NBER WORKING PAPER SERIES

COGNITIVE IMPRECISION AND SMALL-STAKES RISK AVERSION

Mel Win Khaw
Ziang Li
Michael Woodford

Working Paper 24978
http://www.nber.org/papers/w24978

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2018

An earlier version of this work, under the title "Cognitive Limitations and the Perception of Risk," was presented as the 2015 AFA Lecture at the annual meeting of the American Finance Association. We thank Colin Camerer, Tom Cunningham, Daniel Friedman, Xavier Gabaix, Frank Heinemann, Arkady Konovalov, Ifat Levy, Rosemarie Nagel, Charlie Plott, Rafael Polania, Robin Pope, Antonio Rangel, Christian Ruff, Andrei Shleifer, Hrvoje Stojic, Chris Summerfield, Shyam Sunder, Peter Wakker, and Ryan Webb for helpful comments, and the National Science Foundation for research support. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Cognitive Imprecision and Small-Stakes Risk Aversion
Mel Win Khaw, Ziang Li, and Michael Woodford
NBER Working Paper No. 24978
August 2018
JEL No. C91,D03,D81,D87

ABSTRACT

Observed choices between risky lotteries are difficult to reconcile with expected utility maximization, both because subjects appear to be too risk averse with regard to small gambles for this to be explained by diminishing marginal utility of wealth, as stressed by Rabin (2000), and because subjects' responses involve a random element. We propose a unified explanation for both anomalies, similar to the explanation given for related phenomena in the case of perceptual judgments: they result from judgments based on imprecise (and noisy) mental representations of the decision situation. In this model, risk aversion results from a sort of perceptual bias — but one that represents an optimal decision rule, given the imprecision of the mental representation of the situation. We propose a quantitative model of the noisy mental representation of simple lotteries, based on other evidence regarding numerical cognition, and test its ability to explain the choice frequencies that we observe in a laboratory experiment. Our model is more consistent with the laboratory data than random versions of expected utility theory or prospect theory, using both in-sample and out-of-sample tests of model fit.

Mel Win Khaw
Department of Economics
Columbia University
420 W. 118th Street
New York, NY 10027
mwk2126@columbia.edu

Michael Woodford
Department of Economics
Columbia University
420 W. 118th Street
New York, NY 10027
and NBER
mw2230@columbia.edu

Ziang Li
Department of Economics
Columbia University
420 W. 118th Street
New York, NY 10027
zl2505@columbia.edu

Risk-averse choices are conventionally explained as reflecting expected utility maximization (EUM) on the part of decision makers for whom the marginal utility of additional wealth decreases with increases in their wealth. However, the observation that people often decline even very small bets that offer somewhat better than fair odds poses a problem for this theory. In the case of any smooth utility-of-wealth function, choices ought to become nearly risk-neutral in the case of small enough stakes (Arrow, 1971). And while it is always possible to explain rejection of any given bet by assuming sufficient rapidly diminishing marginal utility of wealth, the degree of curvature of the utility function that is required will then imply that the same person should reject even extremely favorable bets when potential losses are moderately large (though in no way catastrophic), as explained by Rabin (2000); this too seems plainly counter-factual.[1]

A well-known response to this difficulty (Rabin and Thaler, 2001) is to propose that people maximize the expected value of a nonlinear utility function, but that this function is *reference-dependent:* it is not a context-invariant function of wealth, but instead depends on how the wealth that may be obtained in different possible states compares to some reference level of wealth. This context-sensitive reference level might be identified with the decision maker's existing wealth at the time of the choice (as in Kahneman and Tversky, 1979), or with a level of wealth that she has reason to expect to achieve (as in Koszëgi and Rabin, 2006, 2007). Under such a generalization of standard EUM, small-stakes risk aversion requires only a sufficiently rapid decrease in marginal utility near the reference level of wealth;[2] if the marginal utility of wealth does not continue to decrease at a similar rate with additional small increases in wealth above the reference level, such an assumption remains consistent with acceptance of larger gambles with only moderately better-than-fair odds.

However, this solution to the puzzle raises the question why the human mind should exhibit this reference-dependence, given that it leads to behavior that would seem not to be in the decision maker's interest.[3] Simply stating that this appears to be what many people prefer — as if they perfectly understand what they are getting from their choices and nonetheless persistently choose that way — is not entirely convincing. We propose instead an alternative interpretation, under which decision makers often fail to accurately choose the option that would best serve their true objectives, because their decision is based not on the exact characteristics of the available options, but rather on an imprecise mental representation of them.

[1] Rabin's argument appeals to introspection. But see Cox *et al.* (2013) for examples of experiments in which subjects make choices with respect to both small and large bets that are inconsistent with EUM under any possible concave utility function. A wider range of experimental anomalies that challenge EUM as a complete theory of choices with regard to risk are reviewed in Pope *et al.* (2007) and Friedman *et al.* (2014).

[2] This might be an actual kink at the reference point, as implied by the hypothesis of loss aversion introduced by Kahneman and Tversky. If the reference point is given by the decision maker's expected wealth level, then loss aversion implies that there exist better-than-fair odds at which a risky gamble will be declined even in the case of arbitrarily small stakes, as discussed by Rabin and Thaler (2001). Risk aversion with arbitrarily small stakes is also possible if the marginal utility of additional wealth becomes unboundedly large at the reference point, as discussed further below.

[3] As Rabin and Thaler (2001) point out, "myopic loss-averters ... make decisions that mean that others can take money from them with very high probability and very low risk." They also note that such exploitation seems all too commonplace. Our point is not to assert that the predictions of such a model must be wrong, but rather to urge that persistent behavior of this kind calls for an explanation.

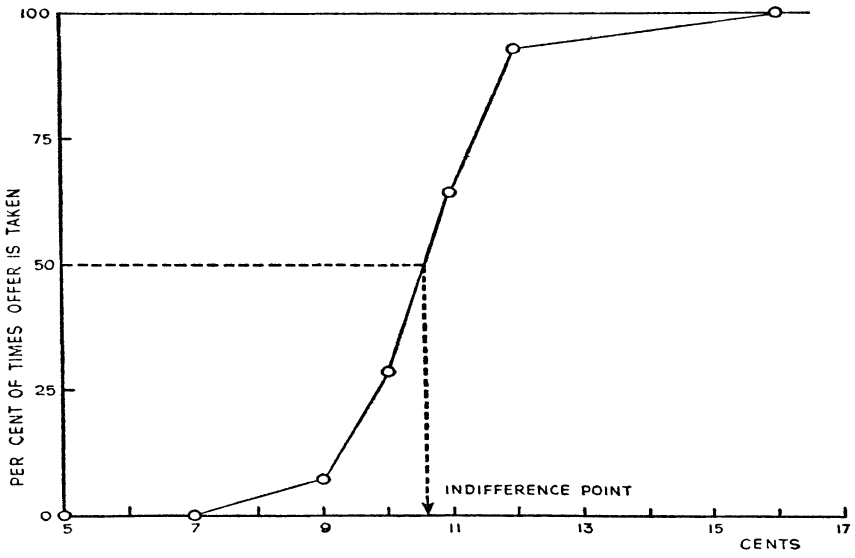


Figure 1: Probability of acceptance of a simple gamble as a function of the amount x that can be won (shown on horizontal axis); from Mosteller and Nogee (1951).

Our alternative explanation has the advantage that it can simultaneously explain another anomalous feature of choice behavior in experimental settings. EUM implies that choice should be a deterministic function of the monetary payoffs offered and their associated probabilities. But in the laboratory, instead, choices appear to be random, in the sense that the same subject will not always make the same choice when offered the same set of simple gambles on different occasions (Hey and Orme, 1994; Hey, 1995, 2001). This was evident (though little remarked upon) already in Mosteller and Nogee (1951), one of the earliest experimental studies of the empirical support for EUM. Figure 1 (reproduced from their paper) plots the responses of one of their subjects to a series of questions of a particular type. In each case, the subject was offered a choice of the form: are you willing to pay five cents for a gamble that will pay an amount X with probability $1/2$, and zero with probability $1/2$? The figure shows the fraction of trials on which the subject accepted the gamble, in the case of each of several different values of X. The authors used this curve to infer a value of X for which the subjects would be indifferent between accepting and rejecting the gamble, and then proposed to use this value of X to identify a point on the subject's utility function.

The fact that the indifference point is at a value of X greater than 10 cents (the case of a fair bet) is taken by the authors to indicate a concave utility function. But in fact, *no* utility function is consistent with the data shown in the figure; for EUM implies that the probability of acceptance should be zero for all values of X below the indifference point, and one for all values above it. Instead one observes probabilistic choice for a range of values of X, with the probability of acceptance increasing monotonically with X. This randomness is often de-emphasized in discussions of the experimental evidence for particular types of preferences over risky gambles, by simply focusing on modal or median responses. For example, Kahneman and Tversky (1979) assume (without much discussion) that prospect theory — presented as a theory that makes deterministic predictions about the valuation of risky prospects — should be understood, for purposes of empirical testing, as making predic-

tions about which of two prospects will be chosen *more often* in a binary choice experiment, rather than predictions about which alternative will *always* be chosen.[4]

There exists a view of the nature of trial-to-trial random variation in responses under which such an approach can be defended, as discussed by Becker *et al.* (1963). Suppose that a deterministic theory assigns a scalar value $v(\chi)$ to any vector χ of lottery characteristics, but that on any individual trial, the value assigned to a lottery with characteristics χ will be given by $v(\chi) + \epsilon$, where ϵ is a draw from some atomless probability distribution F that is independent of χ. In the case of a binary choice between two lotteries with characteristics χ_1 and χ_2 respectively, the former option is chosen if and only if $v(\chi_1) + \epsilon_1 > v(\chi_2) + \epsilon_2$, where ϵ_1, ϵ_2 are two independent draws from the distribution F, as in the additive random utility model of discrete choice (McFadden, 1981). Then option 1 will be chosen more often than option 2 if and only if $v(\chi_1) > v(\chi_2)$.

However, once one admits that the choice process must involve a stochastic element, there is no reason to suppose that randomness enters only in the way proposed in an additive random utility model — that is, that the valuations $v(\chi)$ are computed with perfect precision, with a random component added only after the relevant features of each available option have been summarized by a single scalar index of value. Randomness of responses of the kind illustrated in Figure 1 is commonplace in experimental studies of sensory perception, where such a figure is known as a "psychometric function."[5]

But in the psychophysics literature, it is commonly understood that an important part of the randomness in perceptual judgments results from randomness at relatively early stages of cognitive processing — randomness of the data produced by the sense organs, that must then be interpreted by the observer's nervous system. And while the earliest models in psychophysics posited that perceptual judgments themselves were random draws from a distribution of possible "percepts," the more modern literature (e.g., Green and Swets, 1966) separates the question of how subjects *select a response* on the basis of sensory data from the noise in the sensory data themselves. From this point of view, perceptual judgments represent a form of inference from noisy evidence. Indeed, an important branch of the literature explores the hypothesis that perceptual judgments can be modeled as optimal Bayesian inference, subject to the constraint that they must be based on noisy sensory data.

The hypothesis that perceptual judgments represent optimal Bayesian inference from an imprecise internal representation of the situation does not rule out the possibility of systematic bias in perceptual judgments, as we illustrate in the next section.[6] But it does imply that

[4]This is also the assumption that motivates the method used by Mosteller and Nogee (1951) to test the consistency of their subjects' behavior with EUM, illustrated in Figure 1; Becker *et al.* (1964) call this "the Fechner postulate." Other studies explicitly model the randomness in individual responses (e.g., Loomes and Sugden, 1995; Ballinger and Wilcox, 1997; Holt and Laury, 2002; Loomes, 2005; Wilcox, 2008), but still typically treat the randomness as something that can be specified independently of a "core" deterministic model of preference over lotteries (such as EUM), which is supposed to explain subjects' risk attitudes.

[5]See, for example, Gabbiani and Cox (2010), chap. 25; Gescheider (1997), chap. 3; or Glimcher (2011), chap. 4. It was doubtless due to familiarity with such figures in the literature on sensory perception that Mosteller and Nogee found it natural to plot their data in the way that they did. Note that the method that they use to identify an indifference point for their subject corresponds to a standard way of defining a "point of subjective equality" between two different types of sensory stimuli using a psychometric function (see, e.g., Gescheider, 1997, p. 52).

[6]See also section 1 of Khaw *et al.* (2017) for a detailed discussion of another leading example, the Bayesian

there should be a strong connection between the nature of the *noise* in perceptual judgments and their average *bias:* if there were no random noise in the internal representations on which the judgments are based, there should (if the subject forms optimal Bayesian judgments) be no noise in perceptual judgments, and *no bias either* — thus the degree and nature of the bias should be directly connected to the magnitude and nature of the noise.

We propose that small-stakes risk aversion can be explained in the same way as perceptual biases that result from noise in internal representations.[7] According to our theory, intuitive estimates of the value of risky prospects (not ones resulting from explicit symbolic calculations) are based on mental representations of the magnitudes of the available monetary payoffs that are imprecise in roughly the same way that the representations of sensory magnitudes are imprecise, and in particular are similarly random, conditioning on the true payoffs. Intuitive valuations must be some function of these random mental representations. We explore the hypothesis that they are produced by a decision rule that is optimal, in the sense of maximizing the (objective) expected value of the decision maker's expected wealth, subject to the constraint that the decision must be based on the random mental representation of the situation.

Under a particular model of the noisy coding of monetary payoffs, we show that this hypothesis will imply apparently risk-averse choices: the expected net payoff of a bet will have to be strictly positive for indifference, in the sense that the subject accepts the bet exactly as often as she rejects it (as in Figure 1). Risk aversion of this sort is consistent with a decision rule that is actually optimal from standpoint of an objective (expected wealth maximization) that involves no "true risk aversion" at all; this bias is consistent with optimality in the same way that perceptual biases (such as the oblique bias in the perception of orientation) can be consistent with Bayesian inference from noisy sensory data. And not only can our theory explain apparent risk aversion without any appeal to diminishing marginal utility, but it can also explain why the "risk premium" required in order for a risky bet to be accepted over a certain payoff does not shrink to zero (in percentage terms) as the size of the bet is made small, contrary to the prediction of EUM.

Section 1 reviews evidence regarding the mental representation of numerical magnitudes that motivates our model of noisy coding of monetary payoffs. Section 2 presents an explicit model of choice between a simple risky gamble and a certain monetary payoff, of the kind that occurs in the experiment of Mosteller and Nogee, and derives predictions for the both the randomness of choice and the degree of apparent risk aversion implied by an optimal decision rule. Section 3 describes a simple experiment in which we are able to test some of the specific quantitative predictions of this model. Section 4 discusses further implications of our theory of small-stakes risk attitudes, and concludes.

---

explanation for the well-documented "oblique bias" in perceptions of orientation.

[7] Our theory is thus similar to proposals in other contexts (such as Koszëgi and Rabin, 2008) to interpret experimentally observed behavior in terms of mistakes on the part of decision makers — i.e., a failure to make the choices that would maximize their true preferences — rather than a reflection of some more complex type of preferences. More specifically, we follow Woodford (2012), Steiner and Stewart (2016), Gabaix and Laibson (2017), and Natenzon (2017) in proposing that choice biases can reflect optimal Bayesian decision making on the basis of a noisy representation of the decision problem.

1 Imprecision in Numerical Cognition

An important recent literature on the neuroscience of perception argues that biases in perceptual judgments can actually reflect optimal decisions — in the sense of minimizing average error, according to some well-defined criterion, in a particular class of situations that are possible ex ante — given the constraint that the brain can only produce judgments based on the noisy information provided to it by sensory receptors and earlier stages of processing in the nervous system, rather than on the basis of direct access to the true physical properties of external stimuli (e.g., Stocker and Simoncelli, 2006; Wei and Stocker, 2015). The approach has been used to explain systematic biases in perception in a variety of sensory domains (Petzschner *et al.,* 2015; Wei and Stocker, 2017).

The relevance of these observations about perceptual judgments for economic decision might nonetheless be doubted. Some may suppose that the kind of imprecision in mental coding just discussed matters for the way in which we perceive our environment through our senses, but that an intellectual consideration of hypothetical choices is an entirely different kind of thinking. Moreover, it might seem that typical decisions about whether to accept gambles in a laboratory setting, such as the experiment of Mosteller and Nogee (1951), involve only numerical information that is presented to the subjects in an exact (symbolic) form, offering no obvious opportunity for imprecise perception. However, we have reason to believe that reasoning about numerical information often involves imprecise mental representations of a kind directly analogous to those involved in sensory perception.

1.1 Imprecise Perception of Numerosity

This is clearest (and has been studied most thoroughly) in the case of perceptions of the number of items present in a visual display. For example, quick judgments can be made about the number of dots present in a visual display of a random cloud of dots, without taking the time to actually count them.[8] As with perceptions of physical magnitudes such as length or area, such judgments of numerosity are subject to random error. And just as in the case of sensory magnitudes, the randomness in judgments can be attributed to randomness in the neural coding of numerosity, resulting from the width of the "tuning curves" of neurons that selectively respond to arrays with greater or smaller numbers of dots.[9]

We can learn about how the degree of randomness of the mental representation of a number varies with its size from the frequency distribution of errors in estimation of numerosity. A well-established finding is that when subjects must estimate which of two numerosities is greater, or whether two arrays contain the same number of dots, the accuracy of their judgments is a function of the ratio of the two numbers (but independent of their absolute magnitudes) — a "Weber's Law" for the discrimination of numerosity analogous to the one observed to hold in many sensory domains (Ross, 2003; Cantlon and Brannon, 2006; Nieder and Merten, 2007; Nieder, 2013). Moreover, when subjects must report an estimate of the

[8] One of the earliest published experimental investigations was by Jevons (1871).

[9] The tuning curves of "number neurons" have been measured using single-cell recording techniques in the brains of both cats and macaques (Thompson *et al.,* 1970; Nieder and Merten, 2007; Nieder and Dehaene, 2009). While similar methods cannot be used with humans, more indirect evidence suggests the existence of "number neurons" in the human brain as well (Piazza *et al.,* 2004; Nieder, 2013).

number of dots in a visual array,[10] the standard deviation of the distribution of estimates grows in proportion to the mean estimate, with both the mean and standard deviation being larger when the true number is larger (Izard and Dehaene, 2008; Kramer *et al.,* 2011); similarly, when subjects are required to produce a particular number of responses (without counting them), the standard deviation of the number produced varies in proportion to the target number (and to the mean number of responses produced) — the property of "scalar variability" (Whalen *et al.,* 1999; Cordes *et al.,* 2001).

All of these observations are consistent with a theory according to which such judgments of numerosity are based on an internal representation that can be represented mathematically by a quantity that is proportional to the logarithm of the numerical value that is being encoded, plus a random error the variance of which is independent of the numerical value that is encoded (van Oeffelen and Vos, 1982; Izard and Dehaene, 2008).[11] Let the number n be represented by a real number r that is drawn from a distribution

$$r \;\sim\; N(\log n,\, \nu^2), \tag{1.1}$$

where ν is a parameter independent of n. Suppose furthermore that if two stimuli of respective numerosities n_1 and n_2 are presented, their corresponding internal representations r_1, r_2 are independent draws from the corresponding distributions. Finally, suppose that a subject judges the second array to be more numerous than the first if and only if the internal representations satisfy $r_2 > r_1$.[12] Then a subject is predicted to respond that array 2 is more numerous with probability

$$\text{Prob}[\text{"2 is more"}] \;=\; \Phi\left(\frac{\log(n_2/n_1)}{\sqrt{2}\nu}\right), \tag{1.2}$$

where $\Phi(z)$ is the cumulative distribution function of a standard normal variate z.

Equation (1.2) predicts that "Weber's Law" should be satisfied: the response probability depends only on the ratio n_2/n_1, and not on the absolute numerosity of either array. More specifically, it predicts that the z-transformed response probability ($z(p) \equiv \Phi^{-1}(p)$) should be an increasing linear function of $\log n_2$, with a slope that is independent of the numerosity n_1 of the first array, and a value of zero when $n_2 = n_1$. This is exactly what the discrimination data of Krueger (1984) show.[13]

[10] Here we refer to arrays containing more than five or so dots. As discussed by Jevons (1871) and many subsequent authors, the numerosity of very small arrays can be immediately perceived (without counting) with high accuracy and confidence; the cognitive process used in such cases, termed "subitizing" by Kaufman *et al.* (1949), is quite distinct from the ability to estimate the approximate numerosity of larger arrays, to which the statements in the text refer.

[11] Buckley and Gillman (1974) had earlier proposed a similar model to explain behavior in experiments involving magnitude comparisons between numbers represented by Arabic numerals; these related experiments are discussed below.

[12] This is an optimal decision rule, in the sense of maximizing the frequency of correct answers, in the case of any prior distribution under which (n_2, n_1) has the same prior probability as (n_1, n_2) — that is, the choice of which stimulus to present first is arbitrary.

[13] See Figure 5 of Krueger (1984), in which the three panels correspond to three successively larger values of n_1; each panel plots the z-transformed frequency of judgment that array 2 is more numerous (on the vertical axis) as a function of $\log n_2 - \log n_1$ (measured in 4-percent "steps," on the horizontal axis).

The observed variability of estimates of numerosity is consistent with the same kind of model of noisy coding. Suppose that the subject's estimate $\hat{n}$ of the numerosity of some array must be produced on the basis of the noisy internal representation r hypothesized above. If we approximate the prior distribution from which the true numerosity n is drawn (in a given experimental context) by a log-normal distribution,[14] $\log n \sim N(\mu, \sigma^2)$, then the posterior distribution for n, conditional on an internal representation r drawn from (1.1), will also be log-normal: $\log n|r \sim N(\mu_{post}(r), \sigma^2_{post})$. Here $\mu_{post}(r)$ is an affine function of r, with a slope $0 < \beta < 1$ given by

$$\beta \equiv \frac{\sigma^2}{\sigma^2 + \nu^2}, \tag{1.3}$$

while $\sigma^2_{post} > 0$ is independent of r.[15]

If we hypothesize that the subject's numerosity estimate is optimal, in the sense of minimizing the mean squared estimation error when stimuli are drawn from the assumed prior distribution,[16] then we should expect the subject's estimate to be given by the posterior mean, $\hat{n}(r) = \mathrm{E}[n|r]$. In this case, $\log \hat{n}(r)$ will be an affine function of r, with a slope of β. The same will be true (though the affine function will have a slightly different intercept) if we assume instead that the subject's estimate is given by the posterior mode (a "maximum *a posteriori* estimate," as often assumed in Bayesian models of statistical inference), or that it minimizes the mean squared percentage error.[17] In any of these cases, the fact that $\log \hat{n}(r)$ is an affine function of r, together with (1.1), implies that conditional on the true numerosity n, the estimate $\hat{n}$ will be log-normally distributed: $\log \hat{n} \sim N(\hat{\mu}(n), \hat{\sigma}^2)$, where $\hat{\mu}(n)$ is an affine function of $\log n$ with slope β, and $\hat{\sigma}^2$ is independent of n.

