Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Preregistered Falsification Test of the Decision by Sampling Model and Rank-Order Effect

Mattias Forsgren, Lars Frimanson, Peter Juslin

To cite this article:
Mattias Forsgren, Lars Frimanson, Peter Juslin (2025) A Preregistered Falsification Test of the Decision by Sampling Model and Rank-Order Effect. Management Science 71(10):8218-8229. https://doi.org/10.1287/mnsc.2022.03611

https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 71, No. 10, October 2025, pp. 8218–8229
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

A Preregistered Falsification Test of the Decision by Sampling Model and Rank-Order Effect

Mattias Forsgren,[a,*] Lars Frimanson,[a,b] Peter Juslin[a]

[a] Department of Psychology, Uppsala University, 751 42 Uppsala, Sweden; [b] Department of Business Studies, Uppsala University, 751 20 Uppsala, Sweden

*Corresponding author

Contact: mattias.forsgren@psyk.uu.se, https://orcid.org/0000-0003-0394-1626 (MF); lars.frimanson@fek.uu.se, https://orcid.org/0000-0002-5978-9005 (LF); peter.juslin@psyk.uu.se, https://orcid.org/0000-0001-9594-2153 (PJ)

Received: November 22, 2022
Revised: August 17, 2023; January 4, 2024
Accepted: January 10, 2024
Published Online in Articles in Advance: January 21, 2025

https://doi.org/10.1287/mnsc.2022.03611

Abstract. Many social scientists have assumed that people's preferences can be described by stable and coherent "utility" functions. This notion of stable utility functions has been challenged by cognitive psychologists who suggest that preferences are malleable and constructed in the moment, but neither camp has explained how the subjective valuations underpinning preferences arise. One influential attempt to do so is the Decision by Sampling (DbS) model, which suggests that a quantitative attribute's (e.g., money sum's) subjective value is its rank order in a momentarily activated memory sample. DbS thus implies that manipulating the recently experienced attribute distribution should change people's subsequent valuations of that attribute: for example, from the typically assumed concave shape of the utility function to a convex shape. However, recent studies have pointed out methodological concerns in the evidence previously thought to support this prediction (and thus, DbS). In this preregistered study, we replicate the previous paradigm but address the methodological concerns to test if such a "rank-order" manipulation does change valuations. We derive qualitative predictions from DbS to verify that our conditions yield distinct predictions. We find strong evidence against the DbS's prediction that a "rank-order" manipulation changes what options the participants select and how strongly they prefer the options. We also find extreme evidence in favor of a contextualization effect, implying that people value formally identical gambles differently depending on whether they cue a real-life setting or not. Although we encourage replication by independent laboratories, these results suggest that the DbS is falsified for this binary choice task.

History: Accepted by Yuval Rottenstreich, behavioral economics and decision analysis.

Funding: This research was funded by the Marcus and Amalia Wallenberg Foundation [Grant MAW 2016.0132] and the Swedish Research School of Management and IT.

Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mnsc.2022.03611.

Keywords: decision by sampling • risky choice • subjective utility • open science movement

1. Introduction

Traditional accounts based on the notion of "utility" assume that my preferences—what I like—are based on subjective values of the outcomes of various choices given a priori (Bentham 1789, Fisher 1892, von Neumann and Morgenstern 1944). These values are stable and logically coherent, and making a choice consists of identifying which option has the highest value ("utility maximization"). Cognitive psychologists (Tversky 1975, Tversky and Kahneman 1992, Gigerenzer 2004, Birnbaum and Martin 2012) criticize this account of the decision-making process and argue that preferences are unstable. These cognitive theories are not primarily accounts of the valuations of the attributes that options consist of but of the operations performed on those valuations to form preferences. In other words, they explicate the reasons for the incoherence and instability of preferences, but—just as with expected utility theory—they typically offer no explanation of the perceived value of the quantitative attributes themselves. In a nutshell, what determines if a reward of $5 is perceived as a small fortune or as pocket money or if a probability of 50% of something is a "high" probability or not?

An attempt to explain where the valuations themselves come from (and in extension, the instability of preferences) is the Decision by Sampling (DbS) model (Stewart et al. 2006). The DbS model draws inspiration from range-frequency theory (Parducci 1965), which posits that valuations of a perceptual stimulus along some dimension (e.g., the perceived intensity of a light source) are determined by mentally partitioning the range of recently observed values into equal-sized subranges (the range principle) but then adjusting the end points of these subranges such that about equal proportions of the recently observed values fall within each (the frequency principle). Thus, valuations are perceived on an ordinal scale determined by the range and frequency of the absolute values of stimuli in the immediate context. The DbS model is inspired by the frequency principle but does not include the range principle[1] of range-frequency theory, and it expands it to new stimuli (e.g., money and probability) by claiming that the subjective valuation of an option is based on the rank order of the values of its attributes (e.g., the reward and probability of winning in the case of gambles) within a set of exemplar values sampled from memory (Stewart et al. 2006). What exemplar values are sampled depends on recent experience; stimuli immediately preceding the choice can sway the decision maker one way or the other. For example, if they have been exposed to low-magnitude stimuli, low magnitudes are more likely to be sampled from memory, and a target of moderate magnitude will have a greater subjective value. Conversely, if exposed to high magnitudes, moderate targets will lose more comparisons with sampled memories and have lower subjective value (Stewart et al. 2015). Thus, the moderate target will be valued higher in the first circumstance than in the second. In this way, instability of preferences can be explained as a result of the environment in opposition to the previously mentioned (Tversky 1975, Tversky and Kahneman 1992, Gigerenzer 2004, Birnbaum and Martin 2012) "mentalistic" accounts that focus on the information processing. We will refer to this manipulation of subjective values by varying the preceding stimuli as the "rank-order effect."

