The Spanish Journal of Psychology (2019), **22**, e60, 1–20.
© Universidad Complutense de Madrid and Colegio Oficial de Psicólogos de Madrid
doi:10.1017/sjp.2019.50

CAMBRIDGE UNIVERSITY PRESS

Tutorial: Nuisance or Substance? Leveraging Heterogeneity of Preferences

Michel Regenwetter[1] **and Maria M. Robinson**[2]

[1] *University of Illinois at Urbana-Champaign (USA)*
[2] *University of California, San Diego (USA)*

Abstract. Psychology and neighboring disciplines are currently consumed with a replication crisis. Recent work has shown that replication can have the unintended consequence of perpetuating unwarranted conclusions when repeating an incorrect line of scientific reasoning from one study to another. This tutorial shows how decision researchers can derive logically coherent predictions from their theory by keeping track of the heterogeneity of preference the theory permits, rather than dismissing such heterogeneity as a nuisance. As an illustration, we reanalyze data of Barron and Ursino (2013). By keeping track of the heterogeneity of preferences permitted by Cumulative Prospect Theory, we show how the analysis and conclusions of Barron and Ursino (2013) change. This tutorial is intended as a blue-print for graduate student projects that dig deeply into the merits of prior studies and/or that supplement replication studies with a quality check.

Received 8 February 2019; Revised 10 October 2019; Accepted 21 October 2019

Keywords: heterogeneity of behavior, random preference, replicability, tutorial.

Heterogeneity of Preference: Nuisance or Substance?

It's late in the evening and Mike has lost his driver's license. He believes it fell out of his pocket while he was jogging on the trail that circles the parking lot. He returns to the parking lot and searches underneath the single bright street light at the center of the lot. But all he finds is litter. Mike looks under a light where it is easier to see objects on the ground. It is true that his *average* location during his jog around the lot would be roughly near the light, but it is a logical fallacy to predict that his permit must lie under the light if it fell out of his pocket during his jog. It is no surprise that he finds nothing of value.

This metaphor illustrates how deriving predictions at an aggregate level sometimes lacks a logical foundation and can be counter-productive. The metaphor also suggests that two errors do not make one right. Repeating leaps of scientific reasoning in replication may perpetuate unwarranted scientific conclusions. Just as Mike would be well advised to take his route into account when looking for his belongings, so might behavioral decision researchers benefit from considering scientific predictions that don't contradict the diversity of preferences a theory or hypothesis permits. This diversity is genuinely part of the theory, rather than an error term that one might shrug off as a nuisance.

It is common in behavioral decision research, for example, in papers claiming that Cumulative Prospect Theory (CPT) cannot account for a finding, to reduce that entire theory down to a single preference pattern. For instance, as Regenwetter and Robinson (2017) have pointed out, papers often limit themselves to the single preference pattern associated with Tversky and Kahneman's (1992) median parameter estimates (for high-profile examples, see e.g. Birnbaum, 2008; Brandstätter, Gigerenzer, & Hertwig, 2006). Since these are medians of parameters defined on a continuous domain, there is no reason to expect that anyone satisfying CPT would actually have the preference pattern associated with these particular parameter values (see also Regenwetter & Robinson, 2019; Robinson, Wang, & Regenwetter, 2019). Comparing the pattern of modal choices or tallying the choices consistent with the single preference pattern associated with such aggregate-level

Correspondence concerning this article should be addressed to Michel Regenwetter. University of Illinois at Urbana-Champaign. Department of Psychology. 603 East Daniel St. 61820 Champaign, Illinois (USA).

E-mail: regenwet@illinois.edu

Authorship is alphabetical. This paper grew out of an invited talk given at the *VII Advanced International Seminar – Mathematical Models of Decision Making Processes: State of the Art and Challenges* held at the School of Psychology, Universidad Complutense de Madrid (Spain) in October 2018 (http://eventos.ucm.es/go/DecisionMakingModels). It was supported financially by the U.S. National Science Foundation (NSF). This project used the public-domain QTest software, which was developed with support by NSF, under grants SES–08–20009 and SES–14–59699 (PI: Michel Regenwetter). NSF had no other role other than financial support in this project. We are grateful to Dr. Ying Guo for assistance with some computations.

How to cite this article:
Regenwetter, M., & Robinson, M. M. (2019). Tutorial: Nuisance or substance? Leveraging heterogeneity of preferences. *The Spanish Journal of Psychology*, 22. e60. Doi:10.1017/sjp.2019.50

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press

"prior parameters" is akin to looking only under the bright light where it is easy to see objects on the ground, but where there is, a priori, nothing of value to find. Rejecting such a theory once, and then again in a replication that uses the same approach and the same incorrect line of reasoning, is like finding only litter under two parking lot lights. Just as Mike might find his lost driver's license if he took the heterogeneity in his locations seriously, so can decision research gain tremendous traction by tracking heterogeneity of preference as a precious source of information instead of dismissing and discarding it as a nuisance. The literature has issued numerous warnings about this and related challenges (e.g., Blavatskyy & Pogrebna, 2010; Block & Marschak, 1960; Harless & Camerer, 1994; Hey, 2005; Hey & Orme, 1994; Luce, 1959, 1995; Luce & Suppes, 1965; Loomes & Sugden, 1995; Regenwetter et al., 2014; Stott, 2006; Wilcox, 2008). Building on four in-depth prior papers (Regenwetter et al., 2014; Regenwetter & Robinson, 2017, 2019; Robinson et al., 2019), this tutorial details what one gains by tracking the heterogeneity of preferences that is inherent to the theory.

Illustrative Example

Throughout the tutorial, we revisit a study by Barron and Ursino (2013) as an illustrative example. Barron and Ursino (2013) studied the description-experience (DE) gap, according to which decision making behavior systematically differs under conditions in which people are explicitly provided with information regarding payoffs and their probabilities versus conditions by which they learn payoffs and their probabilities through an experiential process such as free-sampling or repeated choice from an urn. Some, but not all, researchers formally define the DE gap in terms of Cumulative Prospect Theory (henceforth CPT). In its standard form, CPT models decisions using a value function v and probability weighting functions w^+ and w^- for gains and losses, with parameters α, β, γ, δ, λ, of the form

$$v(x)=\begin{cases} x^{\alpha} & \text{if } x\geq 0 \quad (\text{with } 0\leq\alpha\leq 1), \\ -\lambda(-x)^{\beta} & \text{if } x<0 \quad (\text{with } 0\leq\beta\leq 1; 0<\lambda), \end{cases} \tag{1}$$

$$w^{+}(p)=\frac{p^{\gamma}}{\left(p^{\gamma}+(1-p)^{\gamma}\right)^{\frac{1}{\gamma}}} \qquad (\text{with } 0\leq\gamma<1), \tag{2}$$

$$w^{-}(p)=\frac{p^{\delta}}{\left(p^{\delta}+(1-p)^{\delta}\right)^{\frac{1}{\delta}}} \qquad (\text{with } 0\leq\delta<1), \tag{3}$$

where x are monetary amounts and p are probabilities, and where $\gamma, \delta < 1$ captures overweighting of small probabilities. It is common nowadays to consider a broader range for γ, δ and to report that experience-based paradigms yield behavior consistent with underweighting of small probabilities ($\gamma, \delta > 1$), whereas description-based paradigms yield behavior consistent with overweighting of small probabilities ($\gamma, \delta < 1$) in CPT (see, e.g. Hertwig & Erev, 2009).

Some of the debate centers on whether the DE gap reflects a true difference between how people weigh prospects across the two conditions, or is an artifact of under-sampling of low probability events in experience-based groups (Fox & Hadar, 2006; Hadar & Fox, 2009). Barron and Ursino (2013) examined this question by conducting two experiments in which they controlled for bias in sampling and by reanalyzing previous data from a repeated choice (experience) paradigm of Barron and Erev (2003). For the reanalysis, the authors focused on only those participants who had sampled the low probability event sufficiently often to observe the approximate underlying probability distribution (within a margin of error of ± 1 *SD* of the underlying probability). Based on these analyses, the authors reported that for 4 out of 6 pairs of decision problems, the percentage of choices did not differ statistically significantly from those observed in the original paper (Barron & Erev, 2003). For the two remaining gambles, they reported a significant change in choice frequencies. As we discuss in more detail in Appendix A, we are unable to reproduce these findings: There are ambiguities regarding Barron and Ursino's test statistic, and, more importantly their sample size. While they apparently used the number of participants as sample size, one could also take into account that each participant, in turn, generated numerous observations because they made numerous choices. Ignoring the latter gives hard-to-interpret choice proportions and may yield an underpowered test, whereas including it may violate iid assumptions presumed by the statistical test. In our analyses, we will consider both ways of tallying the sample size (and both caveats will apply with these data).

Barron and Ursino (2013) reported that modal choices were 'consistent' with underweighting[1] and concluded (abstract) that the reanalysis "*reveals that the gap persists even when subjects observe the expected frequency of rare events*" and furthermore, that the "*results of the two experiments strengthen the suggestion that descriptive theories of choice that assume overweighting of small probabilities are not useful in describing decisions from experience.*" We start our discussion by showing first that Barron and Ursino (2013) drew these latter conclusions without genuinely considering the heterogeneity of CPT's predictions.

[1]The pattern UN_{18} in Table 2 matches the pattern of modal choices.

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press

Predictions from Cumulative Prospect Theory

Recommendation 1:

> Before running a study aimed at comparing two or more decision theories, it is useful to inspect the domain of permissible preference states under each theory and to understand how much the theories overlap in the preference patterns they allow. That is, it is useful to assess how easily the theories can mimic each other on the stimulus set under consideration. Reducing, instead, each theory to one single preference pattern as a proxy for that theory, can jeopardize the interpretability of the entire study.

For our illustrations, we concentrate on the second half of Barron and Ursino (2013). In their reanalysis of the data from Barron and Erev (2003), they used six decision problems reported also in their Table 5 (see our Table 1). In their own Experiment 2, *"participants chose between two unmarked buttons presented on the screen. Each button was associated with one of two distributions referred to here as S (for safe) and R (for risky). The S distribution provided a certain loss of 3 points while the R distribution provided a loss of 20 points with probability 0.15 and zero otherwise"* (Baron and Ursino, 2013, p. 284). We show these also in Table C1 of Appendix C.

We start by considering the heterogeneous preferences predicted by CPT for the six stimuli in Table 1. We distinguish two cases: CPT with overweighting of small probabilities ($\gamma, \delta < 1$) and CPT with underweighting of small probabilities ($\gamma, \delta > 1$). In an ideal situation, the scholar would determine all possible preference patterns that are permitted by a theory. Following a path of least resistance, we first ran a discrete grid search that considered a large number of parameter combinations in Equations 1–3. We considered a trillion combined parameter values for overweighting and a trillion combined parameter values for underweighting in CPT. Specifically, for α, β, we considered values from .001 to 1.0, using a step size of .001. For δ and γ and overweighting, we considered values from .001 to .999, using a step size of .001. Similarly, for underweighting, we used values for δ and γ with that same step size from 1.001 to 2. Since no decision problems involved both gains and losses, we considered gains and losses independently, fixing λ to 1, with no loss in generality.

Table 1. *Decision Problems of Barron & Erev (2003) used in the Reanalysis by Barron & Ursino (2013) Adapted from their Table 5*

Problem	High EV choice	Low EV choice
4	Gain 4 with probability 0.8	Gain 3 for sure
6	Lose 3 for sure	Lose 4 with probability 0.8
7	Gain 10 with probability 0.9	Gain 9 for sure
8	Lose 10 with probability 0.9	Lose 9 for sure
9	Gain 32 with probability 0.1	Gain 3 for sure
11	Lose 3 for sure	Lose 32 with probability 0.1

Disregarding, for a moment, indifference (which none of the parameter values in our grid search produced), there are $2^6 = 64$ distinct possible preference patterns for six decision problems. Of these, our grid search for CPT with overweighting of small probabilities generated 16, as marked by $OV_1 - OV_{16}$ in the right side of Table 2. On the other hand, the grid search for CPT with underweighting of small probabilities generated 25 different patterns, as marked by $UN_1 - UN_{25}$ in the table. Nine of these preference patterns are consistent with both over- and underweighting of small probabilities.