It then follows from the properties of log-normal distributions that

$$\frac{\mathrm{SD}[\hat{n}]}{\mathrm{E}[\hat{n}]} = \sqrt{e^{\hat{\sigma}^2} - 1} > 0,$$

regardless of the true numerosity n. Thus the property of scalar variability is predicted by a model of optimal estimation.[18]

[14] We adopt this approximation in order to allow a simple analytical calculation of the Bayesian posterior distribution, even though in the experiments referred to here, the value of n is actually always an integer. For more exact models of numerosity estimation, also based on the hypothesis of log-normal coding, see for example van Oeffelen and Vos (1982) or Izard and Dehaene (2008). The calculation presented here is offered as an introduction to the model of noisy coding proposed in section 2, where monetary payments are assumed to be positive real numbers rather than integers.

[15] See the online appendix for details of the calculation.

[16] Such a hypothesis does not imply that subjects in numerosity estimation experiments consciously calculate anything using Bayes' rule; only that, in some way or another, their intuitive judgments have come to be calibrated so as to be optimal for a certain environment. We do not here discuss the question of how much experience should be required in order for subjects' estimates to become well-calibrated to a given context.

[17] Again, see the online appendix for details. Because numerosity estimation experiments are typically not incentivized, it is unclear what objective subjects should be assumed to maximize under an optimizing model of perceptual judgments. Instead, in the case of the choices between simple gambles modeled in the next section, our theory is based on subjects' well-defined financial incentives.

[18] Alternatively, the standard deviation of the distribution of $\log \hat{n}$ should be independent of n. This is found to be roughly the case, when statistics of the distribution of $\log \hat{n}$ are plotted as functions of $\log n$, as in Kramer *et al.* (2011).

A further implication of a Bayesian model of numerosity estimation is that the average subjective estimate $\mathrm{E}[\hat{n}|n]$ should in general differ from the true numerosity n: subjects' estimates should be *biased*. Specifically, the model just proposed implies a power-law relationship,

$$\mathrm{E}[\hat{n}|n] \;=\; An^{\beta} \tag{1.4}$$

for some $A > 0$, where $0 < \beta < 1$ is again defined by (1.3). This implies *over-estimation* of small numerosities (greater than five), but *under-estimation* of larger numerosities, to a progressively greater extent the larger the true numerosity n. The result illustrates our earlier remark that random noise in internal representations results not only in arbitrary randomness in judgments based on those representations, but in biased judgments as well, even when the judgments are optimal (conditional on having to be based on the noisy internal representation).

This kind of "regressive bias" in subjects' estimates of numerosity is characteristic of all experiments in this area, beginning with the classic study of Kaufman *et al.* (1949).[19] In fact, authors often report that average estimates can be fit reasonably well by a concave power law (or log-log plot), of the kind indicated by (1.4).[20] The cross-over point, however, at which the bias switches from over-estimation to under-estimation varies across studies. Over-estimation is found only in the case of numerosities of no more than 10, in the studies of Kaufman *et al.* (1949) and Indow and Ida (1977); but for all numerosities less than 25, in the studies reviewed by Krueger (1984); and for all arrays with less than 130 dots, in the study of Hollingsworth *et al.* (1991). As noted by Izard and Dehaene (2008), the cross-over point seems to depend on the range of numerosities used in the study in question; the validity of this interpretation is indicated by the recent study of Anobile *et al.* (2012), who find different concave mappings[21] from n to $\mathrm{E}[\hat{n}|n]$ in two experiments using similar methodologies, but different ranges for the true numerosities used in the experiment (1 to 30 dots in one case, 1 to 100 dots in the other).

This is just what the Bayesian model proposed above would predict: if we vary μ across experiments, holding the other parameters fixed, the cross-over point is predicted to vary in proportion to the variation in the prior mean of n.[22] The Bayesian model also predicts, for a given prior, that increased imprecision in mental coding (a larger value of ν) should result in a lower value of β, and hence a more concave relationship between the true and estimated numerosities; and this is what Anobile *et al.* (2012) find when subjects' cognitive load is increased, by requiring them to perform another perceptual classification task in addition to estimating the number of dots present. Thus many quantitative features of observed errors in judgments of numerosity are consistent with a model of optimal judgment based on a noisy internal representation of numerosity, and a specific (log-normal) model of the noisy coding of numerical magnitudes in such cases.

[19] It can be seen in the data of Jevons (1871), though not remarked upon by him.

[20] See, e.g., Krueger, 1972, 1984; Indow and Ida, 1977; or Kramer *et al.*, 2011.

[21] See panel B of their Figure 3.

[22] Again, see the online appendix for details. A similar regression bias, with the cross-over point similarly varying with the range of stimulus magnitudes used in a given experiment, is observed in the case of estimates of a variety of sensory magnitudes. See Petzschner *et al.* (2015) for a review, and discussion of how a Bayesian model of perceptual judgments similar to the one proposed here can explain these and other patterns.

1.2 Symbolically Presented Numerical Information

The well-documented imprecision in people's perception of visually presented numerical information might seem, however, to be irrelevant to situations like the experiment of Mosteller and Nogee, in which the relevant monetary amounts are described to the decision maker using number symbols. One might reasonably suppose that symbolically presented numbers are generally understood precisely by the hearer; and to the extent that perceptual errors do occur, they should not generally be expected to conform to Weber's Law, as in the case of sensory magnitudes.[23]

Nonetheless, there is a good deal of evidence suggesting that even when numerical quantities are presented using symbols such as Arabic numerals, the semantic content of the symbol is represented in the brain in a way that is similar to the way in which magnitudes are represented — involving imprecision, just as with the representation of physical magnitudes, and with similar quantities represented in similar ways, so that nearby numerical magnitudes are more likely to be confused with one another (Dehaene, 2011). This is not the *only* way in which numerical information is understood to be represented in the brain; according to the well-known "triple-code model" of Dehaene (1992), numbers are represented in three different ways (three "codes"), in circuits located in different regions of the brain, each with a distinct function. An Arabic code, located in the left and right inferior ventral occipital-temporal areas, is used for explicit multi-digit arithmetic calculations. Simple verbal counting and retrieval of memorized facts of arithmetic are instead executed via a verbal code, subserved by the left perisylvian area.

Yet a third code, the "analog magnitude code," is drawn upon in tasks involving number comparisons and approximation. This is thought to be a "semantic" representation of the size of the quantity represented by a given number — "the abstract quantity meaning of numbers rather than the numerical symbols themselves" (Dehaene *et al.,* 2003, p. 492) — and to be independent of the symbolic form in which the number is presented; neuro-imaging studies suggest that this code is located in the intraparietal sulcus in humans (Piazza *et al.,* 2004). Scalp EEG recordings while subjects process information presented in the form of Arabic numerals also indicate that the neural patterns evoked by particular numbers vary continuously with numerical distance, so that (for example) the neural signals for "3" are more similar to those for "4" than to those for "5" (Spitzer *et al.,* 2017; Teichmann *et al.,* 2018; Luyckx *et al.,* 2018).

The existence of an approximate semantic representation of numerical quantities, even when numbers are presented symbolically, can also be inferred behaviorally from the ability of patients with brain injuries that prevent them from performing even simple arithmetic (using the exact facts of arithmetic learned in school) to nonetheless make fairly accurate approximate judgments (Dehaene and Cohen, 1991). In normal adult humans, this approximate "number sense" seems also to be drawn upon when number comparisons are made very quickly, or when previously presented numerical information that has not been precisely memorized must be recalled.

For example, Moyer and Landauer (1967) presented subjects with two numerals, and required them to press one of two keys to indicate which numeral indicated the larger number.

[23]For example, if it were a simple matter of sometimes mis-hearing numbers stated by an experimenter, one might expect that $34.13 could more easily be mistaken for $44.13 than for $34.89.

They found that both the fraction of incorrect responses and the time required to decide were decreasing functions of the numerical distance between the two numbers referred to by the numerals; these findings are analogous to the way that both error rates and response times vary with the magnitude difference between two sensory stimuli in experiments where a subject must determine which of two stimuli is greater in magnitude (the louder sound, the longer line, and so on). Moyer and Landauer conclude that "the displayed numerals are converted [by the mind] to analogue magnitudes, and a comparison is then made between those magnitudes in much the same way that comparisons are made between physical stimuli" (p. 1520).

Moreover, there is evidence that the mental representation of numerical information used for approximate calculations involves the same kind of logarithmic compression as in the case of non-symbolic numerical information, even when the numerical magnitudes have originally been presented symbolically. Moyer and Landauer (1967), Buckley and Gillman (1974), and Banks *et al.* (1976) find that the reaction time required to judge which of two numbers (presented as numerals) is larger varies with the distance between the numbers on a compressed, nonlinear scale — a logarithmic scale, as assumed in the model of the coding of numerosity sketched above, or something similar — rather than the linear (arithmetic) distance between them.[24]

In an even more telling example for our purposes, Dehaene and Marques (2002) showed that in a task where people had to estimate the prices of products, the estimates produced exhibited the property of scalar variability, just as with estimates of the numerosity of a visual display. This was found to be the case, even though both the original information people had received about prices and the responses they produced involved symbolic representations of numbers. Evidently, an approximate analog representation of the prices remained available in memory, though the precise symbolic representation of the prices could no longer be accessed.[25]

Not only is there evidence for the existence of an approximate semantic representation of numerical information that is presented symbolically; it seems likely that this "analog magnitude code" is the *same* representation of number that is used when numbers are presented non-symbolically. The region in the intraparietal sulcus that is thought to be the locus of the analog magnitude code seems to be activated by the presentation of numerical stimuli, regardless of the format in which the information is presented: written words or Arabic numerals, visual or auditory presentations, symbolic or non-symbolic (Piazza *et al.*, 2004; Brannon, 2006). If this is true, it means that we should expect the quantitative model of imprecise internal representations that explains the perception of numerosity, a context in which the statistical structure of errors has been documented in more detail, to also apply to the imprecise internal representations that are drawn upon when fast, approximate

[24]Buckley and Gillman (1974) develop an extension of the model of noisy logarithmic coding of numerical magnitudes sketched above that explicitly models the dynamic process of comparison between two magnitudes, and show that the model predicts not only that the frequency of correct ranking should depend on the ratio of the two numbers (as discussed above) but that the mean time required to decide should depend on this ratio as well, as they find in their experiment. (See also Dehaene, 2008, for a related model.) The dynamic version of the model is not needed for our purposes here.

[25]This example is of particular relevance for our purposes, as it involves the mental representation of monetary amounts.

judgments are made about symbolically presented numerical information. We shall explore the implications of this hypothesis for risky choice.

More specifically, our hypothesis is that when people must decide whether a risky prospect (offering either of two possible monetary amounts as the outcome) is worth more or less than another monetary amount that could be obtained with certainty, they can make a quick, intuitive judgment about the relative value of the two options using the same mental faculty as is involved in making a quick estimate (without explicit use of precise arithmetic calculations) as to whether the sum of two numbers is greater or less than some other number.

If this is approached as an approximate judgment rather than an exact calculation (as will often be the case, even with numerate subjects), such a judgment is made on the basis of mental representations of the monetary amounts that are approximate and analog, rather than exact and symbolic; and these representations involve a random location of the amount on a logarithmically compressed "mental number line." The randomness of the internal representation of the numerical quantity (or perhaps, of its value to the decision maker) then provides an explanation for the randomness in the data of Mosteller and Nogee (1951); and as we show below, the logarithmic compression provides an explanation for subjects' apparent risk aversion, even in the case of gambles for very small stakes.[26]

Note that we do not assume that all decisions involving money are made in this way. If someone is asked to choose between \$20 and \$22, either of which can be obtained with certainty, we do not expect that they will sometimes choose the \$20, because of noise in their subjective sense of the size of these two magnitudes. The question whether \$20 is greater or smaller than \$22 can instead be answered reliably (by anyone who remembers how to count), using the "verbal code" hypothesized by Dehaene (1992) to represent the numbers, rather than the "analog magnitude code."

Likewise, we do not deny that numerate adults, if they take sufficient care (and consciously recognize the problem facing them as having the mathematical structure of a type of arithmetic problem), are capable of exact calculations of averages or expected values that would not introduce the kind of random error modeled in the next section. Nonetheless, we hypothesize that questions about small gambles in laboratory settings (even when incentivized) are often answered on the basis of an intuitive judgment based on approximate analog representations of the quantities involved. And though our results here cannot prove this, we suspect that many economic decisions in everyday life are made in a similar way, and hence may involve a similar error structure.

2 A Model of Noisy Coding and Risky Choice

We now consider the implications of a model of noisy internal representation of numerical magnitudes for choices between simple lotteries, of the kind that subjects are presented with in experiments like that of Mosteller and Nogee (1951). We assume a situation in which a subject is presented with a choice between two options: receiving a monetary amount $C > 0$

[26] Schley and Peters (2014) also propose that a compressive nonlinear mapping of symbolically presented numbers into mental magnitudes can give rise to additional risk aversion, alongside the risk aversion that can be attributed to diminishing marginal utility; but as we discuss in section 4.2 below, their theory differs in important respects from the one that we propose here.

with certainty, or receiving the outcome of a lottery, in which she will have a probability $0 < p < 1$ of receiving a monetary amount $X > 0$. We wish to consider how decisions should be made if they must be based on imprecise internal representations of the monetary amounts rather than their exact values.

We hypothesize that the subject's decision rule[27] is optimal, in the sense of maximizing the expected value of $U(W)$, subject to the constraint that the decision must be based on an imprecise representation $\mathbf{r}$ of the problem, rather than the true data. Here W is the subject's final wealth at the end of the experiment, and $U(W)$ is an indirect utility function, indicating the (correctly assessed) expected value to the subject of a given wealth (given the ways in which it can subsequently be spent). Note that our conceptual of the subject's objective (from the standpoint of which the decision rule can be said to be optimal) involves no "narrow bracketing" of the gains from a particular decision: it is assumed that only final wealth W matters, and not the sequence of gains and losses by which it is obtained. The expected value is defined with respect to some prior probability distribution over possible decision situations (here, possible values of X and C that might be offered).

Let W^a be the random final wealth if option a is chosen. If we consider only gambles for small amounts of money, we can use the Taylor approximation $U(W^a) \approx U(W_0) + U'(W_0) \cdot \Delta W^a$, where W_0 is the subject's wealth[28] independent of any gain from the experiment, ΔW^a is the random monetary amount gained in the experiment if option a is chosen, and $U'(W_0)$ is positive for all possible values of W_0. If we assume furthermore that the subject's information about W_0 is coded by some internal representation r_0, with a distribution that is independent of the details of the gains offered by the decision problem, while the quantities X and C have internal representations r_x and r_c respectively, that are distributed independently of W_0, then

$$\mathrm{E}[U(W^a)|\mathbf{r}] \;\approx\; \mathrm{E}[U(W_0)|r_0] \;+\; \mathrm{E}[U'(W_0)|r_0] \cdot \mathrm{E}[\Delta W^a|r_x, r_c]$$

will be an increasing function of $\mathrm{E}[\Delta W^a|r_x, r_c]$, regardless of the value of r_0.

It follows that, as long as stakes are small enough, an optimal decision rule is one that chooses the action a for which the value of $\mathrm{E}[\Delta W^a|r_x, r_c]$ is larger; we therefore consider the hypothesis that decisions are optimal in this sense. Note that our theory's predictions are thus consistent with "narrow bracketing": the choice between two risky prospects is predicted to depend only on the distributions of possible net gains associated with those prospects, and not on the level of wealth W_0 that the subject has from other sources. But for us this is a *conclusion* (a property of optimal decision rules) rather than a separate *assumption*. Note also that while we do not deny the reasonableness of assuming that the function $U(W)$ should involve diminishing marginal utility of wealth (in the case of sufficiently large changes in wealth), the degree of curvature of the function $U(W)$ plays no role in our predictions. Thus small-stakes risk aversion is not attributed to nonlinear utility of income or wealth in our theory.

In line with the evidence discussed in the previous section regarding internal representations of numerical magnitudes, we assume more specifically that the representations r_x and

[27]Here we mean a mathematical relationship that describes the systematic pattern in a subject's decisions; reference to a "rule" should not be taken to mean that the subject consciously seeks to conform to the formula.

[28]Technically, this should be an assessment of their anticipated lifetime prospects: a measure of their intertemporal budget, counting all sources that are independent of the choice made in the experiment.

r_c are each a random draw from a probability distribution of possible representations, with distributions

$$r_x \sim N(\log X,\, \nu^2), \qquad r_c \sim N(\log C,\, \nu^2). \tag{2.1}$$

Here $\nu > 0$ is a parameter that measures the degree of imprecision of the internal representation of such quantities (assumed to be the same regardless of the monetary amount that is represented); we further assume that r_x and r_c are distributed independently of one another. We treat the parameter p as known (it does not vary across trials in the experiment described below), so that the decision rule can (and indeed should) depend on this parameter as well.[29]

As in the model of numerosity perception presented in section 1.2, these representations do not themselves constitute perceived values of the monetary amounts; instead, the internal representations must be "decoded" in order to provide a basis for decision, in the case of a given decision problem. The optimal decision in the case of a pair of mental representations $\mathbf{r} = (r_x, r_c)$ depends not only on the specification (2.1) of the noisy coding, conditional on the true magnitudes, but also on the relative ex ante likelihood of different possible decision situations, which we specify by a prior probability distribution over possible values of (X, C). We can then consider the optimal decision rule from the standpoint of Bayesian decision theory. It is easily seen that $\mathrm{E}[\Delta W^a | r_x, r_c]$ is maximized by a rule under which the risky lottery is chosen if and only if

$$p \cdot \mathrm{E}[X|r_x] \;>\; \mathrm{E}[C|r_c], \tag{2.2}$$

which is to say if and only if the expected payoff from the risky lottery exceeds the expected value of the certain payoff.[30]

The implications of our logarithmic model of noisy coding are simplest to calculate if (as in the model of numerosity estimation) we assume a log-normal prior distribution for possible monetary quantities. To reduce the number of free parameters in our model, we assume that under the prior X and C are assumed to be independently distributed, and furthermore that the prior distributions for both X and C are the same (some ex ante distribution for possible payments that one may be offered in a laboratory experiment). It is then necessary only to specify the parameters of this common prior:

$$\log X,\, \log C \;\sim\; N(\mu,\, \sigma^2). \tag{2.3}$$

Under the assumption of a common prior for both quantities, the common prior mean μ does not affect our quantitative predictions about choice behavior; instead, the value of σ does matter, as this influences the ex ante likelihood of X being sufficiently large relative to C for the gamble to be worth taking. The model thus has two free parameters, to be estimated from subjects' behavior: σ, indicating the degree of ex ante uncertainty about what the payoffs might be, and ν, indicating the degree of imprecision in the coding of information that is presented about those payoffs on a particular trial.

[29] See section 4.1 for discussion of an extension of the model in which p is also imprecisely represented.

[30] Note that while the payoff C is certain, rather than random, once one knows the decision situation (which specifies the value of C), it is a random variable ex ante (assuming that many different possible values of C might be offered), and continues to be random even conditioning on a subjective representation of the current decision situation, assuming that mental representations are noisy as assumed here.

2.1 Predicted Frequency of Acceptance of a Gamble

Under this assumption about the prior, the posterior distributions for both X and C are log-normal, as in the model of numerosity estimation in the previous section. It follows that the posterior means of these variables are given by[31]

$$\mathrm{E}[X|\mathbf{r}] \;=\; e^{\alpha+\beta r_x}, \qquad \mathrm{E}[C|\mathbf{r}] \;=\; e^{\alpha+\beta r_c},$$

with β is again defined by (1.3). Taking the logarithm of both sides of (2.2), we see that this condition will be satisfied if and only if

$$\log p \;+\; \beta r_x \;>\; \beta r_c,$$

which is to say, if and only if the internal representation satisfies

$$r_x - r_c \;>\; \beta^{-1} \log p^{-1}. \tag{2.4}$$

Under our hypothesis about the mental coding, r_x and r_c are independently distributed normal random variables (conditional on the true decision situation), so that

$$r_x - r_c \;\sim\; N(\log X/C,\, 2\nu^2).$$

It follows that the probability of (2.4) holding, and the risky gamble being chosen, is given by

$$\text{Prob[accept risky}|X, C] \;=\; \Phi\left(\frac{\log X/C - \beta^{-1}\log p^{-1}}{\sqrt{2}\nu}\right). \tag{2.5}$$

Equation (2.5) is the behavioral prediction of our model. It implies that choice in a problem of this kind should be stochastic, as observed by Mosteller and Nogee (1951). Furthermore, it implies that across a set of gambles in which the values of p and C are the same in each case, but the value of X varies, the probability of acceptance should be a continuously increasing function of X, as shown in Figure 1. Figure 2 gives an example of what this curve is predicted to be like, in the case that $\sigma = 0.25$ and $\nu = 0.08$. Note that these values allow a reasonably close fit to the choice frequencies plotted in the figure from Mosteller and Nogee.

Moreover, the parameter values required to fit the data are fairly reasonable ones. The value $\nu = 0.08$ for the so-called "Weber fraction" is only half as large as the value of 0.17 in the logarithmic coding model that best fits human performance in comparisons of the numerosity of different fields of dots (Dehaene, 2008, p. 540); on the other hand, Dehaene (2008, p. 552) argues that one should expect the Weber fraction to be smaller in the case of numerical information that is presented symbolically (as in the experiment of Mosteller and Nogee) rather than non-symbolically (as in the numerosity comparison experiments). Hence this value of ν is not an implausible degree of noise to assume in the mental representations of numerical magnitudes used in approximate calculations.[32]

[31] See the online appendix for details of the calculation.