1.1. Evidence for the Rank-Order Effect

Research on the DbS model has taken several approaches. One stream of evidence has focused on observational data consistent with the DbS model (e.g., Stewart et al. 2006, Ungemach et al. 2011).[2] In another vein, Noguchi and Stewart (2018) extended the DbS model with an evidence accumulation mechanism and modified the pairwise comparisons to be probabilistic rather than deterministic. This version could explain multialternative choice data from an experiment on par with two previous theories (Roe et al. 2001, Trueblood et al. 2014). Here, we will focus on research on a causal rank-order effect because it is a key prediction of the theory. Canic (2016) reports two unpublished experiments where there appears to be[3] a rank-order effect on "attractiveness ratings" of single gambles on a seven-point Likert scale, but previous work has discussed similar results in terms of more superficial effects on the scale use (Parducci and Perrett 1971) rather than the valuations of stimuli per se. A number of experiments have instead used paradigms where participants make a number of pairwise choices. However, they have manipulated both rank and range simultaneously, thus confounding any potential rank-order effect. For example, participants in study 1b in Ungemach et al. (2011) were exposed to either (i) the prices £0.19 and £3.80 or (ii) the prices £0.74 and £1.07 before making a choice between a 55% chance of £0.50 or a 15% chance of £1.50 (else £0). The prices in (ii) were in between the amounts appearing in the test options, thus increasing the relative rank of £1.50 vis-à-vis £0.50 in the postulated memory sample (the rank-order effect), whereas the prices in (i) were higher and lower than the test option attribute values and thereby, not affecting their relative ranks (no rank-order effect). However, £1.50 also had a considerably higher position in the range in (ii) than in (i). It is therefore not clear that the significant effect of being exposed to (i) with range [£0.19, £3.80] or (ii) with range [£0.74, £1.07] was caused by the difference in the rank or by the different position in the range in the latter context. Frydman and Jin (2021) find results consistent with the DbS model in one range-confounded experiment and one unconfounded experiment, presenting their results as evidence for "efficient coding" (Laughlin 1981) of symbolic, numeric information too. Canic (2016) reports five unpublished range-confounded pairwise choice experiments where the results appear to be mixed; two of five 95% confidence intervals on the odds ratio do not overlap zero. Lastly, Canic (2016) does not find effects on positive affect for an achieved reward under range-confounded designs.[4]

1.2. Three Methodological Concerns

Recent work has highlighted three other methodological concerns regarding evidence for both the rank-order effect and the DbS model in the context of binary risky choice tasks. A first concern is that Walasek and Stewart (2015, 2019) and Rakow et al. (2020) estimate the effects of encountered stimulus magnitudes by fitting computational models to the data and treating the parameters as outcome variables. However, this work was revisited by the same research group (Alempaki et al. 2019, Stewart et al. 2020, Walasek and Stewart 2020) who showed that the parameters of several such models, including prospect theory (Kahneman and Tversky 1979), cannot be reliably recovered. They thus repealed their previous conclusion that Stewart et al. (2015) provided evidence for Decision by Sampling.[5] More generally, "model misspecification" (fitting a

model that does not match the "true" model that generated the data) biases the parameter estimates (Nassar and Gold 2013). This issue may be alleviated by testing statistics of the raw data (such as group mean differences). Similarly, we may visualize people's preferences (e.g., Stewart et al. 2015) using statistics of the data rather than fitted models.

A second concern is that some test stimuli in experiments supporting the rank-order effect in binary risky choice tasks (Walasek and Stewart 2015, Rakow et al. 2020)—even when unconfounded by range (Walasek and Stewart 2019)—violate measurement invariance (Meredith 1993), as highlighted by André and de Langhe (2021b). Measurement invariance refers to having the same dependent measurement(s) for all conditions. If conditions are evaluated on different measurements, we cannot tell if any group differences are because of the experimental manipulation or differences between the measurements. Although previous experiments vary in their details, participants in one condition may, for example, be tested on a stimulus set of predominantly low magnitudes, whereas participants in another condition are tested on a different stimulus set of predominantly high magnitudes, with a smaller number of trials being identical for both groups ("common gambles"). Only the common gambles satisfy measurement invariance why, to clearly see if an effect exists, we should focus on those trials only.

A third concern is that the DbS model only predicts that the rank-order effect should emerge *after* training. Previous studies (Alempaki et al. 2019, André and de Langhe 2021b) seek to address the first and second challenges by picking out the common gambles from data sets and analyzing them separately, finding null or contrary results. Walasek et al. (2021) reanalyzed the common gambles in Walasek and Stewart (2015) and argued that the mean proportions of accepted/rejected monetary gambles between conditions were qualitatively consistent with the DbS model. In contrast, André and de Langhe (2021a) emphasized that a viable model should be consistent with data patterns both across different conditions as well as within them. They argued that in the same common gambles, the accept/reject patterns within conditions were qualitatively inconsistent with the DbS model. However, because the DbS model's proposition is that preferences are constructed by sampling from recent experience, any effect on common gambles may be quite small (and thus, difficult to detect) without sufficient training. One way to ensure sufficient training is to present participants with a dedicated training phase followed by a dedicated test phase.

In sum, one may perform a falsification test of the DbS model that addresses these concerns by making the following design choices: testing (i) group differences in summary statistics of (ii) common gambles from (iii) a test phase preceded by a training phase. Ideally, we would like to (iv) visualize any rank-order effect by model-free plotting of how the preferences differ between test choices. In the present preregistered study, we perform such a test, fulfilling (i)–(iv) and using a large (n = 1,000) sample. We extend previous work by checking if any rank-order effect is attenuated by contextualizing the choices as relating to a specific everyday setting as opposed to decontextualized monetary gambles. Our reasoning here is that the DbS model's reliance on recently encountered stimuli might be a "default" that the mind uses in unfamiliar settings that provide insufficient cues to elicit a more informed strategy. Over time, we might learn to use, for example, heuristics, declarative decision rules, or crystallized knowledge from repeated experience of whether our decisions turned out well. A more representative design (Brunswik 1956), which contextualizes the choices, might elicit the use of such learned tools, leading to more stable preferences for individuals who have substantial experience of said context. In that case, the rank-order effect, if it exists, might not generalize to all everyday environments.

1.3. Purpose of This Experiment

The purpose is to present an experiment that addresses the three methodological concerns above and is sufficient to observe the rank-order effect if it exists. Doing so would support the predictions of DbS, whereas not doing so would support their rejection, at least for this binary choice task. If a rank-order effect exists, we expect it to be attenuated when the task is recontextualized to cue participants' experience.

2. Method

2.1. Overview

A large sample of participants made a number of risky choices between two options, each defined by a monetary value and a probability of receiving that value (e.g., "reward of \$4.00 with 60% probability" versus "reward of \$3.20 with 77% probability"). The complement outcome was always to not receive any reward. Participants first completed a training phase where the monetary values of the risky choices were manipulated between conditions; either the risky choices primarily featured lower monetary values, or they primarily featured higher monetary values. They then quietly and seamlessly proceeded to a test phase where all conditions were presented with the same risky choices ("common gambles"). In these trials, participants also stated their relative preference strength for the option they chose vis-à-vis the alternative. The task was presented as a choice between either monetary gambles or prospective jobs on the crowd-working website that we recruited participants from. This context was manipulated between conditions. Participants were recruited

from two demographics: one with little experience of completing jobs on the crowd-working website and another with substantial experience. Below, we provide details of the method.

