Cognition 229 (2022) 105245

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Biased confabulation in risky choice☆

Alice Mason [a,*], Christopher R. Madan [b], Nick Simonsen [b], Marcia L. Spetch [c], Elliot A. Ludvig [a]

[a] University of Warwick, UK
[b] University of Nottingham, UK
[c] University of Alberta, Canada

ARTICLE INFO

Keywords:
Free recall
Decisions-from-experience
Extreme outcomes
Risky choice
Episodic memory

ABSTRACT

When people make risky decisions based on past experience, they must rely on memory. The nature of the memory representations that support these decisions is not yet well understood. A key question concerns the extent to which people recall specific past episodes or whether they have learned a more abstract rule from their past experience. To address this question, we examined the precision of the memories used in risky decisions-from-experience. In three pre-registered experiments, we presented people with risky options, where the outcomes were drawn from continuous ranges (e.g., 100–190 or 500–590), and then assessed their memories for the outcomes experienced. In two preferential tasks, people were more risk seeking for high-value than low-value options, choosing as though they overweighted the outcomes from more extreme ranges. Moreover, in two preferential tasks and a parallel evaluation task, people were very poor at recalling the exact outcomes encountered, but rather confabulated outcomes that were consistent with the outcomes they had seen and were biased towards the more extreme ranges encountered. This common pattern suggests that the observed decision bias in the preferential task reflects a basic cognitive process to overweight extreme outcomes in memory. These results highlight the importance of the edges of the distribution in providing the encoding context for memory recall. They also suggest that episodic memory influences decision-making through gist memory and not through direct recall of specific instances.

1. Introduction

Many of our everyday choices regarding healthy lifestyles, climate action and personal finances are made on the basis of remembered information. For example, when deciding where to go for your weekly shop, you might draw on past experiences and prices at different stores. To make these decisions we must learn, from experience, the risk and reward associated with our actions. For example, people respond differently when presented with the potential risks and outcomes associated with a decision (e.g., reading about medication side effects), compared to deciding based on experience (e.g., choosing the most effective painkiller for a headache). These experience-based decisions necessarily rely to some degree on memory, yet the nature of this relationship is not well understood (Rakow & Newell, 2010). In decision-making, such memories for past outcomes are known to be biased, with stronger memory for outcomes that are more extreme (Ludvig, Madan, McMillan, Xu, & Spetch, 2018; Madan, Ludvig, & Spetch, 2014), more salient (Tsetsos, Chater, & Usher, 2012), and more strongly preferred (Weilbächer, Kraemer, & Gluth, 2020).

Here we examine the nature of the memory representations that develop during decision-making, evaluating whether people recall specific instances of past episodes or whether they learn a more abstract rule or heuristic from their past experiences. This point distinguishes between important models of choice that are predicated on either episodic/instance-based mechanisms (e.g., Bornstein, Khaw, Shohamy, & Daw, 2017; Gonzalez, Lerch, & Lebiere, 2003; Hotaling, Donkin, Jarvstad, & Newell, 2022) or based on abstract/gist representations (Brainerd, Reyna, & Mojardin, 1999; Nosofsky, 1988; e.g., Tversky & Kahneman, 1973). For example, when considering whether to take climate action, do people recall a specific record-breaking day of the summer or instead use the general rule that summers are getting hotter? If we are looking at ways of influencing people to take climate action, then whether their decision is influenced by memories of individual instances or a general gist impacts how best to potentially shift

☆ This research was supported by a Leverhulme Early Career Fellowship (ECF-2018-408) to AM and Economic and Social Science Research Council grants to AM (ES/T016639/1) and NS (ES/P000711/1). The data for all experiments can be downloaded from the Open Science Framework (https://osf.io/2ey8m/).

* Corresponding author at: Department of Psychology, University of Warwick, Coventry CV4 7AL, UK.
E-mail address: alice.m.a.mason@warwick.ac.uk (A. Mason).

https://doi.org/10.1016/j.cognition.2022.105245
Received 16 December 2021; Received in revised form 13 June 2022; Accepted 27 July 2022

behaviour (e.g., reminders of past episodes are likely to be less effective if decisions are based on the latter memory representations).

In three experiments we directly test how people recall uncertain information and how these memory representations are related to decision-making. We examine the degree to which people recall specific values versus applying a rule-based approach to generate outcomes. We show that, when asked, people did not recall the exact outcomes encountered, but rather confabulated outcomes that aligned with their decision biases.

1.1. *Risky choice and memory*

When making risky decisions, people often make different decisions depending on whether they learn about the odds and outcomes through an explicit description or from their own experience. This description-experience gap has been attributed to a process of drawing samples from memory (Hertwig & Erev, 2009), which might only feature a small subset of past experiences (Plonsky, Teodorescu, & Erev, 2015) or can be explicitly biased (Madan, Ludvig, & Spetch, 2017). In previous work, using a risky-choice paradigm, we have consistently found that people overweight the extreme outcomes (highest and lowest in a context) in memory (Madan et al., 2014; Madan, Spetch, Machado, Mason, & Ludvig, 2021). This overweighting in memory leads to people being more risk-seeking for choices that involve the high extreme and less risk-seeking for choices that involve the low extreme. In these experience-based situations, people are more risk-seeking for relative gains than relative losses (Konstantinidis, Taylor, & Newell, 2018; Madan et al., 2014), showing the opposite pattern to the standard reflection effect in decisions-from-description (Kahneman & Tversky, 1979). For example, in experience, people choose a 50/50 chance of winning 90 or 50 over a certain 70, but choose a certain 30 over a 50/50 chance of winning 50 or 10 points, where 90 is the high extreme and 10 is the low extreme. Recent models of decision-making have modeled this pattern by weighting the probability of including an outcome in a memory sample as a function of how close to the edges of the distribution the outcome falls (Lieder, Griffiths, & Hsu, 2018; Ludvig et al., 2018; Vanunu, Hotaling, & Newell, 2020). This overweighting parallels findings in the memory literature that the best and worst information is better encoded, in the context of both reward outcomes and emotional valence (Madan, 2017; Mason, Farrell, Howard-Jones, & Ludwig, 2017). Similarly, the *peak-end rule* describes how in a range of decision contexts, people tend to overweight the most intense and the most recent events in their overall evaluation of an experience (Redelmeier & Kahneman, 1996).

Thus far, studies examining the relationship between memory for extreme outcomes and risky choice have shown that after completing a risky-choice task participants have better memory for the most extreme outcomes (highest and lowest) in a decision context (Ludvig et al., 2018; Ludvig, Madan, & Spetch, 2014). The small outcome set typically used in decision-from-experience tasks has limited the data that can be collected after the decision-making task and therefore used to examine the relationship. These tasks typically include a limited set of outcomes (usually 3–6 different reward outcomes total), making the memory demands relatively low (Hertwig & Erev, 2009; Ludvig et al., 2018; Madan et al., 2014, 2021; Wulff, Mergenthaler-Canseco, & Hertwig, 2018). In previous experiments, we have used two key memory measures: frequency judgments and first-outcome reported. The frequency judgment presents participants with each of the risky options and outcomes and asks them to state what percentage of the time they saw each of the outcomes. Participants systematically overestimate the frequency of the extreme outcomes (highest and lowest experienced). In the first-outcome-reported test participants are presented with the risky options and asked to state the outcome that they most readily associate with that option. Once again, participants are more likely to report the extreme outcomes (Madan et al., 2014; Madan, Ludvig, & Spetch, 2019). Neither of these measures allow the examination of the precision of the memory representations involved. In the current experiments, each payoff is associated with a distribution of outcomes, meaning that we can collect free-recall data for each of the options. Other decision-making tasks have included a greater range of reward outcomes but have not examined memory effects (Kunar, Watson, Tsetsos, & Chater, 2017; Tsetsos et al., 2012; Vanunu et al., 2020). It is worth noting that Spektor, Gluth, Fontanesi, and Rieskamp (2019) used a decisions-from-experience paradigm with continuous outcomes and asked participants for estimates of the average probability and magnitude of each option but did not ask participants to recall individual items. They found that participants' estimates of probability and magnitude did not correlate with how often they chose the options.

1.2. *Specificity vs generality of memory in risky choice*

Dual-route memory models suggest that memory traces exist at two levels, such as verbatim and gist (Brainerd et al., 1999), item-specific and relational (Hunt & Mitchell, 1982), and form or content-based (Steyvers & Griffiths, 2008). The verbatim, item-specific or form-based representation is close to the raw form of the item and involves processing individual features of the item. The gist, relational or content-based level involves processing shared features of items and is a highly abstracted representation of the past.

In prominent models of memory and decision-making the samples used to predict an upcoming choice represent distinct episodes of past experience and are more aligned with the item-specific representation outlined above (Gonzalez et al., 2003). For example, MINERVA-DM relies on a "database" of memories that are degraded representations of experienced events (e.g., due to lack of attention at encoding) (Dougherty, Gettys, & Ogden, 1999), whereas Decision-by-Sampling (DbS) (Stewart, Chater, & Brown, 2006) assumes that the contents of memory reflect the structure of the world (Chater & Brown, 1999; Stewart, 2009). Exemplar models assume that each item is stored in a unique memory trace (Nosofsky, 1988) and have been used to predict choice in decisions-from-experience paradigms (Hotaling, Donkin, Jarvstad, & Newell, 2022). Whilst such models are able to accurately account for choice patterns, data that would shed light on people's memory representations is not routinely collected.

This instance-based sampling approach has also gained traction in reinforcement learning in cases where the extensive experience required by reinforcement learning is not available (e.g., Lengyel & Dayan, 2007). Standard incremental reinforcement learning models do not maintain a memory of individual outcomes/events (Sutton & Barto, 1998). In contrast, the episodic reinforcement-learning models include a record of individual trials that can be used to enhance the incremental-learning system (Gershman & Daw, 2017). These models are able to readily accommodate the experimental finding that priming participants with either previous wins or losses can shift choice (Bornstein et al., 2017; Gibson & Zielaskowski, 2013; Ludvig, Madan, & Spetch, 2015). Whilst Bornstein et al. (2017) showed a strong influence of past instances on future choices, they did not directly test memory for items or examine the possibility that categorically related primes would produce the same impact on future choices. Murty, FeldmanHall, Hunter, Phelps, and Davachi (2016) found evidence that decision-making was linked to memory for specific item-reward associations. Critically, in their experiment when they tested participants' memory for reward outcomes recalled within a $1 range of the correct item were scored as correct. So although this study is often cited as evidence of contextually detailed episodic memory predicting choice, the specificity of the memory is not known.