[32] In the experiment reported below, our subjects' choices are best fit by values of ν larger than this — in fact, more similar to the Weber fraction obtained in the study of numerosity comparisons.

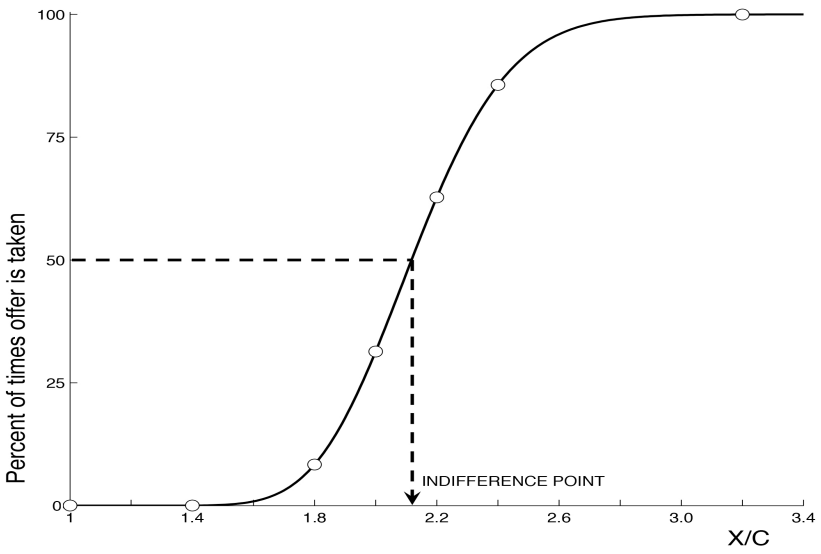


Figure 2: Theoretically predicted probability of acceptance of a simple gamble, as a function of X/C. Circles show the data from Figure 1, in which $C = 5$ cents.

The value of σ for the degree of dispersion of the prior over possible monetary rewards implies that if under the prior, the median value of X was expected to be 10 cents in the experiment, then the subject should have expected X to fall within a range between 6 cents and 16 cents 95 percent of the time — and this is more or less the range of values offered to the subject, as shown in Figure 1. Hence a prior with this degree of uncertainty would be fairly well calibrated to the subject's actual situation.

2.2 Explaining the Rabin Paradox

Our model explains not only the randomness of the subject's choices, but also her apparent risk aversion, in the sense that the indifference point (a value of X around 10.7 cents in Figure 1) corresponds to a gamble that is better than a fair bet. This is a general prediction of the model, since the indifference point is predicted to be at $X/C = (1/p)^{\beta^{-1}} > 1/p$, where the latter quantity would correspond to a fair bet. The model predicts risk neutrality (indifference when $X/C = 1/p$) only in the case that $\beta = 1$, which according to (1.3) can occur only in the limiting cases in which $\nu = 0$ (perfect precision of the mental representation of numerical magnitudes), or σ is unboundedly large (radical uncertainty about the value of the payoff that may be offered, which is unlikely in most contexts).

The model furthermore explains the Rabin (2000) paradox: the fact that the compensation required for risk does not become negligible in the case of small bets. According to EUM, the value of X required for indifference in a decision problem of the kind considered above should be implicitly defined by the equation

$$pU(W_0 + X) + (1 - p)U(W_0) \;=\; U(W_0 + C).$$

For any increasing, twice continuously differentiable utility function $U(W)$ with $U'' < 0$, if $0 < p < 1$, this condition implicitly defines a solution $X(C; p)$ with the property that

$pX(C;p)/C > 1$ for all $C > 0$, implying risk aversion. However, as C is made small, $pX(C;p)/C$ necessarily approaches 1. Hence the ratio pX/C required for indifference exceeds 1 (the case of a fair bet) only by an amount that becomes arbitrarily small in the case of a small enough bet. It is not possible for the required size of pX to exceed the certain payoff even by 7 percent (as in the case shown in Figure 1), in the case of a very small value certain payoff, unless the coefficient of absolute risk aversion $(-U''/U')$ is very large — which would in turn imply an implausible degree of caution with regard to large bets.

In our model, instead, the ratio pX/C required for indifference should equal $\Lambda \equiv p^{-(\beta^{-1}-1)}$, which is greater than 1 (except in the limiting cases mentioned above) by the same amount, regardless of the size of the gamble. As discussed above, the degree of imprecision in mental representations required for Λ to be on the order of 1.07 is one that is quite consistent with other evidence. Hence the degree of risk aversion indicated by the choices in Figure 1 is wholly consistent with a model that would predict only a modest degree of risk aversion in the case of gambles involving thousands of dollars.

It is also worth noting that our explanation for apparent risk aversion in decisions about small gambles does not rely on loss aversion, like the explanation proposed by Rabin and Thaler (2001). Our model of the mental representation of prospective gains assumes that the coding and decoding of the risky payoff X are independent of the value of C, so that small increases in X above C do not have a materially different effect than small decreases of X below C.

Instead, in our theory the EUM result that the compensation for risk must become negligible in the case of small enough gambles fails for a different reason. Condition (2.4) implies that the risky gamble is chosen more often than not if and only if $p \cdot m(X) > m(C)$, where $m(\cdot)$ is a power-law function of a kind that also appears in (1.4). It is *as if* the decision maker assigned a nonlinear utility $m(\Delta W^a)$ to the wealth increment ΔW^a. Our model of optimal decision on the basis of noisy internal representations explains why the ratio $m(X)/m(C)$ is in general not approximately equal to X/C even in the case that X and C are both small.

3 An Experimental Test

A notable feature of the behavioral equation (2.5) is that it predicts that subjects' choice frequencies should be *scale-invariant*, at least in the case of all small enough gambles: multiplying both X and C by an arbitrary common factor should not change the probability of the risky gamble being chosen. This feature of the model makes it easy to see that the Rabin paradox is not problematic for our model. In order to test this predictions of our model, we conducted an experiment of our own, in which we varied the magnitudes of both X and C. We recruited 20 subjects from the student population at Columbia University,[33] each of whom was presented with a sequence of several hundred trials. Each individual trial presented the subject with a choice between a certain monetary amount C and a probability p of receiving a monetary amount X.[34]

[33] Our procedures were approved by the Columbia University Institutional Review Board, under protocol IRB-AAAQ2255.

[34] The experimental design is discussed further in the online appendix.

The probability p of the non-zero outcome under the lottery was 0.58 on all of our trials, as we were interested in exploring the effects of variations in the magnitudes of the monetary payments, rather than variations in the probability of rewards, in order to test our model of the mental coding of monetary amounts. Maintaining a fixed value of p on all trials, rather than requiring the subject to pay attention to the new value of p associated with each trial, also made it more plausible to assume (as in the model above) that the value of p should be known precisely, rather than having to be inferred from an imprecisely coded observation on each occasion.

We chose a probability of 0.58, rather than a round number (such as one-half, as in the Mosteller and Nogee experiment discussed above), in order not to encourage our subjects to approach the problem as an arithmetic problem that they should be able to solve exactly, on the basis of representations of the monetary amounts using the "Arabic code" rather than the "analog magnitude code," in the terminology of Dehaene (1992).[35] We expect Columbia students to be able to solve simple arithmetic problems using methods of exact mental calculation that are unrelated to the kind of approximate judgments about numerical magnitudes with which our theory is concerned, but did not want to test this in our experiment. We chose dollar magnitudes for C and X on all trials that were not round numbers, either, for the same reason.

The value of the certain payoff C varied across trials, taking on the values \$5.55, \$7.85, \$11.10, \$15.70, \$22.20, or \$31.40. (Note that these values represent a geometric series, with each successive amount $\sqrt{2}$ times as large as the previous one.) The non-zero payoff X possible under the lottery option was equal to C multiplied by a factor $2^{m/4}$, where m took an integer value between 0 and 8. There were thus only a finite number of decision situations (defined by the values of C and X) that ever appeared, and each was presented to the subject several times over the course of a session. This allowed us to check whether a subject gave consistent answers when presented repeatedly with the same decision, and to compute the probability of acceptance of the risky gamble in each case, as in the experiment of Mosteller and Nogee. The order in which the various combinations of C and X were presented was randomized, in order to encourage the subject to treat each decision as an independent problem, with the values of both C and X needed to be coded and encoded afresh, and with no expectations about these values other than a prior distribution that could be assumed to be the same on each trial.

Our experimental procedure thus differed from ones often used in decision-theory experiments, where care is taken to present a sequence of choices in a systematic order, so as to encourage the subject to express a single consistent preference ordering. We were instead interested in observing the randomization that, according to our theory, should occur across a series of genuinely independent reconsiderations of a given decision problem; and we were concerned to simplify the context for each decision by eliminating any obvious reason for the data of one problem to be informative about the next.

We also chose a set of possible decision problems with the property that each value of X could be matched with the same geometric series of values for C, and vice versa, so that on each trial it was necessary to observe the values of both C and X in order to recognize the problem, and neither value provided much information about the other (as assumed

[35] See discussion of these alternative systems for the representation of numbers in section 1.3 above.

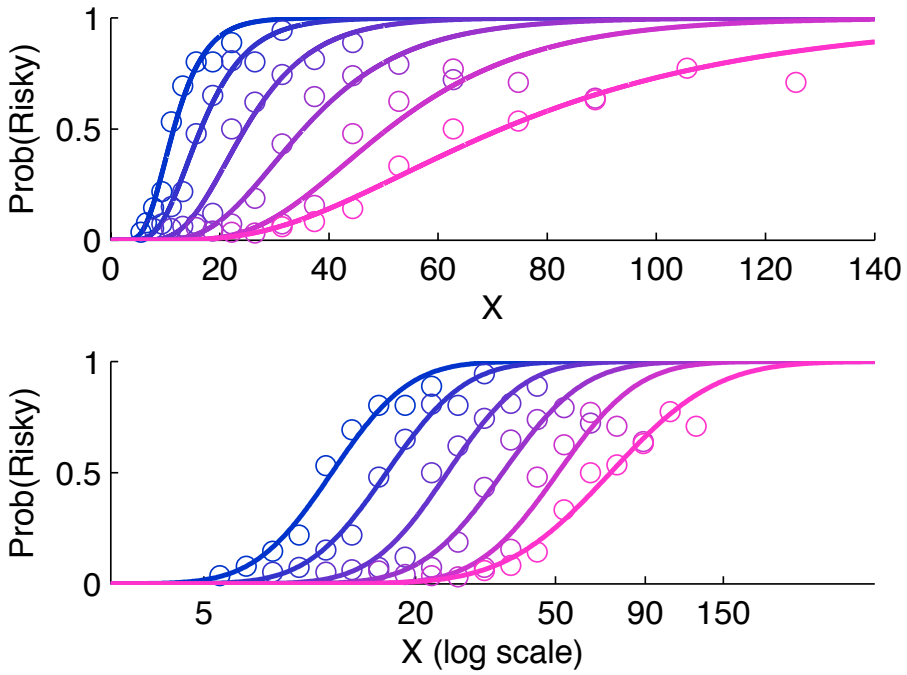


Figure 3: The probability of choosing the risky lottery, plotted as a function of the risky payoff X (data pooled from all 20 subjects). (a) The probability plotted as a function of X, for each of the different values of C (indicated by darkness of lines). (b) The same figure, but plotted against $\log X$ for each value of C.

in our theoretical model). At the same time, we ensured that the ratio X/C, on which the probability of choosing the lottery should depend according to our model, always took on the same finite set of values for each value of C. This allowed us to test whether the probability of choosing the lottery would be the same when the same value of X/C recurred with different absolute magnitudes for X and C.

3.1 Testing Scale-Invariance

Figure 3 shows how the frequency with which our subjects chose the risky lottery varied with the monetary amount X that was offered in the event that the gamble paid off, for each of the five different values of C. (For this first analysis, we pool the data from all 20 subjects.) Each data point in the figure (shown by a circle) corresponds to a particular combination (C, X).

In the first panel, the horizontal axis indicates the value of X, while the vertical axis indicates the frequency of choosing the risky lottery on trials of that kind $[Prob(Risky)]$. The different values of C are indicated by different colors of circles, with the darker circles corresponding to the lower values of C, and the lighter circles the higher values. (The six successively higher values of C are the ones listed above.) We also fit a sigmoid curve to the points corresponding to each of the different values of C, where the color of the curve again

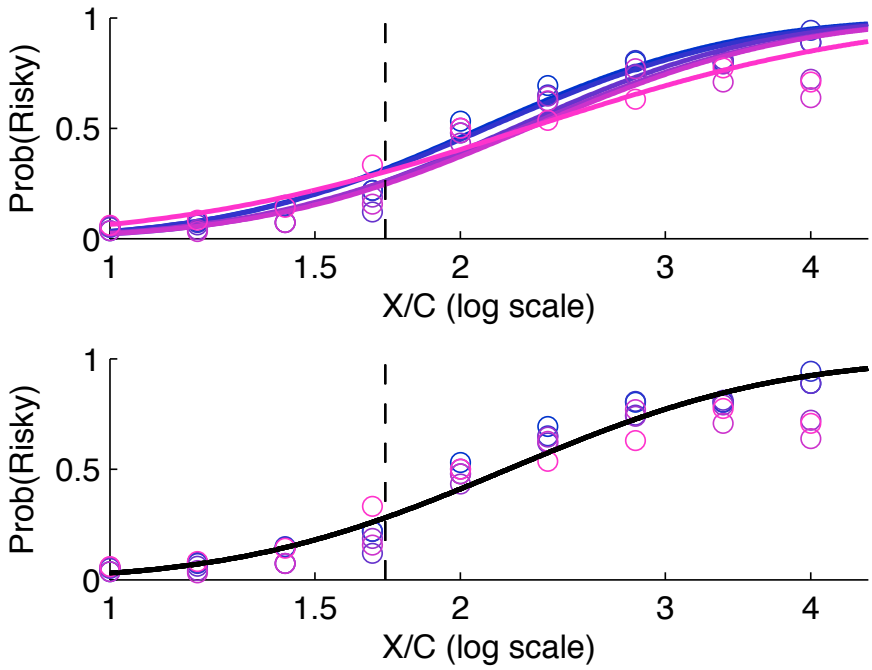


Figure 4: The same data as in Figure 3, now plotted as a function of $\log X/C$. (a) A separate choice curve estimated for each value of C, as in Figure 3. (b) A single choice curve, with parameters estimated to maximize the likelihood of the pooled data.

identifies the value of C. Each curve has an equation of the form

$$\text{Prob(Risky)} \;=\; \Phi(\delta_C + \gamma_C \log X), \tag{3.1}$$

where $\Phi(z)$ is again the CDF of the standard normal distribution, and the coefficients (δ_C, γ_C) are estimated separately for each value of C so as to maximize the likelihood of the data corresponding to that value of C. Note that for each value of C, we obtain a sigmoid curve similar to the one in Figure 2, though the fit is less perfect (at least partly because here, unlike in Figure 2, we are pooling the data from 20 different subjects).

The similarity of the curves obtained for different values of C can be seen more clearly if we plot them as a function of $\log X$, rather than on a scale that is linear in X, as shown in the second panel of Figure 3. (The color coding of the curves corresponding to different values of C is again the same.) The individual curves now resemble horizontal shifts of one another. The elasticity γ_C is similar for each of the values of C (with the exception of the highest value, $C = \$31.40$), and the value of $\log X$ required for indifference increases by a similar amount each time C is multiplied by another factor of $\sqrt{2}$.

These observations are exactly what we should expect, according to our logarithmic coding model. Condition (2.5) implies that a relationship of the form

$$\text{Prob(Risky)} \;=\; \Phi(\delta + \gamma \log(X/C)) \tag{3.2}$$

should hold for all values of C, meaning that in equation (3.1), γ_C should be the same for each value of C, and that the value of $\log X$ required for indifference should equal a constant

plus $\log C$. We can see more clearly the extent to which these precise predictions hold by plotting the curves in Figure 3(b) as functions of $\log(X/C)$, rather than as functions of $\log X$; this is done in the first panel of Figure 4. The six different curves come close to falling on top of one another, as predicted by the model (although, again, the curve for $C = \$31.40$ is somewhat out of line with the others). If we instead simply estimate parameters (δ, γ) to maximize the likelihood of the pooled data under the model (3.2), we obtain the single choice curve shown in the second panel of Figure 4.[36] This fits the data for the different values of X/C slightly worse than the individual choice curves shown in the previous panel, but not by much.

We can consider quantitatively the extent to which our data are more consistent with the more flexible model (3.1) than with the more restrictive predictions of (3.2), in two different ways. First, we consider the *in-sample* fit of the two models by selecting a subset of our observations (the "calibration dataset"), and find the parameter estimates for each model that maximize the likelihood of this dataset. The column labeled $LL^{calibration}$ in Table 1 reports the maximized value of the log-likelihood of the data in the calibration dataset. Of course, this is higher for the more flexible model, since (3.2) is nested within this class of models as a special case.

A more relevant comparison between the in-sample fits of the two models is given by the Bayes information criterion (BIC) statistic, also reported in the table for each model, which penalizes the use of additional free parameters. This is defined as[37] $BIC \equiv -2LL + k \log N_{obs}$, where k is the number of free parameters (adjusted to maximize the likelihood) for a given model, and N_{obs} is the number of observations in the calibration dataset. The data provide more evidence in favor of the model with the lower BIC statistic. In particular, for any two models $\mathcal{M}_1$ and $\mathcal{M}_2$, the *Bayes factor* K defined by

$$\log K_1 \;=\; \frac{1}{2}\left[BIC(\mathcal{M}_2) - BIC(\mathcal{M}_1)\right]$$

is the multiplicative factor by which the relative posterior probability that $\mathcal{M}_1$ rather than $\mathcal{M}_2$ is the correct model of the data is increased by the observations in the calibration dataset.[38]

We can also compare the *out-of-sample* fit of the two models, by reserving some of our observations (the "validation dataset"), and not using them to estimate the model parameters. The column labeled $LL^{validation}$ in Table 1 then reports the log-likelihood of the data in the validation dataset under each model, when the parameter values are used that were estimated using the calibration dataset.[39] If we update the posterior probabilities that the two models $\mathcal{M}_1$ and $\mathcal{M}_2$ are correct after observing the validation dataset as well, we obtain a composite Bayes factor $K = K_1 \cdot K_2$, where

$$\log K_2 \;=\; LL^{validation}(\mathcal{M}_1) \;-\; LL^{validation}(\mathcal{M}_2)$$

[36] The maximum-likelihood parameter estimates for the different choice curves, and the associated likelihoods, are reported in the online appendix.

[37] Here, as elsewhere in the paper, "log" refers to the natural logarithm.

[38] See, for example, Burnham and Anderson (2002), p. 303.

[39] In Table 1, and in the similar out-of-sample prediction exercises reported below and in the online appendix, the first 3/4 of each subject's trials are included in the calibration dataset, and the remaining 1/4 of the trials are used for the validation dataset.

Model	$LL^{calibration}$	BIC	$LL^{validation}$	log K
Pooled Data				
Scale-invariant	-2838.6	5694.6	-914.2	0.0
Unrestricted	-2820.4	5723.4	-912.8	13.0
Heterogeneous Parameters				
Scale-invariant	-1860.2	3946.1	-663.9	0.0
Unrestricted	-1594.9	4037.0	-755.6	137.1

Table 1: In-sample and out-of-sample measures of goodness of fit compared for the scale-invariant model (our logarithmic coding model) and an unrestricted statistical model in which a separate choice curve is estimated for each value of C. In the top panel, each model is fit to the pooled data from all 20 subjects. In the bottom panel, separate model parameters are fit to the data for each subject. (See text for further explanation.)

by Bayes' Rule. The logarithm of the composite Bayes factor K is reported in the final column of the table, as an overall summary of the degree to which the data provide support for each model. (In each case, $\mathcal{M}_1$ is the scale-invariant model, while $\mathcal{M}_2$ is the alternative model considered on that line of the table; thus values $K > 1$ indicate the degree to which the data provide more support for the scale-invariant model than for the alternative.)

In Table 1, we compare two models: our scale-invariant model (3.2) and the unrestricted alternative in which a separate probit model (3.1) is estimated for each of the six values of C, as in Figure 3.[40] In the case of the scale-invariant model, N_{obs} is the total number of observations in the calibration dataset, pooling the data for all six values of C, and there are $k = 2$ free parameters in the single model fit to all of these data. In the case of the unrestricted model, a separate probit model (each with $k = 2$ free parameters) is estimated for each value of C, and a BIC statistic is computed for that model (where N_{obs} is the number of observations in the calibration dataset with that value of C); the BIC reported in the "Unrestricted" row of the table is then the sum of the BIC statistics for these six independent probit models (just as the $LL^{calibration}$ is the sum of the log likelihoods for the six models).[41] In the top panel of the table, the two models are compared when a common set of parameters is used to fit the pooled data from all 20 subjects, as in Figures 3 and 4. In the lower panel, instead, individual model parameters are estimated for each subject, and the statistics reported are sums over all subjects of the corresponding model fit statistics for each subject.

Whether we assume a common set of parameters or subject-specific parameters, we see that the BIC statistic is lower for the scale-invariant model. This means that while the unrestricted model achieves a higher likelihood (necessarily), the data are not fit enough better to justify the use of so many additional free parameters; thus based on the calibration dataset alone, we would have a Bayes factor $K > 1$, meaning an increase in the relative

[40] Note that the scale-invariant model and unrestricted alternative referred to in Table 1 do not correspond precisely to the predictions shown in Figures 3 and 4, since in Figures 3 and 4 the parameters of both models are fit to our entire dataset, while in Table 1 the parameters are estimated using only the calibration dataset. The corresponding statistics for the models plotted in Figures 3 and 4 are given in the online appendix.

[41] In the case of the pooled data, the individual probit models and their associated statistics are described in the online appendix.