2.2. Transparency and Openness

The design, the hypotheses, and all analyses presented below were preregistered at https://doi.org/10.17605/OSF.IO/V5T98. Data were analyzed using JASP (JASP Team 2022) and R version 4.1.1 (R Core Team 2021) with the packages tidyverse (Wickham et al. 2019), readxl (Wickham and Bryan 2019), and openxlsx (Walker and Braglia 2017). All data, code, and research materials are available at the Open Science Framework (OSF) and can be accessed at https://doi.org/10.17605/OSF.IO/T5XQ4. We report below how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. We preregistered the two alternative hypotheses articulated above—that finding an effect would support the DbS model and that not finding an effect would be inconsistent with the DbS model—and the hypothesis that any rank-order effect will be attenuated when the risky choices are contextualized as crowd-working jobs and participants have substantial experience of such.

2.3. Participants and Exclusion Restrictions

We set out to recruit 1,000 participants on Amazon Mechanical Turk (AMT) (Mason and Suri 2012), a crowd-working platform previously used to investigate the rank-order effect (Walasek and Stewart 2015). Because we did not know what effect size to expect, we selected a sample size that was substantially greater than those of previous experiments (max $n = 429$) (Walasek and Stewart 2015) and divisible by our number of conditions. The 1,000-participants stopping rule was automated by AMT, which counted the number of participants who accepted to participate and collected their reward using a code provided at the end of the experiment. Participants were recruited from two demographics (low experience and high experience in performing jobs on AMT) (see Section 2.5) by posting two identical data collections, each visible to only one of the groups. Twenty-one participants from the low-experience demographic seem to have completed the experiment without collecting their reward, and one participant from the high-experience demographic presumably collected the reward fraudulently without completing the experiment. This yielded a total of 1,020 valid participants. The exclusion restrictions for our study were that a participant needed to be located in the United States (as indicated by their internet protocol address), have an AMT job approval rating[6] of at least 95% (to increase data quality) (Peer et al. 2014), and have successfully completed either fewer than 50 jobs on AMT (low-experience demographic) or more than 5,000 jobs on AMT (high-experience demographic) as indicated by the participant's AMT user account's tallied completions. We terminated the participation of any participant who failed an attention or sincerity check (see Section 2.6) and blocked their re-entry using cookies. Lastly, we excluded any participants who, when asked, indicated that they had completed substantially more or fewer AMT jobs, depending on demographic, than required by our exclusion restriction. Such a discrepancy could arise from individuals using someone else's AMT account or having set up several accounts. We set the cut-off points to indicating that they had performed more than 100 or fewer than 1,000 because we do not expect individuals to have perfect recollection of how many AMT jobs they have completed, particularly for very large numbers. After applying this exclusion restriction, we obtained 300 low-experience participants and 428 high-experience participants for a final total of 728 participants. Participants were rewarded with $1.50 for an anticipated 10 minutes of work, which is above the U.S. minimum hourly wage.

2.4. Stimuli

We selected stimuli such that we (i) hold the range of presented monetary values and probabilities constant between conditions (i.e., we isolate the rank-order effect); (ii) hold training phase probabilities constant between conditions, only manipulating the monetary values; (iii) minimize the impact of the training on the valuation of the probabilities used in the test phase, thus maximizing the impact of our manipulation of monetary values; (iv) ensure that the qualitative predictions of DbS, given our stimuli, hold regardless of whether valuations of monetary values and probabilities are integrated using addition, averaging, or multiplication; and (v) ensure that the predictions hold regardless of whether participants encode the distribution of stimulus magnitudes (as in the DbS model) or only the frequencies of unique stimuli. Because one of our manipulations involved presenting the task as choosing between prospective AMT jobs, we (vi) tried to ensure that the ranges of monetary values and probabilities were representative of that context.

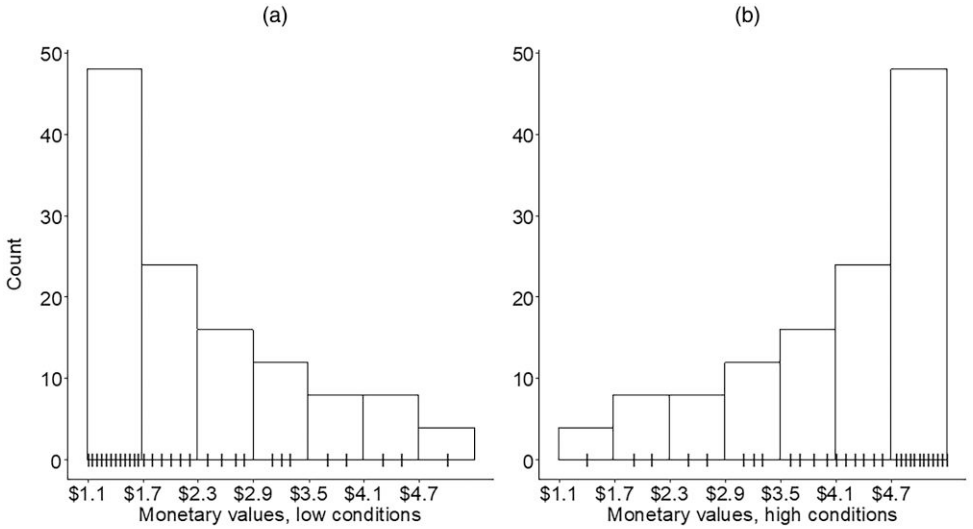
We verified that the resulting training and test trials yield distinct qualitative predictions by the DbS model for the low and high training conditions (see Section 2.5) using a simulation. Specifically, the low-training values conditions should tend to select the risky options more often and have higher mean relative preference strength for the risky options compared with the high training conditions. Graphs of cumulated relative preference strength should tend toward concave for low training and convex for high training (Figure 1 in Online Appendix A). See Online Appendix A for details on the procedure for stimulus selection and the simulation.

2.5. Design

The experiment was a 2 × 2 × 2 between-subjects factorial design with the variables *Training* (low values/high values), *Context* (gambles/AMT jobs), and *Experience* (low experience/high experience). Only the first two were manipulated, whereas the third was demographic. Training consisted of a training phase with stimuli of either mostly low or mostly high monetary values (see Figure 1). Context consisted of presenting the task as a choice between either monetary gambles or what AMT job the participant would prefer to perform.