This heavy overlap threatens one's ability to diagnose whether decision makers over- or underweight small probabilities, especially if we treat heterogeneity of behavior as a nuisance. This is because, in the latter case, there are nine scenarios of a potential shared preference pattern in which these stimuli do not differentiate between over- and underweighting. In this tutorial, we rely on the QTest open-source public-domain software[2] to illustrate a complete quantitative analysis. Figure 1 gives some milestone screenshots of portions of the graphical user interface (GUI) from the process of loading the preference patterns $OV_1 - OV_{16}$ and $UN_1 - UN_{25}$ into QTest. Appendix B provides screenshots of the input files.

Returning to the possibility of indifference, the parameters underlying CPT form a continuum that we approximated with a discrete grid search. Consider, for example, OV_1 and OV_2, which differ only in the preference for Decision Problem 11. In some of the models that follow, we implicitly also include the preference pattern that results from replacing the preference in Decision Problem 11 in OV_1 and OV_2 by a pairwise indifference. From the point of view of CPT, these are knife-edge situations in which Equations 1–3 must generate exactly equal overall utilities for the two choice options in Decision Problem 11.

Having applied Recommendation 1 to the six decision problems of Table 2, we now apply it to the single gamble pair in Experiment 2 of Barron and Ursino (2013). In Appendix C, we show that CPT with underweighting of small probabilities predicts that the decision maker prefers gamble R to gamble S regardless of parameter values. For CPT with overweighting of small probabilities, on the other hand, depending on the parameter values, the decision maker can either prefer R to S (e.g., with $\delta = .69$,

[2]QTest is available at www.regenwetterlab.org. Bayesian analyses require Version 2.0 or later.

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press

Table 2. *Preference Patterns according to CPT with Underweighting (UN) and Overweighting (OV)*

CPT with underweighting							CPT with overweighting						
	Decision Problem							Decision Problem					
Preference pattern	4	6	7	8	9	11	Preference pattern	4	6	7	8	9	11
UN_1	0	0	0	1	0	0	OV_1	0	0	0	1	0	0
							OV_2	0	0	0	1	0	1
							OV_3	0	0	0	1	1	0
							OV_4	**0**	**0**	**0**	**1**	**1**	1
UN_2	0	1	0	0	0	0							
UN_3	0	1	0	0	0	1							
UN_4	0	1	0	1	0	0	OV_5	0	1	0	1	0	0
UN_5	0	1	0	1	0	1	OV_6	0	1	0	1	0	1
							OV_7	0	1	0	1	1	0
							OV_8	0	1	0	1	1	1
UN_6	1	0	0	1	0	0	OV_9	1	0	0	1	0	0
							OV_{10}	1	0	0	1	0	1
UN_7	1	0	0	1	1	0	OV_{11}	1	0	0	1	1	0
							OV_{12}	1	0	0	1	1	1
UN_8	1	0	1	1	0	0							
UN_9	1	0	1	1	1	0							
UN_{10}	1	1	0	0	0	0							
UN_{11}	1	1	0	0	0	1							
UN_{12}	1	1	0	0	1	0							
UN_{13}	1	1	0	0	1	1							
UN_{14}	1	1	0	1	0	0	OV_{13}	1	1	0	1	0	0
UN_{15}	1	1	0	1	0	1	OV_{14}	1	1	0	1	0	1
UN_{16}	1	1	0	1	1	0	OV_{15}	1	1	0	1	1	0
UN_{17}	1	1	0	1	1	1	OV_{16}	1	1	0	1	1	1
UN_{18}	**1**	**1**	**1**	**0**	**0**	**0**							
UN_{19}	1	1	1	0	0	1							
UN_{20}	1	1	1	0	1	0							
UN_{21}	1	1	1	0	1	1							
UN_{22}	1	1	1	1	0	0							
UN_{23}	1	1	1	1	0	1							
UN_{24}	1	1	1	1	1	0							
UN_{25}	1	1	1	1	1	1							

Note: A "1" entry denotes a preference for option H (high expected value), a "0" entry denotes a preference for option L (low expected value). The Decision Problem labels are taken from Barron and Ursino (2013). The boldface pattern UN_{18} appears to be their proxy for underweighting. The boldfaced pattern OV_4, which can be derived from the median parameter estimates of Tversky and Kahneman (1992), is their declared proxy for CPT with overweighting.

$\beta = .88$, Tversky and Kahneman's median parameter estimates) or prefer S to R (e.g., with $\delta = .69$, $\beta = .8$). In a nutshell: Barron and Ursino's R/S stimulus pair is problematic for diagnosing over- versus underweighting, because CPT with overweighting of small probabilities can perfectly mimic CPT with underweighting.

Nuisance Model versus Substance Model

Recommendation 2:

> It is useful, for any given core theory, such as CPT, to distinguish different types of heterogeneity:
>
> 1. Treating heterogeneity of behavior in a population as a nuisance means assuming a single shared (i.e., unanimous) deterministic preference. This is extremely restrictive. We find it very implausible, but, like Barron and Ursino (2013), many papers use methods that are difficult to interpret short of making that assumption.
> 2. While different people may have different preferences consistent with the same core theory, individual preferences may be deterministic.
> 3. Individual preferences may be uncertain and different people may satisfy the same or

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press

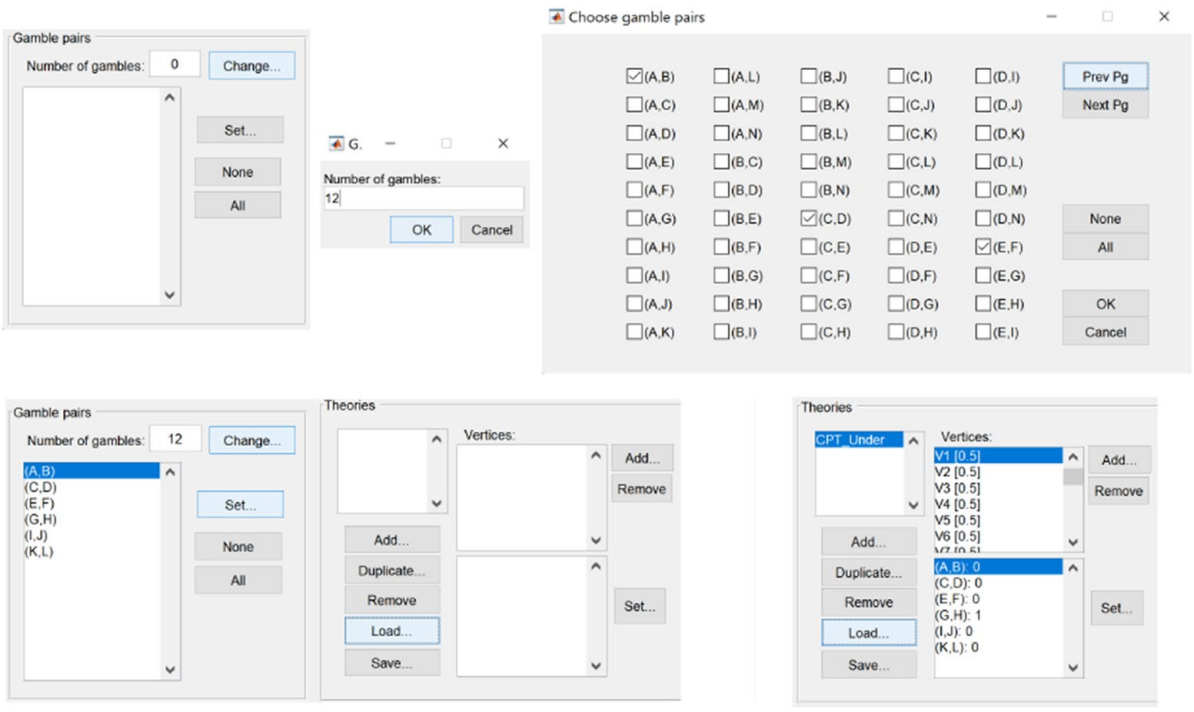


Figure 1. Loading predicted preference patterns into QTest: In the "Gamble pairs" section of the GUI press "Change..." and set the "Number of gambles" to 12, then select "Set" and choose 12 distinct pairs of labels. In the "Theories" section click on "Load..." and direct the software to the file containing all preference patterns (see Figure B1 in Appendix B). The last screenshot shows the "Theories" section of the interface after "CPT_Under" has been loaded.

different probability distributions over permissible preference states.

4. In all of the above scenarios, responses may or may not be error prone.

There are many different possible models to capture these scenarios. Different models often generate profoundly different predictions about behavior (see, e.g., Marley & Regenwetter, 2017; Regenwetter et al., 2014; Regenwetter & Robinson, 2017, for introductions, surveys, and guides). Here, we consider two types of models. We use simple names that reflect that one model class treats heterogeneity as a nuisance, whereas the other treats it as substance.

According to the "Nuisance Model" participants share a single (deterministic but unspecified) preference pattern and any deviations from this shared preference is due to probabilistic response error. Formally, writing P_{xy} to denote the probability that x is chosen over y, according to the *Nuisance Model*

$$P_{xy} \geq 1-\tau \quad \text{if and only if } x \text{ is unanimously preferred to } y. \qquad (4)$$

According to the "Substance Model," the source of heterogeneity in observed choice behavior is uncertainty in decision-makers' preference states, genuine individual differences, compounded with potential errors in responses. In an effort to keep the model parsimonious, we only allow for errors in responses when a "random preference" model, regardless of the distribution over heterogeneous preference states, assigns a pairwise preference probability of one or zero.[3] Formally, writing $P(\succ)$ for the probability of preference pattern $\succ$ and $x \succ y$ to denote that x is preferred to y in preference state $\succ$, let

$$R_{xy} = \sum_{all \succ in CPT with x \succ y} P(\succ). \quad \text{(Random Preference Model)} \qquad (5)$$

According to the *Substance Model*

$$P_{xy} \begin{cases} = R_{xy} & \text{if there exists a distribution } P \text{ with } 0 \neq R_{xy} \neq 1, \\ \leq .25 & \text{if } 0 = R_{xy} \text{ for every distribution } P, \\ \geq .75 & \text{if } R_{xy} = 1 \text{ for every distribution } P. \end{cases} \qquad (6)$$

In the terminology of Regenwetter and Robinson (2017) this is a type of "hybrid model" that incorporates components of a random preference model (5) and an error model (4). The choice of an upper bound on error rates is up to the scholar. Here, since this is an illustration for a tutorial, we consider a single model with an upper bound of 25% on error rates. That design choice is partially inspired by Harless and Camerer's (1994) recommendation of 25% errors in probabilistic error models for preferential choice. In contrast to them,

[3]Another reason to introduce these error bounds is that any model assigning probability zero to an observable discrete event is rejected the moment that one observes even just one instance of that prohibited event.

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press

however, we use 25% as an upper bound rather than a point hypothesis about error rates.

For CPT with overweighting of small probabilities, respectively CPT with underweighting of small probabilities, we consider all conceivable probability distributions over the 16 preference states $OV_1 - OV_{16}$, respectively all distributions over $UN_1 - UN_{25}$. As we discuss in Appendix D, a random preference model for CPT with overweighting leads to deterministic choice for Decision Problems 7 & 8, hence in this case, the Substance Model (6) augments the random preference model by allowing up to 25% error on those decision problems. As we also show in the next section, a random preference model for CPT with underweighting does not imply deterministic choice, hence the Substance Model for CPT with underweighting is the same as the random preference model for that case, and, for the sake of parsimony, it does not include an error term.

The Nuisance Model (unanimous preference, error-prone responses) and the Substance Model (probabilistic preferences, error-prone responses only where the random preference model predicts deterministic choice) are two parsimonious simple examples of a panoply of possible approaches.

Predictions

Nuisance Model: The predicted choice probabilities according to the Nuisance Model are directly given in its definition (4) above. It is worth noting that this model is not mathematically as simplistic as it appears at first sight. This is because, as we would argue to be the case here, the unanimous preference state is normally unknown. Hence, it must be treated as a free parameter in statistical inference. Furthermore, the model is characterized by one or multiple systems of inequality constraints on probabilities. That makes it an "order-constrained model," which requires advanced statistical methods (see, e.g., Davis-Stober, 2009; Klugkist & Hoijtink, 2007; Myung, Karabatsos, & Iverson, 2005; Regenwetter et al., 2014). We provide the main steps for reproducing our analysis using the QTest software that is designed for such inference.