In contrast, rule-based strategies assume the decision-maker learns the underlying function/distribution of the items by abstracting knowledge from the environment, for example by determining how much a cue or feature relates to the decision criterion (DeLosh, Busemeyer, & McDaniel, 1997). Rule-based and exemplar models have been tested and pitted against each other extensively in the categorisation and judgment literature (Hoffmann, von Helversen, & Rieskamp, 2014;

Juslin, Olsson, & Olsson, 2003; Pachur & Olsson, 2012). For example, individual differences in rule-based or exemplar strategies have been linked to memory: a reliance on exemplar-based strategies may be linked to episodic memory abilities, whereas reliance on rule-based strategies may be related to working memory capacity (Hoffmann, Von Helversen, & Rieskamp, 2016; Juslin, Karlsson, & Olsson, 2008). A more combined approach is reflected in a Bayesian model of reconstructive memory (Hemmer & Steyvers, 2009), in which memory errors can be explained by using prior knowledge at the category level to help the recall of instance-specific attributes.

1.3. *Current experiments*

Our three experiments were designed to extend our understanding of the relationship between memory for outcomes and risky decisions from experience. By using continuous outcomes and instance-based memory measures, the experiments aimed to shed light on the trade-off between specificity and generality in memory. In particular, by analyzing the errors made during memory recall and value estimation, we asked to what extent people develop veridical memories versus distorted ones, and whether the pattern of errors suggests confabulation based on category knowledge. By conducting the same memory tests after making preferential choices (Experiments 1a and 1b) or after experiencing the outcomes without making any choices (Experiment 2), we aimed to also shed light on the correlation between memory and risky choice. Specifically, we asked whether people would still show the same pattern of memory results when they have not made any preferential decisions.

Experiments 1a and 1b use a decisions-from-experience task (Ludvig et al., 2014) where participants choose between pairs of risky and safe coloured doors, which lead to different reward outcomes. In addition, trials are either high value or low value, which allows us to examine risky choice for relative gains and losses. In the current experiments, we introduce continuous outcomes for the reward values (Olschewski, Dietsch, & Ludvig, 2019). The outcomes are uniformly sampled from a range of possible outcomes (+/− 45 of the mean value); Table 1 details the exact ranges. For example, in Experiment 1a, for a low-value safe outcome where the mean is 345, the outcome shown ranges from 300 to 390. For the low-value risky option there is a 50:50 chance that the outcome is sampled from either a lower range (100–190) or a higher range (500–590). Experiment 1b uses similar ranges, but with non-overlapping values. At the end of the choice phase, participants were asked to recall as many outcomes associated with each option as they could. They were also asked to state the average value of the door. If people are retrieving specific instances they should accurately recall (some of) the reward outcomes. If instead they have learned the distribution and are subsequently generating outcomes from that distribution, they may confabulate and recall numbers they have not actually encountered but that fall within the experienced range.

The extreme-outcome effect has typically been observed in preferential choice tasks (Konstantinidis, Taylor, & Newell, 2018; Ludvig et al., 2018, 2014). One way to assess the degree to which this choice bias reflects a parallel memory bias is to assess whether the effects still occur in the absence of a preferential choice task. To answer this question, we conducted an additional experiment where participants did not directly choose between the risky and safe options. Instead, they were shown a single option (high or low value, risky or safe) on each trial and presented with the same ranges of outcomes as in the first experiment. Once they experienced all options and outcomes we conducted the same memory and estimation tasks.

Table 1
Safe and risky door options for each experiment.

Value	Option	Range	Expected value
Experiment 1a and Experiment 2			
Low	Safe	300–390	345
Low	Risky (p = .5)	100–190 or 500–590	145 or 545
High	Safe	700–790	745
High	Risky (p = .5)	500–590 or 900–990	545 or 945
Experiment 1b			
Low	Safe	200–290	245
Low	Risky (p = .5)	100–190 or 300–390	145 or 345
High	Safe	600–690	645
High	Risky (p = .5)	500–590 or 700–790	545 or 745

For each of the experiments we pre-registered a series of experimental hypotheses. For Experiments 1a and 1b where the task included choice, we tested the Overweighting-in-Choice hypothesis, which states that people will overweight the extreme outcomes in choice. For example, in Experiment 1a both the high- and the low-value risky options led to a 50% chance of an outcome in the 500–590 range. The high-value option also led with a 50% chance to outcomes in the 900–990 range (the highest possible range). In contrast the low-value risky option also led with a 50% chance to outcomes in the 100–190 range (the lowest possible range). If people overweight the values at the extreme ends (i.e., the 100 s and the 900 s), then they will be more risk-seeking for the high-value choices when the highest range (900 s) will be overweighted than for the lower-value choices when the lowest range (100 s) will be overweighted

Across all experiments, we tested two hypotheses related to memory and estimation. The *Memory-Overweighting* hypothesis states that people will overweight the extreme outcomes in memory and will be more likely to recall outcomes from the more extreme range. The *Estimation-Overweighting* hypothesis states that people will judge the average value of the high-value risky option to be higher and the low-value risky option to be lower if they are overweighting the extremes. For Experiment 2, we additionally tested the *Preferential-Overweighting* hypothesis which states that without a preference task, extreme outcomes will be no more likely to be recalled, nor to influence estimation.

2. Method

Data files, pre-registration documents and materials can be found on the Open Science Framework (https://osf.io/2ey8m/).

For all experiments, participants were recruited via Prolific Academic to participate in the experiment online. To be eligible to take part in the experiments participants needed to be aged 18–65, have English as their first language (self-reported), and have a Prolific Academic approval rating of over 90%. For all experiments, the target sample size was 102, which, with a frequentist approach, would have given 95% power for a one-tailed t-test with a small-medium effect size (Cohen's d = 0.4) and an alpha of 0.01. As per the pre-registered sampling plan, we aimed to recruit 120 participants for each experiment to allow for incomplete data sets and dropout during the experiment.

For Experiment 1a, 123 participants signed up via Prolific Academic before the experiment closed (age range 19–57, M = 30.0, SD = 10.2; 54 female, 69 male). According to the pre-registered exclusion criteria, 21 were excluded for scoring less than 60% correct on the catch trials. The final sample size was 102. Participants were paid £4 for completing the 40-min session and could earn an additional bonus of £1 for every 20,000 points up to a maximum bonus of £7. Participants were told the conversion rate after they had completed the task. Participants earnt between 111,773 and 149,253 points (mean = 140,387). Due to an error in recording the Prolific IDs, all participants were paid the maximum bonus of £7.

In Experiment 1b, 123 participants completed the experiment (age range 18–54, M = 25.2, SD = 7.7; 50 female, 70 male, 3 undisclosed), but 1 participant was excluded as they did not have data on the server. According to the pre-registered exclusion criteria, 18 were excluded for scoring less than 60% correct on the catch trials. The final sample size was 104. The payment structure was identical to Experiment 1a. Participants earnt between 91,468 and 123,645 points (mean = 117,068), and the mean bonus payment was £5.85.

For Experiment 2, 120 participants completed the experiment (age range 19–65, M = 32.6, SD = 11.0; 57 female, 62 male, 1 undisclosed), and three participants were excluded as they did not have data on the server. The final sample size was 117 complete datasets. This experiment did not involve a choice component and therefore there were no exclusions based on catch trials. Participants were paid £2 for completing the 20-min session.

2.1. Procedure

The procedure followed the protocols published in Ludvig et al. (2014). The choice options and two tasks are shown in Fig. 1. There were 5 blocks of 48 trials. Between blocks participants were given a 15-s break.

Table 1 lists the details for the 4 choice options (doors) available in the task: low-value safe, low-value risky, high-value safe and high-value risky. Experiments 1a and 1b were preferential choice tasks, and on each trial, up to 2 options appeared on either side of the screen as pictures of different-coloured doors. Participants started with zero points and selected a door by clicking with their mouse. After the selection, a numerical outcome (i.e., reward amount) drawn from the corresponding range(s) appeared for 1.2 s. Trials were self-paced, and, after each trial, participants pressed a button at the centre of the screen to re-centre the mouse and move on to the next trial. The outcomes were uniformly sampled from a range of outcomes +/− 45 of the mean value. The exact mean number, however, was never shown. A random number from the range (180 numbers for the risky doors and 90 numbers for the safe door) was shown each time the option was selected. For example, in Exp 1a and 2, for the low-value safe option where the mean was 345 the outcome shown ranged from 300 to 390, and 345 never appeared. For the risky options, a number was drawn equiprobably from one of the two possible ranges associated with that option. For example, for the high-value risky option, the number displayed was either drawn from the range 500–590 or from the range 900–990.

In Experiments 1a and 1b, in each block, there were three types of trials. On 24 *Choice* trials, participants selected between safe and risky options of equal expected values (low or high). On 8 *Single-Option* trials, only one door was presented, and participants needed to select that option to continue. These trials ensured that all options and outcomes were experienced on occasion independent of preference. On 16 *Catch* trials, participants selected between a high-value option and a low-value option (either both risky or both safe). These trials allowed us to detect inattentive participants or those who were not motivated to get higher expected values, irrespective of risk preference. As pre-registered, participants who selected the high-value option less than 60% of the time on these trials were excluded from the data analyses reported below.

In Experiment 2, participants were told that they were going to complete a task to learn how much cash people living in a neighbourhood keep at home. On each trial, only one door appeared at a time. Participants were told that when a door appeared on screen they should click on it and the number shown represented how much money (in pence) a person in that neighbourhood had in their house. Participants were told that at the end of the experiment they would be asked some questions about the neighbourhood, but the exact questions were not specified. In Experiment 2, there were 48 Single-Option trials in each of the 5 blocks and participants saw an equal number of high-value and low-value doors and safe and risky doors (each door was shown 12 times each per block).

In all experiments, at the end of this task, participants were shown each of the 4 doors one at a time in a randomised order. They were asked to type as many outcomes as they could recall for each door within two minutes. They were then shown each door again and asked to type what they thought the average value of that door was. They had 30 s to type their answers.

2.2. Data analysis

The inferential framework used was Bayesian statistics. We used R to run the analysis (R Core Team, 2020) and the BayesFactor package to estimate Bayes Factors. For *t*-tests, the analyses used an uninformative Jeffrey's prior on the variance and a standard Cauchy prior of $\sqrt{2}/2$ on the *r* scale value. For a detailed discussion on prior selection, see Rouder, Morey, Verhagen, Province, and Wagenmakers (2016).

The value of the Bayes factors quantifies the strength of evidence in favour of one model with respect to another, given the data obtained. This value indicates how much prior beliefs should shift in response to the data obtained. Although there are no strict cut-offs, we apply the verbal labels used by Kass and Raftery (1995) to describe the results. Although not in our pre-registered plan, for ease of interpretation, we also report the equivalent frequentist statistics. In all cases, the primary analysis followed the pre-registered plan; any deviations or exploratory analyses are clearly marked below.