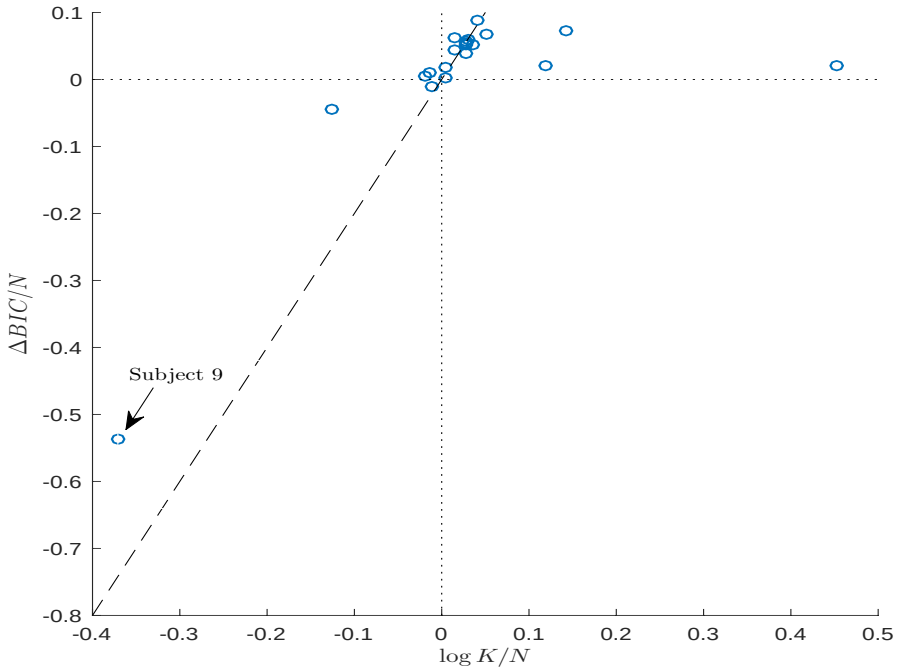


Figure 5: In-sample and out-of-sample model comparison statistics, for each of the 20 individual subjects, when separate parameters are estimated for each subject. (See explanation in text.)

posterior probability of the scale-invariant model (compared to whatever relative probability was assigned to that model in one's prior). When we then consider out-of-sample fit of the two models, if we assume a common set of parameters for all 20 subjects, the out-of-sample fit is slightly better for the unrestricted model. However, the fit is only modestly better, and when one takes into account both the in-sample and out-of-sample fit of the two models, we obtain an overall Bayes factor $K > 400,000$, greatly increasing the relatively probability of the scale-invariant model.

Moreover, the slight inferiority of the scale-invariant model with regard to out-of-sample fit is due primarily to the data for a single subject (subject 9), whose choice curves do not satisfy scale-invariance. If we fit a single set of parameters to the pooled data for all subjects except subject 9, the scale-invariant model fits better both in-sample and out-of-sample, and the overall Bayes factor would be greater than 700,000,000. If we instead fit separate parameters for each subject, then as shown in the bottom panel of Table 1, the aggregate evidence provides more support for the scale-invariant model both in-sample and out-of-sample, even when the data for subject 9 are included in the sample. In this case, the overall Bayes factor is greater than 10^{59}. Thus if we assume that either the scale-invariant model or the more flexible alternative must be the correct model for all subjects (though the parameters may differ across subjects), the evidence is overwhelming in favor of the scale-invariance hypothesis.[42]

In fact, the scale-invariant model fits reasonably well for most of the subjects considered

[42]There is nonetheless even stronger evidence in favor of the more complex hypothesis that choice behavior is scale-invariant for all subjects but subject 9. See the online appendix for details.

individually. Figure 5 shows a scatter plot of the values of the BIC difference, and the overall Bayes factor K, for each individual subject, when separate choice curves are estimated for each subject.[43] Each open dot corresponds to one subject. A location above the horizontal axis indicates that the in-sample fit of the scale-invariant model is better (using the BIC statistics as the basis of the model comparison); a location to the right of the dashed line indicates that the out-of-sample fit of the scale-invariant model is better (higher $LL^{validation}$ for that model); and a location to the right of the vertical axis indicates that the overall Bayes factor favors the scale-invariant model ($K > 1$). While it is not true that $K > 1$ for each individual subject, there are only two subjects (subjects 9 and 14) for whom either the in-sample or out-of-sample model comparisons are too unfavorable to the scale-invariant model.[44]

Our data are nonetheless not perfectly scale-invariant, even when we consider only the pooled data. We see in Figure 4 that the estimated choice curve in the case $C = \$31.40$ is not a perfect horizontal translation of the others, but instead is somewhat flatter.[45] This may indicate inaccuracy of the assumption of a log-normal prior (2.3), used in our theoretical calculations above for convenience. Under the assumption of a log-normal prior, $\log \mathrm{E}[X|r_x]$ is a linearly increasing function of r_x, with constant slope β. But if people instead form correct inferences based on a prior under which monetary payments greater than \$50 are less likely than a log-normal prior would allow (as was actually the case in our experiment, since we never offered lotteries involving $X/C > 4$), then $\log \mathrm{E}[X|r_x]$ would increase less rapidly with further increases in r_x, for values of r_x above $\log 50$. (Under the prior, such large values of r_x would more likely result from mis-coding of a payment of less than \$50 than from a large value of X that has been correctly coded.) This would result in a frequent failure to recognize how attractive the risky lottery truly is when X exceeds \$50, and hence less frequent acceptance of the risky lottery in such cases than the scale-invariant model would predict, as can be observed in Figure 3. (We leave for future work more detailed consideration of the extent to which our data may be better explained by a more subtle account of subjects' prior beliefs.)

Holt and Laury (2002) also obtain nearly perfect scale-invariant choice curves (see their Figure 1), when the amounts offered in hypothetical gambles are scaled up by a factor as large as 90 times those used in their small-stakes gambles. They find, however, that their subjects' apparent degree of risk aversion increases when the scale of the gambles is increased, in the case of gambles for real money (their Figure 2). It is unclear whether this difference from our results (which also involve real money) reflects a difference in the kind of gambles

[43] The vertical axis plots the amount by which the BIC statistic for the unrestricted model is greater than the one for the scale-invariant model (ΔBIC), divided by N, the number of trials for that subject. (Because N is not the same for all subjects, the values scaled by N are more comparable across subjects.) The horizontal axis plots the value of $\log K$, again divided by N. The dashed line identifies points at which $\log K = (1/2)\Delta BIC$, which is to say, points at which there is no difference in $LL^{valuation}$ between the two models. Points to the right of the dashed line are thus those for which $LL^{valuation}$ is higher for the scale-invariant model than for the unrestricted model.

[44] These are the two dots in Figure 5 that are both well below the horizontal axis and well above the diagonal dashed line.

[45] Note however that this curve is also less well estimated than the others shown in Figure 3, as a number of our subjects were not presented with trials including values of C this large, so that the N_{obs} for this case is smaller, as indicated in Table 3 in the online appendix.

Model	$LL^{calibration}$	BIC	$LL^{validation}$	log K
Pooled Data				
Log coding	-2838.6	5694.6	-914.2	0.0
ARUM-Probit	-3027.1	6071.7	-972.1	246.4
ARUM-Logit	-3000.1	6017.7	-967.0	214.3
Heterogeneous Parameters				
Log coding	-1860.2	3946.1	-663.9	0.0
ARUM-Probit	-1935.3	4096.4	-922.5	333.7
ARUM-Logit	-1889.4	4004.5	-721.5	86.8

Table 2: In-sample and out-of-sample measures of goodness of fit for three models: our logarithmic coding model and two additive random-utility models. The format is the same as in Table 1. (See text for further explanation.)

presented to their subjects, or the fact that their large gambles involved greater amounts of money than even our largest gambles (hundreds of dollars rather than mere tens of dollars).[46] Further studies would be desirable to clarify this.

3.2 Comparison with Random Expected Utility Models

As noted in the introduction, both the random variation in subjects' choices between simple gambles and existence of small-stakes risk aversion are often explained, in the experimental economics literature, by positing (i) "narrow bracketing" of the choice problem, so that the small amounts that can be gained in the experiment are not integrated with the subject's overall wealth (or overall lifetime budget constraint), (ii) significant concavity of the utility function that is used to value different possible monetary gains in the experiment, and (iii) a random term in the utility function, so that the expected utility assigned to a given probability distribution over possible gains is not always the same. We have offered an alternative model of both the randomness and the degree of apparent risk aversion in the choices of our subjects that we regard as more theoretically parsimonious, and in our view this theoretical parsimony should be a reason to prefer our interpretation, even if the competing views were equally consistent with the data from a single experiment such as this one. Nonetheless, it is interesting to ask whether our data could not be equally well explained by a more familiar model.

Table 2 compares the fit of our model with two variants of an additive random-utility model. In the case of each of the ARUMs, the subject is assumed to choose the option for which $\mathrm{E}[u(Y)] + \epsilon$ is larger, where Y is the monetary amount gained from the experiment (a random quantity, in the case of a risky prospect), $u(Y)$ is a nonlinear utility function for such gains (valued separately from the subject's other wealth), and ϵ is a random term (drawn at the time of choice) that is independent of the characteristics of the option, and also independent of the corresponding random term in the value assigned to the other option. In the ARUMs considered in the table, $u(Y)$ is assumed to be of the CRRA form, $u(Y) =$

[46]Note that it is perfectly consistent with our model to suppose that diminishing marginal utility of wealth becomes an additional source of risk aversion in the case of large gambles.

$Y^{1-\gamma}/(1-\gamma)$, for some $\gamma \geq 0$.[47] The random term ϵ is assumed either to be normally distributed (the ARUM-Probit model), or to have an extreme-value distribution (the ARUM-Logit model). Thus each of the ARUMs has two free parameters (the coefficient of relative risk aversion γ and the standard deviation of ϵ), like the logarithmic coding model. The ARUMs can also be considered random variants of prospect theory, in which $u(Y)$ is the Kahneman-Tversky value function for gains,[48] but we use the true probabilities of the two outcomes as weights rather than distorted weights of the kind posited by Kahneman and Tversky (1979).[49]

As in the case of Table 1, we consider both in-sample and out-of-sample measures of model fit, where the calibration dataset and validation dataset are the same as in the earlier table. In each case, we find that our model based on logarithmic coding provides a better fit to the experimental data, both in-sample and out-of-sample. The alternative model which comes closest to being a competitor is the ARUM-logit model, when separate parameters are estimated for each subject. Yet even in this case, the implied Bayes factor $K > 10^{37}$. If one of the models considered must represent the correct statistical model of our data, then the evidence overwhelmingly favors the model based on logarithmic coding.[50]

3.3 Heterogeneity in the Precision of Mental Coding

We have shown above (Tables 1 and 2) that our data are better fit by allowing the parameters of the scale-invariant psychometric function to vary across subjects. Here we provide further information about the heterogeneity in the parameters that best fit the behavior of each subject. We focus solely on the scale-invariant model, which as argued above is best supported by our data, and estimate a scale-invariant choice curve for each of the 19 subjects other than subject 9. We omit subject 9 from the discussion in this section, since as shown in Figure 5, this subject's data are not well-described by a scale-invariant model. Our theoretical model implies not only that a scale-invariant curve (3.2) should describe each subject's data, but also that the coefficients for each subject should satisfy the inequalities

$$\gamma \geq 0, \qquad -\frac{\delta}{\gamma} \geq \log p^{-1}, \tag{3.3}$$

which are required in order for there to exist values of σ^2 and ν^2 consistent with those coefficients. Hence in estimating the subject-specific models, we impose the further restriction that the value of γ be non-negative.[51]

[47] In the online appendix, we present results for ARUMs in which $u(Y)$ is allowed to be of the more general HARA form. Allowing for this generalization does not improve the fit of the ARUMs, once the penalty for the additional free parameter is taken into account.

[48] Note that the isoelastic functional form used here for $u(Y)$ is also commonly used in quantitative implementations of prospect theory, following Tversky and Kahneman (1992).

[49] In the online appendix, we show that allowing for a probability weight different from the true probability does not improve the fit of the random version of prospect theory, once the penalty for the additional free parameter is taken into account.

[50] In the online appendix, we also compare the fit of random versions of the model proposed by Bordalo *et al.* (2012), in which risk attitudes result from differences in the salience of alternative outcomes. These models fit our data even less well than do the ARUMs, or random versions of prospect theory.

[51] For all but one subject, the estimated value of γ would be positive, even without imposing the restriction, as is also true when we estimate a scale-invariant psychometric function using the pooled data. Even in the

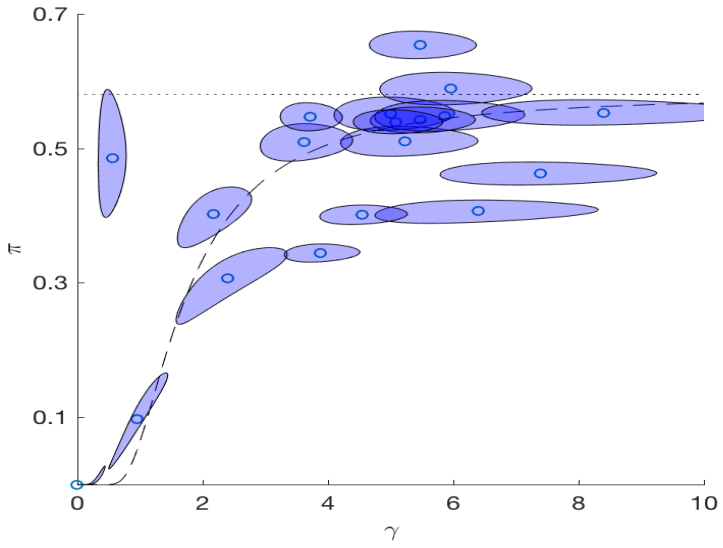


Figure 6: Heterogeneity in subjects' choice curves. Each shaded region indicates the credible region for an individual subject's parameters γ and π, with an open circle at that subject's maximum-likelihood values. The dashed line shows the theoretical relationship between γ and π that should exist if all subjects share a common prior, under which $\sigma = 0.35$.

In comparing the choice curves of the different subjects, it is useful to parameterize them not by γ and δ, but instead by the values of γ and

$$\pi \equiv e^{\delta/\gamma}.$$

In terms of this alternative parameterization, the theoretical constraints (3.3) can be written:

$$\gamma \geq 0, \qquad \pi \leq p. \tag{3.4}$$

Note that when $\delta/\gamma \leq 0$ (as required by our theoretical model), $0 \leq \pi \leq 1$, and π has the interpretation of a "risk-neutral probability": the subject's indifference point is the same as that of a risk-neutral, optimizing decision maker who believes that the probability of the non-zero lottery payoff is π (rather than the true probability p). The prediction that $\pi \leq p$ is another way of saying that our theoretical model predicts apparent risk aversion; and the degree to which a subject's estimated π is less than p provides a simple measure of the degree of apparent risk aversion.

Figure 6 shows the estimated values of γ and π for each of the 19 subjects for whom it is not grossly inaccurate to estimate a scale-invariant choice curve (now imposing the theoretical constraint that $\gamma \geq 0$). For each subject, an open circle indicates the parameter values that maximize the likelihood of the data for that subject (the ML estimate), and the surrounding shaded region indicates the set of parameter values for which the log likelihood

case of the subject for whom the best-fitting parameter values involve $\gamma < 0$, the value of γ is only slightly negative; and since a likelihood of acceptance of the risky lottery that is genuinely decreasing in X would be difficult to interpret, we presume that this represents sampling error, and treat this subject as having a γ of zero.

of the data is no more than 2 points lower than at the maximum. Thus the shaded region indicates a Bayesian credible region for the parameter estimates, under the assumption of a uniform prior for those parameters.[52]

The largest value of π that would be consistent with prediction (3.4) is indicated by the horizontal dotted line in Figure 6; we see that for all but one of the 19 subjects, the credible region includes points consistent with this prediction. Thus the individual choice curves of 18 out of our 20 subjects are reasonably consistent with both the model prediction of scale invariance and with the coefficient constraints (3.4).

Accounting for the choice curves of all subjects in this way, however, requires us to allow different subjects to have different priors (more specifically, different values for σ^2). If instead we assume a common log-normal prior (2.3) for all subjects, but allow the precision of mental coding of monetary amounts (i.e., the parameter ν^2) to vary across subjects, then the values of γ and π estimated for each subject are predicted by the model to be linked by a relationship of the form

$$\pi \;=\; p^{1+(2\sigma^2\gamma^2)^{-1}}, \tag{3.5}$$

where σ^2 is a parameter common to all subjects.[53] This is an upward-sloping relationship, of the kind illustrated by the dashed curve in Figure 6, which graphs equation (3.5) in the case that $\sigma = 0.35$. Here ν^2 is decreasing (the precision of mental coding is increasing) as one moves up and to the right along the dashed curve.

If we estimate a choice curve for each subject without imposing the restriction of a common σ^2, the estimated coefficients do not all line up perfectly on a curve consistent with (3.5); nonetheless, there is a strong positive correlation between the ML estimates of γ and π for the various subjects, as can be seen in Figure 6. That is, the degree of apparent risk aversion (measured by the degree to which π is less than p) is generally greater for those subjects whose choices are less sensitive to variation in X/C (measured by the size of γ). The fact that these two features of behavior go hand in hand is consistent with our theory, which attributes both to greater imprecision in the mental coding of monetary payoffs (a larger value of ν^2). Models such as EUM or prospect theory, extended to allow for stochastic choice as in section 3.2, instead provide no reason to expect such a relationship, since in these theories the degree of randomness of choice and the degree of risk aversion are determined by independent parameters.

4 Discussion

We have shown that it is possible to give a single unified explanation for the observed randomness in choices by subjects evaluating risky income prospects on the one hand, and the apparent risk aversion that they display on average on the other, as natural consequences

[52]The boundary of each maximum-posterior density credible region is chosen according to a criterion which, in the case of a Gaussian posterior for a single variable, would report the interval corresponding to the mean estimate plus or minus two standard deviations.

[53]Equation (3.5) can be derived by using (1.3) and (2.5) to obtain an equation for γ as a function of σ^2 and ν^2; inverting this to obtain the value of ν^2 implied by any subject's value for γ; and then using this result to eliminate ν^2 from the model prediction for the value of π.

of people's intuitions about the value of gambles being based on imprecise internal representations of the monetary amounts that are offered. Our theory explains the possibility of small-stakes risk aversion without implying any extraordinary degree of aversion to larger gambles in other contexts. Moreover, it can also explain the fact (demonstrated in our experiment) that the degree of risk aversion, as measured by the percentage by which the expected value of a random payoff must exceed the certain payoff in order for a subject to be indifferent between them, is relatively independent of the size of the stakes (as long as these remain small), contrary to what should be found if risk aversion were due to diminishing marginal utility.

4.1 Further Implications of the Theory

The "reflection effect." Our model of noisy mental coding of monetary amounts can also account for further anomalous features of subjects' choices with regard to small gambles documented by Kahneman and Tversky (1979). For example, Kahneman and Tversky report that if subjects must choose between a risky loss and a certain loss — with similar probabilities and monetary quantities as in the kind of problem considered above, but with the *signs* of the monetary payoffs reversed — risk *seeking* is observed more often than risk aversion (something they call the "reflection effect"). The coexistence of both risk-averse choices and risk-seeking choices by the same subject, depending on the nature of the small gambles that are offered, is a particular puzzle for the EUM account of risk attitudes, since a subject should be either risk averse or risk seeking (depending whether the subject's utility of wealth is concave or convex) regardless of the sign of the gambles offered.

The explanation of risk aversion for small gambles offered here instead naturally implies that the sign of the bias (i.e., of the apparent risk attitude) should switch if the signs of the monetary payoffs are switched. Consider instead the case of a choice between a risky gamble that offers a probability p of losing an amount X (but losing nothing otherwise), and the option of a certain loss of an amount C. If we assume that the quantities X and C are mentally represented according to the same logarithmic coding model as above,[54] regardless of whether they represent gains or losses, then in the case of losses, the subject's expected wealth is maximized by a rule under which the risky lottery is chosen if and only if

$$p \cdot \mathrm{E}[X|r_x] \;<\; \mathrm{E}[C|r_c], \tag{4.1}$$

reversing the sign in (2.2).

The set of internal representations (r_x, r_c) for which this holds will be the complement of the set discussed earlier, so that the model predicts

$$\text{Prob[accept risky}|X, C] \;=\; \Phi\left(\frac{\beta^{-1}\log p^{-1} - \log X/C}{\sqrt{2}\nu}\right). \tag{4.2}$$

[54]Note that we assume that the absolute value of each of the monetary payoffs is coded, rather than a signed magnitude. This is in accordance with the model of approximate numerical cognition proposed by authors such as Dehaene (2008), which assumes that all numerical quantities are coded as positive amounts, making use of brain circuits originally developed to represent information about the numerosity of sets of items in one's environment. The information whether the quantity in question represents a monetary gain or loss must also be represented, but is assumed to be coded separately, and without error.

Indifference again will require $pX > C$, but this will now count as *risk-seeking* behavior; when $pX = C$, the risky loss should be chosen more often than not.

A framing effect. Kahneman and Tversky (1979) further show that subjects' preferences between a risky and a safe outcome can be flipped, depending whether the options are presented as involving gains or losses. In one of their problems, subjects are asked to imagine being given a substantial monetary amount $2M$,[55] and then being presented with a choice between (a) winning an additional M with certainty, or (b) a gamble with a 50 percent chance of winning another $2M$ and a 50 percent chance of winning nothing. In a second problem, the initial amount was instead $4M$, and the subsequent choice was between (a) losing M with certainty, and (b) a gamble with a 50 percent chance of losing $2M$ and a 50 percent chance of losing nothing.

These two problems are equivalent, in the sense that in each case the subject chooses between (a) ending up with $3M$ more than their initial wealth with certainty, or (b) a gamble under which they have an equal chance of ending up with $2M$ or $4M$ more than their initial wealth. Nonetheless, a substantial majority of their subjects chose (a) in the first problem, while a substantial majority chose (b) in the second. This contradicts any theory (not just EUM) under which people should have a consistent preference ranking of probability distributions over final wealth levels.