If we either hypothesize or assume that the population (unanimously) has the single shared deterministic preference pattern UN_{18}, then the Nuisance Model, with an upper bound $\tau = \frac{1}{2}$ on error probabilities, implies that (up to sampling variability in the sample of respondents) the modal respondent chooses H in Decision Problems 4, 6, 7 and chooses L in Decision Problems 8, 9, and 11. This is the pattern of empirical modal choices that Barron and Ursino (2013) observe in both the original Barron and Erev (2003) data and in their own subset of those data (see also the "perfect fit" of the Nuisance Model for underweighting discussed later). Even though CPT with over- and underweighting share many preference patterns, none of the 16 possible preference patterns for CPT with overweighting, that we found in our grid search, predict that same pattern UN_{18}. At the same time, however, there are 24 other patterns that CPT with underweighting permits. The Nuisance Model for CPT with underweighting encompasses 25 different patterns of modal choice. Hence, the full-fledged Nuisance Model of CPT with underweighting forms a rather complicated collection of sets of constraints on binary choice probabilities (mathematically, it forms a disjoint union of 25 distinct hypercubes of side length $1 - \tau$); and CPT with overweighting likewise forms a rather complicated collection of sets of constraints on binary choice probabilities (mathematically, it forms a disjoint union of 16 distinct hypercubes of length $1 - \tau$, altogether 9 of which are shared with the other model.)

Substance Model: Figuring out the permissible choice probabilities according to the random preference model and, hence, the Substance Model, involves some mathematical work. While we give some specifics, we omit most of the details here (see, e.g., Guo & Regenwetter, 2014; Regenwetter & Cavagnaro, 2019; Regenwetter, Dana, & Davis-Stober, 2011; Regenwetter & Davis-Stober, 2012; Regenwetter et al., 2014; Regenwetter & Robinson, 2017; Zwilling et al., 2019, for detailed examples, related tutorials, and other useful pointers).

Let $P_{HL}(i)$ denote the probability of choosing H rather than L in Decision Problem i. For CPT with underweighting and for these decision problems, binary choice probabilities satisfy the random preference model (and hence the Substance Model, since no choice probabilities are forced to be zero or one) if and only if

$$\left(1 - P_{HL}(8)\right) = P_{LH}(8) \leq P_{HL}(6), \tag{7}$$

$$max\left(P_{HL}(7), P_{HL}(9)\right) \leq P_{HL}(4) \ \ and \tag{8}$$

$$P_{HL}(11) \leq P_{HL}(6). \tag{9}$$

The necessity of these conditions becomes apparent from noticing that, no matter which *UN* preference pattern one considers, whenever L is preferred in Decision Problem 8, then H is preferred in Decision Problem 6, but the converse does not hold. Similarly, no matter which *UN* preference pattern, whenever H is preferred in Decision Problem 7, then H is also preferred in Decision Problem 4, but not vice versa. Likewise, whenever H is preferred in Decision Problem 9, then H is also preferred in Decision Problem 4, but not vice versa. No matter which *UN* preference pattern, whenever H is preferred in Decision Problem 11, then H is also preferred in Decision Problem 6, but the converse does not hold. Those observations imply that, regardless of the probability distribution over

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press

preference states, the above inequality constraints have to hold. Plugging in all possible ways of satisfying these constraints with probabilities zero or one gives back all the permissible preference patterns. This proves the sufficiency of the constraints (7)–(9) to completely characterize the parameter space of the model.

Figure 2 shows how one can derive these constraints by calling the public domain software Porta from QTest. For example, running Porta for the 25 preference patterns of CPT_Under generates a text file like the one partially shown in the right-most screenshot of Figure 2. The first line of the file states that there are 10 inequalities in six variables. The first two lines of the matrix below, together with the first two entries in the column vector at the bottom of the screenshot mean that, with the labels used in QTest,

$$0\times P_{AB} - 1\times P_{CD} + 0\times P_{EF} - 1\times P_{GH} + 0\times P_{IJ} \le -1,$$
$$0\times P_{AB} + 0\times P_{CD} - 1\times P_{EF} + 0\times P_{GH} + 0\times P_{IJ} \le 0,$$

i.e., dropping terms that are multiplied by zero, and using our labels here,

$$-P_{HL}(6) - P_{HL}(8) \le -1,$$
$$-P_{HL}(7) \le 0.$$

This gives Inequality 7 as well as one of the conditions that binary choice probabilities must be in the interval [0, 1]:

$$1 - P_{HL}(8) \le P_{HL}(6),$$
$$0 \le P_{HL}(7).$$

The rest of the Porta output file likewise states an additional eight nonredundant inequality constraints on binary choice probabilities.

We now proceed to overweighting. We show that the stimuli in this study are such that a model of CPT with overweighting that allows for heterogeneity of preferences is extremely restrictive on two stimuli, but vacuous on the other four. Table 2 reveals that, no matter which of preference states OV_1 – OV_{16} we consider, CPT with overweighting of small probabilities predicts that one prefers L in Decision Problem 7 and H in Decision Problem 8. Therefore, if our grid search found all (strict) preference patterns for these two decision problems, then a random preference model of CPT with overweighting predicts $P_{HL} = 0$ in Decision Problem 7 and $P_{HL} = 1$ in Decision Problem 8. In Appendix D we directly prove these two predictions without reliance on a grid search.

Consequently, in the Substance Model, we permit for response error with probability up to .25 in Decision Problems 7 and 8. At the same time, if we concentrate only on Decision Problems 4, 6, 9, and 11, of the $2^4 = 16$ preference patterns conceivable, every single one can be derived from CPT with overweighting with a suitable choice of parameters in CPT. Hence, if a given

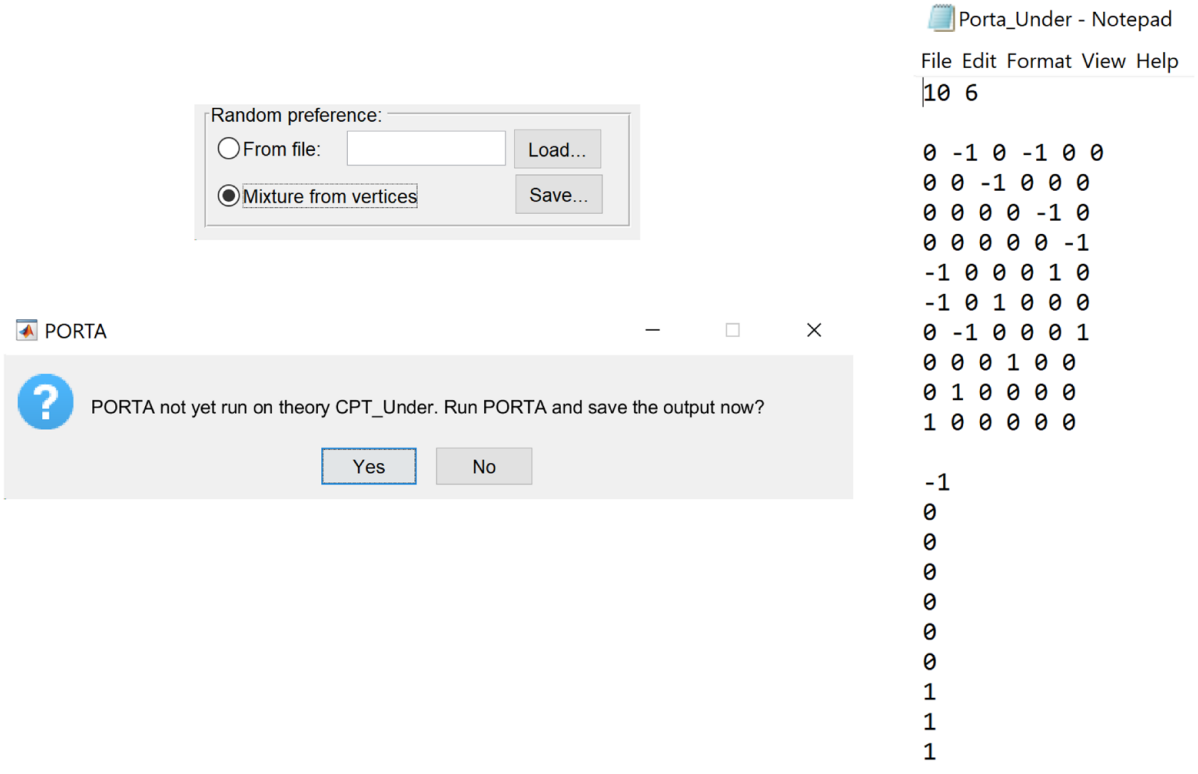


Figure 2. Calling PORTA to characterize the random preference model: In the "Random preference" section of the GUI, select "Mixture from vertices" then click "Save..." and QTest will generate a popup window like the one shown in the lower left screenshot. Completion of this process will generate a text file, part of which is shown in the screenshot on the right.

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press

person satisfying CPT with overweighting is allowed to fluctuate in his or her preference state, or if different individuals satisfying CPT with overweighting are allowed to differ in their preference states, then Decision Problems 4, 6, 9, and 11 do not predict any constraints on choice probabilities according to either a random preference or the Substance Model. In all, the Substance Model for CPT with overweighting places no constraints on four of the six binary choice probabilities and places moderate constraints (error rates $\leq .25$) on the remaining two choice probabilities:

$$0 \leq P_{HL}(7) \leq .25;\ .75 \leq P_{HL}(8) \leq 1;\\ 0 \leq P_{HL}(4), P_{HL}(6), P_{HL}(9), P_{HL}(11) \leq 1. \tag{10}$$

Put simply, four of these Decision Problems are uninformative for evaluating CPT with overweighting in the presence of heterogeneity, whereas the other two Decision Problems are extraordinarily diagnostic.

Insufficiently frequently asked questions (iFAQ)

Regenwetter and Robinson (2017) have shown several reasoning fallacies that permeate behavioral decision research (see also Regenwetter & Robinson, 2019). We first tackle their "Fallacy of Modal Choice."

QUESTION 1.0: *If most decision makers satisfy Theory X, then can we predict that the pattern of modal choices will satisfy Theory X, i.e., that 'the modal decision maker satisfies X?'* In general, this conclusion is invalid. Short of adding very strong additional features to the premise, e.g. that there exists a majority of decision makers who share a single unanimous preference pattern, there is no logical merit to such an inference. We zoom into a specific example for illustration.

QUESTION 1.1 (CPT, with Decision Problems 4, 6, 7, 8, 9, 11): *If most decision makers underweight small probabilities, then can we predict that the pattern of modal choices will be UN_{18}?* The answer is no.

Consider the 11 hypothetical decision makers described in Table 3. To keep matters simple, suppose that they each have a single deterministic preference and that they make no errors in responses. Suppose that four of these decision-makers overweight small probabilities and that their preference patterns are OV_2, OV_2, OV_2 and OV_8, respectively. Suppose that the remaining seven participants, a substantial majority among 11 people, underweight small probabilities, and that their preference patterns are UN_3, UN_3, UN_9, UN_9, UN_9, UN_{13} and UN_{25}, respectively. For such a combination of preferences, the predicted proportions of H choices , for Decision Problems 4, 6, 7, 8, 9, 11, are $\frac{5}{11}, \frac{5}{11}, \frac{4}{11}, \frac{8}{11}, \frac{6}{11}, \frac{8}{11}$, respectively. In Decision Problems 4, 6, 7, most people choose L, whereas in Decision Problems 8, 9, 11, most people choose H. This means that the predicted pattern of modal choices, or what is often misleadingly referred to as "the modal decision maker's choice pattern," is OV_4, and not a pattern permitted by underweighting. Even though the majority of participants underweight rare events, the pattern of predicted modal choices is a pattern consistent only with overweighting. Furthermore, this pattern is exactly the opposite of UN_{18}, which Barron and Ursino (2013) use as a proxy for underweighting of small probabilities. It even happens to be the pattern predicted by CPT (with overweighting) with median parameter values of Tversky and Kahneman (1992).

