As in Experiment 1, to control for mechanical swiftness, we include the log-transformed response time for pressing a space bar before the start of each trial (RT0). Other controls deliver no additional insights, except that left-handed subjects took longer than right-handed ones.

The two regressions in Table 3 correspond to Model 4 in Tables 4 and 6, respectively, and refer to Experiment 2 using expected value differences as a regressor (EV distance). Again, the first is a random-effects probit model on correct answers. When defining errors according to expected values, females did behave differently (males 37.22%, females 33.91%; MWW test,  $N = 93$ ,  $z = 2.399$ ,  $p = 0.0164$ ), as did native speakers (natives 35.68%, others 39.26%; MWW test,  $N = 93$ ,  $z = -1.874$ ,  $p = 0.0609$ ). The second regression is a random-effects model on log-transformed response times. Again, left-handed subjects took longer than right-handed ones.

Table 1: Experiment 1. Random-effects Probit regressions on correct answers and random-effects regressions on log response times.

	Correct	Log RT
Econ. Dist.	1.620*** (0.141)	-0.316*** (0.035)
Num. Dist.	0.582*** (0.093)	-0.209*** (0.039)
Econ. Dist. $\times$ Num. Dist.	-0.010*** (0.004)	0.005*** (0.001)
Sum Won	0.011* (0.006)	0.008 (0.006)
Female	-0.438*** (0.141)	-0.056 (0.078)
Native	0.462*** (0.169)	0.281*** (0.119)
Age	-0.045*** (0.010)	-0.014 (0.009)
Left handed	0.378** (0.194)	-0.014 (0.191)
Round	-0.007** (0.003)	-0.005*** (0.003)
RT0		0.117*** (0.037)
RT1		0.145*** (0.032)
Constant	0.886*** (0.284)	0.853*** (0.253)
Wald test	251.391***	259.572***
Obs.	14880	14880

Robust standard errors in brackets, \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Table 2: Experiment 2. Random-effects Probit regressions on correct answers and random-effects regressions on log response times both for EU distance. Correct answer is defined as passing when  $EU \leq 0$  and betting when  $EU \geq 0$ .

	Correct	Log RT
EU_Dist.	16.277*** (0.639)	-1.154*** (0.206)
Num._Dist.	0.003 (0.044)	-0.136*** (0.019)
Num._Dist. $\times$ EU_Dist.	-5.332*** (0.907)	1.824*** (0.285)
Sum Won	-0.000 (0.000)	0.000 (0.000)
Female	-0.077 (0.050)	0.011 (0.082)
Native	0.069 (0.066)	0.088 (0.116)
Age	-0.016*** (0.005)	-0.009 (0.011)
Left handed	0.128 (0.105)	0.227*** (0.085)
Round	0.003*** (0.000)	-0.003*** (0.000)
RT0		0.200*** (0.018)
Constant	0.536** (0.257)	1.268*** (0.040)
Wald test	1041.316***	579.216 ***
Obs.	15810	15810

Robust standard errors in brackets, \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Table 3: Experiment 2. Random-effects Probit regressions on correct answers and random-effects log response times both for EV distance. Correct answer is defined as passing when  $EV \leq 0$  and betting when  $EV \geq 0$ .

	Correct	Log RT
EV_Dist.	1.113*** (0.070)	0.009 (0.023)
Num._Dist.	0.639*** (0.062)	-0.211*** (0.024)
Num._Dist. $\times$ EV_Dist.	-1.600*** (0.124)	0.145 (0.043)
Sum Won	-0.000 (0.000)	0.000 (0.000)
Female	-0.086** (0.041)	0.011 (0.083)
Native	0.096* (0.054)	0.073 (0.116)
Age	-0.003 (0.005)	-0.009 (0.010)
Left handed	-0.001 (0.070)	0.221** (0.091)
Round	0.002*** (0.000)	-0.003*** (0.000)
RT0		0.196*** (0.018)
Constant	-0.159 (0.258)	1.290*** (0.310)
Wald test	466.95***	687.710***
Obs.	15810	15810

Robust standard errors in brackets, \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

