Spiliopoulos, Leonidas; Hertwig, Ralph

Noisy Retrieval of Experienced Probabilities Underlies Rational Judgment of Uncertain Multiple Events

Journal of Behavioral Decision Making

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Spiliopoulos, Leonidas; Hertwig, Ralph (2024) : Noisy Retrieval of Experienced Probabilities Underlies Rational Judgment of Uncertain Multiple Events, Journal of Behavioral Decision Making, ISSN 1099-0771, Wiley Periodicals, Inc., Hoboken, NJ, Vol. 37, Iss. 5, https://doi.org/10.1002/bdm.70002

Journal of Behavioral Decision Making

WILEY

RESEARCH ARTICLE OPEN ACCESS

Noisy Retrieval of Experienced Probabilities Underlies Rational Judgment of Uncertain Multiple Events

Leonidas Spiliopoulos | Ralph Hertwig

Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

Correspondence: Leonidas Spiliopoulos (spiliopoulos@mpib-berlin.mpg.de)

Received: 22 January 2024 | **Revised:** 2 November 2024 | **Accepted:** 5 November 2024

Funding: Leonidas Spiliopoulos acknowledges support from the Alexander von Humboldt-Stiftung in the form of a renewed research stay.

Keywords: error-based models of rational judgment | learning from experience | overestimation and underestimation | sampling | subjective probability judgment | underextremity

ABSTRACT

Learning the probabilities of multiple events from the environment is an important core competency of any organism. In our within-participant experiment, participants experienced samples from two distributions, or prospects, each comprised of two to four events, and were required to provide simultaneous, rather than sequential, judgment of the likelihood of the complete set of observed events. Empirical calibration curves that map experienced probabilities to subjective probabilities reveal that the degree of underextremity (overestimation of low likelihood events and underestimation of high likelihood events) is strongly conditional on the number of judged events. We uncover two regularities conditional on the number of events that modify (a) the crossover points of the calibration curves with the identity line and (b) the gradient or sensitivity of probability judgments. We present a process model of elicited (subjective) probabilities that captures these empirical regularities. Experienced events recalled from memory may be erroneously attributed to the wrong events based on the similarity of event outcomes. We conclude that the observed miscalibration of probability judgments can be attributed to the noisy retrieval component of a rational process-based decision model. We discuss the implications of our model for the conflicting empirical findings of overweighting and underweighting in the decisions from experience literature. Finally, we show that reliance on small samples can be an ecologically rational strategy for a bounded rational decision-maker (subject to noisy recall), as aggregated subjective probabilities are closer to the ecological probabilities than the experienced (or sampled) probabilities are.

1 | Introduction

Learning the likelihood of real-world events, or ecological probabilities, is a crucial prerequisite and core competency for the adaptation of organisms to their environment that should not be ignored in psychological research (Brunswik 1943). Experiential probability learning is relevant not only to cognitive psychology[1] but also to economics and finance as people must infer the probability of future events based on experience, which can significantly influence beliefs and/or preferences (e.g., Malmendier and Nagel 2011; Lejarraga, Woike, and Hertwig 2016). Peterson and Beach (1967) echoed these sentiments arguing that humans must have evolved competencies that permit the accurate representation of statistical probabilities. In the spirit of the "mind as an intuitive statistician," decision makers' subjective probabilities should be relatively well-calibrated to the true ecological probabilities. However, this does not exclude the possibility of apparent systematic miscalibration that is simply the result of noise. That is, the deterministic component of the procedural model of subjective probability formation may still be perfectly calibrated. Acknowledging that no physical system—including the human brain—can ever function error-free, it would be remiss to call into question decision makers' rationality if

miscalibration is solely the result of an inherent error mechanism. In such cases, subjective probabilities may be inaccurate in terms of levels (or magnitude) compared to the ecological probabilities, but may still be highly positively correlated with them and respectful of the rank-order relationship. On the other hand, systematic miscalibration arising from the deterministic component is valid grounds for questioning the rationality of decision makers (although even this may not necessarily be damning from the perspective of bounded rationality). An example of such cases in probability judgment could arise from decision processes involving representativeness and availability (Tversky and Kahneman 1973; Kahneman and Tversky 1972), which are not guaranteed to be bias-free and could inject systematic miscalibration into probabilistic judgments.