Table 3. *Hypothetical Preference Patterns of 11 Decision Makers, a Majority of whom Underweight Small Probabilities, and whose Pattern of Modal Choices is OV_4. That Pattern is the Opposite of UN_{18} and Incompatible with all 25 Patterns We Found for Underweighting*

Preference	Decision Problem					
Pattern	4	6	7	8	9	11
Decision Maker 1 (OV_2)	0	0	0	1	0	1
Decision Maker 2 (OV_2)	0	0	0	1	0	1
Decision Maker 3 (OV_2)	0	0	0	1	0	1
Decision Maker 4 (OV_8)	0	1	0	1	1	1
Decision Maker 5 (UN_3)	0	1	0	0	0	1
Decision Maker 6 (UN_3)	0	1	0	0	0	1
Decision Maker 7 (UN_9)	1	0	1	1	1	0
Decision Maker 8 (UN_9)	1	0	1	1	1	0
Decision Maker 9 (UN_9)	1	0	1	1	1	0
Decision Maker 10 (UN_{13})	1	1	0	0	1	1
Decision Maker 11 (UN_{25})	1	1	1	1	1	1
Totals:	5	5	4	8	6	8
Pattern OV_4	0	0	0	1	1	1
Pattern UN_{18}	1	1	1	0	0	0

We now turn to the "Fallacy of Choice Tallies" that Regenwetter and Robinson (2017) also showed to be very prevalent in the behavioral choice literature.

QUESTION 2.0: *If all or most decision makers satisfy Theory X, then can we pick one of the preference patterns derived from Theory X and predict that the tally of choices consistent with that pattern, aggregated over trials, stimuli and participants will exceed $\frac{1}{2}$ of all observations?* In general, this conclusion is invalid. Short of adding very strong additional features to the premise, e.g. that there exists a majority of decision makers who share a single unanimous preference pattern, there is no logical merit to such an inference. We zoom into a specific example for illustration.

QUESTION 2.1 (CPT, with Decision Problems 4, 6, 7, 8, 9, 11): *If all or most decision makers underweight small probabilities, then will the tally of choices consistent with UN_{18}, aggregated over trials, stimuli and participants, automatically exceed $\frac{1}{2}$ of all observations?* The answer is no.

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press

QUESTION 2.2 (CPT, with Decision Problems 4, 6, 7, 8, 9, 11): *The preference associated with the median parameter estimates of Tversky and Kahneman (1992) is OV_4. Therefore, if all or most decision makers overweight small probabilities, can we predict that the tally of choices consistent with OV_4, aggregated over trials, stimuli and participants, will exceed $\frac{1}{2}$ of all observations?* The answer is no.

Let u_i, respectively o_j, denote the probability that a randomly sampled person, on a randomly sampled trial, has preference UN_i, respectively OV_j. Suppose that every person at every trial has one of these preference states, that is, suppose that $\sum_i^{25} u_i + \sum_j^{16} o_j = 1$ and, for simplicity, suppose that nobody makes errors. Then, what is the expected proportion of choices consistent with UN_{18}, the pattern that Barron and Ursino (2013) use as proxy for underweighting? For example, suppose that $u_1 = \frac{1}{4}$, $o_3 = \frac{3}{4}$. Since UN_1 matches UN_{18} only for 2 of the six problems and since OV_3 matches UV_{18} on one of the six problems, altogether $u_1 \frac{2}{6} + o_3 \frac{1}{6} = \frac{1}{4} \times \frac{2}{6} + \frac{3}{4} \times \frac{1}{6} = \frac{5}{24}$ of all choices will be consistent with UV_{18}. In general, the expected proportion of choices consistent with UN_{18} is

$$\frac{2\mathbf{u}_1 + 4u_2 + 3u_3 + 3u_4 + 2u_5 + 3u_6 + 2u_7 + 4u_8 + 3u_9 + 5u_{10}}{6}$$
$$+ \frac{4u_{11} + 4u_{12} + 3u_{13} + 4u_{14} + 3u_{15} + 3u_{16} + 2u_{17} + 6u_{18} + 5u_{19} + 5u_{20}}{6}$$
$$+ \frac{4u_{21} + 5u_{22} + 4u_{23} + 4u_{24} + 3u_{25}}{6}$$
$$+ \frac{2o_1 + o_2 + \mathbf{o}_3 + 3o_5 + 2o_6 + 2o_7 + o_8 + 3o_9 + 2o_{10}}{6}$$
$$+ \frac{2o_{11} + o_{12} + 4o_{13} + 3o_{14} + 3o_{15} + 2o_{16}}{6},$$

where the boldfaced terms are those of the example. If <u>all</u> decision makers underweight small probabilities, i.e., $\sum_i^{25} u_i = 1$ and all $o_j = 0$, assuming error-free responses, then this expected proportion can take any value between $\frac{1}{3}$ and 1. Permitting errors in responses broadens that range further. We see no reason to predict that this expected proportion will exceed $\frac{1}{2}$. If <u>most</u> decision makers underweight small probabilities, i.e., $\sum_i^{25} u_i > \frac{1}{2}$ and $\sum_i^{16} o_j < \frac{1}{2}$, assuming error-free responses, then this proportion can take any value exceeding $\frac{1}{6}$. Allowing errors in responses broadens that range further. Here again, we see no reason to predict that this expected proportion will exceed $\frac{1}{2}$.

The expected proportion of choices consistent with OV_4 is

$$\frac{4u_1 + 2u_2 + 3u_3 + 3u_4 + 4u_5 + 3u_6 + 4u_7 + 2u_8 + 3u_9 + u_{10}}{6}$$
$$+ \frac{2u_{11} + 2u_{12} + 3u_{13} + 2u_{14} + 3u_{15} + 3u_{16} + 4u_{17} + u_{19} + u_{20}}{6}$$
$$+ \frac{2u_{21} + u_{22} + 2u_{23} + 2u_{24} + 3u_{25}}{6}$$
$$+ \frac{4o_1 + 5o_2 + 5o_3 + 6o_4 + 3o_5 + 4o_6 + 4o_7 + 5o_8 + 3o_9 + 4o_{10}}{6}$$
$$+ \frac{4o_{11} + 5o_{12} + 2o_{13} + 3o_{14} + 3o_{15} + 4o_{16}}{6}.$$

If <u>all</u> decision makers overweight small probabilities, i.e., $\sum_i^{16} o_j = 1$ and all $u_i = 0$, assuming error-free responses, then this expected proportion can take any value between $\frac{1}{3}$ and 1. Allowing errors in responses broadens that range further. We see no reason to predict that this expected proportion will exceed $\frac{1}{2}$. If <u>most</u> decision makers overweight small probabilities, i.e., $\sum_i^{25} u_i < \frac{1}{2}$ and $\sum_i^{16} o_j > \frac{1}{2}$, assuming error-free responses, then this proportion can take any value exceeding $\frac{1}{6}$. Allowing errors in responses broadens that range further. Once more, we see no reason to predict that this expected proportion will exceed $\frac{1}{2}$.

In Barron and Ursino's Experiment 2, since CPT with underweighting only permits preference of R over S, if <u>all</u> decision makers underweight small probabilities, then, (i) with error free choice, <u>all</u> choices must be R choices and (ii) with error prone choice we expect the proportion of R choices to be 1 - *s*, where *s* is the error rate. Since CPT with overweighting (considering a range of parameter values) permits either option to be preferred, regardless of whether we allow or disallow error, the expected proportion of S choices is not in any way constrained. This means that Table 6 of Barron and Ursino (2013) which reports that, "consistent with underweighting," the proportion of R choices is significantly greater than 0.5, is uninformative because that finding is equally consistent with either under- or overweighting. Testing whether the proportions significantly exceed $\frac{1}{2}$ is not based in testing CPT with over- or underweighting in the presence of heterogeneity of preferences. Barron and Ursino (2013) did not spell out their rationale for carrying out these tests.

Findings

The second author, for a graduate student project on assessing the scientific merit of Barron and Ursino (2013), carried out the mathematical analysis and the order-constrained statistical analysis using the public

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press

domain software QTest (Regenwetter et al., 2014; Zwilling et al., 2019). Table A1 in Appendix A shows our reconstruction of the choice frequencies, based on Table 5 of Barron and Ursino (2013). Figure 3 shows how to enter such data into QTest manually (where necessary, we rounded the frequencies to the nearest integers). Figures 4 – 6 show some key steps in running the analysis in QTest.

Our findings are summarized in Table 4, with the top half assuming a large sample size of 9,600 choice per option pair, for the 2003 master data set, and the bottom half assuming a small sample size of 24 participants, for the 2003 master data set (see Table A1 for specifics). We rely on Bayes Factors (BF, see also Kass & Raftery, 1995) to quantify the evidence in favor or against a given model. For a formal statement of the likelihood function and a discussion of technical matters, including independence, stationarity, prior distribution, and exchangeability assumptions, as well as how they relate to common approaches in the literature, see the four detailed tutorial papers by Regenwetter et al. (2014); Regenwetter and Davis-Stober (2018); Regenwetter and Cavagnaro (2019); Zwilling et al. (2019). We use the standard convention and terminology that a BF < 0.3 indicates "strong evidence against" and BF > 3.2 "strong evidence in favor" of a given model, with intermediate values offering "circumstantial evidence" (Jeffreys, 1961). We have boldfaced the first case, used a larger and underlined font for the second case, and placed the latter case into parenthesis for easy parsing. All Bayes Factors in Table 4 are in comparison to an unconstrained model. We also provide p values of frequentist hypothesis tests, treating each model as a null hypothesis in each case. Small p values are marked in boldface.

The full data from Barron and Erev (2003) included many cases in which respondents experienced biased samples, and hence could not possibly know the 'objective' distributions underlying the lotteries. Short of reconstructing what each of them saw, it is not clear that modeling the full data with Cumulative Prospect Theory is even possible, since the theory needs probabilities as part of its input, and it would not make much

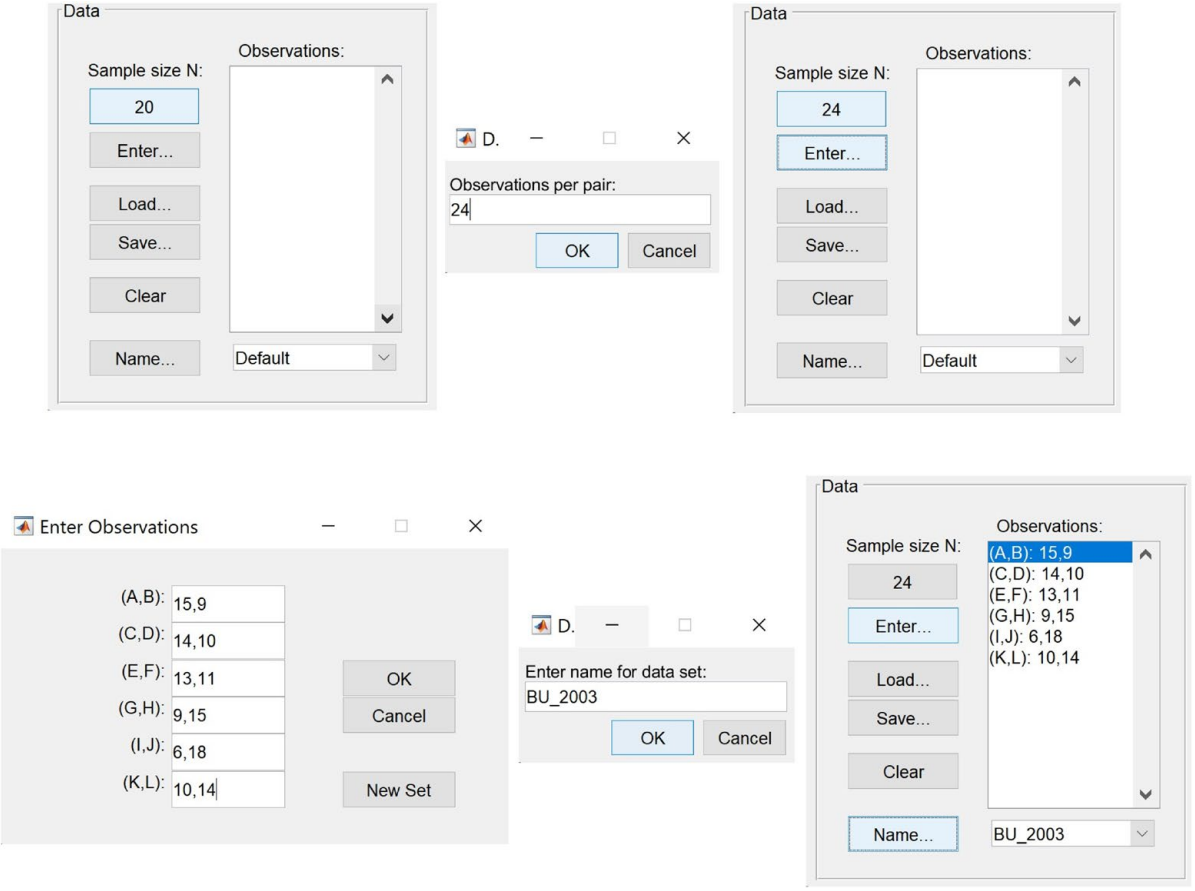


Figure 3. Loading the (rounded) data from Barron and Erev (2003) into QTest with an assumed sample size of 24: In the "Data" section of the GUI, click on the preset value "20" and enter "24" for the number of "Observations per pair." Select "Enter" and type in the inferred frequencies (see Table A1 and the surrounding cautions in Appendix A) and give the dataset a name. The bottom right screenshot shows the "Data" section after the data entry for "BU_2003" is completed.