For the memory test, participants could only be included if they recalled at least one item for each cell in the corresponding analysis. For Experiments 1b and 2, we pre-registered an additional criterion that participants who recalled values or gave estimates outside of the experimental range with a small buffer (i.e., less than 90 or more than 1000) would be excluded, and we applied this criteria to all experiments.

3. Results

3.1. Risky choice

The *Overweighting-in-Choice* hypothesis stated that, in Experiment 1a

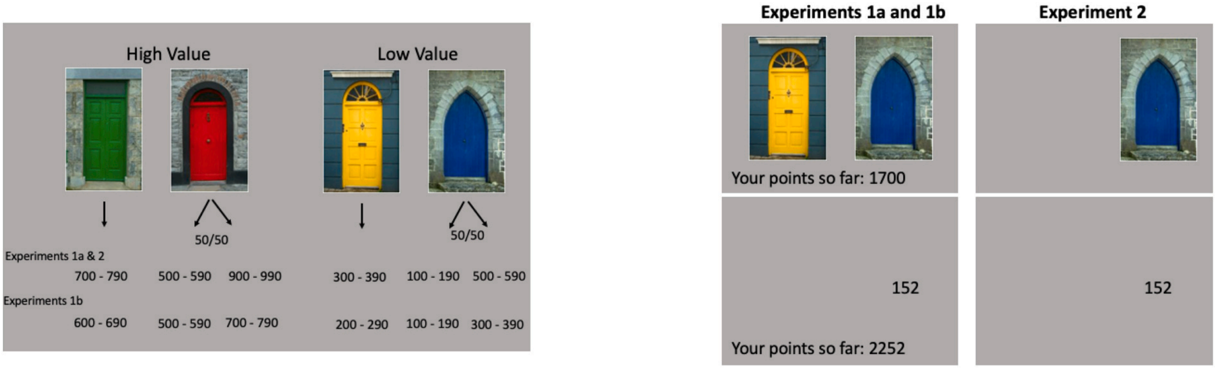

Fig. 1. A) Choice options with outcomes and probabilities across the experiment (not shown to participants). B) The trial structure for Experiments 1a, 1b and 2. In Experiments 1a and 1b participants were asked to choose between two doors and to maximise the points they earned. In Experiment 2, participants were told to click on each door and to observe the outcomes, which were said to represent how much cash (in pence) a person had in their home.

and 1b where people made preferential risky choices, they would overweight the extreme outcomes in choice. As a result, they would be more risk-seeking when deciding between high-value options as compared to when deciding between low-value options. We tested this hypothesis with a Bayesian *t*-test comparing risky choice in the high- and low-value decisions averaged over the last three blocks of the experiment. Fig. 2 shows how in Experiment 1a people were 24.9 ± 6.9% [M ± 95%CI] more risk seeking for the high- compared to the low-value decisions and in Experiment 1b they were 22.9 ± 6.8% more risk seeking for the high- compared to the low-value decisions. The Bayes Factors indicate very strong evidence in favour of this difference (Experiment 1a: $BF = 7.58e{+}06$, [$t(101) = 7.13$, $p < .001$, $d = 0.89$]; Experiment 1b: $BF = 3.78e{+}06$, [$t(103) = 6.58$, $p < .001$, $d = 0.86$]), in line with previous results on the impact of extreme outcomes in decisions-from-experience (Ludvig et al., 2018; Madan et al., 2014).

3.2. Memory

3.2.1. Recall accuracy

Initially in line with the pre-registration for Experiment 1a we examined people's correct recall for each of the outcomes. We calculated the proportion recalled with reference to the total number of outcomes recalled, as opposed to the number of outcomes observed. As is shown in Fig. 3, overall participants had very poor recall of the exact outcomes they saw. In Experiment 1a, of all the outcomes recalled, only 3.27 ± 0.44% [M ± 95%CI] were among the received outcomes. In Experiment 1b, 3.41 ± 0.45% of recalls were received outcomes. Overall recall was slightly higher in Experiment 2, and 5.13 ± 1.00% of participants' recalled outcomes were among the actually received outcomes.

We therefore examined participants' *consistent recalls*. These are recalled outcomes that are within the correct range(s) for the door shown, but did not have to be the exact numbers that the participants saw. Fig. 3 plots the consistent recalls for each of the conditions and experiments. If we examine recalls in terms of whether or not they are in the correct category then the proportion of outcomes recalled was considerably higher, with 72.6 ± 5.0% of recalls being consistent in Experiment 1a, 63.4 ± 5.0% in Experiment 1b and 65.0 ± 6.0% in Experiment 2.

Given that participants had very low precise recall we used the consistent recalls as the alternate basis for our pre-registered analyses in Experiment 1a and pre-registered the consistent-recalls analysis for Experiments 1b and 2.

Fig. 4 shows the general pattern of consistent recalls was very similar across all experiments. For the low-value risky door, participants recalled more outcomes from the extreme range compared to the non-extreme (Experiment 1a: 50.2 ± 10.3%; Experiment 1b: 25.2 ± 6.9%; Experiment 2: 29.2 ± 10.2%). For the high-value risky door, participants also recalled more outcomes from the extreme range compared to the non-extreme (Experiment 1a: 24.0 ± 11.3%; Experiment 1b: 12.7 ± 6.8%; Experiment 2: 20.3 ± 9.5%). There was very strong evidence in all cases that people recalled more outcomes from the extreme ranges (Experiment 1a: $BF_{Low}= 2.59e{+}20$, [$t(90) = 9.54$, $p < .001$, $d = 1.00$], $BF_{High}= 25{,}548$, [$t(92) = 4.17$, $p < .001$, $d = 0.43$]; Experiment 1b: $BF_{Low}= 1.47e{+}05$, [$t(89) = 5.86$, $p < .001$, $d = 0.62$], $BF_{High}= 31.3$, [$t(92) = 7.80$, $p < .001$, $d = 0.36$]; Experiment 2: $BF_{Low}= 5.18e{+}07$, [$t(105) = 5.62$, $p < .001$, $d= 0.55$], $BF_{High}= 5800$, [$t(104) = 4.19$, $p < .001$, $d = 0.41$]).

We conducted an exploratory analysis to examine whether there was a difference in recall of extreme items between low- and high-value options. In Experiment 1a there was strong evidence that people were recalling more extreme items for the low-value compared to the high-value options ($BF_{Diff}= 24.0$, [$t(90) = 2.61$, $p = .011$, $d = 0.35$]. The results from Experiments 1b and 2 were less diagnostic. In Experiment 1b, the evidence in favour of a difference was "not worth more than a bare mention" ($BF_{Diff}= 1.46$, [$t(85) = 2.08$, $p = .041$, $d = 0.22$]). In Experiment 2, the evidence against a difference was also "not worth more than a bare mention" ($BF_{Diff}= 0.30$, [$t(99) = 1.25$, $p = .22$, $d = 0.12$]).

3.2.2. First recall

Previous studies where each door was linked to a maximum of two outcomes have used the first outcome to come to mind as a measure of memory strength. Here we have multiple outcomes per option, which allows us to examine the probability that the outcomes from extreme ranges will be recalled first. Our prediction (the *Memory-Overweighting* hypothesis) was that numbers from the extreme ranges (Experiment 1a and 2: EV = 145 and 945; Experiment 1b: EV = 145 and 745) will be overweighted in memory and more likely to be reported when the risky doors are presented.

As the recall patterns in Fig. 4 show, numbers from the extreme ranges were far more likely to be reported as the first outcome to come to mind. For each experiment a pair of Bayesian contingency tests were performed (one for the low-value risky door, and one for the high-value risky door). In all cases, there was very strong evidence that people reported an extreme outcome more often (Experiment 1a: $BF_{Low}= 51{,}476$ [χ^2 (1, $N = 91$) = 43.62, $p < .001$], $BF_{High}= 10{,}162$ [χ^2 (1, $N = 93$) = 37.43, $p < .001$]; Experiment 1b: $BF_{Low}= 64$ [χ^2 (1, $N = 90$) = 17.78, $p < .001$], $BF_{High}= 37.57$, [χ^2 (1, $N = 92$) = 15.70, $p < .001$]; Experiment 2: $BF_{Low}= 2556$, [χ^2 (1, $N = 101$) = 32.17, $p < .001$], $BF_{High}= 5995$, [χ^2 (1, $N = 104$) = 28.04, $p < .001$]).

3.2.3. Distribution of recalls

To better follow up on how the recalled outcomes were distributed within a given range, we also examined the distribution of recalls. As observed with the recall counts (see Fig. 3), Fig. 5 shows how for the low-value risky doors the frequency of recalls was higher in the lower range (i.e., outcomes in the 100–190 range). For the high-value risky doors, there were more recalls in the higher range (Experiments 1a and 2: 900–990 or Experiment 1b: 700–790). More revealingly, this figure shows how the distribution of outcomes within each range also trended towards the edges in some cases–this pattern is most notable for the low-value risky and low value safe option in Experiment 1a and Experiment 1b and the high-value risky option in Experiment 2. The overall pattern confirms that participants generated items that fall at the edges of the distributions they experienced–both very starkly over the ranges from the whole experiment and also more mildly within several of the individual experienced ranges.

3.3. Estimations

Our third hypothesis (the *Estimation-Overweighting* hypothesis) was that people's judgments of the value of the risky doors would be influenced by overweighting of the extreme outcomes. Accordingly, the high-value risky door would be overestimated (as it might lead to a high extreme), and the low-value risky door would be underestimated (as it might lead to a low extreme). Fig. 6 shows participants' mean estimation of the risky doors, with respect to their estimate for the corresponding safe doors (i.e., by subtracting out the mean estimate for the safe door). As predicted, participants underestimated the value of the low-value risky door by −95.9 ± 32.1 points in Experiment 1a, by −42.4 ± 17.7 points in Experiment 1b and by −54.7 ± 32.8 points in Experiment 2; also as predicted, they overestimated the value of the high-value risky door by 86.3 ± 34.6 points in Experiment 1a, by 35.4 ± 18.2 in Experiment 1b and by 79.3 ± 31.4 points in Experiment 2. For each experiment for the high- and low-value options, we ran a one-sample Bayesian *t*-test comparing participants' estimates of the risky door (with respect to the fixed door with the same expected value) to zero (i.e. no over/underestimation). There was very strong evidence in favour of these differences (Experiment 1a: $BF_{Low}= 1.55e{+}05$, [$t(89) = -5.85$, $p < .001$, $d = 0.62$], $BF_{High}= 3174$, [$t(79) = 4.90$, $p < .001$, $d = 0.55$]; Experiment 1b: $BF_{Low}= 1462$, [$t(74) = -4.70$, $p < .001$, $d = 0.54$], $BF_{High}= 81$, [$t(70) = 3.83$, $p < .001$, $d = 0.45$]; Experiment 2: $BF_{Low}=$