The purpose of this paper is to assess whether decision makers (bounded) rationally learn the probabilities of two to four events only from direct observations of distribution draws without the aid of cues or signals, that is, to understand multiple probability learning as advocated by Vlek (1970). Consider the distinction made by Peterson and Beach (1967, 30), who separate *intuitive statistics* from *psychological decision theory*, where the former concerns gaining knowledge of the environment, and the latter with how to use this knowledge to select courses of action. Analyses based on choices may confound the underlying (separate) perceptual learning process and the decision process, because the latter is also influenced by additional task-related characteristic such as the choice payoffs and specific instructions. In this spirit, we elicit the subjective beliefs of participants, rather than inferring them from decisions derived from them, which may conflate the underlying learning process. Consequently, we can cleanly test the intuitive statistician hypothesis.

The subjective probability judgment (under uncertainty) literature can be broadly classified along two dimensions: the first concerns the source of the uncertainty and the second concerns the type of environmental information that is available to aid judgment. Regarding the source of uncertainty, the bulk of the literature reviews (e.g., see Wallsten and Budescu 1983) deal with *epistemic* uncertainty, that is, events that are in theory knowable (Tversky and Kahneman 1974) and a significantly smaller literature deals with *aleatory* uncertainty, that is events that are inherently stochastic in nature.[2] Our study is interested in subjective probability judgment primarily under aleatory uncertainty with *reducible* epistemic uncertainty. Decision makers experience the likelihood of multiple events by sampling from the true stochastic distribution of the events; as sampling progresses, the epistemic uncertainty of the task is reduced, however the aleatory uncertainty is irreducible. For simple tasks consisting of two mutually exclusive events, the decision-maker often knows a priori the size of the event-space as it is often based on whether a statement is true or false, or whether a specific event will occur or not. However, in our case the size of the event-space (ranging from two to four) is unknown and only revealed through sampling. Consequently, there exist two sources of epistemic uncertainty that are reduced through sampling: (a) the size of the event-space and (b) the true likelihood of events.[3] Such a frequentist sampling process is more closely aligned to how people learn in the real world; this is important as biases in tasks with descriptive probabilities are substantially reduced or even eliminated when they are described as frequencies of events (e.g., Gigerenzer and Hoffrage 1995; Gigerenzer 1996; Hertwig and Gigerenzer 1999).

A key question is how well calibrated are peoples' subjective probabilities? The JDM literature typically concludes that significant biases exist in judgment (e.g., Tversky and Kahneman 1973; Kahneman and Tversky 1972) whereas the cognitive psychology (CP) literature concludes that decision makers are typically well calibrated (Zacks and Hasher 2002; Sedlmeier and Betsch 2002; Betsch et al. 2010; Kelly and Martin 1994; Hintzman 1976; Underwood 1969). We note that these opposing conclusions may be influenced by two important differences in terms of tasks and modeling. The JDM community focuses on tasks of epistemic uncertainty modeled with judgmental heuristics as exemplified by representativeness, anchoring, and availability (e.g., Tversky and Kahneman 1973; Kahneman and Tversky 1972). The CP community instead focuses on tasks of aleatory uncertainty and process-based models involving memory encoding/retrieval and attention. For example, error-based models explain deviations from perfect calibration, such as overestimation of low probabilities, as the consequence of noisy retrieval and/or encoding of experiences (leading to a regression to the mean effect), while the underlying processes are essentially rational (e.g., Erev, Wallsten, and Budescu 1994; Costello and Watts 2014; Fiedler and Unkelbach 2014).