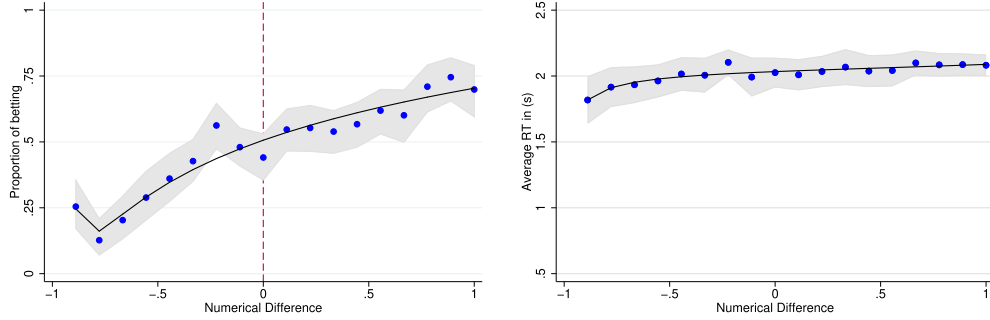


Figure 1: Experiment 2. Proportion of betting decisions (left) and average response times (right) as functions of numerical differences. Gray areas indicate 95% confidence intervals.

Figure 1 displays the analogous plots to the ones in Figures 5 and 6 when considering payoff-irrelevant, numerical differences in the stimuli. The left panel shows the proportion of betting decisions. As commented in the main text, although there appears to be a trend relating payoff-irrelevant numerical differences and error rates, the relation is weaker than the one observed for expected utility differences (recall the discussion for Experiment 1). The right panel shows the average response times, illustrating that there is no relation between those and numerical differences.

## C Robustness Analysis: CRRA

In this section, we conduct a further robustness check. Specifically, we repeat the RUM-based estimation exercise described in Appendix A using the following constant relative risk aversion (CRRA) utility function

$$(9) \quad u(x \mid r) = \begin{cases} \frac{x^{1-r}}{1-r}, & \text{if } r \neq 1 \\ \ln(x), & \text{if } r = 1 \end{cases}$$

instead of the CARA function used for the estimation reported in the main text. We then reproduce the analysis of Experiment 2 reported in the main text with this alternative estimation. As we detail below, this robustness check confirms that the results reported in the main text do not hinge on the CARA specification of the utility function, but remain robust when the CRRA specification above is used instead.

We re-define errors using the estimated CRRA function. The mean error rate across participants was 27.41%, with a median of 27.65% (SD = 8.40, max 57.64%, min 3.53%). Figure 2 (left panel) plots the frequency of betting decisions as a function of expected utility differences. The shaded areas correspond to errors with the definition above. We confirm that the relation between betting frequency and expected utility differences has a sigmoidal shape resembling a cumulative normal distribution or a logistic curve.

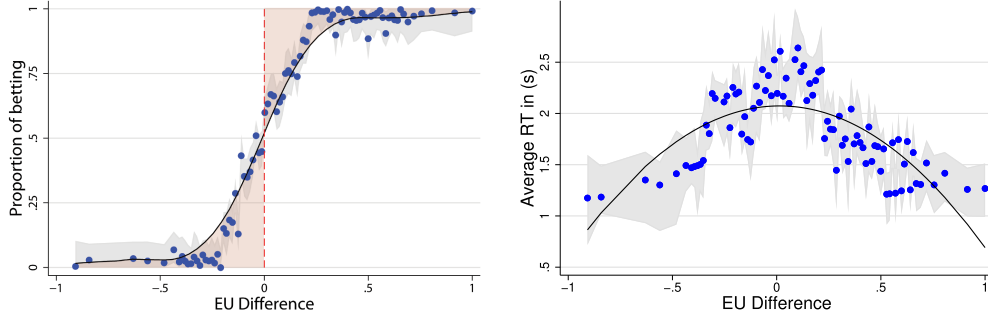


Figure 2: Experiment 2. Proportion of betting decisions as a function of expected utility differences (left panel). Average response times on EU Difference (right panel).

We ran random-effects panel Probit regressions where the dependent variable is 1 in case of a correct answer. Table 4 is the equivalent of Table 3 when CRRA is used. In Model 1 of both tables we see that larger economic distances lead to less errors, confirming the main results. We further replicate that EU distance has a coefficient larger than EV distance (almost 28 times). To conduct a proper comparison, we recalculated the relative elasticities. A percentage variation in EU distance increases the probability of a correct answer by an average of 18.82%, while the analogous percentage for EV distance increases is only 11.98%.