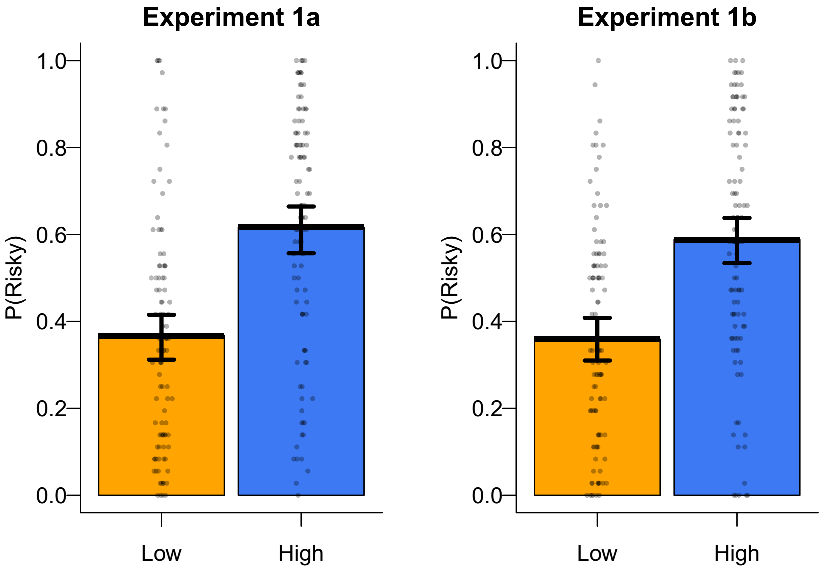

Fig. 2. Proportion of risky choices for each of the high- and low-value decision trials averaged across the last 3 blocks in Experiment 1a and Experiment 1b. Participants were more risk seeking for the high- compared to the low-value trials. Each grey dot represents an individual participant, when more than one participant has the same value the dots become darker.

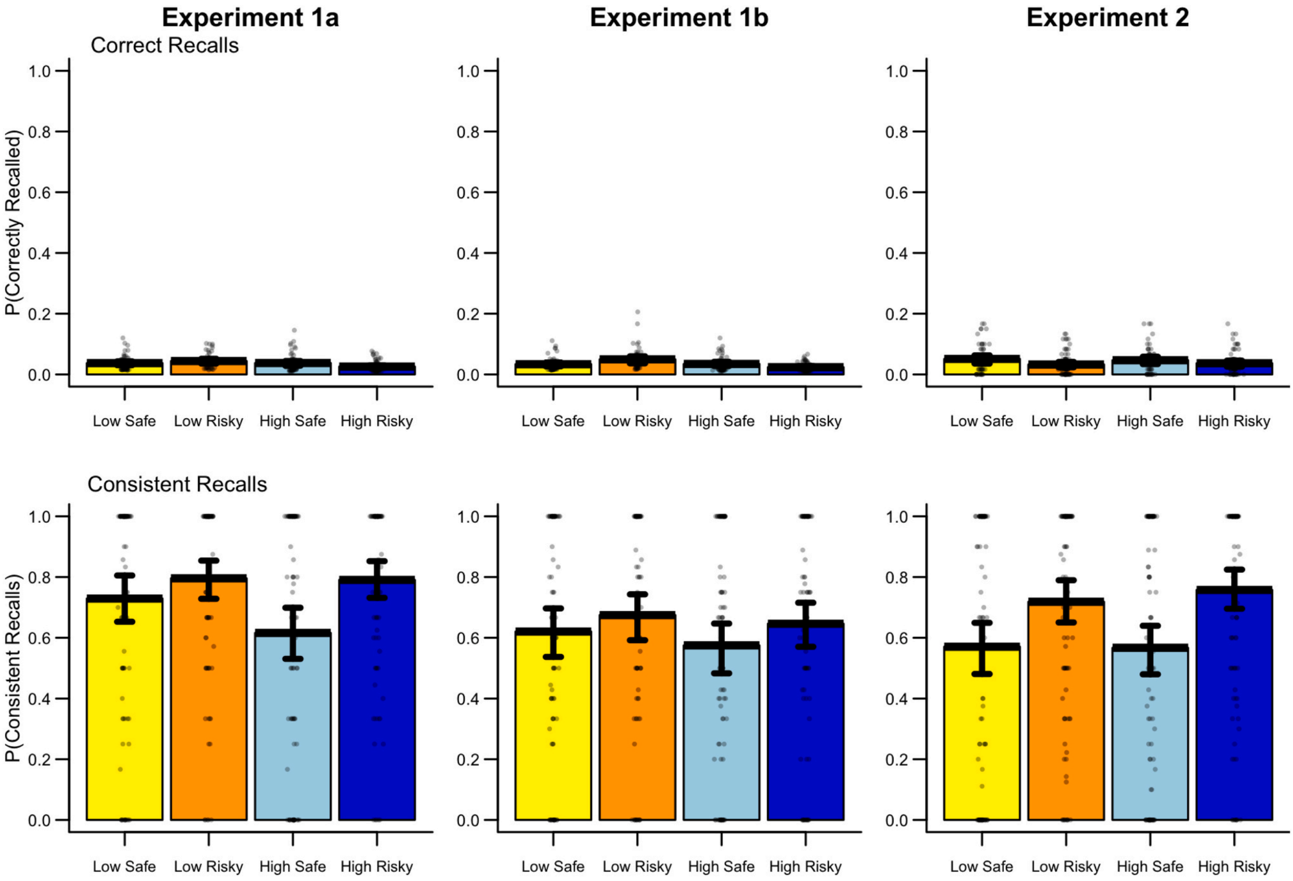

Fig. 3. The first row shows the correct recalls which is the proportion of outcomes correctly recalled in Experiments 1a, 1b and 2. In all experiments participants had very poor recall of the precise outcomes that they had seen. The second row plots the proportion of recalls that were consistent with the possible ranges of outcomes for a given door (see Table 1). These are recalled outcomes that are within the correct range for the door shown, but did not have to be the exact numbers that the participants saw. For example, for the high-risky door in Experiments 1a and 2, consistent recalls would have been within the ranges of 500–590 and 900–990. Across all experiments, participants generated a high proportion of recalls consistent with the cue.

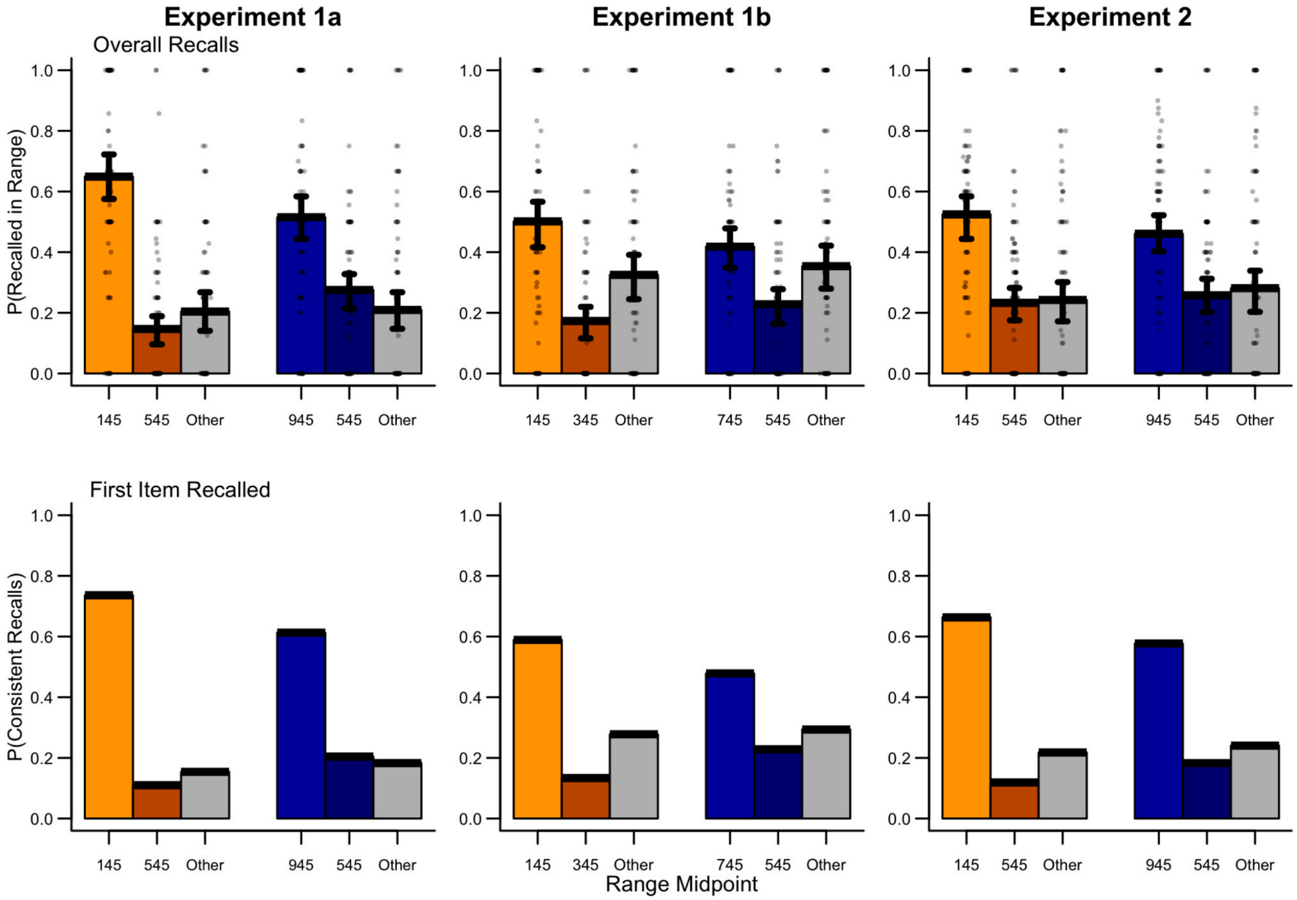

Fig. 4. The top panel shows the proportion of outcomes recalled for the extreme and the non-extreme ranges for the high and low-value risky doors. For each door, the consistent recalls were classified as extreme (EV = 945 and 145 in Exp 1a and 2; EV = 745 and 145 in Exp 1b) or non-extreme (EV = 545 in Exp 1a and 2; EV = 545 and 345 in Exp 1b) and recalled outcomes outside of the door ranges were classified as other. In all experiments, participants generated a greater number of outcomes for the extreme ranges compared to the non-extreme range. For each of the risky doors, we examined the first item that each participant recalled. The second row plots the proportion of participants who reported an extreme, non-extreme, or other outcome first (e.g. an intrusion from a different door). Participants were more likely to report an extreme outcome.