In this study, we will examine how participants learn probabilistic information when observing a sequence or sample of events drawn (freely and without consequence) from the true or objective distribution. We use the dataset from the experiment in Spiliopoulos and Hertwig (2023), who elicited participants' subjective beliefs about the likelihood of *all* the events they had observed. Note, that there exist many studies examining multiple probability judgment *without* learning in *non*-aleatory tasks such as research on conjunction and disjunction effects for multiple events (e.g., Stolarz-Fantino et al. 2003; Costello 2009), but only a handful of studies involving aleatory tasks and learning from sampling. We report below these previous laboratory studies with elicited beliefs but note that—in contrast to our study—they did not implement within-participant tasks concurrently with significant variance in the number of outcomes per prospect (more than two). Fox and Hadar (2006) elicited the subjective probabilities for a small number of lotteries consisting of sure and two-outcome prospects and reported a close correspondence (median correlation of 0.97) between subjective and experienced probabilities. When using the subjective probabilities in a two-stage model of decisions under uncertainty, they found *over*-weighting of rare events (relative to the subjective probabilities) even though as-if underweighting (relative to the objective sampling probabilities) could be inferred from choices. This difference was attributed to the sampling error arising from the use of small samples. Ungemach, Chater, and Stewart (2008, 477) analyzed the observed deviations between objective and subjective probabilities, and uncovered small deviations consistent with overestimation of low probabilities and underestimation of high probabilities. Barron

and Yechiam (2009) examined a set of lotteries consisting of a safe prospect and a two-outcome risky prospect and elicited subjective beliefs in a within-participant design (unlike the between-participant design of other manuscripts). They also concluded that overestimation of experienced probabilities exists simultaneously with underweighting of rare events in choice, a finding confirmed by other subsequent studies as well (Plonsky, Teodorescu, and Erev 2015; Szollosi et al. 2019, fn. 25). Ert and Trautmann (2014, Fig. 2) reported that subjective probabilities overestimated both the experienced and observed probabilities for tasks where participants had to choose between two-outcome risky and ambiguous prospects.

The overestimation of low probability events in decisions from experience (relative to the sampled or experienced probabilities) is consistent with the findings in the wider probability judgment literature covering a very broad range of different tasks. The canonical findings point to a pattern of overestimation of low probabilities and underestimation of high probabilities, the joint pattern referred to as *underextremity* (see Griffin and Brenner 2004, for an extensive review of the literature). More recent research has explored a broader range of tasks involving probability judgment, generally confirming the pattern of underextremity: in beliefs about opponents' behavior in repeated games (Spiliopoulos 2012), when guessing the relative number of black and white dots in an image (Zhang, Ren, and Maloney 2020), in tasks of aleatory and epistemic judgments (Tannenbaum, Fox, and Ülkümen 2017) and temporally distal experience of the likelihood of card combinations in poker (Zhu et al. 2022, see also Wagenaar and Keren 1985).

Our experiment includes payoff distributions with two, three, and four outcomes and within-participant elicitation of the full probability distribution[4] for 240 prospects. This allows us to systematically document subjective probability functions (or calibration curves) and how they depend on the number of events whose likelihood is to be judged. Due to limited sampling by participants, the objective[5] probabilities (OP) will differ from the experienced probabilities (EP) that are realized by the sampling process. Consequently, we distinguish between two types of calibration that are relevant to probability learning from limited experience or samples: (a) *e*-calibration, which refers to the calibration function mapping experienced probabilities to subjective probabilities (SP) and (b) *o*-calibration, which refers to the calibration function mapping objective probabilities to subjective probabilities. As shown below in Figure 1, the driver of the difference between objective and experienced probabilities is the external sampling of events, and the driver of the difference between experienced and subjective probabilities is the (internal) encoding and retrieval processes arising in the mind. Note that an alternative yardstick for rationality, instead of the objective probabilities, can be derived from Bayesian inference, such as Laplace's Rule of Succession (Costello and Watts 2019).[6] We chose to focus on the former for three reasons. First, since we are interested in the ecological rationality of probability judgment, the appropriate metric is the objective probabilities that determine the likelihood of payoffs to a decision-maker. Second, we are interested in examining the implications of our cognitive model for the description-experience gap, which is calculated with reference to the objective probabilities, not those derived from Bayesian inference. Third, when the number of events is a priori unknown as in this study, the principle of indifference upon which the uninformative priors in the Rule of Succession is predicated on is problematic.