Model 2 in both tables introduces numerical distance as an additional control. In the presence of EU distance, numerical similarity does not seem to play a role. The effect is also insignificant when controlling for the interaction between numerical distance and EU distance (Model 3), and when adding further controls (Model 4).

In all models we further control for learning effects. Participants appear to improve during the course of the experiment. As in the CARA case, native speakers (77) did not perform significantly differently from other participants (natives 26.62%, others 31.21%; MWW test,  $N = 93$ ,  $z = -0.662$ ,  $p = 0.5133$ ). When using CRRA, there are marginally significant gender differences in errors defined according to EU distance, as confirmed by a non-parametric test (females 28.47%, males 24.68%; MWW test,  $N = 93$ ,  $z = 2.146$ ,  $p = 0.0321$ ). This is a different result compared to Appendix B, where using CARA we did not find significant gender differences. However, this difference across estimations might be due to measurement errors and should not be overstated (Gillen, Snowberg, and Yariv, 2019).

Figure 2 (right panel) plots average response times as a function of expected utility differences. Response times and EU distance clearly show an inverted U-shaped relation. Table 5 reports the corresponding regressions, analogously to Table 5. We confirm that response times are significantly shorter for larger EU distances across all models. This fundamental effect is robust to controlling for numerical distance, accumulated earnings, gender, native language, and other controls. Additionally, we replicated that numerical

Table 4: Experiment 2. Random-effects Probit regressions on correct answers for EU distance. Correct answer is defined as passing when  $EU \leq 0$  and betting when  $EU \geq 0$ .

Correct	Model 1	Model 2	Model 3	Model 4
EU_Dist.	12.941*** (0.724)	12.863*** (0.750)	13.606*** (0.843)	13.583*** (0.844)
Num._Dist.		-0.021 (0.052)	-0.023 (0.052)	-0.027 (0.053)
Num._Dist. $\times$ EU_Dist.			-1.780** (0.848)	-1.480* (0.835)
Sum Won				-0.001*** (0.000)
Female				-0.112*** (0.055)
Native				0.110 (0.068)
Age				-0.003 (0.006)
Left handed				-0.065 (0.078)
Round	0.003*** (0.000)	0.003*** (0.000)	0.003*** (0.000)	0.003*** (0.000)
Constant	-0.229*** (0.039)	-0.225*** (0.036)	-0.226*** (0.037)	0.359 (0.279)
Log L.	-8510	-8509	-8507	-8445
Wald test	411.431***	415.257***	415.727***	422.769***
Obs.	15810	15810	15810	15810

Robust standard errors in brackets, \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

distance does have an effect on response times, validating the view from psychophysics (Moyer and Landauer, 1967; Dehaene, Dupoux, and Mehler, 1990) that even payoff-irrelevant perceptual differences might influence actual choice difficulty.

Table 5: Experiment 2. Random-effects regressions on log response times, EU distance.

Log RT	Model 1	Model 2	Model 3	Model 4
EU_Dist.	−0.893*** (0.258)	−1.063*** (0.243)	−0.895*** (0.239)	−0.903*** (0.240)
Num._Dist.		−0.045*** (0.012)	−0.142*** (0.022)	−0.142*** (0.022)
Num._Dist. × EU_Dist.			1.957*** (0.383)	1.956*** (0.382)
Sum Won				0.000 (0.000)
Female				0.018 (0.081)
Native				0.074 (0.116)
Age				−0.010 (0.011)
Left handed				0.230*** (0.089)
Round	−0.002*** (0.000)	−0.002*** (0.000)	−0.002*** (0.000)	−0.003*** (0.000)
RT0	0.198*** (0.018)	0.199*** (0.018)	0.200*** (0.018)	0.200*** (0.018)
Constant	1.303*** (0.041)	1.310*** (0.041)	1.316*** (0.040)	1.296*** (0.304)
$R^2$ overall	0.131	0.131	0.137	0.156
Wald test	458.704***	458.844 ***	618.502***	642.214***
Obs.	15810	15810	15810	15810

Robust standard errors in brackets, \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

## D Robustness Analysis: RPM Estimation

Random utility models have been recently criticized (see Wilcox, 2008, 2011; Bruner, 2017; Apesteguía and Ballester, 2018; Vieider, 2018) and Random Preference Models (RPM) have been suggested as an alternative (e.g., Loomes and Sugden, 1995), since they are immune to some of the critiques. The RPM approach is based on the idea that a subject's risk parameter is not fixed but varies randomly. The drawback is that this also renders it arguably less appropriate in our context since our goal is to obtain a (fixed) measure of individual risk attitudes. Nevertheless, we also consider an RPM as an additional robustness check to further ensure that the results obtained in the main text do not hinge on the specifics of the estimation procedure.