15.9, [$t(83) = -3.27, p = .001, d = 0.36$], BF_{High}= 3861, [$t(80) = 4.95, p < .001$, $d = 0.55$]).

3.4. Memory recall and choice

For Experiments 1a and 1b, we conducted partial correlations[1] to examine the relationship between memory recall and risky choice. To control for variation in the different outcomes individuals experienced, we first binned the experienced outcomes according to the range they were sampled from (e.g., 100–190) and then calculated the proportion of outcomes in this range. For each participant we entered this proportion as a control variable in the partial correlation (see Madan et al. (2014)). For the high-value decisions, there was a positive but non-significant relationship between risky choice and recall of high-value items (Experiment 1a: $R_p(74) = 0.16$, $p = .17$; Experiment 1b: $R_p(67) = 0.16, p = .21$). For the low-value decisions in Experiment 1a, there was a strong relationship between recall and risky choice: the more low-value outcomes people reported in the recall task the less risk seeking they were ($R_p(82) = -0.53, p < .001$). In Experiment 1b the sample size was smaller and there was a non-significant relationship between risky choice and recall of the low-value items ($R_p(67) = -0.15$, $p = .19$).

3.5. Memory recall and estimation

We conducted an exploratory analysis, using another set of partial correlations that controlled for individual experience as above, to examine whether there was a relationship between recall and estimation. Only participants who provided both an estimated average for the doors and at least one memory recall could be included in this analysis. Overall the results from all three experiments provide support in favour of the *Estimation-Overweighting* Hypothesis. For the low-value risky doors, the higher the proportion of recalls in the extreme range, the lower their estimation of that door. There was a strong, significant negative correlation between estimation of the low-value risky door and recall of outcomes in the extreme range of the lower value in all experiments (Experiment 1a: $R_p(85)$= -0.55, $p < .001$; Experiment 1b: $R_p(70)$= -0.33, $p = .006$; Experiment 2: $R_p(79)$= -0.39, $p < .001$). In other words, if people recalled more numbers from the lowest range, they were more likely to underestimate the value of the low-value door. For the high-value risky doors, there was a positive relationship between estimation of the risky-door value and the proportion of recalls in the extreme range (Experiment 1a: $R_p(74)$= 0.29, $p = .013$; Experiment 1b: $R_p(67)$= 0.24, $p = .047$; Experiment 2: $R_p(73)$= 0.52, $p < .001$).

3.6. Exploratory modelling

One possibility for the observed patterns is that people are recalling items by randomly sampling from the ranges/categories. Though people

[1] This was pre-registered as a Bayesian partial correlation; however, the package we planned on using for this analysis was subsequently removed from the online repositories and therefore unavailable.

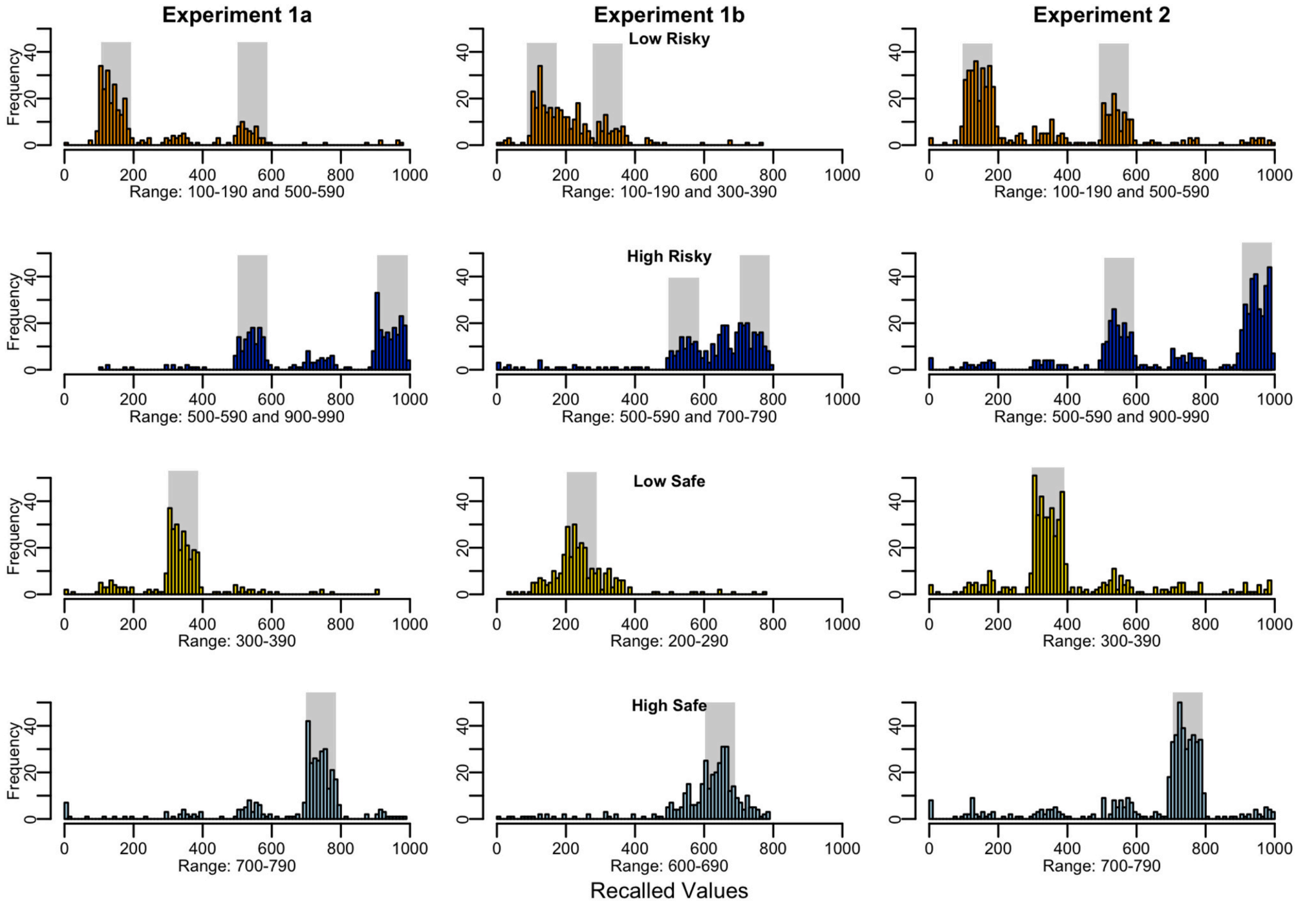

Fig. 5. Distribution of all recalls for each of the risky doors. Participants recalled a greater number of outcomes in the extreme ranges. For example, in Experiment 1a, for the low-value risky option, there were more recalls in the 100–190 outcome range, whereas for the high-value risky option, there were more in the 900–990 range. For both risky doors, the majority of errors were in the ranges of the safe doors (300–390 and 700–790).

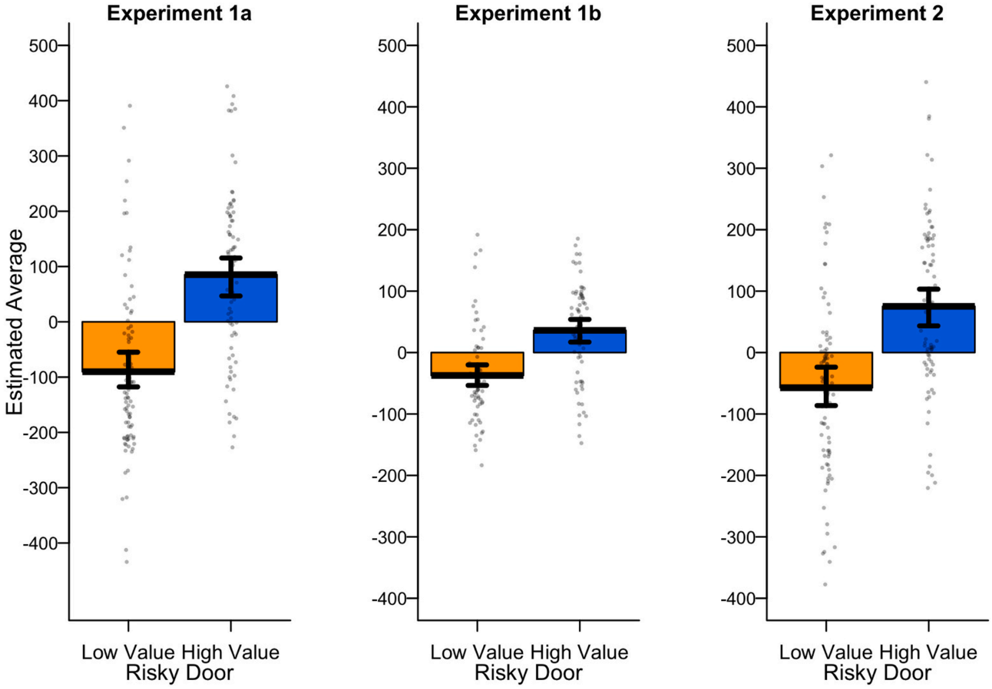

Fig. 6. Participants' mean estimation of the risky doors with reference to the safe door (i.e., after subtracting out participants' mean estimate for the safe door). In all experiments participants overestimated the high-value door and underestimated the low-value door.

clearly did not recall the exact items they encountered, another possibility is that they may have had noisy recall of those items. To distinguish between these two possibilities we tested a *Range-Sampling* model of recall and a *Noisy-Item* model of recall. In both cases we modeled the consistent recalls. Both models generate a probability of recall of all the possible outcomes. Fig. 7 shows an individual participant's probability of recalling each number in the outcome range according to both models.

The Range-Sampling model predicts that, for each option, all numbers within the range (e.g., 300–390) have an equal chance (1/91) of being recalled. For the risky doors there are twice as many possible numbers, and therefore each number has a 1/182 probability of being recalled. The Noisy-Item model assumes that when a number is encountered in the task the probability of recalling this number and nearby numbers increases.

To calculate the probability of recalling each number in a given range (e.g. 100–190), we used a normal distribution around each item that was encountered. The normal distribution had a standard deviation of 2 so as to reflect noisy recall of specific items. For items encountered at the ends of the range (e.g., 102 or 190), we used a truncated normal distribution so that all possible recalls were within the range of numbers encountered (an assumption of both models). The probability of recalling each number can then be determined from the combined 'activation' of each number that was encountered as an outcome in the task. Fig. 7 shows this probability distribution for one participant. We also report the model fits when wider standard deviations are used (5 and 10), which reduced the effect of individual items and makes the probability of recall more similar to the Range-Sampling model. Table 2 shows the negative log likelihood for each model summed across participants.