The degree of underextremity can be broken down into two components arising from the curvature of the calibration curve and the elevation. Our first contribution is to empirically validate that both *e*- and *o*-calibration curves differ significantly in both elevation and curvature conditional on the number of events. Prior studies have observed that the crossover point (elevation) of the *e*-calibration curve (w.r.t. the perfectly calibrated 45° line) is approximately equal to the inverse of the number of events; the general effect was hypothesized by Fox and Rottenstreich (2003) and See, Fox, and Rottenstreich (2006). For example, the crossover occurs close to 1/26 when estimating the likelihood that words start with one of the 26 letters of the English alphabet (Attneave 1953) and close to 1/4 when estimating the likelihood of four differently colored dots presented visually (Zhang and Maloney 2012). We dub this the 1/N *crossover effect*. Previous studies did not systematically manipulate the number of events within-participants and over a wide range of probability distributions to ascertain the generality and robustness of this effect. We confirm that the 1/N crossover effect is robustly present in our data and also discover that the gradient of *e*-calibration curves, or the sensitivity of subjective probabilities to experienced probabilities falls with the number of events, henceforth dubbed the *inverse-sensitivity effect*.

Our second contribution involves cognitively modeling the observed 1/N crossover and inverse-sensitivity effects. We propose and estimate the similarity-based error diffusion model of probability judgment (SEDM) that implements differential diffusion of recall errors, conditional on a dimension of similarity that operates over the *outcome values* of events. We show that the SEDM best captures the behavior of 30%–40% of our participants, while the remaining participants' behavior is captured by a special case of our model, where similarity does not play a role in recall (the error-diffusion model, EDM). The EDM is consistent with both the crossover and inverse-sensitivity effects. The full SEDM suggests that small deviations from the crossover effect occur due to the dependence on outcome values; however, averaging over different tasks, the expected crossover point is still very close to 1/N, with deviations for individual tasks. Both the SEDM and EDM predict the inverse-sensitivity effect if the

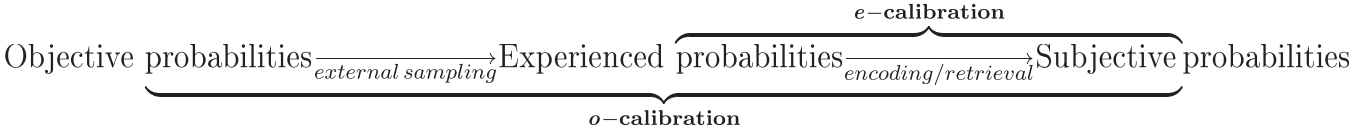

FIGURE 1 | Two measures of subjective probability calibration.

recall error rate increases quickly enough with the number of events to be estimated.

Moving on, we will discuss the strong connections of our model to the decision-experience gap and show how our analysis of *o*-calibration (which is driven by the interaction of external sampling and internal encoding/retrieval processes) relates to the mixed results observed in the literature regarding whether probabilities tend to be overweighted or underweighted with respect to objective probabilities.