### D.1 Description of RPM procedure

For the RPM estimation, we use the same setup as for RUM, and the CARA utility function given by Equation (1). In contrast to the RUM approach, the RPM assumes that a subject's risk parameter is not fixed across trials but varies randomly between trials. Specifically, we assume that subject  $i$ 's risk parameter in trial  $t$  is distributed according to  $r_{it} \sim N(m_i, \sigma^2)$  where  $m_i$  is subject  $i$ 's mean risk attitude. Assuming Expected Utility maximization, in this setup subject  $i$  with utility function  $u_i$  chooses  $A_t$  over  $B$  if and only if  $\Delta_t(r_{it}) > 0$ , where  $\Delta_t(r_{it})$  is as given in equation (2).

Let  $r_t^*$  be the risk parameter that would make a subject exactly indifferent between betting and passing in task  $t$ , that is,  $\Delta_t(r_t^*) = 0$ . Then,

$$\Delta_t(r_{it}) > 0 \quad \Leftrightarrow \quad r_{it} > r_t^*.$$

Again using  $\gamma_{it} \in \{1, -1\}$  as a binary indicator that  $A_t$  is chosen by subject  $i$  in trial  $t$ , the probability of a choice conditional on a subject's mean risk attitude  $m_i$  is given by

$$P(\gamma_{it}|m_i) = P(\gamma_{it}(r_{it} - r_t^*) > 0|m_i) = \Phi\left(\gamma_{it}\frac{m_i - r_t^*}{\sigma}\right)$$

where  $\Phi$  is the standard normal cumulative distribution function. As in Appendix A, in order to introduce between-subject heterogeneity we let the individual mean risk attitude vary at the population level. In particular, we assume that

$$m_i \sim N(\mu, \eta^2).$$

Hence, the log-likelihood for a sample consisting of  $T$  trials and  $N$  subjects given by the matrix  $\Gamma = (\gamma_{it})$  is

$$(10) \quad \log L = \sum_{i=1}^N \ln \int_{-\infty}^{\infty} \prod_{t=1}^T \Phi\left(\gamma_{it}\frac{m - r_t^*}{\sigma}\right) f(m | \mu, \eta) dm$$

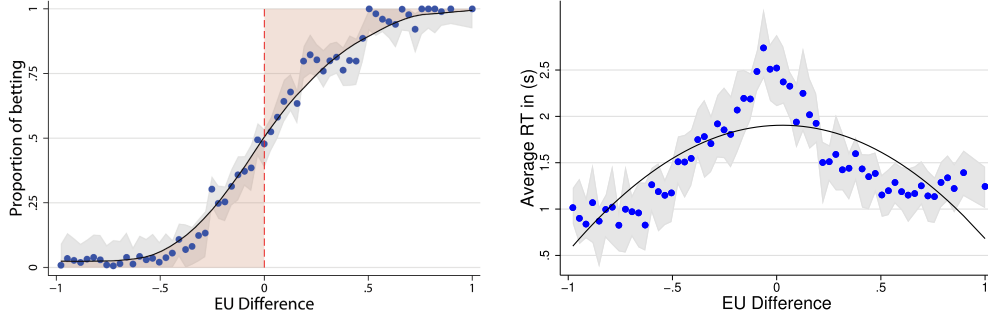


Figure 3: Experiment 2. Proportion of betting decisions as a function of expected utility differences (left panel). Average response times on EU Difference (right panel).

where  $f(m \mid \mu, \eta)$  is the density function of the mean risk attitude  $m$ . This function is then approximated following the MSL approach as in Appendix A to obtain the parameter estimates  $(\hat{\mu}, \hat{\eta}, \hat{\sigma})$ . Given those estimates, we compute the posterior expectation of a subject's mean risk attitude  $\hat{m}_i$  conditional on the observed  $T$  choices analogously to equation (8).

Given the estimated individual mean risk parameter  $\hat{m}_i$ , we obtain

$$\hat{u}_i(x) = \frac{1 - e^{-\hat{m}_i x}}{1 - e^{-\hat{m}_i x_{\max}}} \text{ for } \hat{m}_i \neq 0$$

as the estimated utility function of subject  $i$ .