The Range-Sampling Model was the best-fitting model across all participants, also when fitted to individual participants, including all 323 participants in the three experiments. Thus, a simple memory-plus-noise model does not suffice to capture the pattern of recall beyond random sampling from the range. As the standard deviation of the Noisy-Item model increases, the model becomes more similar to the Range-Sampling model and the fit improves.

Table 2
Negative log likelihood of each model (summed across participants) for the three experiments. The lower the value the better the fit of the model. The best-fitting model is in bold.

Experiment	Noisy-item model			Range-sampling model
	SD = 2	SD = 5	SD = 10	
1a	321,424	286,473	282,067	**279,973**
1b	327,583	292,895	288,595	**286,642**
2	332,770	322,292	320,474	**319,292**

4. Discussion

Together these experiments establish that reported memories for experienced outcomes may be confabulated and can drive a strong decision bias. People primarily recalled outcomes that fit the experienced distributions, but those outcomes were ones that they had not actually experienced. In addition, across all three Experiments we found evidence to support the *Memory-Overweighting*, *Estimation-Overweighting* and *Overweighting-in-Choice* hypotheses. Memory, estimation and choice (Experiments 1a and 1b) exhibited an overweighting of extreme outcomes, even when those outcomes were drawn from a continuous distribution. We did not find support for the *Preferential-Overweighting* hypothesis. Instead, Experiment 2 highlights how this memory bias emerges even in the absence of preferential choice, suggesting that the choice biases are driven by these memory biases and are not the cause of the memory biases (see also Vanunu et al. (2020) and Olschewski, Newell, Oberholzer, and Scheibehenne (2021)). We did not make specific predictions regarding the correlations between memory and choice or memory and estimation. This is because a much larger sample size would be required to make firm conclusions on the basis of these correlations. Nonetheless, in the preferential tasks, we observed a consistent relationship between under/over estimation and risky choice. In Experiment 1a for low-value items there was also a positive relationship between recall of low value items and risky choice. Notably, the memory recalls were not veridical, nor evenly distributed. Instead, people

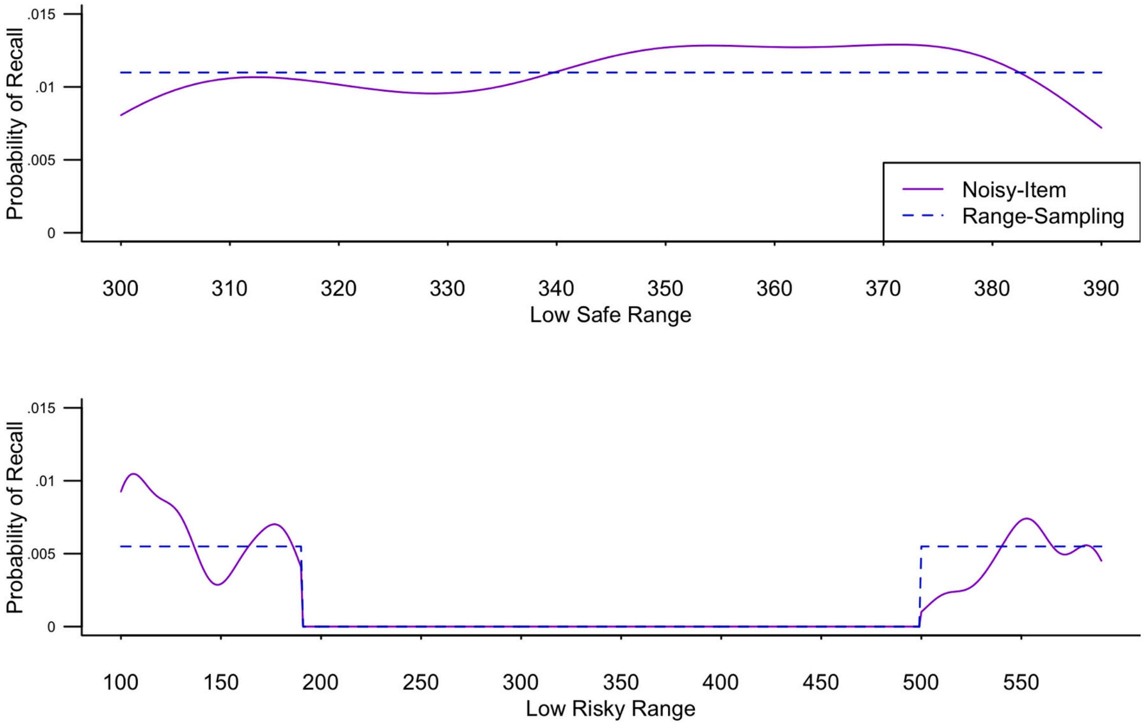

Fig. 7. The two plots show the probability of recall of all possible items according to the Noisy-Item and the Range-Sampling models for a single participant (plot based on Experiment 1a - Participant 11).

generated recalls according to an overweighting of the extreme ranges.

Across all experiments, people were more likely to recall extreme outcomes first, extending earlier work with binary gambles (Ludvig et al., 2018; Madan et al., 2014). In addition, people overweighted the extremes when estimating the mean of the presented options, overestimating high-value risky options and underestimating low-value risky options. Across experiments, people had very low proportions correct for memory recalls (see Fig. 3). Participants, however, effectively encountered a very long list of items (240 total outcomes across all doors), so we would expect low accuracy in a free-recall task, given that accuracy decreases with list length (Ward & Tan, 2004). We therefore examined the errors or intrusions that people made. There is a large literature examining memory for numbers including number representation (Thevenot & Barrouillet, 2006), recognition memory for decision outcomes (Sobkow, Olszewska, & Traczyk, 2020), and memory for probability distributions (Goldstein & Rothschild, 2014). There is, however, scant evidence on free recall of numeric stimuli (but see Dale and Baddeley (1966) and Mason, Brown, Ward, and Farrell (2019)). In similar word-learning tasks, people are more likely to recall semantically related items from lists other than the target list (Miller, Weidemann, & Kahana, 2012). Our results show a surprising pattern of memory intrusions, as we see increased recall of plausible but confabulated outcomes from the same source or “list”.

One interesting possibility is that people are engaging in a form of generalization as seen in category and function learning (DeLosh et al., 1997) or gist abstraction (Feld & Born, 2017). This generalization enables them to confabulate recalls for each of the cues in line with the features of the category (e.g., a number between 100 and 190). This possibility poses the interesting question of how people learn these categories and whether they focus on the edges or boundary conditions. Gist abstraction refers to the process by which people learn a rule or category and generate new samples accordingly and is typically studied in relation to sleep and memory consolidation. Following periods of sleep, compared to wakefulness, people sometimes show better memory for related items than the studied items themselves (Stickgold & Walker, 2013). For example, if a list of words included “snow” and “frost”, people would show increased memory for “cold” (although see Pardilla-Delgado and Payne (2017)). This tendency is similar to the distorted recalls for semantically-related word lists found in the false-memory literature (Reyna, Corbin, Weldon, & Brainerd, 2016). In the false-memory literature, confabulations are false statements, made without the intention to deceive, which can have similar qualities to real memories (Johnson & Raye, 1998). Similarly, the word confabulation has been used to describe ad-hoc generated explanations of one's own behaviour (Bergamaschi Ganapini, 2020; Coltheart, 2017). This view is in line with our use of the term, specifically when participants report outcomes they did not experience but that align with their general experiences.

The present results provide evidence that people are using a higher-level, abstract representation to support decision-making. Such a gist-like trace features prominently in Fuzzy Trace Theory (Brainerd et al., 1999); according to Fuzzy Trace Theory, the verbatim, or item-specific, trace fades faster than the gist trace. Presumably if the experiment were repeated with a shorter delay between items and test, we would observe better recall for specific instances. Based on previous work, reliance on a different type of memory representation could shift risk preference. St-Amand, Sheldon, and Otto (2018) directly examined how gist-based or episodic-memory strategies influence risk-seeking behaviour, using a variation of the current task. Participants were trained to either use episodic-memory recollection or general impression formation before completing a risky-choice task. They found that participants trained to use episodic memory strategies were more risk seeking. Their decision-making task had a limited set of decision outcomes and did not directly test memory accuracy, making it more difficult to determine the extent to which participants were relying on recall of specific instances. Interestingly, the risk preferences (and memory recall patterns) in the current study are consistent with prior findings that use a smaller set of outcomes (Konstantinidis, Taylor, & Newell, 2018; Madan et al., 2021). Based on our results and previous findings, we would predict that when choice involves repeated exposure to similar outcomes people rely on gist-based memory strategies but that when particular outcomes are made salient it is possible to shift risk preferences (Bornstein et al., 2017; Cherkasova et al., 2018; Ludvig et al., 2015; Spetch, Madan, Liu, & Ludvig, 2020).

The precision or coarseness with which outcomes are encoded, represented or retrieved (and reconstructed) provide the raw materials for recall responses (Goldsmith, Koriat, & Weinberg-Eliezer, 2002). In a recent study examining how encoding and retrieval affect risky choices, we found that the encoding context determines how items are used in the decision-making process (Madan et al., 2021). Therefore, one possibility is that the encoding and subsequent representation of the outcomes themselves is noisy or fuzzy (Reyna, 2012). This suggestion is related to the idea that recalling outcomes from memory when making decisions involves reconstructing the original as opposed to making a carbon copy of them (Weber & Johnson, 2006). In Bayesian approaches to reconstructive memory, model noise is introduced as a weighted function of the memory itself and the assumed prior distribution of similar objects or categories in memory (Hemmer & Steyvers, 2009). The prior information is one way that errors in reconstruction can occur, but additionally, samples are drawn from memory with a Gaussian noise distribution centered on the original value of the studied item, similar to the Noisy-Item model of recall we tested. This noisy sampling is akin to decay or interference in other memory-based models (Lehman & Malmberg, 2009; Stewart et al., 2006). In situations where memories are noisy, the prior is weighted more heavily. In our experiments, the repeated exposure to relatively indiscriminate individual events may lead people to rely more heavily on their prior beliefs that extreme outcomes are more informative.

An interesting issue is how well existing models of memory and decision-making would handle the current data. The first issue, which is not new, is how they can account for extreme overweighting, for example by including a sampling bias so that extreme events are sampled more often from memory (Lieder et al., 2018; Vanunu et al., 2020; Vanunu, Hotaling, Le Pelley, & Newell, 2021). A similar bias towards items at the extremes is found in models of estimation (Tsetsos et al., 2012). The current data, however, allows us to further probe what information or representation is being sampled from memory. Here, people confabulated outcomes that came from the correct ranges, but were not the exact outcomes they had experienced. This pattern of sampling from memory is reflected in the Range-Sampling model, which outperformed a Noisy-Item model and assumes that people have learnt the ends of the range. Decision-by-sampling similarly assumes that people have knowledge of the underlying distributions of outcomes and draw a small sample of these from memory to make choices (Stewart, 2009). Our results provide support that people are indeed using knowledge about the range of possible outcomes to support decision-making.