Finally, we compare our model to other theories that, while originally designed for judgments of uncertain epistemic events rather than probability learning of stochastic events, could be modified for our task, namely the simultaneous judgment of an exhaustive set of events (rather than individual judgments). These other theories fall into two categories. The first category is the most similar to the SEDM as they are error-based models (the PT + N model by Costello and Watts 2014, and the noisy retriever model, NRM, by Marchiori, Guida, and Erev 2015). The SEDM can be viewed as extending the PT + N model previously applied to tasks of epistemic uncertainty in the following directions: (a) to tasks of irreducible aleatory uncertainty and reducible epistemic uncertainty through sampling, (b) to *multiple exhaustive* events (more than two) requiring some form of normalization, and (c) to similarity-based retrieval from memory. Also, the SEDM shares features with the NRM with respect to (a) modeling errors in the retrieval process and (b) considering the effects of similarity. In the NRM, similarity comparisons result from *between-task* exposure to other similar tasks, whereas in the SEDM they result from similar *within-task* comparisons based on the similarity of the magnitude of payoffs *between-events*. The second category consists of models involving priors (the Bayesian sampler model by Zhu, Sanborn, and Chater 2020, 2022, and partition-dependent support theory by Fox and Clemen 2003, 2005). We question the cognitive foundations of these prior-based alternative models for our specific task, which although relevant for the original tasks these models were proposed for, do not seem conceptually plausible for ours. A special case of our SEDM can be shown to be isomorphic to some of these models from both categories if free parameters are allowed to vary across the number of events—we will discuss this in more detail later. Crucially, none of these alternative models capture the outcome dependence arising from similarity-based retrieval that the SEDM does for a subset (30%–40%) of our participants.

2 | Methods

This study analyzes a subset of data (the elicited beliefs) generated by a previously published experiment (Spiliopoulos and Hertwig 2023), whose data was made publicly available online at the time of its publication https://osf.io/p924g/. We report again the methods of the whole experiment below, as it is crucial in fostering understanding of the complex experimental design and how this relates to the modeling and analysis of the elicited belief data. We note that the elicited beliefs were not modeled or analyzed at all in Spiliopoulos and Hertwig (2023), and they were only used directly as inputs in models of choice behavior as observable data instead of inferring beliefs as latent variables in the estimation procedure.

2.1 | Participants

The experiment was conducted at the laboratory of the Center of Adaptive Rationality at the Max Planck Institute for Human Development with an average participant age of 26 years (s.d. = 4.2, min. = 18, max. = 39) including both students (from all disciplines) and non-students. The data from 96 participants are analyzed after excluding a few participants for specific reasons spelled out in the Supporting Information. The data and code can be found online at OSF repository (https://osf.io/ywkvn/). Approval for the experiment was granted by the Ethics Committee of the Max Planck Institute for Human Development (ARC 2016/37). This study was not preregistered.

2.2 | Procedure

The experiment (first reported in Spiliopoulos and Hertwig 2023) consisted of three within-participants treatments, performed on separate days, typically with a gap of two to three days, designed for another study on decision making from description and experience. From the first two treatments (randomized in order), one consisted solely of decisions from description and the other solely of decisions from experience. Each treatment consisted of the same 120 lotteries comprised of two prospects each. The number of outcomes of each prospect ranged from one to four; here, we are interested in the prospects with two or more outcomes. The lotteries were constructed by quasi-random sampling in the following fashion to ensure that they covered the whole probability and outcome space after fixing the number of outcomes in the lotteries. Twenty (out of 120 lotteries) involved a sure outcome versus a 2-outcome prospect, 40 lotteries involved two 2-outcome prospects, 20 lotteries a sure outcome versus a 4-outcome prospect, and 40 lotteries were comprised of two 4-outcome prospects. Given this fixed structure, payoffs and outcomes were independently drawn uniformly from the discrete probability simplex $\{p_i \in [0{,}0.05, \ \dots, 1]: \ \sum p_i = 1\}$ and uniformly from the range [0,200] in multiples of 10 (without replacement within a single lottery), respectively. A final step removed lotteries with a stochastically dominated prospect and those where the difference in the expected value between the two prospects was greater than 10% of the expected value of the prospect with the highest expected value. The complete experimental instructions can be found in the Supporting Information.