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press

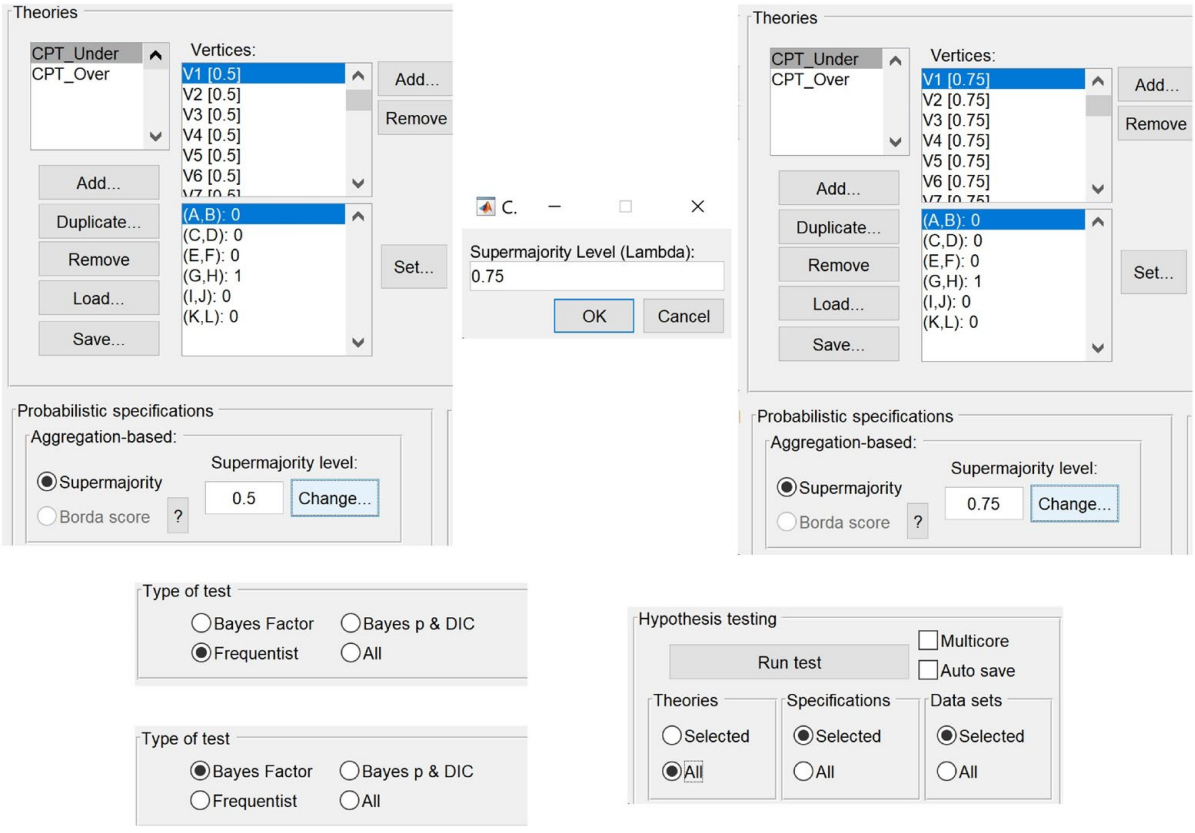


Figure 4. Order-constrained inference for the Nuisance Model in QTest: With the preference patterns loaded, the user selects "Supermajority" in the GUI section titled "Probabilistic Specifications" and "Aggregation-based" and sets the "Supermajority Level (Lambda)" to 1 - τ. In the "Type of Test" section of the GUI, the user chooses "Frequentist" or "Bayes Factor" and tells the program to "Run test" for "All" Theories in the "Hypothesis testing" section of the GUI.

sense to input hidden probabilities that respondents could not possibly know. Therefore, we start with the subsample of cases in which Barron and Ursino (2013) considered the experience to be representative of the actual distribution. These analyses are marked 2013 in Table 4.

Nuisance Model: For large sample size, even when we permit very high error rates in responses (i.e., up to 50% when $\tau = \frac{1}{2}$), the Bayesian analysis yields strong evidence against overweighting (BF < 10^{-100}) and the frequentist hypothesis test is significant. However, with small sample size, the Bayesian evidence becomes circumstantial and the result of the frequentist hypothesis test becomes nonsignificant. For underweighting, restated in our terminology, as Barron and Ursino (2013) reported, the pattern of modal choices fits preference pattern UN_{18} of CPT with underweighting perfectly. Indeed, writing $Q_{HL}(i)$ to denote the proportion of H choices in Decision Problem i,

$$Q_{HL}(1) = 0.64 \geq \frac{1}{2},\ Q_{HL}(2) = 0.56 \geq \frac{1}{2},\ Q_{HL}(3) = 0.63 \geq \frac{1}{2}$$
$$Q_{HL}(4) = 0.47 \leq \frac{1}{2},\ Q_{HL}(5) = 0.43 \leq \frac{1}{2},\ Q_{HL}(6) = 0.33 \leq \frac{1}{2}.$$

This *perfect fit* of the Nuisance Model for CPT with underweighting and high error rates is also reflected in the p value of 1: No matter what significance level, we could not have sufficient evidence to reject the Null Hypothesis. A perfect fit in no way proves the Null Hypothesis to be true. Interestingly, the Bayesian analysis in Table 4 only yields circumstantial evidence in favor of underweighting with a BF = 2.56. While there is a perfect fit from a frequentist point of view, that modest Bayes Factor reflects the penalty for complexity because CPT with underweighting permits 25 different possible scenarios for the unanimous preference pattern. When the error rate is bounded at $\tau = \frac{1}{4}$, we find strong evidence against CPT with both overweighting (BF < 10^{-100}) and underweighting (BF < 10^{-100}) and we find the frequentist test to yield rejections of the models.

Recall that there is some ambiguity as to how to treat sample size in the 2003 data set. Assuming 9,600 observations per binary choice may be problematic because those observations may not be exchangeable[4] and may

[4]The term "exchangeable" is a generalization of "stochastically independent" used in Bayesian statistics.

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press

Table 4. *QTest Analysis of Barron & Erev's (2003) Data that Barron & Ursino (2013) Re-analyzed*

Large Sample Size (9,600 binary choices in 2003 data set)						
Model	Nuisance Model		Nuisance Model		Substance	
CPT	Error Rate $\leq \frac{1}{2}$		Error Rate $\leq \frac{1}{4}$		Model	
2013	BF	*p* values	BF	*p* value	BF	*p* value
Overweighting	**$<10^{-100}$**	**$<10^{-15}$**	**$<10^{-100}$**	**$<10^{-15}$**	**$<10^{-100}$**	**$<10^{-100}$**
Underweighting	(2.56)	1	**$<10^{-100}$**	**$<10^{-100}$**	7.7	1
2003	BF	*p* values	BF	*p* value	BF	*p* value
Overweighting	**$<10^{-100}$**	**$<10^{-100}$**	**$<10^{-100}$**	**$<10^{-100}$**	**$<10^{-100}$**	**$<10^{-100}$**
Underweighting	(2.56)	1	**$<10^{-100}$**	**$<10^{-100}$**	**$<10^{-4}$**	**$<10^{-100}$**
Small Sample Size (24 participants in 2003 data set)						
Model	Nuisance Model		Nuisance Model		Substance	
CPT	Error rate $\leq \frac{1}{2}$		Error rate $\leq \frac{1}{4}$		Model	
2013	BF	*p* value	BF	*p* value	BF	*p* value
Overweighting	**0.15**	.33	**$<10^{-8}$**	**$<10^{-5}$**	**$<10^{-5}$**	**$<10^{-5}$**
Underweighting	(1.4)	1	**$<10^{-4}$**	**.008**	(1.45)	0.32
2003	BF	*p* values	BF	*p* value	BF	*p* value
Overweighting	**0.15**	.19	**$<10^{-9}$**	**$<10^{-6}$**	**$<10^{-6}$**	**$<10^{-5}$**
Underweighting	(1.9)	1	**$<10^{-5}$**	**.002**	(2.40)	.27

not have arisen from a stationary distribution, hence raising questions about robustness in statistical inference. On the other hand, treating the data as originating from 24 observations implies nonsensical choice proportions in the original paper (see Appendix A) and jeopardizes statistical power. Unsurprisingly, as we compare the large sample and small sample analyses, with the exception of a perfect fit, evidence becomes less compelling. However, the results tell virtually the same story. The only place where a substantive conclusion is affected is in the frequentist analysis of the Nuisance Model with error rate $\leq \frac{1}{2}$, which is rejected with large sample size, but not rejected with small sample size.

These results suggest that, with a unanimous preference and a high response error rate, these data may be more consistent with underweighting of rare events than overweighting. However, when allowing only a moderate error rate, we find strong evidence also against underweighting of rare events in decisions from experience. While we do not report the details in Table 4, we have also considered each single preference pattern as a possible unanimous preference, yielding 64 different nested submodels of the Nuisance Model with $\tau = \frac{1}{4}$. By treating each of these 64 cases as a separate model on its own, we strongly boost each one for parsimony. Nonetheless, each of these models has a minuscule Bayes Factor (BF < 10^{-100}). This gives very strong evidence against *any* unanimous preference pattern with moderate error rates. In other words, we have very strong evidence that, unless one permits a huge proportion of responses to be mistakes, heterogeneity is not a mere nuisance in these data. This result mirrors a similar finding of Regenwetter and Robinson (2017) on different data.

*Substance Model***:** Our analysis yields strong evidence against the Substance Model of CPT with overweighting (BF < 10^{-100}). For CPT with underweighting, substituting the choice proportions Q_{HL} (i) for P_{HL} (i) in Conditions 7–9, for the 2013 data, gives

$$Q_{LH}(8) = 0.53 \leq 0.56 = Q_{HL}(6),$$
$$max\left(Q_{HL}(7), Q_{HL}(9)\right) = max(0.63, 0.43) \leq 0.64 = P_{HL}(4) \quad and$$
$$Q_{HL}(11) = 0.33 \leq 0.56 = Q_{HL}(6).$$

Since the proportions satisfy the inequalities required of the probabilities, the best fitting probabilities are the same as the observed proportions. This explains the *p* value of 1 in the frequentist hypothesis tests of the random preference model of CPT with underweighting (hence the Substance Model), for the large sample size analysis. However, as we move to the small sample size analysis, where the reported proportions do not generate integer valued frequencies

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press

(see Table 5), rounding to integer valued frequencies leads to a small change in the effective proportions. This change is enough to violate one of the conditions above: Because, in this interpretation of the data, 13 people (out of 20) chose H over L in Problem 7 and 6 people (out of 10) did so in Problem 4, these numbers provide evidence against the prediction that $P_{HL}(7) \leq P_{HL}(4)$.