In the gambles context, options were specified by their potential monetary outcomes (in dollars) and the probability of that outcome. In the AMT context, options were specified by their potential monetary reward for completing the job (in dollars) and the percentage of completed instances of that job which were accepted by the "requester" (approval rates) (see Figure 2). On AMT, "requesters" can list jobs for users to perform and can choose if they pay ("approve") the user after the job is completed. Whether a job will be approved or not is often experienced as nontransparent by users (McInnis et al. 2016), making it appear stochastic. We specified in the instructions that the options were based on real-world AMT jobs of about 10 minutes and that the approval rates that we presented had been cleared from all predictable sources of nonapproval (such as the job not being fully completed). This way, we stated to the participants, the supplied approval rate could be thought of as the probability of a correctly completed job being approved. The complementary outcome of every option across conditions was receiving nothing, which was specified in the instructions. Participants were divided into a low-experience sample and a high-experience sample by publishing two data collections on AMT: one that could only be accessed by accounts

Figure 1. Monetary Values in the Training Phase and Test Phase Options

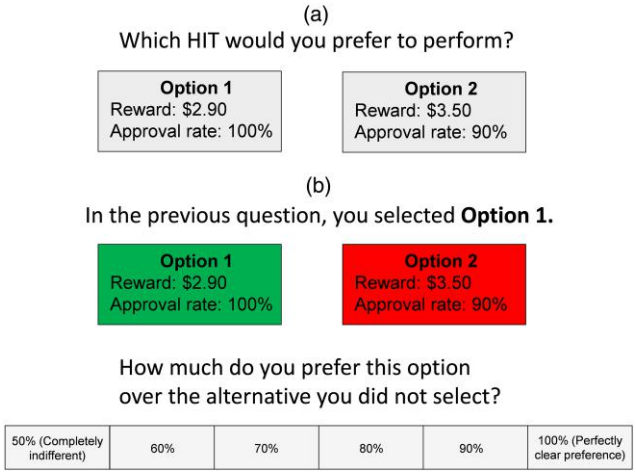


(c)

	Safe option	Risky option
Test trial 1	$1.1 × 100%	$1.7 × 90%
Test trial 2	$1.7 × 100%	$2.3 × 90%
Test trial 3	$2.3 × 100%	$2.9 × 90%
Test trial 4	$2.9 × 100%	$3.5 × 90%
Test trial 5	$3.5 × 100%	$4.1 × 90%
Test trial 6	$4.1 × 100%	$4.7 × 90%

Notes. (a) Monetary values used to generate the training stimuli for the low conditions. (b) Monetary values used to generate the training stimuli for the high conditions. Vertical bars on the horizontal axes indicate unique values. Every unique value was presented four times. Thus, the manipulation holds both if we assume frequency encoding (as in DbS) or if we assume that only unique occurrences are encoded. (c) Options of the test phase choices. These were common to all conditions.

Figure 2. (Color online) Cartoons of Experiment Task



Notes. (a) Realistic cartoon of test trial with AMT jobs context prior to selecting an alternative. The training phase trials were visually identical. (b) Realistic cartoon of test trial after an option has been selected. A preference judgement is queried using a half-range scale. The selected option is colored green, and the other option is colored red. The training phase did not include any preference judgements. We use cartoons instead of a print screen picture to increase readability. "HIT" (Human Intelligence Task) is what AMT jobs are called on the crowd-working website.

that had completed fewer than 50 accepted jobs on AMT and one that could only be accessed by accounts that had completed more than 5,000 accepted jobs. The dependent variables of interest were (1) the proportion of risky options chosen and (2) the average preference strengths for the risky relative to the safe option.

2.6. Procedure

Before starting the experiment, the participants had to indicate their informed consent. The participants then received written instructions, where we also thanked them to increase completion rates (Berinsky et al. 2016), and they were shown an example trial. The instructions and example were adapted depending on context (see https://doi.org/10.17605/OSF.IO/T5XQ4). We then presented a "screener question" (Berinsky et al. 2016) to check that participants were reading the instructions sincerely. They then carried out the training phase of 60 trials before seamlessly moving on to the test phase of 6 trials.[7] After each test phase trial, the selected (nonselected) option was colored green (red), and the participants got to indicate their strength of preference for the chosen option on a six-point half-range scale from "50%, completely indifferent" to "100%, perfectly clear preference." They were then asked two questions related to a different study before being moved to a demographics phase. It contained two attention/sincerity checks: what number the participants would call if they had an emergency (to check that they were in the United States) (Agley et al. 2022) and their age and year of birth (not asked consecutively). If their age and year of birth did not match, they were concluded to be insincere. This section also contained questions about their gender and how many jobs they had performed on AMT throughout their life (which was used to exclude participants of an inappropriate experience level; see above). The median completion time was 10.4 minutes (interquartile range = 5.2).

2.7. Analyses

The effects of *Training*, *Context*, and *Experience* on the chosen option and the preference strength were estimated using three-way Bayesian analyses of variance (BANOVAs) with default JASP priors.[8] As mentioned, previous

Table 1. Descriptive Statistics by Condition

Training	Context	Experience	Mean	Mean by training	SD	95% Credible interval
Proportion of risky options selected						
High	AMT jobs	High	0.451	0.442	0.443	0.365–0.538
		Low	0.594		0.411	0.496–0.693
	Gambles	High	0.457		0.407	0.377–0.537
		Low	0.278		0.374	0.195–0.362
Low	AMT jobs	High	0.500	0.452	0.424	0.417–0.583
		Low	0.639		0.392	0.550–0.728
	Gambles	High	0.391		0.377	0.323–0.458
		Low	0.296		0.367	0.211–0.380
Strength of preference for risky options						
High	AMT jobs	High	42.006	42.590	36.222	34.927–49.086
		Low	54.976		32.689	47.123–62.829
	Gambles	High	45.132		34.027	38.415–51.849
		Low	29.283		33.163	21.855–36.711
Low	AMT jobs	High	48.186	43.373	33.108	41.683–54.689
		Low	60.065		33.432	52.477–67.653
	Gambles	High	37.240		29.391	31.972–42.509
		Low	29.667		32.504	22.188–37.145

Notes. Descriptive statistics for each condition by dependent measure are shown. Mean by training is the marginal mean proportion and strength of preference, respectively, for high and low training conditions. SD, standard deviation.