Our theory easily explains this finding. If the initial gift is denoted G, and the monetary amounts G, X, and C defining the decision problem must each be independently represented in the fashion postulated above, then in the first problem, an expected wealth-maximizing decision rule will choose (b) if and only if

$$\mathrm{E}[G|r_g] \,+\, p \cdot \mathrm{E}[X|r_x] \;>\; \mathrm{E}[G|r_g] \,+\, \mathrm{E}[C|r_c],$$

which is equivalent to (2.2), while in the second problem it will choose (b) if and only if

$$\mathrm{E}[G|r_g] \,-\, p \cdot \mathrm{E}[X|r_x] \;>\; \mathrm{E}[G|r_g] \,-\, \mathrm{E}[C|r_c],$$

which is equivalent to (4.1). We then get different probabilities of choosing (b) in the two cases, given by equations (2.5) and (4.2) respectively.

Note that our theory assumes that the decision rule is in all cases the one that maximizes expected final wealth, so that only the sum of the initial gift and the additional gain or loss from the further option is assumed to matter to the decision maker; there is no intrinsic interest assumed in gains or losses relative to what one had in the past or what one expected to have. The relevance of the sequence of gains and losses by which one arrives at a given final wealth comes not from the decision maker's assumed objective, but from the need to mentally represent the quantities used in the description of the options, in the form in which they are presented, before integrating the separate pieces of information in further calculations. If this representation were possible with infinite precision (and subsequent operations could also be perfectly precise), then different ways of presenting information that imply the same possible final wealth levels would indeed be equivalent, and lead to the same choices. But when the precision with which each monetary amount can be represented

[55] In their experiment, conducted in the 1970s, $2M$ was equal to 1000 Israeli shekels, a substantial fraction of a typical monthly income at the time.

is limited, mathematically equivalent problems are not processed in identical ways, and the resulting behavior can be different as a result, despite its optimality in each case (conditional on the mental representation).

Imprecise representation of probabilities. Kahneman and Tversky (1992) document other respects in which subjects make both risk-averse choices and risk-seeking choices with respect to small gambles, depending of the nature of the problem. For example, their subjects are more often risk-seeking when choosing between a small certain gain and a small probability of a considerably larger gain; but they are more often risk-averse when choosing between a modest certain loss and a small probability of a considerably larger loss. Prospect theory explains such cases by postulating that in the case of a risky prospect, the values assigned to the various possible gains or losses are weighted not in proportion to each outcome's probability of occurrence (as required by EUM), but rather using weights that represent a nonlinear transformation of the true probabilities.

Choices of this kind are consistent with our model, if the model is generalized to assume (in the case of simple gambles of the kind discussed above) that the probability p of the non-zero outcome also has an internal representation r_p that is probabilistic. In the analysis above, we have assumed for simplicity that the exact value of p is available as an input to the decision rule. This is not inconsistent with our general view of the internal representation of numerical information; for in the situation considered in our theoretical model (and in our experiment), there is a single value of p that is used in all trials (though X and C vary from trial to trial). Thus if we assume that the prior for which the subject's decision rule has been optimized represents the distribution of possible decision situations *in this specific experiment,* the prior (in this type of experiment) would allow for only a single value of p — and the Bayesian posterior would similarly have this single value of p in its support, regardless of the noisy representation r_p.

Nonetheless, it is clearly of interest to consider the case in which the prior is non-degenerate (either because p varies from trial to trial, or because the decision maker has not had enough experience with a particular context for her decision rule to be adapted to the precise statistics of that context). Let us again assume for simplicity that under the prior, the quantities p, X and C are distributed independently of one another; that the distribution of the representation r_p depends only on p (not on the values of X or C), and that the conditional distribution of r_p is independent of the realizations of r_x or r_c; and similarly for the other components of the internal representation $\mathbf{r}$. Then condition (2.2) for an optimal decision rule takes the more general form

$$\mathrm{E}[p|r_p] \cdot \mathrm{E}[X|r_x] \;>\; \mathrm{E}[C|r_c],$$

or alternatively (given the model proposed above for the noisy coding of monetary amounts),

$$r_x - r_c \;>\; \beta^{-1}\rho, \qquad \rho \;\equiv\; -\log \mathrm{E}[p|r_p], \tag{4.3}$$

generalizing (2.4). Here ρ is a random variable (because it depends on the realization of r_p), conditional on the true probability p.

This generalization of the decision rule (2.2) results in an additional source of *randomness* in choice (namely, random variation in ρ); but in general, it also results in an additional

source of *bias* (because ρ also differs from $\log p$ on average). As a simple example, suppose that the noise in the internal coding of X and C is negligible (ν is extremely small), but that the internal representation r_p is drawn from a distribution

$$r_p \;\sim\; N(z(p), \nu_p^2),$$

where $z(p) \equiv \log(p/1-p)$ are the log odds of the two outcomes, and ν_p is non-negligible. Suppose furthermore that the log odds are normally distributed under the prior,

$$z(p) \;\sim\; N(\mu_p, \sigma_p^2)$$

. Then the posterior log odds, conditional on the representation r_p, are also normally distributed, with a mean $\bar{m}(r_p)$ that is a weighted average of r_p and the prior mean log odds, and a variance $\bar{\sigma}^2$ that is independent of r_p.

In this case, (4.3) requires that the value of $\log(X/C)$ required for indifference (acceptance of the gamble exactly half the time) will be equal to the median value of ρ, which (given that r_p has a symmetric distribution) is the value of ρ when r_p is equal to the true log odds. Alternatively, the value of the ratio C/X required for indifference is given by a function $w(p)$, where

$$w(p) \;\equiv\; \mathrm{E}[F(z(p), \epsilon)], \qquad F(z, \epsilon) \;\equiv\; \frac{\exp[\bar{m}(z) + \epsilon]}{1 + \exp[\bar{m}(z) + \epsilon]},$$

and ϵ is a random variable distributed $N(0,\, \bar{\sigma}^2)$. This function plays a role similar to the probability weighting function of Kahneman and Tversky (1979). And as long as the variance $\bar{\sigma}^2$ is not too great, the model implies that $w(p)$ will have the inverse-S shape assumed by Kahneman and Tversky.

In particular, if we fix the ratio ν_p/σ_p but make both σ_p and ν_p small, then in the limit as $\sigma_p, \nu_p \to 0$, we obtain an analytical solution of the form[56]

$$w(p) \;=\; \frac{\alpha p^\beta}{(1-p)^\beta + \alpha p^\beta}, \tag{4.4}$$

for certain coefficients $\alpha > 0, 0 < \beta < 1$, which depend on the ratio ν_p/σ_p and the prior mean log odds μ_p. This function has the inverse-S shape assumed by Kahneman and Tversky; indeed, the two-parameter family of weighting functions (4.4) has often been assumed in econometric implementations of prospect theory.[57] In this case, the model predicts risk-seeking in the case of gains for low values of p, but risk-aversion for larger values of p, and risk-aversion in the case of losses for low values of p, but risk-seeking for larger values of p, all as found by Kahneman and Tversky (1979). We leave further analysis of this extension of our model for a future study.

Effects of varying cognitive load. Thus far, we have discussed implications of our model, taking the precision of coding (parameterized by ν) to be fixed. But the model also makes predictions about the effects of varying ν, which might be subject to predictable variation for a variety of reasons. For example, one might well suppose that increased time pressure,

[56] See the online appendix for details of the calculation.

[57] See Stott (2006) for a review.

distraction or cognitive load should reduce the cognitive resources used to represent the monetary magnitudes that define a particular decision problem, and that this should correspond, in our model, to an increase in ν. According to our model, this should result in both decreased sensitivity of a subject's decisions to variations in the risky payoff X that is offered (i.e., a lower value of γ) and increased apparent risk aversion (a value of π that is lower relative to p).[58] This is an example of a prediction of our theory that is not made by theories like prospect theory, that attribute departures from the predictions of EUM to (presumably stable) distortions in the way that subjects evaluate monetary gains or probabilities.

In fact, a number of authors have found that increasing cognitive load (for example, by requiring subjects to concurrently maintain a list of random letters or numbers in memory) causes subjects to make more risk-averse choices (Whitney *et al.,* 2008; Benjamin *et al.,* 2013; Deck and Jahedi, 2015; Gerhardt *et al.,* 2016).[59] This is often interpreted as support for a "dual systems" view of decision making, in which increased cognitive load makes it harder for subjects to employ a deliberative system that would be called upon under other circumstances, so that emotional reactions or simpler heuristics are relied upon instead. Our theory provides an alternative interpretation, in which the same cognitive mechanism might be employed in both cases, but it relies upon an imprecise analog representation with a degree of precision that depends on the number of other claims on working memory. The fact that our subjects display a range of degrees of apparent risk aversion, as well as a range of degrees of randomness in their choices, as shown in Figure 6 — rather than simply two clusters corresponding to the users of two very different mental systems — is more easily explained under the theory that we propose.

4.2 Comparison with Related Theories

Schley and Peters (2014) offer an explanation for apparent risk aversion which is based on the idea that the perception of the numerical magnitudes of prospective monetary payoffs is biased, and more specifically that perceived magnitudes are an increasing, strictly concave function of the magnitudes. Like us, they base their proposal on limitations on people's general ability to accurate represent numbers mentally, rather than on the true utility obtained from having more money (as in the EUM explanation of risk aversion) or a theory of distorted valuations that is specific to the domain of value-based decision making (as with prospect theory). In support of this proposal, they show that subjects who less accurately represent numbers for other purposes also exhibit greater apparent diminishing marginal utility of income and greater apparent risk-aversion in choices between risky gambles.[60]

[58] Recall that the dashed curve in Figure 6 shows the effect on both γ and π of varying ν, while holding fixed the prior distribution over possible values of X and C.

[59] Olschewski *et al.* (2018) instead find that increased cognitive load increases the randomness of choice, but only increases risk aversion by a small (statistically insignificant) amount. Their analysis of the effect on risk aversion, however, is based on the estimated coefficients of a structural model of the "ARUM-probit" type discussed in section 3.2. As we note there, this specification is not consistent with the predictions of our model, and fits our experimental data less well. It would be interesting to examine the question further within a broader class of stochastic choice models.

[60] More precisely, they fit each of their subjects' choices to a prospect-theoretic valuation formula, where the value function for either gains or losses is assumed to be of the power-law form (1.4), and estimate a value of the exponent β for each subject. They find that subjects who score higher on a test of ability to

However, their discussion assumes that less capacity for symbolic number mapping results in a deterministic distortion of perceived numerical magnitudes (a true quantity X is always perceived as exactly $\hat{X} = AX^{\beta}$), rather than in a more random mental representation as in our theory. This means that they do not explore the connection between the randomness of subjects' choices and apparent risk aversion, as we do here; and their theory provides no explanation for why people should value lotteries according to the average value of the perceived payoffs $\hat{X}_i$ instead of, say, according to the average value of $\hat{X}_i^{1/\beta}$ — a criterion that would reliably maximize expected wealth, taking into account the perceptual distortion.

A theory more similar to ours is the model of risk-sensitive foraging by animals (such as starlings) proposed by Kacelnik and Abreu (1998). These authors are concerned with how animals choose between options that they repeatedly face (alternative possible locations for foraging), on the basis of past experience of the probability distribution of possible outcomes associated with each, and their theory accordingly turns on the way that previously experienced outcomes are represented in memory, and the way in which the distribution represented in memory is drawn upon at the time of a new prospective choice; it is not a theory of the representation of numerical descriptions of available options (which are not available to foraging starlings). Moreover, the variability in the choices made across repeated presentations of the same options is attributed to variation in the random samples drawn on each occasion from a fixed mental representation of the distribution of possible outcomes under each option (with the situation being recognized by the organism as a repetition of the same situation as before), rather than randomness in new representations of the payoffs that are assumed in our theory to be formed each time the decision problem is presented again (and not recognized as having been previously encountered).

Nonetheless, the implications of their theory — which like ours is based on randomness in the representation of rewards that conforms to "Weber's Law," and derives predictions for both choice probabilities and apparent risk aversion — are similar to those of ours, while not mathematically identical.[61] The fact that a similar model can successfully explain animal behavior of the kind that Kacelnik and Abreu review provides further reason, in our view, to consider our proposed explanation for intuitive judgments by humans a plausible one.

Finally, the model of Woodford (2012) is similar to ours, in that it derives both risk aversion with respect to gains and risk seeking with respect to losses from a model of noisy coding of prospective net gains, with a decision rule that maximizes the subject's expected wealth. However, the model of noisy coding is different: net gain is coded as a single variable (which may be of either sign), rather than gains and losses being coded separately, and the assumed inhomogeneity in the precision of coding of net gains makes the mean Bayesian estimate of net gain an S-shaped function of the true net gain, rather than there being separate concave functions for the mean estimates of gains and losses as above.[62] We feel

accurately locate symbolically presented numbers on a spatial number line have values of β closer to 1.

[61] A key mathematical difference is that Kacelnik and Abreu do not model random coding in conformity with Weber's Law in the same way that we do; they assume a truncated normal distribution for the mental representations, with a standard deviation proportional to the mean, rather than a log-normal distribution.

[62] Other arguments for a subjective representation of net gain that is an S-shaped function of the actual net gain, as an optimal form of mental coding under a constraint on the feasible overall precision of cognitive representations, include those of Friedman (1989), Robson (2001), Rayo and Becker (2007), and Netzer (2009).

that the model of mental coding proposed here is more realistic, because in the experiments that we seek to explain, the prospective outcomes are described to subjects in terms of positive quantities of money that can be gained or lost, rather than in terms of a signed net gain. Whether the kind of inhomogeneity in the precision of coding relied upon here can be justified as an efficient use of finite processing resources, as in the model proposed in Woodford (2012), is an important topic for further investigation.

References

[1] Anobile, Giovanni, Guido Marco Cicchini, and David C. Burr, "Linear Mapping of Numbers onto Space Requires Attention," *Cognition* 122: 454-459 (2012).

[2] Arrow, Kenneth J., *Essays in the Theory of Risk-Bearing,* Chicago: Markham Publishing Co., 1971.

[3] Ballinger, T. Parker, and Nathaniel T. Wilcox, "Decisions, Error, and Heterogeneity," *Economic Journal* 107: 1090-1105 (1997).

[4] Banks, William P., Milton Fujii, and Fortunee Kayra-Stewart, "Semantic Congruity Effects in Comparative Judgments of Magnitudes of Digits," *Journal of Experimental Psychology: Human Perception and Performance* 2: 435-447 (1976).

[5] Becker, Gordon M., Morris H. Degroot, and Jacob Marschak, "Stochastic Models of Choice Behavior," *Systems Research and Behavioral Science* 8: 41-55 (1963).

[6] Becker, Gordon M., Morris H. Degroot, and Jacob Marschak, "Measuring Utility By a Single-Response Sequential Method," *Systems Research and Behavioral Science* 9: 226-232 (1964).

[7] Benjamin, Daniel J., Sebastian A. Brown, and Jesse M. Shapiro, "Who is 'Behavioral'? Cognitive Ability and Anomalous Preferences," *Journal of the European Economics Association* 11: 1231-1255 (2013).

[8] Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, "Salience Theory of Choice Under Risk," *Quarterly Journal of Economics* 127: 1243-1285 (2012).

[9] Brannon, Elizabeth M., "The Representation of Numerical Magnitude," *Current Opinion in Neurobiology* 16: 222-229 (2006).

[10] Buckley, Paul B., and Clifford B. Gillman, "Comparisons of Digits and Dot Patterns," *Journal of Experimental Psychology* 103: 1131-1136 (1974).

[11] Burnham, Kenneth P., and Anderson, David R., *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach,* 2d. ed., New York: Springer, 2002.

[12] Cantlon, Jessica F., and Elizabeth M. Brannon, "Shared System for Ordering Small and Large Numbers in Monkeys and Humans," *Psychological Science* 17: 401-406 (2006).

[13] Cordes, Sara, Rochel Gelman, Charles R. Gallistel, and John Whalen, "Variability Signatures Distinguish Verbal from Non-Verbal Counting for Both Large and Small Numbers," *Psychonomic Bulletin and Review* 8: 698-707 (2001).

[14] Cox, James C., Vjollca Sadiraj, Bodo Vogt, and Utteeyo Dasgupta, "Is There a Plausible Theory for Decision Under Risk? A Dual Calibration Critique," *Economic Theory* 54: 305-333 (2013).

[15] Deck, Cary, and Salar Jahedi, "The Effect of Cognitive Load on Economic Decision Making: A Survey and New Experiments," *European Economic Review* 78: 97-119 (2015).

[16] Dehaene, Stanislas, "Varieties of Numerical Abilities," *Cognition* 44: 1-42 (1992).

[17] Dehaene, Stanislas, "Symbols and Quantities in Parietal Cortex: Elements of a Mathematical Theory of Number Representation and Manipulation," in P. Haggard, Y. Rossetti, and M. Kawato, eds., *Sensorimotor Foundations of Higher Cognition,* Oxford: Oxford University Press, 2008.

[18] Dehaene, Stanislas, *The Number Sense,* revised and updated edition, Oxford: Oxford University Press, 2011.

[19] Dehaene, Stanislas, and Laurent Cohen, "Two Mental Calculation Systems: A Case Study of Severe Acalculia with Preserved Approximation," *Neuropsychologia* 29: 1045-1074 (1991).

[20] Dehaene, Stanislas, and J. Frederico Marques, "Cognitive Euroscience: Scalar Variability in Price Estimation and the Cognitive Consequences of Switching to the Euro," *Quarterly Journal of Experimental Psychology A* 55: 705-731 (2002).

[21] Dehaene, Stanislas, Manuela Piazza, Philippe Pinel, and Laurent Cohen, "Three Parietal Circuits for Number Processing," *Cognitive Neuropsychology* 20: 487-506 (2003).

[22] Friedman, Daniel, "The S-Shaped Value Function as a Constrained Optimum," *American Economic Review* 79: 1243-1248 (1989).

[23] Friedman, Daniel, R. Mark Isaac, Duncan James, and Shyam Sunder, *Risky Curves: On the Empirical Failure of Expected Utility,* London: Routledge, 2014.

[24] Gabaix, Xavier, and David Laibson, "Myopia and Discounting," NBER Working Paper no. 23254, March 2017.

[25] Gabbiani, Fabrizio, and Steven J. Cox, *Mathematics for Neuroscientists,* Amsterdam: Academic Press, 2010.

[26] Gerhardt, Holger, Guido P. Biele, Hauke R. Heekeren, and Harald Uhlig, "Cognitive Load Increases Risk Aversion," SFB 649 Discussion Paper no. 2016-011, Humboldt University Berlin, March 2016.

[27] Gescheider, George A., *Psychophysics: The Fundamentals,* 3d ed., Mahwah, NJ: Lawrence Erlbaum Associates, 1997.

[28] Glimcher, Paul W., *Foundations of Neuroeconomic Analysis,* Oxford: Oxford University Press, 2011.

[29] Green, David M., and John A. Swets, *Signal Detection Theory and Psychophysics,* New York: Wiley, 1966.

[30] Hey, John D., "Experimental Investigations of Errors in Decision Making under Risk," *European Economic Review* 39: 633-640 (1995).

[31] Hey, John D., "Does Repetition Improve Consistency?" *Experimental Economics* 4: 5-54 (2001).

[32] Hey, John D., and Chris Orme, "Investigating Parsimonious Generalizations of Expected Utility Theory using Experimental Data," *Econometrica* 62: 1291-1329 (1994).

[33] Hollingsworth, Walter H., J. Paul Simmons, Tammy R. Coates, and Henry A. Cross, "Perceived Numerosity as a Function of Array Number, Speed of Array Development, and Density of Array Items," *Bulletin of the Psychonomic Society* 29: 448-450 (1991).

[34] Holt, Charles A., and Susan K. Laury, "Risk Aversion and Incentive Effects," *American Economic Review* 92: 1644-1655 (2002).

[35] Indow, Tarow, and Masashi Ida, "Scaling of Dot Numerosity," *Perception and Psychophysics* 22: 265-276 (1977).

[36] Izard, Véronique, and Stanislas Dehaene, "Calibrating the Mental Number Line," *Cognition* 106: 1221-1247 (2008).

[37] Jevons, W, Stanley, "The Power of Numerical Discrimination," *Nature* 3: 281-282 (1871).

[38] Kacelnik, Alex, and Fausto Brito e Abreu, "Risky Choice and Weber's Law," *Journal of Theoretical Biology* 194: 289-298 (1998).

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

[40] Kaufman, E.L., M.W. Lord, T.W. Reese, and J. Volkmann, "The Discrimination of Visual Number," *American Journal of Psychology* 62: 498-525 (1949).

[41] Khaw, Mel W., Ziang Li, and Michael Woodford, "Risk Aversion as a Perceptual Bias," NBER Working Paper no. 23294, March 2017.

[42] Koszëgi, Botond, and Matthew Rabin, "A Model of Reference-Dependent Preferences," *Quarterly Journal of Economics* 121: 1133-1165 (2006).

[43] Koszëgi, Botond, and Matthew Rabin, "Reference-Dependent Risk Attitudes," *American Economic Review* 97: 1047-1073 (2007).

[44] Koszëgi, Botond, and Matthew Rabin, "Revealed Mistakes and Revealed Preferences," in A. Caplin and A. Schotter, eds., *The Foundations of Positive and Normative Economics,* Oxford: Oxford University Press, 2008.

[45] Kramer, Peter, Maria Grazia De Bono, and Marco Zorzi, "Numerosity Estimation in Visual Stimuli in the Absence of Luminance-Based Cues," *PLoS ONE* 6(2): e17378 (2011).

[46] Krueger, Lester E., "Perceived Numerosity," *Perception and Psychophysics* 11: 5-9 (1972).

[47] Krueger, Lester E., "Perceived Numerosity: A Comparison of Magnitude Production, Magnitude Estimation, and Discrimination Judgments," *Perception and Psychophysics* 35: 536-542 (1984).

[48] Loomes, Graham, "Modeling the Stochastic Component of Behaviour in Experiments: Some Issues for the Interpretation of Data," *Experimental Economics* 8: 301-323 (2005).

[49] Loomes, Graham, and R. Sugden, "Incorporating a Stochastic Element into Decision Theories," *European Economic Review* 39: 641-648 (1995).