The average number of samples (total from both prospects) drawn by each person was distributed with mean 19.4 (s.d. = 11.2) and median 18 (5% and 95% percentile, 6 and 41, respectively), consistent with the level of sampling in other DfE studies (Wulff, Mergenthaler-Canseco, and Hertwig 2018). Due to sampling, the experienced probabilities differed from the objective probabilities that were specified in the manner set out above. Figure 2 shows the realized probabilities that participants experienced for prospects with N events, confirming that the whole probability space was adequately covered.

The data for this study arises from the third and (always) final treatment of the experiment, the belief elicitation (BE) treatment that did not involve a further decision task based on the experienced probabilities. During the earlier experience treatment, we recorded the exact sampling sequence and resulting outcomes

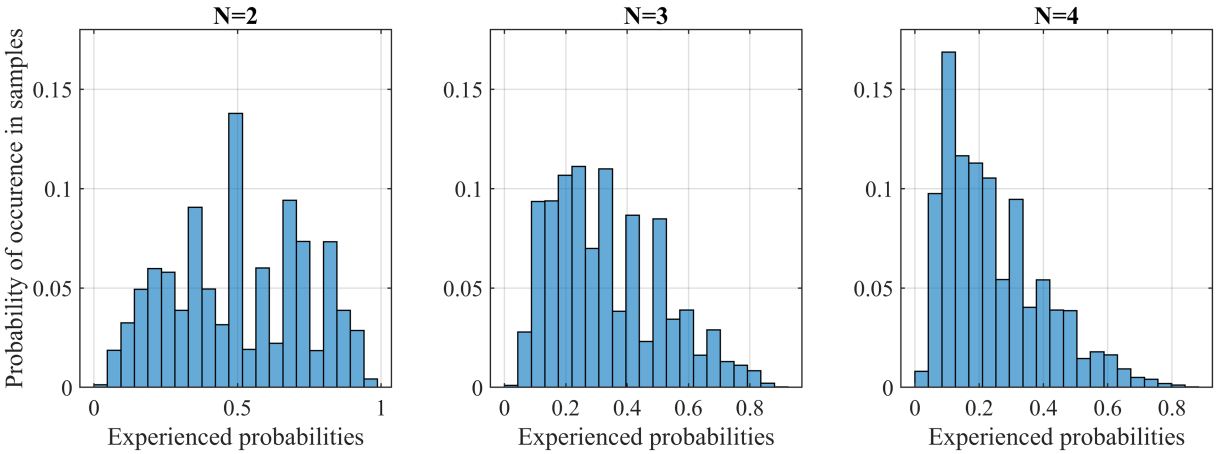

FIGURE 2 | Distribution of experienced probabilities for prospects with two, three, and four outcomes.

observed by each participant. Note, that the outcomes of two prospects may be intermingled through the sampling, as the participants were free to sample from either one in any fashion they desired. Participants would have to keep track of up to eight outcomes (a maximum of four from each prospect) simultaneously throughout the sampling process. In the belief elicitation treatment, we presented each participant with the saved sequences of sampled outcomes they had experienced and asked them to report their estimates of the probability of each of the *observed* outcomes (which could be less than the true number of events in a prospect). They were not informed that these were the samples that they had generated in the prior treatment. An onscreen table was presented where they could record their beliefs with the restriction that probabilities added up to one for each prospect. Participants did not receive any feedback about realized or foregone payoffs or outcomes and could not know how many outcomes existed in each prospect as this could be learned imperfectly only by sampling (some events may not have been sampled). As the order of the lotteries was randomized, they could not infer the number of outcomes in a specific prospect; at best, they could learn from the description treatment that the prospects ranged from one to four outcomes.

That is, we elicited what are sometimes referred to in the literature as relative probabilities; we consider these to be subjective probabilities that are simultaneously elicited, rather than sequentially (or separately) as is often the case. The simultaneous elicitation of the probability distribution of the whole set of events requires elicited beliefs should sum to one, not just in expectation (over many responses/tasks), but also for each individual distribution they were requested to estimate.