Table 2. Bayesian Three-Way ANOVAs on the Two Dependent Measures

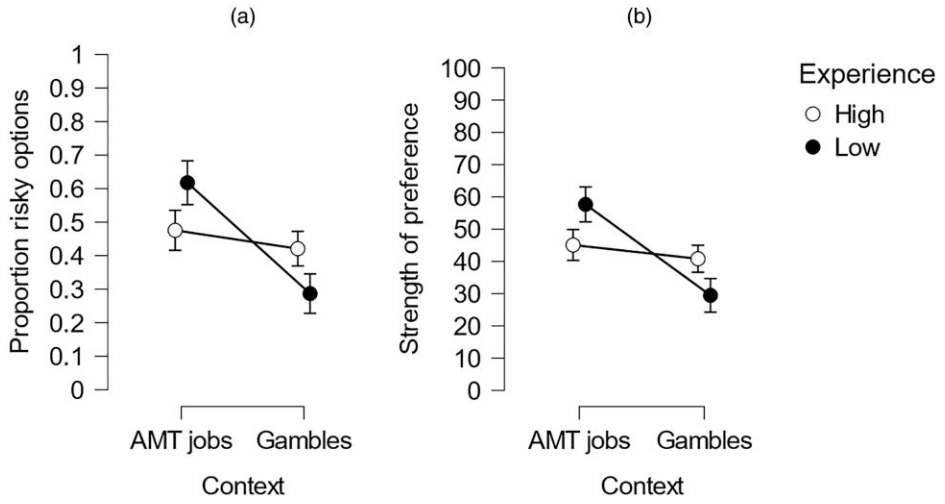
Independent variable	P(incl)	P(excl)	P(incl | data)	P(excl | data)	Incl. BF	Partial η^2
Proportion of risky options selected						
Training	0.263	0.263	0.073	0.892	0.082	<0.001
Context	0.263	0.263	0.005	1.832 e-8	287,899.499	0.052
Training × Context	0.263	0.263	0.023	0.084	0.270	0.002
Experience	0.263	0.263	3.923 e-4	0.005	0.079	<0.001
Training × Experience	0.263	0.263	0.014	0.092	0.156	<0.001
Context × Experience	0.263	0.263	0.994	3.981 e-4	2,497.010	0.028
Training × Context × Experience	0.053	0.053	5.221 e-4	0.003	0.177	<0.001
Strength of preference for risky options						
Training	0.263	0.263	0.067	0.845	0.079	<0.001
Context	0.263	0.263	0.002	4.201 e-9	448,009.512	0.054
Training × Context	0.263	0.263	0.074	0.077	0.956	0.005
Experience	0.263	0.263	1.796 e-4	0.002	0.099	<0.001
Training × Experience	0.263	0.263	0.020	0.131	0.152	<0.001
Context × Experience	0.263	0.263	0.995	1.828 e-4	5,441.252	0.032
Training × Context × Experience	0.053	0.053	0.003	0.009	0.363	0.001

Notes. "Incl. BF" is an abbreviation for inclusion Bayes factor. The inclusion Bayes factor compares models that contain the independent variable with equivalent models stripped of the independent variable. It is thus the relative evidence for including the measure in the final model compared with not including it. An inclusion Bayes factor smaller than one indicates evidence against an effect of the variable, and values greater than one indicate evidence in favor of an effect of the variable. An inclusion Bayes factor close to one indicates inconclusive evidence. Higher-order interactions are excluded. This analysis was suggested by Sebastiaan Mathôt. Partial η^2 cannot be calculated for the Bayesian model that we implement in JASP. We, therefore, calculate them separately using type III sum of squares on a frequentist ANOVA model (see the materials in the OSF repository at https://doi.org/10.17605/OSF.IO/T5XQ4). Inclusion of an effect size measure was suggested by a reviewer and was thus not preregistered. That the Bayes factors in our analyses typically deviate substantially from one, even in cases where the effect size is small, suggests a high statistical sensitivity to detect possible effects.

studies (e.g., Stewart et al. 2015) have also visualized the effects by showing how the value functions estimated by their models change shape between conditions. We want to do this too but in an as assumption-free way as possible, without relying on computational models with unrecoverable parameters (Alempaki et al. 2019, Stewart et al. 2020, Walasek and Stewart 2020) or model misspecification bias (Nassar and Gold 2013).

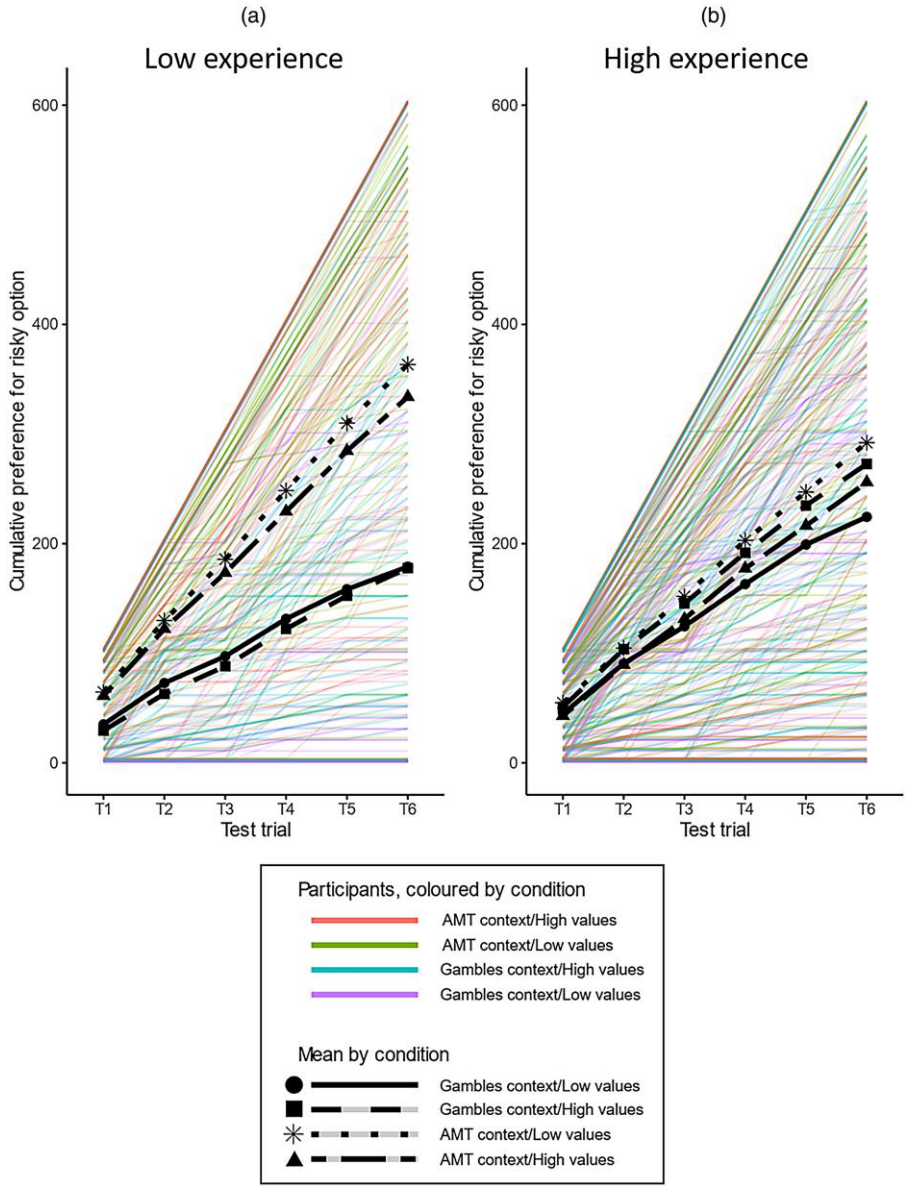
After each test choice, we collected assessments of the relative strength of the preference for the chosen option vis-à-vis the alternative from "50%, completely indifferent" to "100%, perfectly clear preference" (see panel (b) of Figure 2). We shifted these scores such that a perfect preference for the safer option is 0 and a perfect preference for the riskier option is 100. In what follows, the preference scores thus measure preference for the risky option. We then plot these scores for each test choice by participant so that we get individual-level graphs of the empirical relative preferences (analogous to the simulated relative valuations in Figure 1

Figure 3. Effects of Context and Experience



Notes. Means and 95% credible intervals (indicated by whiskers) are shown for the contrasts for which there is evidence in favor of an effect. (a) Proportion of risky options selected. (b) Mean strength of the preference for the risky options.