The Bayesian analysis provides strong evidence in favor (BF = 7.7) of the Substance Model for CPT with underweighting on the 2013 data, rewarding this model for its greater parsimony than the Nuisance Model with high response error rates. The Bayes Factor between the two models is $\frac{7.7}{2.56} = 3.01$. This means that, under the assumption that one or the other model holds, we have circumstantial evidence in favor of the Substance Model over the Nuisance Model with high permissible error rates. For the full data (marked **2003** in the lower panel of Table 4) that Barron and Ursino (2013) took from Barron and Erev (2003), the results for the Nuisance Models and the result for the Substance Model of CPT with overweighting match those for the 2013 subsample. Interestingly, the Substance Model of CPT with underweighting of rare events does not fare well on the full 2003 data. The choice proportions violate Condition 7 and the analysis yields a small Bayes Factor ($BF < 10^{-4}$), providing strong evidence against the model.

Comparing our results for the 2013 data with those of the 2003 data speaks to the discussion that motivated Barron and Ursino (2013) to consider only participants who had experienced representative samples from the underlying distributions. Barron and Ursino (2013) succeeded in extracting a subset of data for which CPT with underweighting provides an explanation of heterogeneous behavior. In Appendix A, we show that, when using the large sample size approach, the choice proportions for the 2013 data and the 2003 data are significantly different from each other. While some authors have questioned whether decision makers only appear to underweight objective probabilities when, in actuality they might overweight probabilities that are biased due to unrepresentative experienced samples, our reanalysis supports a diametrically opposite story in the case of a large sample size: In the 2003 data, which included experiences that Barron and Ursino (2013) deemed as not representative of the objective probabilities in the decision problems, heterogeneous preferences with underweighting (through the lens of CPT using the objective, but in many cases hidden, probabilities) fail to account for the data, whereas, in the 2013 data, which concentrated on cases that Barron and Ursino (2013) deemed representative of the underlying distributions, heterogeneous CPT with weighting parameters $\gamma, \delta > 1$ accounts well for the data. However, for a small sample size interpretation of the data, none of the Bayes factors provide better than circumstantial evidence in favor of any of the models and some frequentist p values increase (except one case that we discussed earlier).

Discussion

This article is intended as a tutorial and as a blue-print for graduate student projects that aim to dig deeply into the merits of past work and that contemplate running replication studies of decision-making experiments.

Recommendations: We recommend that one should thoroughly scrutinize studies for their scientific merit before considering replication studies. For studies that suffer from unwarranted scientific reasoning, such as the Fallacy of Choice Tallies or the Fallacy of Modal Choice, it makes sense to replace, or at least augment, replications with improved methods. Secondly, we recommend that scholars should aim to understand the genuine heterogeneity of preferences permitted by each theory before attempting to let theories compete against each other. Understanding this heterogeneity is key towards designing diagnostic stimuli. Indiscriminately reducing a theory, like CPT or, say, a heuristic toolbox, to a single preference pattern as a proxy for the entire theory readily jeopardizes the interpretability of an entire study. Another recommendation gives an overview of different types of heterogeneity and cautions that, as discussed extensively in the literature, the many different models that formally capture these different sources of heterogeneity make profoundly different predictions about behavior.

Reassessing Barron and Ursino (2013): We developed Nuisance Models and a Substance Model that formally derived predictions from explicitly stated assumptions, for Barron and Ursino's stimuli that they borrowed from Barron and Erev (2003). We tested these models on the data that Barron and Ursino (2013) extracted from Barron and Erev (2003) for their own re-analysis. The analysis with small assumed sample size leads to circumstantial evidence at best and strong evidence against some models at worst. The analysis using an assumed large sample size leads to more interesting findings, which we summarize next.

The analysis of the Nuisance Model with high response error rates provides strong evidence against overweighting and moderate evidence in favor of underweighting. However, when response error rates are limited to moderate levels, no model that treats heterogeneity of behavior as a nuisance, fares well at all.

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press

The Substance Model provides strong evidence against CPT with overweighting under the assumption that response error rates are no higher than 25% for Decision Problems 7 and 8, where a random preference model would predict deterministic choice. This result is striking, given that four out of six stimuli are undiagnostic and the model places no constraints on the corresponding four choice probabilities. The Substance Model for CPT with underweighting, which is also a random preference model and which does not involve any response errors, provides a perfect fit for the 2013 data by frequentist standards and is supported with strong evidence through the lens of a Bayes Factor analysis. Strikingly, and in direct contrast with the literature that motivated Barron and Ursino's study, the Substance Model for CPT with either over- or underweighting does not actually account well for the 2003 data. This ultimately strengthens Barron and Ursino's core message about the high value of experience designs where the decision maker has, at least in principle, the opportunity to learn the underlying probability distribution. For those participants whose data are excluded in the 2013 data, we do not know the (biased) proportions of wins they experienced. In all, it is not particularly surprising if CPT with either over or underweighting of the objective probabilities fails to account for the full 2003 data. While we pointed out various major problems with Barron and Ursino's approach, we support some of the conclusions they drew from re-analyzing the data of Barron and Erev (2003).

On the other hand, unfortunately, Barron and Ursino's Experiment 2 used a pair of gambles that do not constrain the preference of a decision maker who satisfies CPT with overweighting. If they had observed 100% choices of R, this would have supported underweighting as the more parsimonious explanation for this stimulus set. But any other proportion of R choices creates an inherent ambiguity as to which is the better explanation for the data. Hence, we find that their Experiment 2 results are uninterpretable.

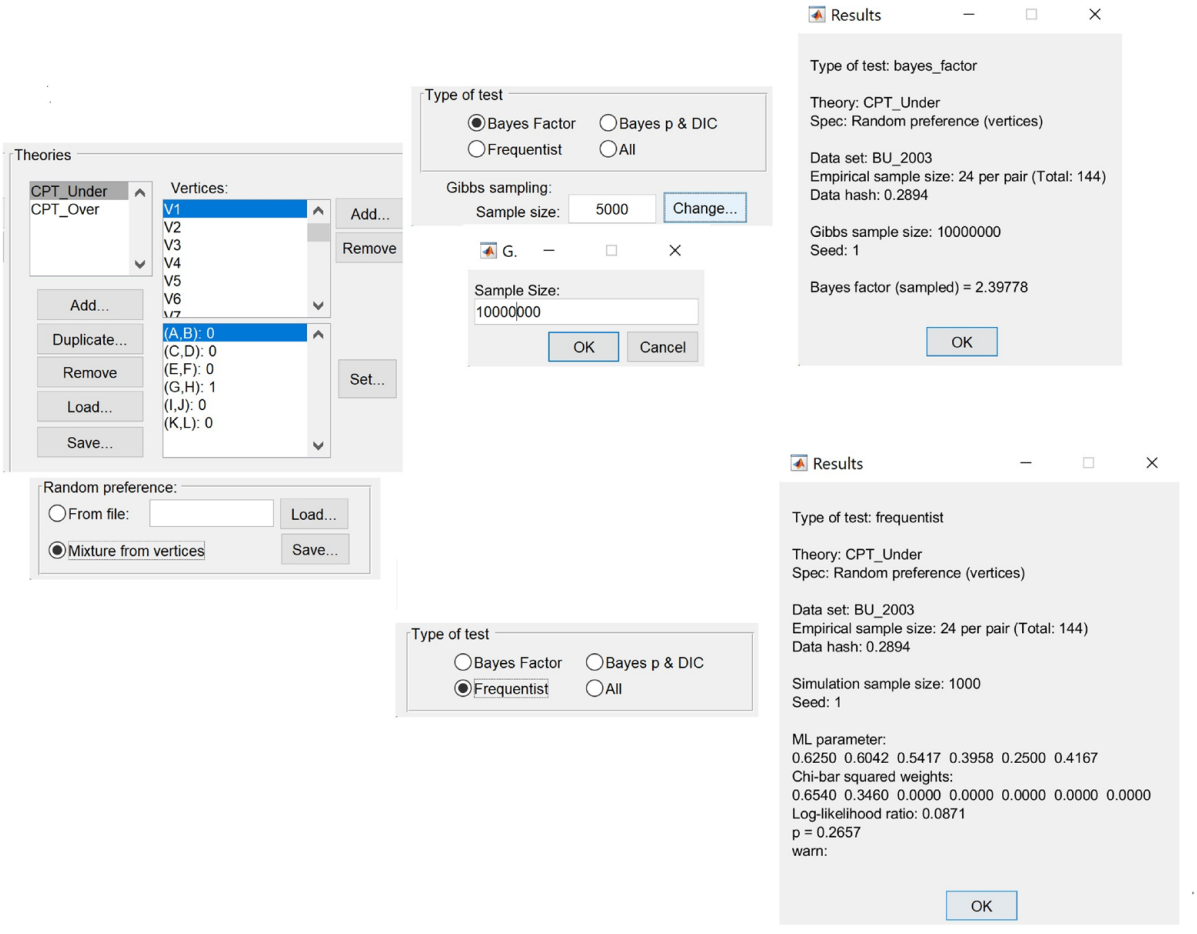


Figure 5. Quantitative Analysis of the Substance Model (Random Preference model) for Underweighting: Select the vertex description for CPT with underweighting, then in the "Random preference" section, select "Mixture from vertices" and select the desired test in the "Type of test" section of the GUI. For the Bayes factor, change the "Sample size" for "Gibbs Sampling" to a very large number before running the analysis. The right most screenshots show the QTest output for these analyses.

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press

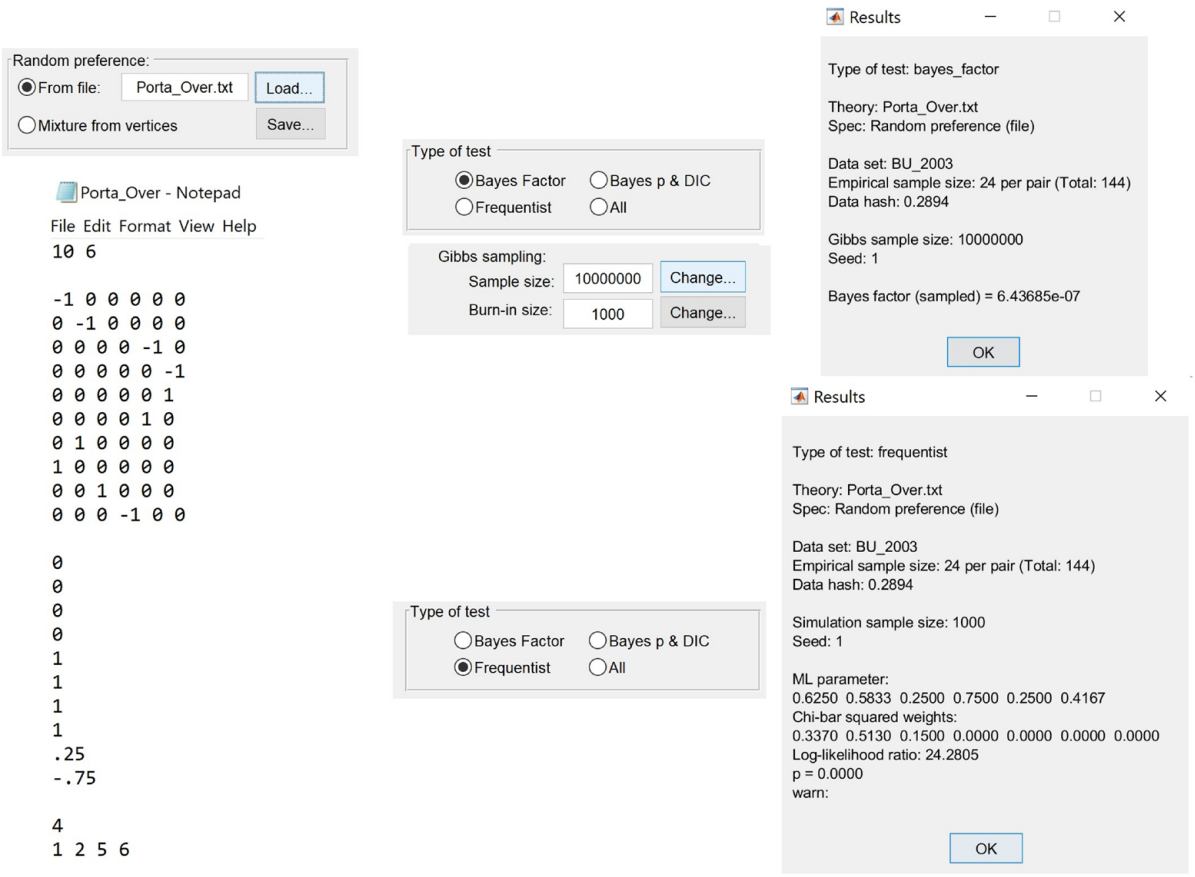


Figure 6. Quantitative Analysis of the Substance Model for Overweighting: In the "Random preference" section of the GUI, select "From File:" then navigate to the text file whose screenshot is shown on the lower left, and click "Load..." (see also Condition 10 in Section on Predictions). For the Bayes factor and the *p* value, proceed with the analysis in the same way as for the Substance Model for Underweighting.