In the absolute identification and production literature, more extreme stimuli are responded to more accurately and faster than central stimuli (Marley & Cook, 1984; Stewart, Brown, & Chater, 2005; Zotov, Shaki, & Marley, 2010). In models of these tasks, the general context of a stimulus set is set by anchors at either end of the range (Braida et al., 1984). The stimuli become more discriminable at the ends of the range, and this effect increases with set size (Lacouture & Marley, 1995). Such a model would provide a good account of the gradual edge effects observed in recalls and the idea that items are generated from within a known experimental range. This approach contrasts to exemplar models that also account for edge effects by assuming that items at the end of the range have fewer neighbors with which to be confused (Brown, Neath, & Chater, 2002; Nosofsky, 1986).

The recall of new or confabulated values is problematic for exemplar representation models (Hotaling, Donkin, Jarvstad, & Newell, 2022) and episodic reinforcement learning models (Bornstein et al., 2017), and the results suggest participants have knowledge instead of the range of outcomes. One possible caveat here is that our task design used outcomes drawn from a set of well-defined ranges. Past memory research, however, indicates that the edges are overweighted even with more continuous outcomes (Madan & Spetch, 2012). Similarly, although not continuous, the ranges used in Experiment 1b are closer than those used in Experiments 1a and 2. On this basis we would predict the same memory and decision biases would be seen if all outcomes occurred over a truly continuous set of outcomes, with no clearly defined ranges. The biases may not be as strong, given that in Experiment 1b, where the ranges are closer together, there were more recalls outside of the prescribed ranges and more confusions between nearby ranges. More generally, the relationship between the operant-like learning involved in decisions-from-experience tasks and the role of episodic memory is not yet well understood (Madan, 2020; Mason, Ludvig, & Madan, 2021). Both our behavioural and modelling results shed light on this relationship and provide evidence against direct episodic recalls of events and casts doubt on models that exclusively rely on individual samples of items from memory (Bornstein et al., 2017; Gonzalez et al., 2003; Hotaling, Donkin, Jarvstad, & Newell, 2022).

An open question is then, under what conditions items are veridically sampled from memory or generated according to a rule when making decisions-from-experience. The design of previous experiments has not allowed for the two approaches to be compared. In other cognitive domains, such as vision, language and reasoning, Bayesian models have been developed where some items are sampled directly from memory or perceptual content and others are simulated/generated (Zhu, Sanborn, & Chater, 2020). This two-pronged approach provides knowledge of the posterior distribution without requiring the distribution to be explicitly calculated. Critically, in complex situations, this approach does not require knowledge of an entire distribution to make effective inferences (Sanborn & Beierholm, 2016). These considerations suggest that both approaches to sampling outcomes (from memory and through a generative process) may be important in decisions-from-experience as well.

A recent study compared how extreme values are used in preferential (averaging tasks) and perceptual (risky choice) judgments (Vanunu et al., 2020). In both tasks, outcomes were displayed simultaneously on screen as an array of numbers. The authors found a bias towards sampling of high extreme numbers regardless of task. In our experiments, we included a perceptual (averaging) task in all experiments, but Experiments 1a and 1b involved choice, whereas Experiment 2 did not. These results support the notion that there are some bottom-up features that determine the weighting of stimuli (Kunar et al., 2017). Interestingly in our experiments the relationship between memory and estimation was stronger in the non-preferential compared to the preferential task, although this may be due to the simplified nature of the non-preferential task and consequent better memory recall overall. Olschewski et al. (2021) found that people consistently under-weighted the average of continuous outcomes in an estimation task. In our task, we demonstrate that people show under- or over-weighting depending on whether a risky option contains the high- or low-value extremes and that the effects of under-weighting are stronger than over-weighting. In both our tasks participants completed the recall task prior to the estimation task, and we therefore cannot exclude the possibility that the order of the tasks influenced the level of over/under estimation as they were more likely to recall extreme values. For the current experiments the recall analysis was central to our research question but future experiments could examine whether task-order effects influence the degree to which participants engage in memory-based strategies during evaluations (Hastie & Park, 1986).

Overall, we have shown that in both preferential and non-preferential experience-based tasks, people did not exhibit veridical recall, but rather confabulated outcomes. These confabulated outcomes were not random, however, and were clearly biased towards being correctly drawn from the more extreme ranges of experienced outcomes. These results suggest that the observed memory biases towards extremes are responsible for the concomitant decision biases and, more generally, provide evidence against decision-making models that rely solely on episodic retrieval of individual instances of past outcomes.

CRediT authorship contribution statement

Alice Mason: Conceptualization, Methodology, Investigation, Formal analysis, Data curation, Writing – original draft, Project administration, Funding acquisition. **Christopher R. Madan:** Conceptualization, Methodology, Writing – review & editing. **Nick Simonsen:** Conceptualization, Methodology, Writing – review & editing. **Marcia L. Spetch:** Conceptualization, Methodology, Writing – review & editing. **Elliot A. Ludvig:** Conceptualization, Methodology, Writing – review & editing, Funding acquisition.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105245.

References

Bergamaschi Ganapini, M. (2020). Confabulating reasons. *Topoi, 39*(1), 189–201. https://doi.org/10.1007/s11245-018-09629-y

Bornstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. D. (2017). Reminders of past choices bias decisions for reward in humans. *Nature Communications, 8*(1), 15958. https://doi.org/10.1038/ncomms15958

Braida, L. D., Lim, J. S., Berliner, J. E., Durlach, N. I., Rabinowitz, W. M., & Purks, S. R. (1984). Intensity perception. XIII. Perceptual anchor model of context-coding. *The Journal of the Acoustical Society of America, 76*(3), 722–731. https://doi.org/10.1121/1.391258

Brainerd, C. J., Reyna, V. F., & Mojardin, A. H. (1999). Conjoint recognition. *Psychological Review, 106*(1), 160–179. https://doi.org/10.1037/0033-295X.106.1.160

Brown, G. D. A., Neath, I. A. N., & Chater, N. (2002). Ratio model of scale invariant memory and identification. In *Memory Lab Technical Report* (pp. 1–94). https://www.ncbi.nlm.nih.gov/pubmed/17638496A.

Chater, N., & Brown, G. D. A. A. (1999). Scale-invariance as a unifying psychological principle. *Cognition, 69*(3), 17–24. https://doi.org/10.1016/S0010-0277(98)00066-3

Cherkasova, M. V., Clark, L., Barton, J. J. S., Schulzer, M., Shafiee, M., Kingstone, A., … C. A.. (2018). Win-concurrent sensory cues can promote riskier choice. *Journal of Neuroscience, 38*(48), 10362–10370. https://doi.org/10.1523/JNEUROSCI.1171-18.2018

Coltheart, M. (2017). Confabulation and conversation. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 87*, 62–68. https://doi.org/10.1016/j.cortex.2016.08.002

Dale, H. C., & Baddeley, A. D. (1966). Remembering a list of two-digit numbers. *The Quarterly Journal of Experimental Psychology, 18*(3), 212–219. https://doi.org/10.1080/14640746608400032

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(4), 968–986. https://doi.org/10.1037//0278-7393.23.4.968

Dougherty, M. R. P. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review, 106*(1), 180–209. https://doi.org/10.1037/0033-295X.106.1.180

Feld, G. B., & Born, J. (2017). Sculpting memory during sleep: Concurrent consolidation and forgetting. *Current Opinion in Neurobiology, 44*, 20–27. https://doi.org/10.1016/j.conb.2017.02.012

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology, 68*(1), 101–128. https://doi.org/10.1146/annurev-psych-122414-033625

Gibson, B., & Zielaskowski, K. (2013). Subliminal priming of winning images prompts increased betting in slot machine play. *Journal of Applied Social Psychology, 43*(1), 106–115. https://doi.org/10.1111/j.1559-1816.2012.00985.x

Goldsmith, M., Koriat, A., & Weinberg-Eliezer, A. (2002). Strategic regulation of grain size in memory reporting. *Journal of Experimental Psychology: General, 131*(1), 73–95. https://doi.org/10.1037//0096-3445.131.1.73

Goldstein, D. G., & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision making, 9*(1), 1–14.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science, 27*(4), 591–635. https://doi.org/10.1016/S0364-0213(03)00031-4

Hastie, R., & Park, B. B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review, 93*(3), 258–268. https://doi.org/10.1037/0033-295X.93.3.258

Hemmer, P., & Steyvers, M. (2009). A Bayesian account of reconstructive memory. *Topics in Cognitive Science, 1*(1), 189–202. https://doi.org/10.1111/j.1756-8765.2008.01010.x

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

Hoffmann, J. A., von Helversen, B., & Rieskamp, J. (2014). Pillars of judgment: How memory abilities affect performance in rule-based and exemplar-based judgments. *Journal of Experimental Psychology: General, 143*(6), 2242–2261. https://doi.org/10.1037/a0037989

Hoffmann, J. A., Von Helversen, B., & Rieskamp, J. (2016). Similar task features shape judgment and categorization processes. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 42*(8), 1193–1217. https://doi.org/10.1037/xlm0000241

Hotaling, J. M., Donkin, C., Jarvstad, A., & Newell, B. (2022). MEM-EX: An exemplar memory model of decisions from experience. *PsyArxiv*. https://doi.org/10.31234/osf.io/fjhr9. Submitted for publication.

Hunt, R. R., & Mitchell, D. B. (1982). Independent effects of semantic and nonsemantic distinctiveness. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*(1), 81–87. https://doi.org/10.1037/0278-7393.8.1.81

Johnson, M. K., & Raye, C. L. (1998). False memories and confabulation. *Trends in Cognitive Sciences, 2*(4), 137–145. https://doi.org/10.1016/S1364-6613(98)01152-8

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition, 106*(1), 259–298. https://doi.org/10.1016/j.cognition.2007.02.003

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General, 132*(1), 133–156. https://doi.org/10.1037/0096-3445.132.1.133

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

Kass, B., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795.