[50] Luyckx, F., H. Nili, B. Spitzer, and C. Summerfield, "Nueral Structure Mapping in Human Probabilistic Reward Learning," *bioRXiv* preprint [doi: http://dx.doi.org/10.1101/366757], posted July 10, 2018.

[51] McFadden, Daniel, "Econometric Models of Probabilistic Choice," in C. Manski and D. McFadden, eds., *Structural Analysis of Discrete Data with Economic Applications,* Cambridge, MA: MIT Press, 1981.

[52] Mosteller, Frederick, and Philip Nogee, "An Experimental Measurement of Utility," *Journal of Political Economy* 59: 371-404 (1951).

[53] Moyer, Robert S., and Thomas K. Landauer, "Time Required for Judgements of Numerical Inequality," *Nature* 215: 1519-1520 (1967).

[54] Natenzon, Paulo, "Random Choice and Learning," working paper, Washington University, October 2017. (Forthcoming, *Journal of Political Economy.*)

[55] Netzer, Nick, "Evolution of Time Preferences and Attitudes Toward Risk," *American Economic Review* 99: 937-955 (2009).

[56] Nieder, Andreas, "Coding of Abstract Quantity by 'Number Neurons' in the Primate Brain," *Journal of Comparative Physiology A* 199: 1-16 (2013).

[57] Nieder, Andreas, and Stanislas Dehaene, "Representation of Number in the Brain," *Annual Review of Neuroscience* 32: 185-208 (2009).

[58] Nieder, Andreas, and Katharina Merten, "A Labeled-Line Code for Small and Large Numerosities in the Monkey Prefrontal Cortex," *Journal of Neuroscience* 27: 5986-5993 (2007).

[59] Olschewski, Sebastian, Jörg Rieskamp, and Benjamin Scheibehenne, "Taxing Cognitive Capacities Reduces Choice Consistency Rather than Preference: A Model-Based Test," *Journal of Experimental Psychology: General* 147: 462-484 (2018).

[60] Petzschner, Frederike H., Stefan Glasauer, and Klaas E. Stephan, "A Bayesian Perspective on Magnitude Estimation," *Trends in Cognitive Sciences* 19: 285-293 (2015).

[61] Piazza, Manuela, Véronique Izard, Philippe Pinel, Denis Le Bihan, and Stanislas Dehaene, "Tuning Curves for Approximate Numerosity in the Human Intraparietal Sulcus," *Neuron* 44: 547-555 (2004).

[62] Pope, Robin, Johannes Leitner, and Ulrike Leopold-Wildburger, *The Knowledge Ahead Approach to Risk: Theory and Experimental Evidence*, Berlin: Springer, 2007.

[63] Rabin, Matthew, "Risk Aversion and Expected-Utility Theory: A Calibration Theorem," *Econometrica* 68: 1281-1292 (2000).

[64] Rabin, Matthew, and Richard H. Thaler, "Anomalies: Risk Aversion," *Journal of Economic Perspectives* 15(1): 219-232 (2001).

[65] Rayo, Luis, and Gary S. Becker, "Evolutionary Efficiency and Happiness," *Journal of Political Economy* 115: 302-337 (2007).

[66] Robson, Arthur, "The Biological Basis of Economic Behavior," *Journal of Economic Literature* 39: 11-33 (2001).

[67] Ross, John, "Visual Discrimination of Number without Counting," *Perception* 32: 867-870 (2003).

[68] Schley, Dan R., and Ellen Peters, "Assessing 'Economic Value': Symbolic-Number Mappings Predict Risky and Riskless Valuations," *Psychological Science* 25: 753-761 (2014).

[69] Spitzer, Bernhard, Leonhard Waschke, and Christopher Summerfield, "Selective Overweighting of Larger Magnitudes During Noisy Numerical Comparison," *Nature Human Behavior* 1, art. 0145 (2017).

[70] Steiner, Jakub, and Colin Stewart, "Perceiving Prospects Properly," *American Economic Review* 106: 1601-1631 (2016).

[71] Stocker, Alan A., and Eero P. Simoncelli, "Noise Characteristics and Prior Expectations in Human Visual Speed Perception," *Nature Neuroscience* 9: 578-585 (2006).

[72] Stott, Henry P., "Cumulative Prospect Theory's Functional Menagerie," *Journal of Risk and Uncertainty* 32: 101-130 (2006).

[73] Teichmann, A. Lina, Tijl Grootswagers, Thomas Carlson, and Anina N. Rich, "Decoding Digits and Dice with Magnetoencephalography: Evidence for a Shared Representation of Magnitude," *bioRXiv* preprint [doi: http://dx.doi.org/10.1101/249342], posted January 23, 2018.

[74] Thompson, Richard F., Kathleen S. Mayers, Richard T. Robertson, and Charlotte J. Patterson, "Number Coding in Association Cortex of the Cat," *Science* 168: 271-273 (1970).

[75] Tversky, Amos, and Daniel Kahneman, "Advances in Prospect Theory: Cumulative Representation of Uncertainty," *Journal of Risk and Uncertainty* 5: 297-323 (1992).

[76] van Oeffelen, Michiel P., and Peter G. Vos, "A Probabilistic Model for the Discrimination of Visual Number," *Perception and Psychophysics* 32: 163-170 (1982).

[77] Wei, Xue-Xin, and Alan A. Stocker, "A Bayesian Observer Model Constrained by Efficient Coding Can Explain 'Anti-Bayesian' Percepts," *Nature Neuroscience* 18: 1509-1517 (2015).

[78] Wei, Xue-Xin, and Alan A. Stocker, "Lawful Relation Between Perceptual Bias and Discriminability," *Proceedings of the National Academy of Sciences USA* 114: 10244-10249 (2017).

[79] Whalen, J., Charles R. Gallistel, and Rochel Gelman, "Non-Verbal Counting in Humans: The Psychophysics of Number Representation," *Psychological Science* 10: 130-137 (1999).

[80] Whitney, Paul, Christa A. Rinehart, and John M. Hinson, "Framing Effects Under Cognitive Load," *Psychonomic Bulletin and Review* 15: 1179-1184 (2008).

[81] Wilcox, Nathaniel T., "Stochastic Models for Binary Discrete Choice Under Risk: A Critical Primer and Econometric Comparison," in J.C. Cox and G.W. Harrison, eds., *Research in Experimental Economics, vol. 12: Risk Aversion in Experiments,* Bingley, UK: Emerald Group Publishing, 2008.

[82] Woodford, Michael, "Prospect Theory as Efficient Perceptual Distortion," *American Economic Review* 102(3): 1-8 (2012).

ONLINE APPENDIX

Khaw, Li, and Woodford, "Cognitive Imprecision and Small-Stakes Risk Aversion"

A A Bayesian Model of Numerosity Estimation

Suppose that a stimulus of numerosity n results in an internal representation r that is drawn from a distribution

$$r \sim N(\log n, \nu^2),$$

where the noise parameter ν is independent of n. Then if the prior distribution from which n is drawn is approximated by a log-normal distribution,

$$\log n \sim N(\mu, \sigma^2),$$

as proposed in the text (section 1.1), the pair of random variables $(\log n, r)$ have a joint distribution of the *bivariate Gaussian* family. It follows from this that the distribution of $\log n$ conditional on the value of r will be a Gaussian distribution,

$$\log n|r \sim N(\mu_{post}(r), \sigma^2_{post}), \tag{A.1}$$

where the mean $\mu_{post}(r)$ is an affine function of r, and the variance σ^2_{post} is the same for all r. This will give the *posterior* distribution for $\log n$ (and hence a posterior distribution for n) that is implied by Bayes' Rule, starting from the Gaussian prior for $\log n$ and updating on the basis of the noisy evidence r. Since the posterior distribution for $\log n$ is normal, the posterior distribution for n is log-normal, as stated in the text.

It further follows from the properties of a bivariate Gaussian distribution that the conditional mean (mean of the posterior distribution) of $\log n$ is given by the linear projection

$$\mu_{post}(r) = \mathrm{E}[\log n|r] = \mu + \beta \cdot (r - \mu), \tag{A.2}$$

where μ is the unconditional mean of both $\log n$ and r, and the slope coefficient (linear regression coefficient) is given by

$$\beta \equiv \frac{\mathrm{cov}(\log n, r)}{\mathrm{var}(r)} = \frac{\sigma^2}{\sigma^2 + \nu^2}, \tag{A.3}$$

as stated in the text at (1.3). The conditional variance of $\log n$ is then given by

$$\begin{aligned} \sigma^2_{post} &= \mathrm{var}(\log n - \mu_{post}(r)) = \mathrm{var}(\log n - \beta r) \\ &= \mathrm{var}((1-\beta)\log n) + \mathrm{var}(\beta r|n) = (1-\beta)^2\sigma^2 + \beta^2\nu^2 \\ &= \frac{\nu^4\sigma^2}{(\sigma^2+\nu^2)^2} + \frac{\sigma^4\nu^2}{(\sigma^2+\nu^2)^2} = \frac{\sigma^2\nu^2}{\sigma^2+\nu^2}. \end{aligned} \tag{A.4}$$

The predicted distribution of numerosity estimates then depends on how we assume that the subject's estimate of the stimulus numerosity relates to the posterior distribution over

possible numerosities implied by the internal representation r. Consider first the hypothesis that the subject's numerosity estimate $\hat{n}$ is optimal, in the sense of minimizing the mean squared estimation error, $MSE \equiv \mathrm{E}[(n-\hat{n})^2]$, among all possible estimation rules under which $\hat{n}$ is some function of r. The rule that is optimal from the standpoint of this objective would be the one under which $\hat{n}(r) = \mathrm{E}[n|r]$ for all r.[63]

It follows from the properties of a log-normal distribution that if the posterior distribution for n is given by (A.1), the posterior mean will be given by

$$\mathrm{E}[n|r] \;=\; \exp(\mu_{post} + (1/2)\sigma^2_{post}).$$

Hence in this case, the Bayesian model predicts that

$$\begin{aligned} \log \hat{n}(r) \;&=\; \log \mathrm{E}[n|r] \;=\; \mu_{post}(r) + (1/2)\sigma^2_{post} \\ &=\; \mu \;+\; \beta \cdot (r-\mu) \;+\; (1/2)\sigma^2_{post}. \end{aligned}$$

Thus as stated in the text, $\log \hat{n}(r)$ is predicted to be an affine function of r with slope β.

Since r is a random variable, conditional on the numerosity n of the stimulus, it follows that $\hat{n}(r)$ is also a random variable conditional on n. More specifically, since r is normally distributed, conditional on n, and $\log \hat{n}(r)$ is an affine function of r, $\log \hat{n}$ will be normally distributed conditional on n:

$$\log \hat{n} \;\sim\; N(\hat{\mu}(n), \hat{\sigma}^2), \tag{A.5}$$

as stated in the text. The mean and variance of this conditional distribution are given by

$$\begin{aligned} \hat{\mu}(n) \;&\equiv\; \mathrm{E}[\log \hat{n}|n] \;=\; \mu + \beta \cdot (\mathrm{E}[r|n] - \mu) + (1/2)\sigma^2_{post} \\ &=\; \mu + \beta \cdot (\log n - \mu) + (1/2)\sigma^2_{post}, \\ \hat{\sigma}^2 \;&\equiv\; \mathrm{var}(\log \hat{n}|n) \;=\; \beta^2 \mathrm{var}(r|n) \\ &=\; \beta^2 \nu^2 \;=\; \frac{\sigma^4 \nu^2}{(\sigma^2+\nu^2)^2}. \end{aligned}$$

Thus as stated in the text, $\hat{\mu}(n)$ is an affine function of $\log n$ with slope β, and $\hat{\sigma}^2$ is independent of n.

Conditional on n, $\hat{n}$ is log-normally distributed with the parameters just stated. It then follows from the properties of a log-normal distribution that

$$\mathrm{E}[\hat{n}|n] \;=\; \exp(\hat{\mu}(n) + (1/2)\hat{\sigma}^2), \tag{A.6}$$

$$\mathrm{var}[\hat{n}|n] \;=\; [\exp(\hat{\sigma}^2) - 1] \cdot \exp(2\hat{\mu}(n) + \hat{\sigma}^2).$$

Hence

$$\frac{\mathrm{SD}[\hat{n}|n]}{\mathrm{E}[\hat{n}|n]} \;=\; \sqrt{e^{\hat{\sigma}^2} - 1} \;>\; 0 \tag{A.7}$$

regardless of the value of n, as stated in the text. This delivers the property of *scalar variability* discussed in the text.

[63] The calculations in this case coincide with the ones needed for the Bayesian model of optimal choice between lotteries, presented in section 2, even though the reason why it is optimal to base the subject's decision on an estimate of this kind is different in that context.

One also observes that (A.6) implies that

$$\log \mathrm{E}[\hat{n}|n] \;=\; \hat{\mu}(n) + (1/2)\hat{\sigma}^2 \;=\; \log A + \beta \log n,$$

where

$$A \;\equiv\; \exp((1-\beta)\mu + (1/2)\sigma_{post}^2 + (1/2)\hat{\sigma}^2) \;>\; 0.$$

This yields the power-law relationship stated as (1.4) in the text for the mean estimated numerosity as a function of the true numerosity.

As discussed in the text, this implies a "regressive bias." Specifically, $\mathrm{E}[\hat{n}|n] > n$ for all $n < n^*$, while $\mathrm{E}[\hat{n}|n] < n$ for all $n > n^*$, where the "cross-over point" n^* is given by

$$\log n^* \;\equiv\; \frac{\log A}{1-\beta} \;=\; \mu + c, \tag{A.8}$$

using the expression

$$c \;\equiv\; \frac{1}{2}\frac{\sigma_{post}^2 + \hat{\sigma}^2}{1-\beta} \;>\; 0$$

for a quantity that depends on σ and ν, but is independent of the prior mean μ. Thus if in different experiments, the degree of prior uncertainty about the stimulus numerosity is the same in percentage terms (that is, the value of σ is the same), while the average numerosity presented is different (implying a different value of μ), the model implies that the cross-over point n^* should vary in proportion to e^{μ}. Alternatively, n^* should vary in proportion to the prior mean numerosity $\mathrm{E}[n]$ (which is equal to e^{μ} times a constant that depends only on σ), as stated in the text.

If instead we assume that the prior is fixed across experiments, but that ν is varied (for example, by varying cognitive load, as in the experiments of Anobile *et al.*, 2012), then both of the coefficients A and β in relation (1.4) are predicted to change across experiments. When ν is larger (internal representations are less precise), the model predicts that β will be smaller (though still positive), so that $\mathrm{E}[\hat{n}|n]$ will be a more concave function of n, as stated in the text.

These qualitative conclusions about subjective estimates of numerosity do not depend on assuming that the subject's estimate must equal the posterior mean, conditional on the internal representation r. If we assume instead that the subject's estimate minimizes the mean squared *percentage* error in the estimates, $\mathrm{E}[(\log \hat{n} - \log n)^2]$, then the Bayesian estimate $\hat{n}(r)$ should satisfy

$$\log \hat{n}(r) \;=\; \mathrm{E}[\log n|\, r] \;=\; \mu_{post}(r).$$

From the above characterization of the posterior distribution, we would again find that $\log \hat{n}(r)$ is predicted to be an affine function of r, with a slope of β; only the intercept of the function is different in this case. The same argument as above then once again implies (A.5), where $\hat{\mu}(n)$ is again an affine function of $\log n$ with a slope of β, though with a different intercept than the one derived above.

Alternatively, if we assume that the subject's estimate is given by the posterior mode (or "maximum *a posteriori* estimate"), then the properties of a log-normal distribution imply that

$$\log \hat{n}(r) \;=\; \log \mathrm{mode}[n|r] \;=\; \mu_{post}(r) - \sigma_{post}^2.$$

Thus once again, $\log \hat{n}(r)$ is predicted to be an affine function of r with slope β, though with yet another value for the intercept term. This again allows us to derive (A.5), in which $\hat{\mu}(n)$ is again an affine function of $\log n$ with a slope of β.

In each of these cases, the same derivations as above allow us to obtain the predictions (A.7) and the power law (1.4). Again the equation for the cross-over point is of the form (A.8); only the expression for the constant c is different in each case. Thus in any of these cases, we obtain the following common predictions: (i) $\log \mathrm{E}[\hat{n}|n]$ should be an affine function of $\log n$ [a log-log plot should be affine] with a slope $0 < \beta < 1$; only the intercept of the log-log plot should be different in the three cases. (ii) Fixing σ and ν, but allowing μ to vary across experiments, the cross-over point n^* should be a constant multiple of the prior mean $\mathrm{E}[n]$; only the positive multiplicative factor is different in the three cases. (iii) In any given experiment, the standard deviation of $\hat{n}$ should grow in proportion to the mean estimate as n is increased [the property of scalar variability]. As discussed in the text, there is support for all of these predictions in experiments on estimation of numerosity (as well as a number of other sensory contexts, as reviewed in Petzschner *et al.,* 2015).

B A Bayesian Model of Lottery Choice

As explained in the text, we assume that the quantities X and C are respectively represented by quantities r_x and r_c, independent random draws from the conditional distributions

$$r_x \;\sim\; N(\log X,\, \nu^2), \qquad r_c \;\sim\; N(\log C,\, \nu^2),$$

where the precision parameter $\nu > 0$ is the same for both monetary amounts. We assume that the subject's decision rule is optimized (that is, that it maximizes the subject's expected financial gain from the decision) for an environment in which the true values (X, C) defining a given decision problem are assumed to drawn from a prior distribution under which X and C are distributed independently of each other, and each have a common (log-normal) marginal distribution

$$\log X, \log C \;\sim\; N(\mu, \sigma^2).$$

Under these assumptions, the posterior distribution for X conditional on the internal representation $\mathbf{r} \equiv (r_x, r_c)$ depends only on r_x, and the posterior distribution for C similarly depends only on r_c.

We wish to determine which of the two options available to the subject on the given trial would maximize the expected financial gain $\mathrm{E}[\Delta W^a|\mathbf{r}]$, given that the decision must be based on the imprecise internal representation $\mathbf{r}$ of the decision problem. In the case of the risky option, the expected financial gain is[64]

$$\mathrm{E}[\Delta W^{risky}|\mathbf{r}] \;=\; p \cdot \mathrm{E}[X|r_x],$$

[64]Here we assume that the zero financial gain in the case of the zero outcome is internally represented as precisely zero. This is consistent with our logarithmic coding model (which implies that extremely small financial gains have virtually zero probability of being mistaken for a financial gain of one cent or more). As discussed in the text, we also assume here that the probabilities of the different outcomes are represented with perfect precision; this last assumption is relaxed in section E below.

while in the case of the certain option, it is

$$\mathrm{E}[\Delta W^{certain}|\mathbf{r}] \;=\; \mathrm{E}[C|r_c].$$

Hence the risky option is predicted to be chosen if and only if

$$p \cdot \mathrm{E}[X|r_x] \;>\; \mathrm{E}[C|r_c], \tag{B.1}$$

as stated in the text at (2.2).

Furthermore, for either of the monetary amounts considered individually ($Y = X$ or C), the model just proposed implies that the joint distribution of $\log Y$ and r_y is a bivariate Gaussian distribution, of the same form as the joint distribution of $\log n$ and r in the model of numerosity estimation. Just as in the calculations in the previous section of this appendix, the distribution of $\log Y$ conditional on the value of r_y will be a Gaussian distribution,

$$\log Y|r_y \;\sim\; N(\mu_{post}(r_y), \sigma^2_{post}), \tag{B.2}$$

where the mean $\mu_{post}(r_y)$ is an affine function of r_y, defined by the same equation (A.2) as above; and the variance σ^2_{post} is the same for all r_y, and again given by (A.4). Thus the conditional distribution for Y will be log-normal.

It then follows (just as in the model of numerosity estimation) that the conditional expectation of either monetary amount will be given by

$$\begin{aligned}
\mathrm{E}[Y|r_y] &= \exp[\mu_{post}(r_y) + (1/2)\sigma^2_{post}] \\
&= \exp[\mu \;+\; \beta \cdot (r_y - \mu) \;+\; (1/2)\sigma^2_{post}] \\
&= \exp[\alpha + \beta r_y],
\end{aligned}$$

as stated in the text, where

$$\alpha \;\equiv\; (1-\beta)\mu \;+\; (1/2)\sigma^2_{post},$$

and β is again defined in (A.3). Substituting this expression for the conditional expectations in (B.1), and taking the logarithm of both sides of the inequality, we find that the risky option should be chosen if and only if the internal representations satisfy

$$\log p \;+\; \beta r_x \;>\; \beta r_c,$$

as stated in the text.

This in turn implies that the risky option should be chosen if and only if $r_x - r_c$ exceeds the threshold stated in (2.4). The proposed model of noisy coding implies that, conditional on the true data (X, C) defining the decision problem, $r_x - r_c$ will be a Gaussian random variable,

$$r_x - r_c \;\sim\; N(\log X - \log C,\, 2\nu^2).$$

Hence the transformed variable

$$z \;\equiv\; \frac{(r_x - r_c) \;-\; \log(X/C)}{\sqrt{2} \cdot \nu}$$

will have a distribution $N(0,1)$, and a cumulative distribution function $\Phi(z)$. In terms of this transformed variable, the condition (2.4) for acceptance of the risky option can be expressed as

$$z \ > \ z^{crit} \ \equiv \ \frac{\beta^{-1} \log p^{-1} - \log(X/C)}{\sqrt{2} \cdot \nu}.$$

The probability of this occurring is $\Phi(-z^{crit})$, as stated by equation (2.4) in the text.

C Testing Scale-Invariance: Additional Statistics

We begin with a further discussion of the degree of scale-invariance of the choice curves (psychometric functions) for the different values of C shown in Figures 3 and 4. The maximum-likelihood parameter estimates for the different choice curves (estimates of (3.1) for each of the individual values of C, and the estimate of (3.2) using the pooled data) are shown in Table 3. For each estimated model, the table also indicates the number of observations N_{obs} used to estimate the parameters, and the maximized value of the log-likelihood of the data, LL. We can use this information to compute a Bayes information criterion (BIC) statistic for each model, defined as[65]

$$BIC \ \equiv \ -2LL \ + \ 2 \log N_{obs},$$

since each model has two free parameters.