References

Barron G., & Erev I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, *16*, 215–233. https://doi.org/10.1002/bdm.443

Barron G., & Ursino G. (2013). Underweighting rare events in experience based decisions: Beyond sample error. *Journal of Economic Psychology*, *39*, 278–286. https://doi.org/10.1016/j.joep.2013.09.002

Birnbaum M. H. (2008). New paradoxes of risky decision making. *Psychological Review*, *115*, 463–501. https://doi.org/10.1037/0033-295X.115.2.463

Blavatskyy P., & Pogrebna G. (2010). Models of stochastic choice and decision theories: Why both are important for analyzing decisions. *Journal of Applied Econometrics*, *25*, 963–986. https://doi.org/10.1002/jae.1116

Block H. D., & Marschak J. (1960). Random orderings and stochastic theories of responses. In I. Olkin, S. Ghurye, H. Hoeffding, W. Madow, & H. Mann (Eds.), *Contributions to probability and statistics* (pp. 97–132). Stanford, CA: Stanford University Press.

Brandstätter E., Gigerenzer G., & Hertwig R. (2006). The Priority Heuristic: Making choices without trade-offs. *Psychological Review*, *113*, 409–432. https://doi.org/10.1037/0033-295X.113.2.409

Davis-Stober C. P. (2009). Analysis of multinomial models under inequality constraints: Applications to measurement theory. *Journal of Mathematical Psychology*, *53*, 1–13. https://doi.org/10.1016/j.jmp.2008.08.003

Fox C. R., & Hadar L. (2006). "Decisions from experience" = sampling error + prospect theory: Reconsidering Hertwig, Barron, Weber, & Erev (2004). *Judgment and Decision Making*, *1*, 159–161.

Guo Y., & Regenwetter M. (2014). Quantitative tests of the Perceived Relative Argument Model: Comment on Loomes (2010). *Psychological Review*, *121*(4), 696–705. https://doi.org/10.1037/a0036095

Hadar L., & Fox C. R. (2009). Information asymmetry in decision from description versus decision from experience. *Judgment and Decision Making*, *4*, 317–225.

Harless D. W., & Camerer C. F. (1994). The predictive value of generalized expected utility theories. *Econometrica*, *62*, 1251–1289.

Hertwig R., & Erev I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Science*, *13*, 517–523. https://doi.org/10.1016/j.tics.2009.09.004

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press

Hey J. D. (2005). Why we should not be silent about noise. *Experimental Economics*, *8*, 325–345. https://doi.org/10.1007/s10683-005-5373-8

Hey J. D., & Orme C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica*, *62*, 1291–1326. https://doi.org/10.2307/2951750

Jeffreys H. (1961). *Theory of probability* (3rd Ed.). Oxford, UK: Oxford University Press.

Kass R. E., & Raftery A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. https://doi.org/10.1080/01621459.1995.10476572

Klugkist I., & Hoijtink H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, *51*, 6367–6379. https://doi.org/10.1016/j.csda.2007.01.024

Loomes G., & Sugden R. (1995). Incorporating a stochastic element into decision theories. *European Economic Review*, *39*, 641–648. https://doi.org/10.1016/0014-2921(94)00071-7

Luce R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York, NY: John Wiley.

Luce R. D. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Review of Psychology*, *46*, 1–27. https://doi.org/10.1146/annurev.ps.46.020195.000245

Luce R. D., & Suppes P. (1965). Preference, utility and subjective probability. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. III, pp. 249–410). New York, NY: Wiley.

Marley A., & Regenwetter M. (2017). Choice, preference, and utility: Probabilistic and deterministic representations. In W. Batchelder, H. Colonius, E. Dzhafarov, & I. Myung (Eds.), *New handbook of Mathematical Psychology: Foundations and methodology* (Vol. 1). Cambridge, UK: Cambridge University Press.

Myung J. I., Karabatsos G., & Iverson G. J. (2005). A Bayesian approach to testing decision making axioms. *Journal of Mathematical Psychology*, *49*, 205–225. https://doi.org/10.1016/j.jmp.2005.02.004

Regenwetter M., & Cavagnaro D. R. (2019). Tutorial on removing the shackles of regression analysis: How to stay true to your theory of binary response probabilities. *Psychological Methods*, *24*(2), 135–152.

Regenwetter M., Dana J., & Davis-Stober C. P. (2011). Transitivity of preferences. *Psychological Review*, *118*, 42–56. https://doi.org/10.1037/a0021150

Regenwetter M., & Davis-Stober C. P. (2012). Behavioral variability of choices versus structural inconsistency of preferences. *Psychological Review*, *119*(2), 408–416.

Regenwetter M., & Davis-Stober C. P. (2018). The role of independence and stationarity in probabilistic models of binary choice: A summary and review. *Journal of Behavioral Decision Making*, *31*, 100–114. https://doi.org/10.1002/bdm.2037

Regenwetter M., Davis-Stober C. P., Lim S. H., Guo Y., Popova A., Zwilling C., … Messner W. (2014). QTest: Quantitative testing of theories of binary choice. *Decision*, *1*(1), 2–34. https://doi.org/10.1037/dec0000007

Regenwetter M., & Robinson M. M. (2017). The construct-behavior gap in behavioral decision research: A challenge beyond replicability. *Psychological Review*, *124*(5), 533–550. https://doi.org/10.1037/rev0000067

Regenwetter M., & Robinson M. M. (2019). The construct-behavior gap revisited: Reply to Hertwig and Pleskac (2018). *Psychological Review*, *126*, 451–454.

Robinson M., Wang C., & Regenwetter M. (2019). *Are you an exception to Cumulative Prospect Theory?* [Manuscript submitted for publication].

Stott H. P. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty*, *32*, 101–130. https://doi.org/10.1007/s11166-006-8289-6

Tversky A., & Kahneman D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323. https://doi.org/10.1007/BF00122574

Wilcox N. (2008). Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison. In J. Cox & G. Harrison (Eds.), *Risk aversion in experiments* (Vol. 12, pp. 197–292). Bingley, UK: Emerald, Research in Experimental Economics.

Zwilling C. E., Cavagnaro D. R., Regenwetter M., Lim S. H., Fields B., & Zhang Y. (2019). QTest 2.1: Quantitative testing of theories of binary choice using Bayesian inference. *Journal of Mathematical Psychology*, *91*, 176–194. https://doi.org/10.1016/j.jmp.2019.05.002

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press

Appendix A

Hypothesis tests in Table 5 of Barron & Ursino (2013).

It is unclear what test Barron and Ursino (2013) used in their Table 5. Writing p_1 and p_2 for two probabilities to be compared, $\hat{p}_1$ and $\hat{p}_2$ for the observed proportions, Y_1 and Y_2 for the corresponding observed frequencies, $\hat{p}$ for the overall proportion, and n_1 and n_2 for the corresponding sample sizes, the standard hypothesis test for $H_0 : p_1 = p_2$ gives the z-score

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \textit{where} \quad \hat{p} = \frac{Y_1 + Y_2}{n_1 + n_2}.$$

Barron and Ursino (2013) appear to use n_1 = 24 and n_2, according to Column 6 of their Table 5 (reproduced here as n_1 and n_2 in the upper half of our Table A1). They report that the proportions significantly differ in Problems 8 and 9 ($p < .05$), a finding that we cannot reproduce. We find all two-sided p values to exceed .34 when using the sample sizes in the top half of the table. That analysis leaves out that the observed proportions were, in actuality, calculated from thousands of observations. Hence, it may be severely underpowered. It is also unclear how, for example, the proportion of choices of H in Decision Problem 1 (in their Column 5, and our Column 6) could be 0.64 when the sample size is 10: No matter how many people out of N = 10 chose H, this proportion cannot equal 0.64, even with rounding. It does not make sense to treat that proportion as arising from 10 observations. (Notice also how our reconstructed choice frequencies Y_1 and Y_2 are noninteger.) In our analyses of the probabilistic choice models, we rounded noninteger choice frequencies to the nearest integers.

Barron and Ursino (2013) reported that the proportions in their table were "*aggregated over subjects and over 400 trials*" (p. 283). The lower half of our Table A1 shows the results for the same six hypothesis tests, assuming that the proportions are aggregated over 24 people and 400 trials in Column 4 of their Table 5 and aggregated over N people and 400 trials in their Column 5, with N given in their Column 6 (see our Table for the corresponding values of n_1 and n_2.).

With that assumption our reconstructed choice frequencies Y_1 and Y_2 are, indeed, integers. With these reconstructed data, we find all but one p value to be minuscule. A downside of this analysis is that the iid assumptions underlying this test may not hold. Regardless, we are unable to reproduce the results of Barron and Ursino (2013), according to which the proportions significantly differ in Problems 8 and 9 at a 5% significance level. In the main text, we report analyses under both sets of reconstructed data given in Table A1.

Table A1. *Tests of Differences of Proportions based on Information from Table 5 of Barron and Ursino, 2013*

$\hat{p}_1$	$\hat{p}_2$	Y_1	Y_2	n_1	n_2	$\hat{p}$	z	p value
.63	.64	15.12	6.4	24	10	.6329	–0.06	.96
.6	.56	14.4	8.4	24	15	.5846	0.26	.81
.56	.63	13.44	12.6	24	20	.5918	–0.47	.64
.37	.47	8.88	7.99	24	17	.4115	–0.64	.52
.27	.43	6.48	4.73	24	11	.3203	–0.94	.34
.4	.33	9.6	3.63	24	11	.378	0.4	.69
$\hat{p}_1$	$\hat{p}_2$	Y_1	Y_2	n_1	n_2	$\hat{p}$	z	p value
.63	.64	6,048	2,560	9,600	4,000	.6329	–1.102	.27
.6	.56	5,760	3,360	9,600	6,000	.5846	4.9	$< 10^{-6}$
.56	.63	5,376	5,040	9,600	8,000	.5918	–9.4	$< 10^{-15}$
.37	.47	3,552	3,196	9,600	6,800	.4115	–12.8	$< 10^{-15}$
.27	.43	2,592	1,892	9,600	4,400	.3203	–18.8	$< 10^{-15}$
.4	.33	3,840	1,452	9,600	4,400	.378	7.9	$< 10^{-15}$

Note: p_1 and Y_1 are the proportions, respectively frequencies, of H choices in the full 2003 data. p_2 and Y_2 are the proportions, respectively frequencies, of H choices in the 2013 subset of data.

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press

Appendix B

Input Files for QTest

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3																
4	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
5	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0
9	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3																									
4	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
7	1	1	1	0	0	1	1	1	0	0	1	1	1	0	0	1	1	1	0	0	1	1	1	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	1	1	1	1	1
9	0	0	1	0	1	0	0	1	0	1	0	0	1	0	1	0	0	1	0	1	0	0	1	0	1

Figure B1. Input files for QTest giving the 16 preference patterns for CPT with overweighting and the 25 preference patterns for CPT with underweighting (see also Figure 2).