Konstantinidis, E., Taylor, R. T., & Newell, B. R. (2018). Magnitude and incentives: Revisiting the overweighting of extreme events in risky decisions from experience. *Psychonomic Bulletin and Review, 25*(5), 1925–1933. https://doi.org/10.3758/s13423-017-1383-8

Kunar, M. A., Watson, D. G., Tsetsos, K., & Chater, N. (2017). The influence of attention on value integration. *Attention, Perception, & Psychophysics, 79*(6), 1615–1627. https://doi.org/10.3758/s13414-017-1340-7

Lacouture, Y., & Marley, A. A. J. (1995). A mapping model of Bow effects in absolute identification. *Journal of Mathematical Psychology, 39*(4), 383–395. https://doi.org/10.1006/jmps.1995.1036

Lehman, M., & Malmberg, K. J. (2009). A global theory of remembering and forgetting from multiple lists. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 35*, 970–988. https://doi.org/10.1037/a0015728

Lengyel, M., & Dayan, P. (2007). Hippocampal contributions to control: The third way. *Advances in Neural Information Processing Systems, 20*, 889–896.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review, 125*(1), 1–32. https://doi.org/10.1037/rev0000074

Ludvig, E. A., Madan, C. R., McMillan, N., Xu, Y., & Spetch, M. L. (2018). Living near the edge: How extreme outcomes and their neighbors drive risky choice. *Journal of Experimental Psychology: General, 147*(12), 1905–1918. https://doi.org/10.1037/xge0000414

Ludvig, E. A., Madan, C. R., & Spetch, M. L. (2014). Extreme outcomes sway risky decisions from experience. *Journal of Behavioral Decision Making, 27*(2), 146–156. https://doi.org/10.1002/bdm.1792

Ludvig, E. A., Madan, C. R., & Spetch, M. L. (2015). Priming memories of past wins induces risk seeking. *Journal of Experimental Psychology: General, 144*(1), 24–29. https://doi.org/10.1037/xge0000046

Madan, C. R. (2017). Motivated cognition: Effects of reward, emotion, and other motivational factors across a variety of cognitive domains. *Collabra: Psychology, 3*(1), 24. https://doi.org/10.1525/collabra.111

Madan, C. R. (2020). Rethinking the definition of episodic memory. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, 74*(3), 183–192. https://doi.org/10.1037/cep0000229

Madan, C. R., Ludvig, E. A., & Spetch, M. L. (2014). Remembering the best and worst of times: Memories for extreme outcomes bias risky decisions. *Psychonomic Bulletin & Review, 21*(3), 629–636. https://doi.org/10.3758/s13423-013-0542-9

Madan, C. R., Ludvig, E. A., & Spetch, M. L. (2017). The role of memory in distinguishing risky decisions from experience and description. *Quarterly Journal of Experimental Psychology, 70*(10), 2048–2059. https://doi.org/10.1080/17470218.2016.1220608

Madan, C. R., Ludvig, E. A., & Spetch, M. L. (2019). Comparative inspiration: From puzzles with pigeons to novel discoveries with humans in risky choice. *Behavioural Processes, 160*, 10–19. https://doi.org/10.1016/j.beproc.2018.12.009

Madan, C. R., & Spetch, M. L. (2012). Is the enhancement of memory due to reward driven by value or salience? *Acta Psychologica, 139*(2), 343–349. https://doi.org/10.1016/j.actpsy.2011.12.010

Madan, C. R., Spetch, M. L., Machado, F. M. D. S., Mason, A., & Ludvig, E. A. (2021). Encoding context determines risky choice. *Psychological Science, 32*(5), 743–754. https://doi.org/10.1177/0956797620977516

Marley, A. A. J., & Cook, V. T. (1984). A fixed rehearsal capacity interpretation of limits on absolute identification performance. *British Journal of Mathematical and Statistical Psychology, 37*(2), 136–151. https://doi.org/10.1111/j.2044-8317.1984.tb00797.x

Mason, A., Brown, G. D. A., Ward, G., & Farrell, S. (2019). Memory-based and online strategies in retrospective evaluations. *PsyArXiv*. https://doi.org/10.31234/osf.io/uaqs5

Mason, A., Farrell, S., Howard-Jones, P., & Ludwig, C. J. H. (2017). The role of reward and reward uncertainty in episodic memory. *Journal of Memory and Language, 96*, 62–77. https://doi.org/10.1016/j.jml.2017.05.003

Mason, A., Ludvig, E. A., & Madan, C. R. (2021). Conditioning and associative learning. In *The Oxford Handbook of Human Memory*.

Miller, J. F., Weidemann, C. T., & Kahana, M. J. (2012). Recall termination in free recall. *Memory & Cognition, 40*(4), 540–550. https://doi.org/10.3758/s13421-011-0178-9

Murty, V. P., FeldmanHall, O., Hunter, L. E., Phelps, E. A., & Davachi, L. (2016). Episodic memories predict adaptive value-based decision-making. *Journal of Experimental Psychology: General, 145*(5), 548–558. https://doi.org/10.1037/xge0000158

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(1), 54–65. https://doi.org/10.1037/0278-7393.14.1.54

Olschewski, S., Dietsch, M., & Ludvig, E. A. (2019). Competitive motives explain risk aversion for others in decisions from experience. *Judgment and Decision making, 14*(1), 58–71. https://doi.org/2019-08941-006.

Olschewski, S., Newell, B. R., Oberholzer, Y., & Scheibehenne, B. (2021). Valuation and estimation from experience. *Journal of Behavioral Decision Making*. https://doi.org/10.1002/bdm.2241 (September 2020), bdm.2241.

Pachur, T., & Olsson, H. (2012). Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology, 65*(2), 207–240. https://doi.org/10.1016/j.cogpsych.2012.03.003

Pardilla-Delgado, E., & Payne, J. D. (2017). The impact of sleep on true and false memory across long delays. *Neurobiology of Learning and Memory, 137*, 123–133. https://doi.org/10.1016/j.nlm.2016.11.016

Plonsky, O., Teodorescu, K., & Erev, I. (2015). Reliance on small samples, the wavy recency effect, and similarity-based learning. *Psychological Review, 122*(4), 621–647. https://doi.org/10.1037/a0039413

R Core Team. (2020). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rakow, T., & Newell, B. R. (2010). Degrees of uncertainty: An overview and framework for future research on experience-based choice. *Journal of Behavioral Decision Making, 23*(1), 1–14. https://doi.org/10.1002/bdm.681

Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain, 66*(1), 3–8. https://doi.org/10.1016/0304-3959(96)02994-6

Reyna, V. F. (2012). New intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and Decision making, 7*(3), 332–359. https://www.ncbi.nlm.nih.gov/pubmed/25530822A.

Reyna, V. F., Corbin, J. C., Weldon, R. B., & Brainerd, C. J. (2016). How fuzzy-trace theory predicts true and false memories for words, sentences, and narratives. *Journal of Applied Research in Memory and Cognition, 5*(1), 1–9. https://doi.org/10.1016/j.jarmac.2015.12.003

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science, 8*(3), 520–547. https://doi.org/10.1111/tops.12214

Sanborn, A. N., & Beierholm, U. R. (2016). Fast and accurate learning when making discrete numerical estimates. *PLoS Computational Biology, 12*(4), 1–28. https://doi.org/10.1371/journal.pcbi.1004859

Sobkow, A., Olszewska, A., & Traczyk, J. (2020). Multiple numeric competencies predict decision outcomes beyond fluid intelligence and cognitive reflection. *Intelligence, 80* (April), Article 101452. https://doi.org/10.1016/j.intell.2020.101452

Spektor, M. S., Gluth, S., Fontanesi, L., & Rieskamp, J. (2019). How similarity between choice options affects decisions from experience: The accentuation-of-differences model. *Psychological Review, 126*(1), 52–88. https://doi.org/10.1037/rev0000122

Spetch, M. L., Madan, C. R., Liu, Y. S., & Ludvig, E. A. (2020). Effects of winning cues and relative payout on choice between simulated slot machines. *Addiction, 115*(9), 1719–1727. https://doi.org/10.1111/add.15010

St-Amand, D., Sheldon, S., & Otto, A. R. (2018). Modulating episodic memory alters risk preference during decision-making. *Journal of Cognitive Neuroscience, 30*(10), 1433–1441. https://doi.org/10.1162/jocn_a_01253

Stewart, N. (2009). EPS prize lecture: Decision by sampling: The role of the decision environment in risky choice. *Quarterly Journal of Experimental Psychology, 62*(6), 1041–1062. https://doi.org/10.1080/17470210902747112

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review, 112*(4), 881–911. https://doi.org/10.1037/0033-295X.112.4.881

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology, 53*(1), 1–26. https://doi.org/10.1016/j.cogpsych.2005.10.003

Steyvers, M., & Griffiths, T. L. (2008). Rational analysis as a link between human memory and information retrieval. In *The probabilistic mind* (pp. 329–350). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199216093.003.0015.

Stickgold, R., & Walker, M. P. (2013). Sleep-dependent memory triage: Evolving generalization through selective processing. *Nature Neuroscience, 16*(2), 139–145. https://doi.org/10.1038/nn.3303

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction.* Cambridge, MA: MIT Press.

Thevenot, C., & Barrouillet, P. (2006). Encoding numbers: Behavioral evidence for processing-specific representations. *Memory and Cognition, 34*(4), 938–948. https://doi.org/10.3758/BF03193439

Tsetsos, K., Chater, N., & Usher, M. (2012). Salience driven value integration explains decision biases and preference reversal. *Proceedings of the National Academy of Sciences, 109*(24), 9659–9664. https://doi.org/10.1073/pnas.1119569109

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Vanunu, Y., Hotaling, J. M., Le Pelley, M. E., & Newell, B. R. (2021). How top-down and bottom-up attention modulate risky choice. *Proceedings of the National Academy of Sciences of the United States of America, 118*(39). https://doi.org/10.1073/PNAS.2025646118

Vanunu, Y., Hotaling, J. M., & Newell, B. R. (2020). Elucidating the differential impact of extreme-outcomes in perceptual and preferential choice. *Cognitive Psychology, 119*, Article 101274. https://doi.org/10.1016/j.cogpsych.2020.101274

Ward, G., & Tan, L. (2004). The effect of the length of to-be-remembered lists and intervening lists on free recall: A reexamination using overt rehearsal. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(6), 1196–1210. https://doi.org/10.1037/0278-7393.30.6.1196

Weber, E. U., & Johnson, E. J. (2006). Constructing preferences from memory. In S. Lichtenstein, & P. Slovic (Eds.), *The construction of preference* (pp. 397–410). New York NY: Cambridge University Press. https://doi.org/10.2139/ssrn.1301075.

Weilbächer, R. A., Kraemer, P. M., & Gluth, S. (2020). The reflection effect in memory-based decisions. *Psychological Science, 31*(11), 1439–1451. https://doi.org/10.1177/0956797620956315

Wulff, D. U., Mergenthaler-Canseco, M., & Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological Bulletin, 144*(2), 140–176. https://doi.org/10.1037/bul0000115

Zhu, J. Q., Sanborn, A. N., & Chater, N. (2020). The Bayesian sampler: Generic Bayesian inference causes incoherence in human probability judgments. *Psychological Review*. https://doi.org/10.1037/rev0000190

Zotov, V., Shaki, S., & Marley, A. A. J. (2010). Absolute production as a - possible - method to externalize the properties of context dependent internal representations. *Proceedings of Fechner Day*, 197–202.