Figure 4. (Color online) Relative Preference for the Risky Option for All Test Trials by Participant and Condition



Notes. Graphs of cumulated relative preference for the risky option across test trials are shown. We rescaled the preference judgements (panel (b) of Figure 2) to [0, 100] such that 0 corresponds to a perfectly clear preference for the safe option and 100 corresponds to a perfectly clear preference for the risky option. The graphs display these rescaled judgements cumulated over the six test trials. Six hundred thus corresponds to a perfectly clear preference for the risky option in every test trial. Zero corresponds to a perfectly clear preference for the safe option in every test trial. As per our preregistration, we display both individual- and group-level data. Thin colored lines are individual participants, with color by condition. Thick black lines are mean relative preference, with line type by condition. See Online Appendix A for details. The only clear difference in the graphs between conditions is that for low-experience participants (in panel (a)), the AMT context conditions (asterisks and triangles) have stronger preferences for the risky options than gambles context conditions (circles and squares).

in Online Appendix A). We repeat this for average preferences by condition. These graphs are analogous to the value functions presented in previous studies.

3. Results

See Table 1 for descriptive statistics. As per our preregistration, we present three-way Bayesian ANOVAs on our two dependent measures: (i) the proportion of risky options selected and (ii) the strength of preference for the risky option. Contrary to the predictions of the Decision by Sampling model, we find strong evidence against a main effect of *Training*; participants who were exposed to "low" and "high" values in the training phase were equally likely to select risky options (*Inclusion Bayes factor (BF)* = 0.082) and had equally strong preferences for the risky options (*Inclusion BF* = 0.079) (see Table 2). We also find moderate evidence against all interactions

between *Training* and the other variables for selected options (*Inclusion BF* = [0.156, 0.270]). For preference strengths, there is moderate evidence against an interaction of *Training* and *Experience* (*Inclusion BF* = 0.152) and the three-way interaction (*Inclusion BF* = 0.363), but there is inconclusive evidence (*Inclusion BF* = 0.956) against an interaction between *Training* and *Context*. In other words, being exposed to "low" or "high" values affects neither choice nor strength of preferences for the risky options. Our most important result is thus that there does not seem to exist any rank-order effect in this task. As a consequence, we are unable to confirm our second preregistered hypothesis that high experience of the AMT jobs context attenuates it. We also find strong evidence against a main effect of *Experience* on both selected option (*Inclusion BF* = 0.079) and preference for the risky option (*Inclusion BF* = 0.099), indicating that neither demographic group is more likely overall to select the risky options or have stronger preferences for them. We do, however, observe extreme evidence in favor of a main effect of *Context* (*Inclusion BF* > 287,000 and 448,000) and an interaction between *Context* and *Experience* (*Inclusion BF* > 2,400 and 5,400) on both dependent measures. These effects are such that participants are overall slightly more likely to select and have slightly stronger preferences for risky options when they are contextualized as AMT jobs, but this is particularly the case for participants with low experience of performing such jobs (see Figure 3). In Online Appendix B, we show that these results are robust to other priors than those we preregistered.

When plotting the empirical relative preferences (Figure 4), all average preferences seem almost linear. We see that there is substantial individual variability (Figure 4, colored lines). These graphs indicate that the rank-order manipulation does not appear to "change their shapes" (Stewart et al. 2015).[9] All in all, in light of the BANOVAs and Figure 4, the only substantial effect in these data is an interaction between *Context* and *Experience*. *Context* has a very limited effect for participants with high experience of performing jobs on Amazon Mechanical Turk but a larger effect for participants with low experience of performing jobs on Amazon Mechanical Turk. Low-experience participants select and have a stronger preference for the risky option more often when the options are framed as AMT jobs rather than monetary gambles.

4. Conclusions

In this preregistered study, we have attempted to alleviate three methodological concerns about previous studies; we (i) evaluate the effects of our independent variables using group differences in summary statistics, (ii) use test items that are common across conditions, and (iii) enforce a dedicated training phase followed by a dedicated test phase. We have controlled for range of stimuli, thus isolating any rank-order effect. An effect was predicted regardless of whether valuations of monetary values and probabilities are integrated additively, by averaging, or multiplicatively and whether participants encode each observed stimulus (as in the DbS model) or only each unique stimulus value. Doing so, we find consistent evidence against the Decision by Sampling model's prediction of a rank-order effect for a binary choice decision-making task. These results are robust to swapping our preregistered priors for extreme effect size priors. If our findings can be replicated by independent laboratories, the main conclusion is that the Decision by Sampling model and the rank-order effect have been falsified, at least for this binary choice task. Because we did not identify a rank-order effect in the first place, we also failed to confirm our second preregistered hypothesis: that any rank-order effect would be attenuated for high-experience participants when the risky choices are contextualized as crowd-working jobs.

The high heterogeneity in judgements (Figure 4, colored lines) might invite the concern that the study is underpowered. However, unlike *p*-values, Bayes factors allow one to assess the evidence in favor of the null hypothesis. The fact that we find evidence against an effect (and not merely inconclusive Bayes factors close to one) indicates that the null result was not because of insufficient power. Our simulation suggests that there should be larger differences in subjective preference for some of our test items than for others (Figure 1 in Online Appendix A). One might, therefore, be worried that the statistical tests that we preregistered, which are based on mean differences across all items, will fail to indicate support for a rank-order effect if it only existed for specific items. However, as is visible from the almost linear cumulated relative preferences (Figure 4), the difference in mean preference strength is highly similar between groups for all items. Thus, there does not appear to be any tendency toward a rank-order effect even if you single out the items where the Decision by Sampling model predicts the largest difference.

We find extreme evidence in support of an effect of *Context*. Specifically, participants preferred the risky option more when it was framed as a prospective job on Amazon Mechanical Turk as opposed to a monetary gamble. This effect was much greater for low-experience participants compared with high-experience participants.