## D.2 Results based on RPM estimation

We re-define errors according to the utilities estimated above. The mean error rate across participants was 29.57%, with a median of 30.59% (SD = 9.08, max 54.71%, min 7.06%). Figure 3 (left panel) plots the frequency of betting decisions as a function of expected utility differences. The shaded areas correspond to errors with the definition above. We confirm that the relation between betting frequency and expected utility differences retains a sigmoidal shape as in the main text (using a RUM with CARA functions) or in Appendix C (using a RUM with CRRA functions).

We ran random-effects panel Probit regressions where the dependent variable is 1 in case of a correct answer. Table 6 is the equivalent of Table 3 when an RPM is used. In Model 1 we see that larger economic distances lead to less errors, confirming the basic prediction. We reproduce the observation that EU distance has a larger coefficient than EV distance (almost 16 times). To conduct a proper comparison, we calculated the relative elasticities. A percentage variation in EU distance increases the probability of a correct answer by an average of 16.15%, while the analogous percentage for EV distance increases is only 11.98%.

Table 6: Experiment 2. Random-effects Probit regressions on correct answers for EU distance. Correct answer is defined as passing when  $EU \leq 0$  and betting when  $EU \geq 0$ .

Correct	Model 1	Model 2	Model 3	Model 4
EU_Dist.	7.567*** (0.418)	7.533*** (0.438)	8.712*** (0.473)	8.683*** (0.474)
Num._Dist.		-0.010 (0.044)	-0.024 (0.043)	-0.026 (0.043)
Num._Dist. $\times$ EU_Dist.			-2.735*** (0.558)	-2.749*** (0.554)
Sum Won				-0.001*** (0.000)
Female				-0.021 (0.035)
Native				0.088** (0.043)
Age				-0.012** (0.005)
Left handed				0.073 (0.067)
Round	0.003*** (0.000)	0.003*** (0.000)	0.003*** (0.000)	0.003*** (0.000)
Constant	-0.163*** (0.040)	-0.161*** (0.038)	-0.163*** (0.038)	0.765*** (0.237)
Log L.	-8931	-8931	-8921	-8859
Wald test	347.236***	347.995***	402.792***	488.139 ***
Obs.	15810	15810	15810	15810

Robust standard errors in brackets, \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Model 2 in both tables introduces numerical distance as an additional control. In the presence of EU distance, numerical similarity does not seem to play a role. The effect is also insignificant when controlling for the interaction between numerical distance and EU distance (Model 3), and when adding further controls (Model 4).

In all models we further control for learning effects. Participants appear to improve during the course of the experiment. As in Appendix B, there are no gender differences in errors defined according to EU distance, as confirmed by a non-parametric test (females 30.29%, males 27.69%; MWW test,  $N = 93$ ,  $z = 1.469$ ,  $p = 0.1432$ ). However, under this specification native speakers (77) performed marginally significantly better than other participants (natives 28.77%, others 33.42%; MWW test,  $N = 93$ ,  $z = -2.312$ ,  $p = 0.0199$ ).

Figure 3 (right panel) plots average response times as a function of expected utility differences. Response times and EU distance clearly show an inverted U-shaped relation. Table 7 reports the analogous regressions to Table 5. We confirm that response times are significantly shorter for larger EU distances across all models in Table 7. This

Table 7: Experiment 2. Random-effects regressions on log response times, EU distance.

Log RT	Model 1	Model 2	Model 3	Model 4
EU_Dist.	−0.585*** (0.178)	−0.532*** (0.180)	−0.635*** (0.194)	−0.635*** (0.193)
Num._Dist.		−0.128*** (0.018)	−0.144*** (0.019)	−0.144*** (0.019)
Num._Dist. × EU_Dist.			0.245 (0.185)	0.240 (0.185)
Sum Won				0.000 (0.000)
Female				0.010 (0.081)
Native				0.078 (0.116)
Age				−0.009 (0.011)
Left handed				0.223*** (0.089)
Round	−0.002*** (0.000)	−0.003*** (0.000)	−0.003*** (0.000)	−0.003*** (0.000)
RT0	0.198*** (0.018)	0.200*** (0.018)	0.200*** (0.018)	0.200*** (0.018)
Constant	1.296*** (0.041)	1.347*** (0.043)	1.354*** (0.043)	1.309*** (0.308)
$R^2$ overall	0.134	0.137	0.137	0.155
Wald test	439.833***	575.801 ***	572.307 ***	589.463***
Obs.	15810	15810	15810	15810

Robust standard errors in brackets, \*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

fundamental effect is robust to controlling for numerical distance, accumulated earnings, gender, native language, and other controls. Additionally, we again replicate the observation that numerical distance does have an effect on response times, validating the view from psychophysics (Moyer and Landauer, 1967; Dehaene, Dupoux, and Mehler, 1990) that even payoff-irrelevant perceptual differences might influence actual choice difficulty.