We can consider quantitatively the extent to which our data are more consistent with the more flexible model (3.1) than with the more restrictive predictions of our theory, using the BIC to penalize the use of additional free parameters. If we consider as one possible model of our complete data set a theory according to which there is a curve of the form (3.1) for each value of C, with parameters that may differ (in an unrestricted way) for different values of C, then the BIC associated with this theory (with 12 free parameters) is the sum of the BIC statistics shown in the last column of Table 3 for the individual values of C, equal to 7545.5.[66] The BIC associated with our more restrictive theory (with only two free parameters) is instead only 7521.0, as reported in the bottom row of the table.[67]

The more restrictive model is therefore preferred under the BIC: it leads to a lower value of the BIC, since the increase in the log-likelihood of the data allowed by the additional free parameters is not large enough to offset the penalty for additional free parameters. In fact, the BIC for our more restrictive model is lower by 24.6 points, corresponding to a *Bayes factor* of $K = e^{12.3}$, as discussed in the text. Thus the data increase the relative posterior

[65] Here, as elsewhere in the paper, "log" refers to the natural logarithm.

[66] Here the BIC is equal to minus 2 times the log-likelihood of the complete data set under the optimized parameters, plus a penalty of $N_{obs}(\theta)$ for each free parameter θ, where $N_{obs}(\theta)$ is the number of observations that are fit using the parameter θ. In the present application, this is the sum of the BICs reported for the models fit to the data for individual values of C.

[67] Our theory implies not only that choice probabilities should be given by a relationship of the form (3.2), but also that the parameters must satisfy conditions (3.3) stated below. However, the unrestricted maximum of the likelihood is attained by parameter values (shown in the bottom line of Table 3) that satisfy these restrictions, so that the best-fitting parameter estimates consistent with our theory, and the associated BIC, are the ones given in the table.

C	N_{obs}	δ	γ	LL	BIC
\$5.50	1476	-6.16	2.52	-681.3	1377.1
\$7.85	1476	-6.93	2.47	-685.8	1386.3
\$11.10	1476	-8.15	2.54	-654.6	1323.8
\$15.70	1476	-8.56	2.40	-674.6	1363.8
\$22.20	1476	-9.52	2.42	-666.8	1348.2
\$31.40	696	-7.87	1.84	-366.7	746.4
All	8076	-1.88	2.39	-3751.5	7521.0

Table 3: Maximum-likelihood estimates of choice curves for each of the values of C considered separately, and when data from all values of C are pooled. (In each case, data from all 20 subjects are pooled.)

probability of the restrictive model being the correct one, over whatever prior probability may have been assigned to this, by a factor of more than 200,000.

As indicated in the text, the individual choice curves of one subject, subject 9, are much farther from exhibiting scale-invariance than those of the other subjects. If we instead pool the data of all subjects except subject 9, the sum of the BIC statistics from the choice curves for individual values of C would equal 7143.4, while the BIC statistic for the restricted (scale-invariant) model equal only 7104.1.[68] Thus in this case, the BIC for our more restrictive model would be lower by 39.3 points, corresponding to a Bayes factor in favor of the scale-invariant model that is greater than 300 million.

This is of course purely a test of in-sample fit of the scale-invariant model. In the text, we instead emphasize tests in which the parameters of each model are estimated using only the first 3/4 of each subject's trials (the "calibration dataset"), and the remaining data (the "validation dataset") are used for an out-of-sample test of fit. In the first panel ("Pooled Data") of Table 1 in the text, the statistics reported in the first two columns correspond to the statistics presented in Table 3, except that the statistics correspond to choice curves estimated using only the "calibration dataset." (The values given for both the log likelihood and the BIC statistic are smaller in Table 1 because of the smaller sample.) In this case, the overall Bayes factor in favor of the scale-invariant model is not as large (when the pooled data are used), as shown in Table 1; but the scale-invariant model is still strongly favored. When we instead estimate separate choice curves for each subject, and then pool the log likelihoods and corresponding BIC statistics across subjects, the scale-invariant model is much more strongly favored, as shown in the bottom panel of Table 1 in the text.

The first two panels of Table 4 repeat the analyses shown in the corresponding panels of Table 1, but using only the data for the 19 subjects other than subject 9. The corresponding statistics when only the data for subject 9 are used are shown in the bottom panel of the table. (Note that if one sums each of the entries in the bottom two panels of Table 4, one obtains the statistics given in the bottom panel of Table 1 in the text.)

If the data for subject 9 are excluded, then even when a common set of parameters is estimated for all of the other 19 subjects, one finds that the scale-invariant model fits better,

[68] With the smaller number of subjects, the parameter estimates for the restricted model are $\delta = -1.95, \gamma = 2.47$, rather than the values shown on the bottom line of Table 3.

Model	$LL^{calibration}$	BIC	$LL^{validation}$	$\log K$
Pooled Data [all but Subject 9]				
Scale-invariant	-2678.8	5374.9	-865.3	0.0
Unrestricted	-2665.9	5414.0	-866.2	20.4
Heterogeneous Parameters [all but Subject 9]				
Scale-invariant	-1723.2	3661.5	-621.5	0.0
Unrestricted	-1564.4	3902.5	-741.7	240.7
Subject 9 Only				
Scale-invariant	-137.0	284.6	-42.4	0.0
Unrestricted	-30.5	134.5	-13.9	-103.6

Table 4: In-sample and out-of-sample measures of goodness of fit compared for the scale-invariant model (our logarithmic coding model) and an unrestricted statistical model, using the same format as in Table 1 in the text. In the top panel, each model is fit to the pooled data from all 19 subjects other than subject 9. In the middle panel, separate model parameters are fit to the data each of the 19 subjects other than subject 9. In the bottom panel, separate model parameters are fit to the data for subject 9.

both in-sample *and* out-of-sample. (In the first panel of Table 1, instead, $LL^{validation}$ is higher for the unrestricted model, meaning that this model fits slightly better out-of-sample, even though the in-sample fit is better for the scale-invariant model, once one penalizes the additional free parameters of the unrestricted model, and the overall Bayes factor in support of the scale-invariant model is relatively large.) The degree to which the overall Bayes factor favors the scale-invariant model is also considerably larger when subject 9 is excluded: we now find that $\log K = 20.4$, meaning that $K > 700$ million, as stated in the text.

If instead we estimate separate choice curves for each subject, then as in Table 1, the degree to which the model comparison favors the scale-invariant model is even greater; but comparison of the middle panel of Table 4 with the bottom panel of Table 1 shows that the conclusion is even more strongly supported if the data for subject 9 are excluded. In this case, the BIC statistic is lower for the scale-invariant model by 241.0 points (implying a Bayes factor $K_1 > 10^{52}$ in favor of the scale-invariant model, if we assume that one model or the other must be correct for all 19 of the non-excluded subjects), while the log-likelihood of the validation sample is also higher for the scale-invariant model by 120.2 points (implying a Bayes factor $K_2 > 10^{52}$ as well). Combining the two sorts of evidence, we obtain a Bayes factor $K > 10^{104}$ in favor of the scale-invariant model as the correct model for these 19 subjects. Thus the hypothesis that the scale-invariant model is correct for all subjects other than subject 9 (though the unrestricted model is correct for subject 9) is favored overwhelmingly over the hypothesis that the unrestricted model is correct for all subjects: by a Bayes factor $K > 10^{104}$. (This is the "even stronger evidence" referred to in footnote 43.)

The bottom panel of Table 4 shows instead that the scale-invariant model fits very badly for subject 9, both in-sample and out-of-sample. The strong evidence against scale-invariance on both counts in the case of subject 9 can also be seen from the location of the dot for subject 9 in Figure 5. We do not model the behavior observed in the case of subject 9.

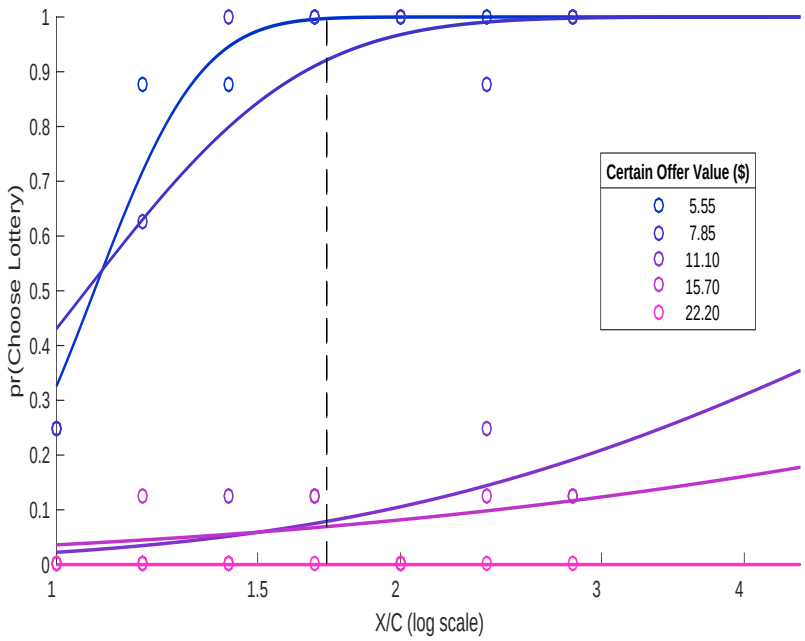


Figure 7: Choice curves for subject 9, for each of five different values of C, plotted as functions of $\log X/C$, as in the first panel of Figure 4. Note that this subject was not presented any trials in which $C = \$31.40$.

However, the direction of deviation from scale-invariance in the case of this subject is clear: the apparent degree of risk-aversion of this subject increases notably as the size of the certain payment C is increased, as shown in Figure 7.[69]

In the context of our Bayesian model, the sharp increase in risk aversion for larger values of C could be interpreted as optimal behavior on the part of a subject with a prior regarding the possible values of X that implies greater skepticism about the likelihood of larger payments than the log-normal prior assumed in our model would imply (perhaps because of prior experience with the amounts paid in campus decision-making experiments). Alternatively, it might reflect a subject for whom the amounts potentially earned in the experiment were of sufficient immediate usefulness for there to be a significant degree of diminishing marginal utility.

[69] In fact, subject 9 is risk-seeking (in the sense that the risky option is accepted more than half the time even for values of $X/C < 1/p$) when C equals \$5.55 or \$7.85, but risk-averse (in the sense that the risky option is declined more than half the time even for values of $X/C > 1/p$) when C equals \$11.10 or more. When C equals \$22.20, subject 9 never chose the risky option, for any of the values of X/C used in the experiment.

D Comparison with Alternative Models: Additional Alternatives

D.1 More General Random Expected-Utility Models

In the text, we consider two kinds of additive random-utility models (ARUMs). In each model, the subject is assumed to choose the option for which $\mathrm{E}[u(Y)] + \epsilon$ is larger, where Y is the monetary amount gained from the experiment and ϵ is a random term (drawn at the time of choice) that is independent of the characteristics of the option. In the "ARUM-Probit" model, the random term ϵ is assumed to be drawn from a normal distribution, whereas in the "ARUM-Logit" model, ϵ is assumed to be drawn from an extreme-value distribution. In each case, the nonlinear utility function $u(Y)$ is assumed to be of the CRRA form, $u(Y) = Y^{1-\gamma}/(1-\gamma)$, for some $\gamma \geq 0$.

Here we consider whether the fit of a nonlinear expected utility model might be improved by allowing a more general form of utility function, specifically, a function in the HARA class. This is the two-parameter family of utility functions $u(Y; \alpha, \beta)$ for which the Arrow-Pratt coefficient of absolute risk aversion $\rho(Y) \equiv -u''(Y)/u'(Y)$ is a hyperbolic function of Y: $\rho(Y) = (\alpha + \beta Y)^{-1}$, for some constant coefficients α, β. We assume that these coefficients are such that $\alpha + \beta Y > 0$ for all values of Y in the interval $(0, \bar{Y})$, where $\bar{Y} = \$125.60$ is the largest monetary amount used in the experiment, so that $u(Y)$ is an increasing, concave function over the range of values $0 < Y < \bar{Y}$. (This assumption requires that $\alpha \geq 0$ and that β not be too negative; it is satisfied if $\alpha, \beta \geq 0$ and at least one of them is positive.) This family of functions nests the CRRA family (the case in which $\alpha = 0, \beta > 0$), but also allows other cases, including the familiar CARA family (the case in which $\alpha > 0, \beta = 0$). Again we can consider both Probit and Logit cases of the random-expected-utility model with HARA utility. In each case, we have a three-parameter model, in which the free parameters are the preference parameters α, β, and the standard deviation of ϵ.[70]

If we estimate a single utility function for all 20 subjects, as in the upper panel of Table 2 in the text, then the generalization to HARA utility makes no difference for our results, for in fact the best-fitting member of the HARA family of utility functions is a member of the CARA family. (That is, when we optimize the parameters α and β, the optimal parameter values are on the boundary of the admissible region where $\alpha = 0$. Because α is constrained, there are also no more free parameters than in the CRRA case, and the BIC statistics also remain those reported in Table 2.) This is true for both the Probit and Logit versions of the ARUM; regardless of whether we estimate the common model parameters pooling the data of all 20 subjects, or only using the data for the 19 subjects other than subject 9; and regardless of whether the model parameters are estimated using the entire dataset, or only the smaller "calibration dataset" used when we wish to test out-of-sample fit of the models.

If instead we estimate a separate utility function for each subject, as in the lower panel of Table 2, then for some individual subjects the best-fitting HARA utility function involves

[70] The parameters α and β only identify the function $u(Y)$ up to an arbitrary affine transformation; the exact function can be pinned down by further specifying two values such as $u(\bar{Y})$ and $u'(\bar{Y})$. These latter parameters can be chosen arbitrarily. The value of $u(\bar{Y})$ has no consequences for predicted choices (since it does not affect the expected utility *difference* between any two lotteries); the value of $u'(\bar{Y})$ amounts to a choice of the units in which the standard deviation of ϵ is measured.

	Sum of BIC over Subjects	
Model	All Subjects	All but S9
Log coding	5223.0	4854.1
Random Expected-Utility Models		
CRRA - Probit	5526.0	5183.7
CRRA - Logit	5359.0	5015.6
HARA - Probit	5571.1	5229.5
HARA - Logit	5411.0	5067.6
Random Prospect-Theory Models		
PT - Probit	5535.0	5218.9
PT - Logit	5377.0	5058.4
Random Salience-Theory Models		
Salience - Probit	5822.3	5506.1
Salience - Logit	5616.0	5297.4

Table 5: Model comparison statistics for alternative models in which separate parameters are estimated for each subject. (The alternative models are explained in the text.) In each case, the statistic given is the sum of the BIC statistics for the models estimated for the individual subjects.

$\alpha > 0$, so that the generalization to HARA utility does allow a better in-sample fit (in the sense of a higher value for the likelihood). However, if we use the BIC criterion to penalize the additional free parameters in the case of the more flexible family of utility functions, the restricted CRRA model still fits the data better, at least if we sum the BIC statistics of the different subjects (so as to compare the two hypotheses under which either CRRA is the right model for all subjects, or the HARA model is). The BIC statistics for these model comparisons are shown in Table 5.

In each row of Table 5, the statistics given show the sum of the BIC statistics for the models estimated for the individual subjects. (The first column shows the sum of these statistics for all 20 subjects; the second column shows the sum for all subjects other than subject 9.) A comparison of the BIC statistics between any two rows can then be used to judge the relative fit of the two models in question; in particular, the difference in the BIC statistics in any two rows (of the same column) can be used to compute a Bayes factor K for comparison of the two hypotheses that one model or the other is the correct model for all subjects (all of those considered in that column).

For reference, the first row of the table shows the BIC statistics for our logarithmic coding model.[71] The second and third rows show the corresponding statistics for the CRRA-Probit and CRRA-Logit models, the models called "ARUM-Probit" and "ARUM-Logit" in the text. The fourth and fifth rows then show these statistics for the cases in which the function $u(Y)$ is allowed to be any member of the HARA family. We see that regardless of the assumed distribution for the noise term ϵ, and regardless of whether we use all 20 subjects or we

[71] The BIC statistics shown in the first column differ from those in the bottom of panel of Table 2 in the text, because here we fit the models to the complete dataset, rather than only to the "calibration dataset" as in Table 2.

exclude subject 9, the conclusion is the same: allowing the more general form of utility function raises the log likelihood (not shown in the table), but also results in a *larger* BIC statistic.

In each case, we would therefore have a Bayes factor K much greater than 1 in favor of the CRRA specification. (For example, in the case of the logit specification and using the data of all 20 subjects, the BIC for the CRRA case is 5359.0, while for the HARA case it is 5411.1. The BIC statistic is larger by 52.1 points, so that $\log K = 26.0$, implying a Bayes factor greater than 200 billion.) Of course, as discussed in the text, the logarithmic coding model fits better than either of the ARUMs based on CRRA utility; regardless of whether we exclude subject 9, the BIC statistics in the first row are lower than those in any other row.

D.2 Stochastic Versions of Prospect Theory

In order to test the ability of prospect theory to explain our data — at least in a way that puts the theory on the same footing as the others considered here, and allows likelihood-based model comparisons — it is necessary to extend the original theory of Kahneman and Tversky (1979) to make it stochastic. A standard approach in econometric tests of prospect theory (as reviewed in Stott, 2006) is to add a random term ϵ to the valuation of each risky prospect that would be specified by Kahneman and Tversky. It is then assumed that the option chosen on a given trial will be the one with the higher value of

$$\sum_i w(p_i)v(Y_i) \; + \epsilon,$$

where i indexes the possible outcomes under the risky prospect; p_i is the objective probability of outcome i and Y_i the associated net monetary gain; $w(p)$ is the Kahneman-Tversky probability "weighting function" and $v(Y)$ their "value function"; and ϵ is a random term, drawn independently for each prospect.[72] Once again, we can consider both the case in which ϵ is assumed to be drawn from a normal distribution and the case in which it has an extreme-value distribution.

As noted in the text, the ARUMs based on a CRRA utility function (considered in Table 2, and in Table 5 above) can also be considered to represent stochastic versions of prospect theory, in which the weighting function is assumed to be linear ($w(p) = p$) and the value function is of the CRRA form. This form of value function is in fact the one most commonly used in empirical tests of prospect theory (following Tversky and Kahneman, 1992; again, see Stott, 2006, for a review of the literature). But the assumption that $w(p) = p$ is of course contrary to what Kahneman and Tversky propose; and one might wonder whether a stochastic version of prospect theory that incorporates a nonlinear probability weighting function would better explain our experimental data than the random expected-utility models considered above.

[72] Except for the presence of the ϵ term, this is the model of the valuation of risky prospects proposed in Kahneman and Tversky (1979). Because in this paper we consider only simple gambles in which there are at most two possible outcomes i, the further refinement of cumulative prospect theory, introduced in Tversky and Kahneman (1992), is not relevant here.

Since in our experiment, the probability p of the non-zero outcome under the risky alternative is always equal to 0.58, it only matters what numerical value we propose for the probability weight $\tilde{p} \equiv w(0.58)$.[73] Kahneman and Tversky assume an "inverse-S" shape for $w(p)$, implying that there exists an intermediate probability $\hat{p}$ such that $p < w(p) < 1$ for all $0 < p < \hat{p}$ but $0 < w(p) < p$ for all $\hat{p} < p < 1$. They further assume a "sub-additivity" property for the weighting function, that requires that $\hat{p} < 1/2$ (so that in the case of two equally likely outcomes, $w(1/2) < 1/2$). Since in our experiment, $p = 0.58 > 1/2$, the assumptions of Kahneman and Tversky would imply that $0 < \tilde{p} < 0.58$.

Our "noisy prospect-theory models" therefore assume that

$$v(Y) = Y^{1-\gamma}/(1-\gamma)$$

and $w(0.58) = \tilde{p}$, where the parameters γ and $\tilde{p}$ satisfy the theoretical restrictions $\gamma \geq 0$ and $0 \leq \tilde{p} \leq 0.58$. There are two versions of the model, "PT-Probit" in which ϵ is drawn from a normal distribution, and "PT-Logit" in which it is drawn from an extreme-value distribution. In each case, the noisy prospect-theory model has three free parameters: the values of $\gamma, \tilde{p}$, and the standard deviation of ϵ.

If we fit the noisy prospect-theory models to our data assuming common parameter values for all subjects, then the results are the same as in the case of the CRRA random expected-utility models discussed in the text. For we find that the best-fitting parameter values involve $\tilde{p} = 0.58$, so that the upper bound is a binding constraint. In this case, the noisy prospect-theory model reduces to a model that is mathematically equivalent to the CRRA random expected-utility model; there is also the same number of free parameters (given that $\tilde{p}$ is constrained), and hence the same BIC statistic as for the CRRA-ARUM.

If instead we fit a separate noisy prospect-theory model to the data for each subject, then we do find that the data can be better fit by a model with $0 < \tilde{p} < 0.58$ for some subjects. However, as in the case of the more flexible class of utility functions considered in the previous subsection, the degree to which the likelihood of the data is increased is not great enough to justify the inclusion of the additional free parameters, if free parameters are penalized according to the BIC criterion. The sixth and seventh rows of Table 5 show the BIC statistics (again, summed over the subjects) for the "PT-Probit" and "PT-logit" models. In each case, the BIC statistics are higher for the less restrictive model (in which $\tilde{p}$ is allowed to be less than 0.58) than if the theoretical assumption that $\tilde{p} = 0.58$ is imposed. Thus these models do not fit better than the random expected-utility models already considered, and *a fortiori* are not competitive with the logarithmic coding model as an explanation for our experimental data.