Appendix C

Proof

CPT with underweighting of small probabilities implies preference for Gamble R over Gamble S in Barron and Ursino's Experiment 2 (see Table C1). More precisely, we prove that, according to CPT with $1 < \delta \leq 2$ and $0 \leq \beta \leq 1$

$$w^{-}(.15)\times v(-20)+(1-w^{-}(.15))\times v\,(0) > w^{-}(1)\times v\,(-3)+(1-w^{-}(1))\times v\,(0). \tag{11}$$

Because $w^{-}(1) = 1$ for all values of δ, since $v(x) = -\lambda(-x)^{\beta}$, hence $v(0) = 0$ for all values of β, and, since, as we assumed, $\lambda = 1$ without loss in generality, we can rewrite Inequality 11 as

$$\underbrace{w^{-}(.15)\times\left(-\left(20^{\beta}\right)\right)}_{\text{Left-hand side of Inequality 12 (LHS}_{12}\text{)}} > \underbrace{-\left(3^{\beta}\right)}_{\text{Right-hand side of Inequality 12 (RHS}_{12}\text{)}}. \tag{12}$$

From now on, we treat both LHS_{12} and RHS_{12} as functions of β on the range [0, 1] and we consider these for every fixed value of δ on the range (1, 2], and hence also for each resulting (fixed) value of $w^{-}(.15)$.

We begin by noting that $w^{-}(p) > 0$ for any $p > 0$, which means that both LHS_{12} and RHS_{12} are power functions with a negative coefficient and an exponent β that varies on the range [0, 1]. It is known that power functions of this type are strictly decreasing and continuous, which makes both LHS_{12} and RHS_{12} strictly decreasing and continuous on $0 \leq \beta \leq 1$ for any fixed value of δ. Given these properties of the LHS_{12} and RHS_{12}, in order to prove Inequality 11, it is sufficient to prove the following two lemmas.

Lemma 1. The maximum value of LHS_{12}, for $0 \leq \beta \leq 1$ and $1 < \delta \leq 2$, is strictly greater than the maximum value of RHS_{12}, for $0 \leq \beta \leq 1$ and $1 < \delta \leq 2$.

Lemma 2. No matter what the values of $0 \leq \beta \leq 1$ and $1 < \delta \leq 2$ are, LHS_{12} cannot equal RHS_{12}.

Proof of Lemma 1:

Since $-(20^{\beta})$ and $-(3^{\beta})$ are both strictly decreasing functions of β, for $0 \leq \beta \leq 1$, they both attain their maximum values when $\beta = 0$, namely $-(20^{0}) = -(3^{0}) = -1$. Thus, Lemma 1 follows if we can show that $w^{-}(.15) \times (-1) > -1$. In other words, we need to show that $w^{-}(.15) < 1$, regardless of $1 < \delta \leq 2$. Instead, we even prove that $w^{-}(.15) < 0.15$ for $1 < \delta \leq 2$, and later use that result to prove Lemma 2.

First, we write out the inequality $w^{-}(.15) < 0.15$ using the formula for the weighting function with δ, namely

$$\frac{\left(\frac{15}{100}\right)^{\delta}}{\left(\left(\frac{15}{100}\right)^{\delta}+\left(\frac{85}{100}\right)^{\delta}\right)^{(1/\delta)}} < \frac{15}{100}. \tag{13}$$

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press

Reducing all component ratios in Inequality 13 gives

$$\frac{\left(\frac{3}{20}\right)^{\delta}}{\left(\left(\frac{3}{20}\right)^{\delta}+\left(\frac{17}{20}\right)^{\delta}\right)^{(1/\delta)}}<\frac{3}{20}. \tag{14}$$

Factoring out 20 on the left turns Inequality 14 into

$$\frac{1}{20^{\delta-1}}*\frac{3^{\delta}}{\left(3^{\delta}+17^{\delta}\right)^{(1/\delta)}}<\frac{3}{20}. \tag{15}$$

Factoring out 3 on the left of Inequality 15 yields

$$\frac{3^{\delta-1}}{20^{\delta-1}}\times\frac{3}{\left(3^{\delta}+17^{\delta}\right)^{(1/\delta)}}<\frac{3}{20}. \tag{16}$$

Multiplying both sides of Inequality 16 by $\left(\frac{3}{20}\right)^{1-\delta}$ produces

$$\frac{3}{\left(3^{\delta}+17^{\delta}\right)^{(1/\delta)}}<\frac{3}{20}\times\left(\frac{3}{20}\right)^{1-\delta}. \tag{17}$$

Exponentiating both sides of Inequality 17 by δ gives

$$\underbrace{\frac{3^{\delta}}{3^{\delta}+17^{\delta}}}_{LHS_{18}}<\underbrace{\left(\frac{3}{20}\right)^{(2-\delta)*\delta}}_{RHS_{18}}. \tag{18}$$

Note that LHS_{18} is strictly decreasing as a function of $1 < \delta \leq 2$, whereas the RHS_{18} is strictly increasing as a function of $1 < \delta \leq 2$. Generally, $LHS_{18} = RHS_{18}$ would imply that $\delta = 1$. Since CPT with underweighting requires $1 < \delta$, therefore, $LHS_{18} < RHS_{18}$ and, hence, $w^{-}(.15) < 0.15$ for $1 < \delta \leq 2$, which completes the proof of Lemma 1.

Proof of Lemma 2:

We use a proof by contradiction. Suppose that

$$w^{-}(.15)\times-\left(20^{\beta}\right)=-\left(3^{\beta}\right). \tag{19}$$

Rearranging and simplifying terms gives

$$w^{-}(.15)=\frac{3^{\beta}}{20^{\beta}}=0.15^{\beta} \tag{20}$$

We have already established that $w^{-}(.15) < 0.15$ for $1 < \delta \leq 2$. Since $0.15 \leq 0.15^{\beta}$ for $0 \leq \beta \leq 1$ (by a line of argument similar to those earlier), it follows that Equation 20 cannot hold. This completes the proof of Lemma 2.

Table C1. *Stimuli from Experiment 2 in Barron and Ursino (2013)*

Gamble S	Gamble R
x, p	x, p
–3, 1	–20, .15

Appendix D

Proof

CPT with overweighting of small probabilities implies preference for Gamble L over H in Decision Problem 7 and Gamble H over L in Decision Problem 8 (see Table 6). More precisely, we prove that, for Decision Problem 7 according to CPT with $0 < \gamma < 1$ and $0 \leq \beta \leq 1$

$$\begin{aligned} &w^{+}(.9)\times v(10)+\left(1-w^{+}(.9)\right)\times v(0)\\ &\quad<w^{+}(1)\times v(9)+\left(1-w^{+}(1)\right)\times v(0). \end{aligned} \tag{21}$$

Our rationale for Decision Problem 8 is the same because the losses case is a mirror image of the gains case.

Because $w^{+}(1) = 1$ for all values of γ, and since $v(x) = (x)^{\beta}$ hence $v(0) = 0$ for all values of β, we can rewrite Inequality 21 as

$$\underbrace{w^{+}(.9)\times\left(10^{\beta}\right)}_{\text{Left-hand side of Inequality 22}(LHS_{22})}<\underbrace{9^{\beta}}_{\text{Right-hand side of nequality 22}(RHS_{22})}. \tag{22}$$

From now on, we treat both LHS_{22} and RHS_{22} as functions of β on the range [0, 1] and we consider these for every fixed value of γ on the range (0, 1), and hence also for each resulting (fixed) value of $w^{+}(.9)$.

We begin by noting that $w^{+}(p) > 0$ for any $p > 0$, which means that both LHS_{22} and RHS_{22} are power functions with a positive coefficient and an exponent β that varies on the range [0, 1]. It is known that power functions of this type are strictly increasing and continuous, which makes both LHS_{22} and RHS_{22} strictly increasing and continuous on $0 \leq \beta \leq 1$ for any fixed value of γ. Given these properties of the LHS_{22} and RHS_{22}, in order to prove Inequality 21, it is sufficient to prove the following two lemmas.

Lemma 3. The maximum value of LHS_{22}, for $0 \leq \beta \leq 1$ and $0 < \gamma < 1$, is strictly smaller than the maximum value of RHS_{22}, for $0 \leq \beta \leq 1$ and $0 < \gamma < 1$.

Lemma 4. No matter what the values of $0 \leq \beta \leq 1$ and $0 < \gamma < 1$ are, LHS_{22} cannot equal RHS_{22}.

Proof of Lemma 3:

Since 10^{β} and 9^{β} are both strictly increasing functions of β, for $0 \leq \beta \leq 1$, they both attain their maximum values when $\beta = 1$, namely $10^{\beta} = 10$ and $9^{\beta} = 9$. Thus, Lemma 3 follows if we can show that $w^{+}(.9) \times 10 < 9$. In other words, we need to show that $w^{+}(.9) < .9$, regardless of $0 < \gamma < 1$.

First, we write out the inequality $w^{+}(.9) < 0.9$ using the formula for the weighting function with γ, namely

$$\frac{\left(\frac{90}{100}\right)^{\gamma}}{\left(\left(\frac{90}{100}\right)^{\gamma}+\left(\frac{10}{100}\right)^{\gamma}\right)^{(1/\gamma)}}<\frac{90}{100}. \tag{23}$$

Reducing all component ratios in Inequality 23 gives

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press

$$\frac{\left(\frac{9}{10}\right)^{\gamma}}{\left(\left(\frac{9}{10}\right)^{\gamma}+\left(\frac{1}{10}\right)^{\gamma}\right)^{(1/\gamma)}}<\frac{9}{10}. \tag{24}$$

Factoring out 10 on the left turns Inequality 24 into

$$\frac{1}{10^{\gamma-1}}*\frac{9^{\gamma}}{\left(9^{\gamma}+1^{\gamma}\right)^{(1/\gamma)}}<\frac{9}{10}. \tag{25}$$

Factoring out 9 on the left of Inequality 25 yields

$$\frac{9^{\gamma-1}}{10^{\gamma-1}}\times\frac{9}{\left(9^{\gamma}+1^{\gamma}\right)^{(1/\gamma)}}<\frac{9}{10}. \tag{26}$$

Multiplying both sides of Inequality 26 by $\left(\frac{9}{10}\right)^{1-\gamma}$ produces

$$\frac{9}{(9^{\gamma}+1^{\gamma})^{(1/\gamma)}}<\frac{9}{10}\times\left(\frac{9}{10}\right)^{1-\gamma}. \tag{27}$$

Exponentiating both sides of Inequality 27 by γ gives

$$\underbrace{\frac{9^{\gamma}}{9^{\gamma}+1^{\gamma}}}_{LHS_{28}}<\underbrace{\left(\frac{9}{10}\right)^{(2-\gamma)*\gamma}}_{RHS_{28}}. \tag{28}$$

Note that LHS_{28} is strictly increasing as a function of $0<\gamma<1$, whereas the RHS_{28} is strictly decreasing as a function of $0<\gamma<1$. Generally, $max_{\gamma}\,(LHS_{28}) = min_{\gamma}\,(RHS_{28}) = 0.9$ would imply that $\gamma = 1$. Since CPT with overweighting requires $\gamma < 1$, therefore, $LHS_{28} < RHS_{28}$ and, hence, $w^{+}\,(.9) < 0.9$ for $0<\gamma<1$, which completes the proof for Lemma 3.

Proof of Lemma 4:

We use a proof by contradiction. Suppose that

$$w^{+}(.9)\times\left(10^{\beta}\right)=9^{\beta}. \tag{29}$$

Rearranging and simplifying terms gives

$$w^{+}(.9)=\frac{9^{\beta}}{10^{\beta}}=0.9^{\beta} \tag{30}$$

We have already established that $w^{+}\,(.9) < 0.9$ for $0 < \gamma < 1$. Since $0.9 \le 0.9^{\beta}$ for $0 \le \beta \le 1$ (by a line of argument similar to those earlier), it follows that Equation 30 cannot hold. This completes the proof of Lemma 4.

Table D1. *Decision Problems 7 and 8 from Barron and Erev (2003)*

Decision Problem	Gamble H	Gamble L
	x, p	x, p
7	10, .9	9, 1
8	–10, .9	–9, 1

https://doi.org/10.1017/sjp.2019.50 Published online by Cambridge University Press