One potential explanation of this is that valuations are indeed based on memory sampling but that the exemplars are drawn from long-term memory only (thus, not producing any rank-order effect) and that cueing some context primes (e.g., Schacter 1987, Janiszewski and Wyer 2014) retrieval of related memory samples: a special case of the anchoring-as-activation effect (Chapman and Johnson 1999). This should be possible to model within existing sampling-based

theories (e.g., Sanborn et al. 2021) by presuming that the semantic information sets the starting position of the sampling "walk." It is not clear to us why such a semantic priming effect would be so greatly exacerbated by low experience (Figure 3 and panel (a) of Figure 4) (we would think that it should be exacerbated by high experience because of experienced individuals having more finely branched memory networks). It is an open question if that fact could be explained within a sampling model. If it can, the Decision by Sampling model's most radical suggestion—that preferences are ultimately created by the stimuli that we are exposed to in our environment—could be consistent with what we present here. If replications of the present work verify that the rank-order effect does not exist, however, preferences are clearly more stable than suggested by the Decision by Sampling model.

An opposing potential explanation is that the context affected the subjective interpretation of the stated probabilities (see Teigen and Brun 1995, 1999, 2003). The instructions emphasized that the approval rates provided under the AMT jobs context were calculated based only on rejections because of unpredictable reasons and thus, from the perspective of the AMT user, represented stochastic probabilities of being paid. In real life, AMT users can influence their probability of being paid: for example, by performing the job very carefully or contacting the person listing the job to talk them into overturning the rejection. Our participants might be aware of this and adjust their perceptions accordingly, which would explain the (slight) main effect of context. Such an "illusion of control" (Langer 1975) would mean that the subjective probabilities that participants acted on when selecting between AMT jobs were higher than the stated probabilities, making the risky options more attractive. This "illusion" might decrease with experience as the individual learns that there exists some stochasticity that they truly cannot affect.

Finally, a reviewer suggested that it could be that there existed a rank-order effect that only persisted on the first trial in the test phase. The argument is that the attribute values from the common test items may displace those from the training phase, and so, any rank-order effect should be washed away on all but the first test trial. This is not possible to test in the present data because we did not record the (randomized) order of presentation for the test items. A systematic replication could do so by presenting a single training trial followed by a single test trial and repeating this procedure for many trials. Here, we note some implications for the DbS theory of assuming that attribute values are displaced from the memory sample after a single trial.

The more restrictions we place on which recent experiences that can influence subjective valuations, the more fleeting and capricious any rank-order effect would be. This (potential) effect would be possible to eliminate by exposing a participant to a single item irrelevant to the target choice, distracting them for a moment (see Canic 2016, pp. 66–73), or introducing more than two test options on the same trial (e.g., Bird and Harris 2018) such that their attribute values displace even the immediately preceding item. In sum, this version of a DbS theory would suggest that people's subjective valuations are only swayed by recent experiences under very specific conditions and that the rank-order effect is a possible but unlikely occurrence in everyday life.

Acknowledgments

The authors are grateful to Neil Stewart, Quentin André, and another anonymous reviewer for their considered and thoughtful comments. Code, data, research materials, and the preregistration for this study can be found at https://doi.org/10.17605/OSF.IO/T5XQ4. Author contributions—Mattias Forsgren: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing – Original draft, Writing – Review & Editing, Visualization, Project administration. Lars Frimanson: Conceptualization, Methodology, Writing – Review & Editing, Supervision, Funding acquisition. Peter Juslin: Conceptualization, Methodology, Writing – Review & Editing, Supervision, Funding acquisition.

Endnotes

[1] See Brown and Walasek (2023) for a discussion and argument as to why.

[2] But, see Matthews (2012) for failed conceptual replications of study 3 in Ungemach et al. (2011).

[3] No hypothesis tests are reported for the experiments in Canic (2016) reviewed here, so our interpretations are based on observing the 95% confidence intervals.

[4] To preview, the present experiment will be unconfounded by range differences. However, we realized in hindsight that there were slight differences in the minimum ($1.1 versus $1.4) and maximum ($5.0 versus $5.3) of the training values. This does not affect our ability to test the DbS model—it predicts an effect regardless—but we note it here for transparency.

[5] See also André and de Langhe (2021b) for evidence that the model-based results in Walasek and Stewart (2015) are not uniquely predicted by the DbS and fail "placebo analyses" (Athey and Imbens 2017).

[6] If an AMT job has not been completed in a satisfactory way, the so-called "requester" who posted the job to AMT may reject the submission, thereby refusing to pay the participant and reposting the job to AMT for someone else to complete. If a job is not rejected, it is approved, and the participant is paid. A high approval rating (i.e., percentage of completed jobs subsequently approved) thus implies that the participant tends to complete jobs in a satisfactory way.

[7] Previous studies (Stewart et al. 2015, Walasek and Stewart 2015, Rakow et al. 2020) have used 64 trials or more. Because we are worried about participant fatigue, we did not greatly exceed this lower bound.

[8] The default model prior is a uniform distribution. The effect size prior (specified in the unit distance from grand mean divided by the

pooled standard deviation) is $N(0, g)$, where $g \sim \text{Inverse } \chi^2(1, h^2)$. The default h (called "r scale" in the JASP interface) is 0.5. The effect size prior thus ranges from zero to ∞ but with the bulk of its density collected close to zero. See Rouder et al. (2017) for details.

[9] Again, please note that the original evidence for that claim has been reconsidered in later work by the same research group (Alempaki et al. 2019, Stewart et al. 2020).

References

Agley J, Xiao Y, Nolan R, Golzarri-Arroyo L (2022) Quality control questions on Amazon's Mechanical Turk (mTurk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7. *Behav. Res. Methods* 54(2022):885–897.

Alempaki D, Canic E, Mullett TL, Skylark WJ, Starmer C, Stewart N, Tufano F (2019) Reexamining how utility and weighting functions get their shapes: A quasi-adversarial collaboration providing a new interpretation. *Management Sci.* 65(10):4841–4862.

André Q, de Langhe B (2021a) How (not) to test theory with data: Illustrations from Walasek, Mullett, and Stewart (2020). *J. Experiment. Psych. General* 150(12):2671–2674.

André Q, de Langhe B (2021b) No evidence for loss aversion disappearance and reversal in Walasek and Stewart (2015). *J. Experiment. Psych. General* 150(12):1–24.

Athey S, Imbens GW (2017) The state of applied econometrics: Causality and policy evaluation. *J. Econom. Perspect.* 31(2):3–32.

Bentham J (1789) *An Introduction to the Principles of Morals and Legislation. Printed in the Year 1780, and Now First Published. By Jeremy Bentham, of Lincoln's Inn, Esquire* (T. Payne and Son, London).

Berinsky AJ, Margolis MF, Sances MW (2016) Can we turn shirkers into workers? *J. Experiment. Soc. Psych.* 66(2016):20–28.

Bird S, Harris AJ (2018) Robust, domain-specific effects of prior context in risk preferences for pension choice. *J. Behav. Decision Making* 31(5):605–618.

Birnbaum MH, Martin T (2012) *Generalization Across People, Procedures, and Predictions: Violations of Stochastic Dominance and Coalescing* (Cambridge University Press, Cambridge, UK), 84–107.