D.3 Stochastic Versions of Salience Theory

Bordalo *et al.* (2012) propose an alternative theory of risk attitudes, according to which deviations from risk-neutral choice result from differential weighting of the different possible outcomes of a gamble according to their degree of "salience." In the case of a simple comparison between a certain outcome and a risky option with two possible outcomes, of the

[73] Kahneman and Tversky assume that $w(0) = 0$ and $w(1) = 1$. The only other probability that occurs in our examples is 0.42, the probability of the zero outcome under the risky alternative. However, Kahneman and Tversky also assume that $v(0) = 0$, so that the value of $w(p_i)$ for this alternative does not matter.

kind considered here, their theory is relatively simple. There are only two possible outcomes to take into account, the good outcome for the risky option (the one in which X is received) and the bad outcome (the one in which zero is received). It is assumed that the relative weight placed on the good outcome as opposed to the bad outcome is greater than the relative probability of that state if and only if the outcome in which the risky option yields X is the more salient of the two possible outcomes.

Algebraically, the theory predicts that the risky option should be chosen if and only if[74]

$$\sum_i g_i p_i Y_i \ > \ C,$$

where $g_i > 0$ is the "salience weight" associated with outcome i ($i = hi, lo$). The salience weights are normalized so that $\sum_i g_i p_i = 1$ (which is why the salience weights do not appear on the right-hand side of the above inequality). It is further assumed that $g_i = \delta g_j$ if outcome i is less salient than outcome j, where $0 < \delta < 1$ is a parameter that indicates the degree to which choice is biased by salience. (The theory reduces to expected-value maximization when $\delta = 1$.)

It follows that

$$\begin{aligned} g_{hi} &= \frac{1}{p + (1-p)\delta} \qquad \text{if } hi \text{ is the more salient outcome,} \\ g_{hi} &= \frac{\delta}{p\delta + (1-p)} \qquad \text{if } lo \text{ is the more salient outcome.} \end{aligned}$$

The deterministic theory proposed by Bordalo *et al.* implies that the risky option should be chosen if and only if

$$g_{hi} \cdot pX \ > \ C,$$

where g_{hi} is defined above. We can make the theory stochastic by proposing instead that the risky option is chosen if and only if

$$g_{hi} \cdot pX \ + \ \epsilon_x \ > \ C \ + \ \epsilon_c,$$

where ϵ_x, ϵ_c are two independent draws of the random variable ϵ. As usual, we can assume that ϵ is drawn either from a normal distribution or an extreme-value distribution, giving rise to two alternative stochastic versions of the theory, that we call "Salience-Probit" and "Salience-Logit."

It remains to specify which of the two outcomes should be more salient in our experiment. Bordalo *et al.* assume that the relative salience of the two outcomes depends on a comparison of the payoff difference $|X - C|$ in the hi outcome to the payoff difference $|0 - C|$ in the lo outcome, with an increase in the absolute payoff difference in either state making that outcome more salient. We adopt a parsimonious specification of their model by assuming that

[74]The theory of Bordalo *et al.* (2012) also allows the utilities associated with different outcomes to be different from the net financial gains in each case, as assumed here. However, none of the interpretations of experimental findings with regard to risk attitudes in laboratory settings proposed in their paper depends on assumption of nonlinear utility. So here we test a more parsimonious specification in which utility is assumed to be linear, for small gambles of the kind presented in our experiment.

the salience function is a homogeneous degree zero function of the two payoff differences,[75] so that the relative salience of the two outcomes should depend only on the ratio X/C.

The theory of Bordalo *et al.* posits that the relative salience of the *hi* outcome should be an increasing function of X/C for all $X \geq C$ (since a larger value of X/C corresponds to a larger ratio of $|X - C|$ to $|0 - C|$). It is then only necessary to specify the critical fraction π such that *hi* will be the more salient outcome if and only if $X/C > 1/\pi$. The theory of Bordalo *et al.* further implies that the critical fraction must satisfy $0 < \pi < 1/2$, so that the *lo* outcome is more salient whenever $|0 - C| \geq |X - C|$.[76] We treat this as a free parameter that can be fit to our data. Each of our random salience-theory models then has three free parameters: δ, π and the standard deviation of ϵ. We impose as theoretical restrictions that $0 < \delta \leq 1$ and $0 < \pi < 1/2$.

When we fit these random salience-theory models to the experimental data, assuming a common set of parameters for all subjects, we find that the best-fitting parameter values involve a value of δ between 0.52 and 0.56 (depending on the precise sample used) and a value of π less than 0.25. That is, it does not improve the fit of the model to assume that the relative salience of the two outcomes switches on trials when X/C is larger: instead, the best-fitting model implies that the *lo* outcome is always more salient (resulting in consistently risk-averse behavior, in the situations that occur in our experiment). The value of δ much less than 1 implies a substantial degree of risk aversion: in the absence of the random terms, the fitted model would imply choice of the risky option only when X/C is greater than 2.3,[77] whereas a risk-neutral decision maker would only require X/C to be greater than 1.7.

However, models of this kind fit our data less well than do the random expected-utility models discussed in the main text, let alone the logarithmic coding model. If we assume a common set of parameters for all subjects, and fit to the entire dataset for all 20 subjects, the BIC statistic for the Salience-Probit model is greater than that for the CRRA-Probit model by 526.8 points, implying a Bayes factor in favor of the random expected-utility model of $K > 10^{114}$. The Salience-Logit model fits better, but still, the BIC statistic for the Salience-Logit model is greater than that for the CRRA-Logit model by 462.7 points, implying a Bayes factor in favor of the random expected-utility model of $K > 10^{100}$. Even the better-fitting of the random salience-theory models (Salience-Logit) has a BIC statistic that is greater than that for the logarithmic coding model by 891.1 points, implying a Bayes factor in favor of the logarithmic coding model of $K > 10^{193}$.

Our conclusion is the same if we estimate separate model parameters for each subject. The bottom two rows of Table 5 show the BIC statistics for the two random salience-theory models in this case (both when the BIC statistics of all 20 subjects are summed, and when we exclude subject 9). Again we find in each case that the random salience-theory models fit less well than the CRRA-ARUM models, and less well *a fortiori* than the logarithmic coding model. We also reach a similar conclusion when the model parameters are fit to

[75] Given the relatively scale-invariant behavior of most of our subjects, this simplification would not seem to bias our test against the salience models.

[76] When $X = 2C$, so that $|X - C| = |0 - C|$ exactly, Bordalo *et al.* assume that the *lo* outcome should be more salient, because of their principle of "diminishing sensitivity" (increasing the quantities $(0, C)$ to $(C, 2C)$ makes the difference between them seem smaller).

[77] Note that this model of risk-averse choice would be scale-invariant, thus providing a solution to the Rabin paradox.

the "calibration dataset" and we then test out-of-sample fit using the "validation dataset." The random salience-theory models fit worse than the CRRA-ARUM models, and worse *a fortiori* than the logarithmic coding model both in-sample and out-of-sample.

E Imprecise Representation of Probability

The model of choice between lotteries presented in section B of this appendix can be generalized to allow for noisy coding of the probability p as well. As explained in the text, if we assume a noisy internal representation r_p of the probability, the distribution of which depends only on the true probability p described to the subject, then the threshold for acceptance of the risky option becomes

$$r_x - r_c \;>\; \beta^{-1}\rho, \qquad \rho \;\equiv\; -\log \mathrm{E}[p|r_p], \tag{E.1}$$

as stated in the text at (4.3).

Suppose that the internal representation r_p is drawn from a distribution

$$r_p \;\sim\; N(\log(p/1-p), \nu_p^2), \tag{E.2}$$

where ν_p is non-negligible. Suppose furthermore that the log odds $z \equiv \log(p/1-p)$ of the two outcomes are normally distributed under the prior,

$$z \;\sim\; N(\mu_p, \sigma_p^2).$$

Then the joint distribution of z and r_p will be a bivariate Gaussian distribution, and calculations similar to those referenced in section B of this appendix can again be used to compute conditional distributions.

In particular, the posterior log odds, conditional on the representation r_p, will also have a Gaussian distribution,

$$z|r_p \;\sim\; N(\bar{m}(r_p), \bar{\sigma}^2), \tag{E.3}$$

where

$$\bar{m}(r_p) \;\equiv\; \mathrm{E}[z|r_p] \;=\; \mu_p \;+\; \beta_p \cdot (r_p - \mu_p)$$

using the notation

$$\beta_p \;\equiv\; \frac{\mathrm{cov}(z,\, r_p)}{\mathrm{var}(r_p)} \;=\; \frac{\sigma_p^2}{\sigma_p^2 + \nu_p^2},$$

and

$$\bar{\sigma}^2 \;\equiv\; \frac{\sigma_p^2 \nu_p^2}{\sigma_p^2 + \nu_p^2}.$$

Note that $\bar{m}(r_p)$ is a weighted average of r_p and μ_p, and $\bar{\sigma}^2$ is independent of r_p, as stated in the text.

Since the probability p of the non-zero payoff can be reconstructed from the log odds as $p = e^z/(1+e^z)$, we obtain

$$\rho(r_p) \;=\; -\log \mathrm{E}[p|r_p] \;=\; -\log \mathrm{E}\left[\frac{e^z}{1+e^z}|r_p\right],$$

where z is distributed in accordance with (E.3). Note that we can alternatively write this function as

$$\rho(r_p) \;=\; -\log \mathrm{E}[F(r_p,\, \epsilon)], \tag{E.4}$$

where

$$F(r_p,\, \epsilon) \;\equiv\; \frac{\exp[\bar{m}(r_p) + \bar{\sigma}\epsilon]}{1 + \exp[\bar{m}(r_p) + \bar{\sigma}\epsilon]} \tag{E.5}$$

and the expectation is over realizations of the random variable ϵ, which has a standard normal distribution (and is distributed independently of the value of r_p).

Conditional on a given true value of p, the internal representation r_p is a random variable with distribution (E.2). The variable ρ is then a nonlinear transformation of r_p defined by (E.4), and so also a random variable with a distribution conditional on the true value of p. The distribution of ρ is complicated to characterize, but it is easy to see that the introduction of noise into the encoding of the log odds by r_p results not only in random variation in the value of ρ (which would instead be a constant, equal to $-\log p$ as assumed in equation (2.4), if p were encoded with perfect precision), but also in a median value for ρ that is generally different from the value it would have in the absence of coding noise. Because $\bar{m}(r_p)$ is a monotonically increasing function of r_p, $F(r_p, \epsilon)$ is an increasing function of r_p for each value of ϵ; it then follows from (E.4) that $\rho(r_p)$ must be a monotonically decreasing function of r_p. The median value of ρ is then the value of the function $\rho(r_p)$ when r_p takes its median value, so that

$$\mathrm{median}[\rho|p] \;=\; \rho(z(p)).$$

Since $e^{-\rho}$ is a monotonically decreasing function of ρ, we similarly have

$$w(p) \;\equiv\; \mathrm{median}[e^{-\rho}|p] \;=\; e^{-\rho(z(p))}.$$

Recalling the definition of $\rho(r_p)$, we can alternatively define this function as

$$w(p) \;=\; \mathrm{E}[F(z(p), \epsilon)]. \tag{E.6}$$

Although for any p, $w(p) \to p$ as the variance ν_p^2 of the coding noise is made arbitrarily small, $w(p)$ is generally not equal to p (so that correspondingly, the mean value of ρ is not equal to $-\log p$) when the variance of the coding noise is positive. In fact, we can show that $w(p)$ has many of the properties that Kahneman and Tversky (1979) assume for their probability weighting function.

First, we observe that for any $0 < p < 1$, we have $0 < F(z(p), \epsilon) < 1$ for all ϵ, so that the expected value of F must lie between these extremes as well. Then (E.6) implies that $0 < w(p) < 1$ for all $0 < p < 1$. Next, we also observe that the derivative of the function is given by

$$\begin{aligned} w'(p) \;&=\; \mathrm{E}\left[\frac{\partial F(z(p), \epsilon)}{\partial z}\right] \cdot \frac{dz}{dp} \\ &=\; \beta_p(e^z + 2 + e^{-z})\, \mathrm{E}[(e^{\bar{m}(z)+\epsilon} + 2 + e^{-\bar{m}(z)-\epsilon})^{-1}]. \end{aligned} \tag{E.7}$$

Because this is the expected value of a variable that is always positive, $w'(p) > 0$, and we observe that $w(p)$ must be a monotonically increasing function of p over its entire range.

We further observe that for any ϵ, $F(z(p),\epsilon) \to 0$ as $p \to 0$, and $F(z(p),\epsilon) \to 1$ as $p \to 1$. The convergence is also uniform enough in each case to allow one to show that (E.6) implies that $w(p) \to 0$ as $p \to 0$, and $w(p) \to 1$ as $p \to 1$. Finally, (E.7) implies that for small p ($z << 0$),

$$w' \;\sim\; \beta_p \, e^{(1-\beta_p)(\mu_p - z)} \, \mathrm{E}[e^{\epsilon}],$$

while for large p ($z >> 0$),

$$w' \;\sim\; \beta_p \, e^{-(1-\beta_p)(\mu_p - z)} \, \mathrm{E}[e^{-\epsilon}].$$

From this we see that $w'(p) \to +\infty$ as $p \to 0$, and similarly that $w'(p) \to +\infty$ as $p \to 1$.

This implies that $w(p) > p$ for all small enough $p > 0$, while $w(p) < p$ for all large enough p. Thus if we interpret $w(p)$ as the "average perceived probability" when the true probability is p, the model implies over-estimation of small positive probabilities, and under-estimation of large probabilities less than 1, as with the probability weighting function of Kahneman and Tversky (1979).

We can observe more about the global shape of the $w(p)$ function if we consider the limiting case in which σ_p and ν_p are both small, but we fix the ratio ν_p/σ_p at some finite positive value γ as $\sigma_p, \nu_p \to 0$. In this limiting case, the value of β_p remains fixed at the value $\beta_p = 1/(1+\gamma^2) < 1$, while the value of $\bar{\sigma}$ approaches zero at the same rate as σ_p and ν_p. We then observe from (E.5) that for each value of ϵ,

$$F(z,\epsilon) \;\to\; \frac{\alpha_p e^{\beta_p z}}{1 + \alpha_p e^{\beta_p z}} \qquad \text{as } \bar{\sigma} \to 0,$$

where $\alpha_p \equiv e^{(1-\beta_p)\mu_p} > 0$, and the convergence is sufficiently uniform in ϵ to ensure that

$$\lim_{\bar{\sigma} \to 0} \mathrm{E}[F(z,\epsilon)] \;=\; \frac{\alpha_p e^{\beta_p z}}{1 + \alpha_p e^{\beta_p z}}$$

for all z. Thus in the limiting case we obtain

$$w(p) \;=\; \frac{\alpha_p p^{\beta_p}}{(1-p)^{\beta_p} + \alpha_p p^{\beta_p}}, \tag{E.8}$$

where the parameters satisfy $\alpha_p > 0, 0 < \beta_p < 1$.

As noted in the text, this function has the "inverse-S" shape assumed for the probability weighting function in prospect theory; indeed, this two-parameter family of probability weighting functions has sometimes been used in quantitative implementations of prospect theory. The function defined in (E.8) has the property that $w(p) > p$ for all $0 < p < p^*$, while $w(p) < p$ for all $p^* < p < 1$, where the critical probability p^* is given by

$$p^* \;\equiv\; \frac{e^{\mu_p}}{1 + e^{\mu_p}}.$$

If $\mu_p < 0$, implying that under the prior, $p < 1/2$ is more likely than $p > 1/2$, then $\alpha_p < 1$, and as a consequence $p^* < 1/2$, and the function $w(p)$ has the property of "subadditivity" assumed by Kahneman and Tversky (1979).

The nonlinearity of the function $w(p)$ in the case of noisy coding of the probability biases choice in our model in a similar way as the nonlinear probability weighting function in

prospect theory. Consider, as an example, the case in which the noise in the internal coding of X and C is negligible (ν is extremely small), while ν_p remains non-trivial (more precisely, the ratio ν_p/σ_p is not too small). If the monetary amounts X and C are represented with high precision, condition (E.1) for choice of the risky option reduces to

$$\rho \ < \log X - \log C.$$

Here the randomness in r_x and r_c have been suppressed as negligible, and β has been replaced by its limiting value of 1; the probability of acceptance of the risky option is then just the probability of a realization of ρ greater than $\log(X/C)$.

For any given value of p, this model predicts (like the model presented in section 2 of this paper) that the probability of acceptance of the risky option should depend only on the ratio X/C, and should be monotonically increasing in X/C; thus one should obtain a smooth psychometric function of the kind shown in Figure 2 of the text. But unlike the model in section 2, this model also predicts that the subject's apparent risk attitude should vary depending on the size of p.

If, as in the experiment discussed in Tversky and Kahneman (1992), we define the "certainty equivalent" C^* for each risky prospect (p, X) as the value of C for which the subject is indifferent between the risky prospect and obtaining C with certainty (i.e., the value of C for which the probability of choosing the risky option is exactly 1/2), then our model of random choice implies that C^* should be implicitly defined by

$$\text{median}[\rho|p] \ = \ \log(X/C^*),$$

so that

$$C^* \ = \ X \cdot e^{-\rho(z(p))} \ = \ X \cdot w(p).$$

Thus a plot of C^*/X as a function of p should show an increasing function of p (relatively independent of the size of X) with the "inverse-S" shape seen in the figures in Tversky and Kahneman (1992). In particular, this simple example suffices to show that our theory is capable of explaining the complete "four-fold pattern of risk attitudes" displayed in those figures.

F Experimental Design: Further Details

F.1 Participants

Twenty adults (10 female, ages 18-28 years) participated in the experiment after giving informed consent. Participants were recruited from the Columbia University community via on-campus informational fliers. Participants completed the experimental procedure in a private room on a single computer station. All procedures involving human subjects were approved by the Institutional Review Board of Columbia University (protocol #IRB-AAAQ2255).

F.2 Experimental Task

Participants completed a task in which they were instructed to choose between a certain monetary payment and a risky payment. Participants were presented with a series of options on the screen and submitted their choices by pressing the left or right arrow keys on the keyboard.

Figure 1 illustrates the screen observed by one of our subjects on a single trial. The two sides of the screen indicate the two options available on that trial; the subject must indicate whether she chooses the left or right option (by pressing the corresponding key). On the left side of the screen, the dollar amount shown is the quantity C that can be obtained with certainty if left is chosen. The right side of the screen shows the possible payoffs if the subject chooses the risky lottery instead. The amounts at the top and bottom of the right side indicate the two possible monetary prizes; the colored rectangular regions in the center indicate the respective probabilities of these two outcomes, if the lottery is chosen. The relative areas of the two rectangular regions provide a visual indication of the relative probabilities of the two outcomes; in addition, the number printed in each region indicates the probability (in percent) that that outcome will occur. (Thus on the trial shown, the subject must choose between a certain payment of $5.55 and a lottery in which there would be a 58 percent chance of receiving $15.70, but a 42 percent chance of receiving nothing.)

Figure 8: The computer screen during a single trial of our experiment. The two sides of the screen show the two options available on this trial.

Participants completed sequences lasting between 280 and 648 choices. The average completion time ranged between 14 minutes (280 choices) and 37 minutes (648 choices). In addition, participants began with instructional and practice session that lasted about 10 minutes. Participants were prompted to take a break after every 100 choices, upon which they could resume the experiment at the press of a button. The experiment interface was created with in-house code designed to run on the Psychophysics Toolbox-3 stimuli presentation package for MATLAB.

F.3 Experimental Task Instruction Slides

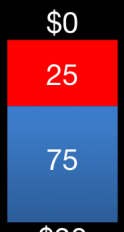
Welcome!

You are participating in an experiment on economic decision making and will be asked to make a number of choices. The study will last about an hour.

Your choices are very important in this task and will determine your final payment.

Understanding the lottery display

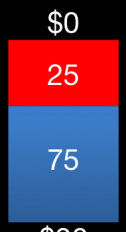
In this experiment, you will see lottery images like this one.



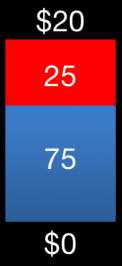
Each lottery represents the possibility of two different outcomes.

The size of the colored areas and the numbers written inside them represent the probability of each outcome.

In this example, the lottery presents a $20 reward at 75% probability and $0 with 25% probability.

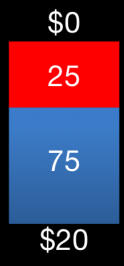


In each lottery one color will be associated with a zero amount and the other color with a positive amount.



The colors help you disambiguate the outcomes.

They are not always attached to the "good" or "bad" outcome in every case.



Understanding the task

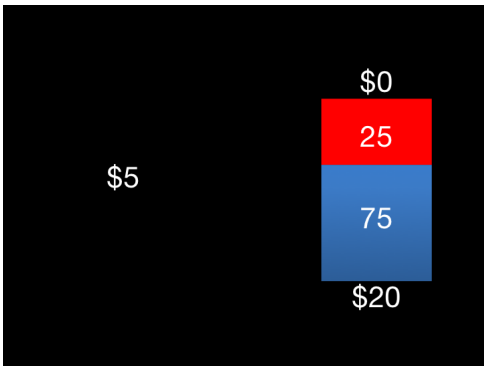
The experiment consists of a series of trials.

On each trial, you will be asked to choose between a fixed monetary amount for sure and playing a lottery.

Each trial will start like this:

+

Your options will be then be presented this way:



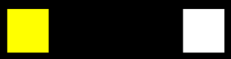
After a delay, the options will be followed by the following screen:

You will have **10 seconds** to make your choice.

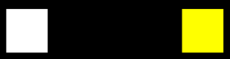
Press "🡐" for the option on the left or "🡒" for the option on the right.

After you have selected one of the options, you will see that choice displayed.

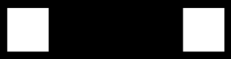
If you chose the option on the left, you would see this.



If you chose the option on the right, you would see this.



If you did not choose within **10 seconds**, you will see this.



Make sure you choose carefully on each trial!

One of the trials will be randomly selected at the end of the experiment, and your choice on this trial will determine your earnings.

Understanding payment

At the end of the experiment, the computer will draw a random number to determine the trial for which you will be paid.

If in the selected trial you chose the fixed amount, you will receive that amount for sure. If you selected the lottery, you will have the opportunity to play it (the computer will run the lottery).

If you did not make a choice in this trial, you will not earn anything.

On top of your earnings, you will receive $10 for participating in this session.

Practice

You will now complete 1 block of 10 practice trials. These trials **WILL NOT** count for payment.

This is a good time to ask questions if anything remains unclear.