## References

- Apesteeguía, José and Miguel A. Ballester. 2018. “Monotone Stochastic Choice Models: The Case of Risk and Time Preferences.” *Journal of Political Economy* 126 (1):74–106.
- Bellemare, Charles, Sabine Kröger, and Arthur van Soest. 2008. “Measuring Inequity Aversion in a Heterogeneous Population Using Experimental Decisions and Subjective Probabilities.” *Econometrica* 76 (4):815–839.
- Bruner, David M. 2017. “Does Decision Error Decrease with Risk Aversion?” *Experimental Economics* 20 (1):259–273.
- Cappellari, Lorenzo and Stephen P. Jenkins. 2003. “Multivariate Probit Regression Using Simulated Maximum Likelihood.” *The Stata Journal* 3 (3):278–294.
- Conte, Anna, John D. Hey, and Peter G. Moffatt. 2011. “Mixture Models of Choice Under Risk.” *Journal of Econometrics* 162 (1):79–88.
- Dehaene, Stanislas, Emmanuel Dupoux, and Jacques Mehler. 1990. “Is Numerical Comparison Digital? Analogical and Symbolic Effects in Two-Digit Number Comparison.” *Journal of Experimental Psychology: Human Perception and Performance* 16 (3):626–641.
- Gillen, Ben, Erik Snowberg, and Leeat Yariv. 2019. “Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study.” *Journal of Political Economy* 127 (4):1826–1863.
- Halton, John H. 1960. “On the Efficiency of Certain Quasi-Random Sequences of Points in Evaluating Multi-Dimensional Integrals.” *Numerische Mathematik* 2 (1):84–90.
- Harless, David W. and Colin F. Camerer. 1994. “The Predictive Utility of Generalized Expected Utility Theories.” *Econometrica* 62 (6):1251–1289.
- Harrison, Glenn and Elisabet Rutstrom. 2008. “Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods.” chap. 81. Elsevier, 1 ed., 752–767.
- Hey, John D. and Chris Orme. 1994. “Investigating Generalizations of Expected Utility Theory Using Experimental Data.” *Econometrica* 62 (6):1291–1326.
- Loomes, Graham and Robert Sugden. 1995. “Incorporating a Stochastic Element into Decision Theories.” *European Economic Review* 39 (3–4):641–648.
- Moffatt, G. Peter. 2015. *Experiments: Econometrics for Experimental Economics*. London: Palgrave Macmillan.
- Moffatt, Peter G. 2005. “Stochastic Choice and the Allocation of Cognitive Effort.” *Experimental Economics* 8 (4):369–388.
- Moyer, Robert S. and Thomas K. Landauer. 1967. “Time Required for Judgements of Numerical Inequality.” *Nature* 215 (5109):1519–1520.
- Train, Kenneth E. 2003. *Discrete Choice Methods with Simulation*. New York: Cambridge University Press.
- Vieider, Ferdinand M. 2018. “Violence and Risk Preference: Experimental Evidence from Afghanistan, Comment.” *American Economic Review* 108 (8):2366–2382.

- Von Gaudecker, Hans-Martin, Arthur Van Soest, and Erik Wengström. 2011. "Heterogeneity in Risky Choice Behavior in a Broad Population." *American Economic Review* 101 (2):664–694.
- Wilcox, Nathaniel T. 2008. "Stochastic Models for Binary Discrete Choice Under Risk: A Critical Primer and Econometric Comparison." In *Risk Aversion in Experiments, Research in Experimental Economics*, vol. 12, edited by James C. Cox and Glenn W. Harrison. Bingley, UK: Emerald, 197–292.
- . 2011. "Stochastically More Risk Averse: A Contextual Theory of Stochastic Discrete Choice Under Risk." *Journal of Econometrics* 162 (1):89–104.