Brown GDA, Walasek L (2023) Homo ordinalus and sampling models: The past, present, and future of decision by sampling. Fiedler K, Juslin P, Denrell J, eds. *Sampling in Judgment and Decision Making* (Cambridge University Press, Cambridge, UK), 35–65.

Brunswik E (1956) *Perception and the Representative Design of Psychological Experiments*, 2nd ed. (University of California Press, Berkeley).

Canic E (2016) Value is context dependent: On comparison processes and rank order in choice and judgment. Unpublished doctoral thesis, University of Warwick, Coventry, UK.

Chapman GB, Johnson EJ (1999) Anchoring, activation, and the construction of values. *Organ. Behav. Human Decision Processes* 79(2):115–153.

Fisher I (1892) *Mathematical Investigations in the Theory of Value and Prices* (Yale University Press, New Haven, CT).

Frydman C, Jin LJ (2021) Efficient coding and risky choice. *Quart. J. Econom.* 137(1):161–213.

Gigerenzer G (2004) *Striking a Blow for Sanity in Theories of Rationality* (MIT Press, Cambridge, MA), 389–409.

Janiszewski C, Wyer RS (2014) Content and process priming: A review. *J. Consumer Psych.* 24(1):96–118.

JASP Team (2022) JASP. Version 0.16.1. https://jasp-stats.org/.

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47(2):263–292.

Langer EJ (1975) The illusion of control. *J. Personality Soc. Psych.* 32(2):311–328.

Laughlin S (1981) A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung C* 36(9–10):910–912.

Mason W, Suri S (2012) Conducting behavioral research on Amazon's Mechanical Turk. *Behav. Res. Methods* 44(1):1–23.

Matthews WJ (2012) How much do incidental values affect the judgment of time? *Psych. Sci.* 23(11):1432–1434.

McInnis B, Cosley D, Nam C, Leshed G (2016) Taking a HIT: Designing around rejection, mistrust, risk, and workers' experiences in Amazon Mechanical Turk. Kaye J, Druin A, Lampe C, Morris D, Hourcade JP, Terveen L, Morris S, eds. *Proc. 2016 CHI Conf. Human Factors Comput. Systems* (ACM, New York), 2271–2282.

Meredith W (1993) Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58(4):525–543.

Nassar MR, Gold JI (2013) A healthy fear of the unknown: Perspectives on the interpretation of parameter fits from computational models in neuroscience. *PLoS Comput. Biol.* 9(4):e10003015.

Noguchi T, Stewart N (2018) Multialternative decision by sampling: A model of decision making constrained by process data. *Psych. Rev.* 125(4):512–544.

Parducci A (1965) Category judgment: A range-frequency model. *Psych. Rev.* 72(6):407–418.

Parducci A, Perrett LF (1971) Category rating scales: Effects of relative spacing and frequency of stimulus values. *J. Experiment. Psych.* 89(2):427–452.

Peer E, Vosgerau J, Acquisti A (2014) Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behav. Res. Methods* 46(4):1023–1031.

Rakow T, Cheung NY, Restelli C (2020) Losing my loss aversion: The effects of current and past environment on the relative sensitivity to losses and gains. *Psych. Bull. Rev.* 27(6):1333–1340.

R Core Team (2021) R: A language and environment for statistical computing. https://www.r-project.org/.

Roe RM, Busemeyer JR, Townsend JT (2001) Multialternative decision field theory: A dynamic connectionst model of decision making. *Psych. Rev.* 108(2):370–392.

Rouder JN, Morey RD, Verhagen J, Swagman AR, Wagenmakers EJ (2017) Bayesian analysis of factorial designs. *Psych. Methods* 22(2):304–321.

Sanborn A, Zhu JQ, Spicer J, Sundh J, León-Villagrá P, Chater N (2021) *Sampling as the Human Approximation to Probabilistic Inference* (Oxford University Press, Oxford, UK), 430–448.

Schacter DL (1987) Implicit memory: History and current status. *J. Experiment. Psych. Learn. Memory Cognition* 13(3):501–518.

Stewart N, Canic E, Mullett T (2020) On the futility of estimating utility functions: Why the parameters we measure are wrong, and why they do not generalize. Preprint, submitted January 26, http://dx.doi.org/10.31234/osf.io/qt69m.

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

Stewart N, Reimers S, Harris AJL (2015) On the origin of utility, weighting, and discounting functions: How they get their shapes and how to change their shapes. *Management Sci.* 61(3):687–705.

Teigen KH, Brun W (1995) Yes, but it is uncertain: Direction and communicative intention of verbal probabilistic terms. *Acta Psychologica* 88(3):233–258.

Teigen KH, Brun W (1999) The directionality of verbal probability expressions: Effects on decisions, predictions, and probabilistic reasoning. *Organ. Behav. Human Decision Processes* 80(2):155–190.

Teigen KH, Brun W (2003) Verbal probabilities: A question of frame? *J. Behav. Decision Making* 16(1):53–72.

Trueblood JS, Brown SD, Heathcote A (2014) The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psych. Rev.* 121(2):179–205.

Tversky A (1975) A critique of expected utility theory: Descriptive and normative considerations. *Erkenntnis* 9(2):163–173.

Tversky A, Kahneman D (1992) Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertainty* 5(1992):297–323.

Ungemach C, Stewart N, Reimers S (2011) How incidental values from the environment affect decisions about money, risk, and delay. *Psych. Sci.* 22(2):253–260.

von Neumann J, Morgenstern O (1944) *Theory of Games and Economic Behavior* (Princeton University Press, Princeton, NJ).

Walasek L, Stewart N (2015) How to make loss aversion disappear and reverse: Tests of the decision by sampling origin of loss aversion. *J. Experiment. Psych. General* 144(1):7–11.

Walasek L, Stewart N (2019) Context-dependent sensitivity to losses: Range and skew manipulations. *J. Experiment. Psych. Learn. Memory Cognition* 45(6):957–968.

Walasek L, Stewart N (2020) You cannot accurately estimate an individual's loss aversion using an accept–reject task. *Decision* 8(1):2–15.

Walasek L, Mullett TL, Stewart N (2021) Acceptance of mixed gambles is sensitive to the range of gains and losses experienced, and estimates of lambda (λ) are not a reliable measure of loss aversion: Reply to André and de Langhe (2021). *J. Experiment. Psych. General* 150(12):2666–2670.

Walker A, Braglia L (2017) openxlsx: Read, write and edit xlsx files. https://CRAN.R-project.org/package=openxlsx.

Wickham H, Bryan J (2019) readxl: Read excel files. https://readxl.tidyverse.org/.

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, et al. (2019) Welcome to the Tidyverse. *J. Open Source Software* 4(43):1686.