3 | Empirical Results

3.1 | Aggregate *e*-Calibration

Figure 3 (left subfigure) plots the relationship between the mean elicited (subjective) probabilities and experienced probabilities averaged over all participants' decisions conditional on the number of experienced outcomes. The size of the markers is proportional to the number of observations for each group of expected probabilities. Note that we have binned the experienced probabilities by rounding to the nearest 0.05 increment, so that we can then compute the mean elicited probabilities for each bin. We exclude the cases where experienced probabilities were equal to zero or one.[7] The right subfigure presents the fitted subjective probabilities using a model that is linear in experienced probabilities.

From Figure 3 it is clear that the participants exhibited underextremity (overestimation for low and underestimation for high experienced probabilities), a regressive pattern often found in the relevant literature on probability judgment (e.g., Edwards 1968; Rapoport and Wallsten 1972; Erev, Wallsten, and Budescu 1994; Spiliopoulos 2012). Upon inspection of the graph, the three relationships (for different N) all seem to be very close to linear, particularly in the regions where many observations are found and the estimates are relatively precise. We formally tested this, comparing a linear model in experienced probabilities to models to quadratic and cubic functions, and concluded that the linear model performed best out-of-sample using cross-validation (see Appendix S2). We note that our conclusion regarding linearity contrasts that of many other studies that typically find an inverse-S shape (e.g., Varey, Mellers, and Birnbaum 1990; Erev, Wallsten, and Budescu 1994; Zhang and Maloney 2012; Zhang, Ren, and Maloney 2020). While directly measuring probability judgments (not inferred judgments made from *choices*), these studies do not involve learning likelihoods by *sequential sampling*, therefore they are not necessarily directly comparable to our tasks. One explanation is that linearity is appropriate or an excellent approximation between the values of 0.05 and 0.95 in the tasks that we investigate, whereas detectable nonlinearities may occur closer to the endpoints.

3.1.1 | The 1/N Crossover and Inverse-Sensitivity Effects

What are the empirically derived crossover points for three and four outcomes? For experienced probabilities of exactly 1/3 and 1/4, the mean subjective probabilities are 0.334 and

0.2491, respectively.[8] We conclude that there is strong evidence for the 1/N crossover effect at the aggregate level. The estimated gradient of the linear fitted models (Figure 3, right panel) for N =2, 3, and 4 is 0.636 [95% CI: 0.627–0.644], 0.471 [0.457–0.485], and 0.428 [0.412–0.445], respectively. The joint hypothesis that these coefficients are equal across N is rejected, $F(2,\ 46,415) = 357.31, p < 0.001$. The estimated sensitivity is therefore consistent with the inverse-sensitivity effect, that is, sensitivity decreases as the number of events N increases.

We conclude that strong evidence exists regarding the regularity of the crossover and inverse-sensitivity effects for aggregate *e*-calibration. However, the analysis on aggregated or pooled data may mask underlying systematic deviations from these effects and may be biased due to ignored heterogeneity. We will revisit the existence of these effects more rigorously at the individual level after proposing a cognitive model of subjective probability formation.

3.2 | Aggregate *o*-Calibration

Figure 4 presents the aggregate *o*-calibration curves, that is, the mapping from objective to the mean subjective probabilities conditional on $N = 2$ and $N = 4$ and averaged over all participants. Recall, that while there existed prospects with three experienced outcomes in the previous analysis, the true prospects had either two or four outcomes; three outcomes only happened when one outcome in a four-outcome prospect was not observed. The crossover point for $N = 4$ calculated as the mean subjective probabilities when the objective probability is 0.25 is equal to 0.263, and when calculated as the median subjective probability, it is 0.25. Consequently, we conclude that the 1/N crossover effect exists not just in *e*-calibration but also *o*-calibration—the latter has important implications for the decision-experience gap that we will return to later. The estimated gradient of the linear fitted models for $N = 2$ and 4 is 0.815 [0.805–0.826] and 0.745 [0.734–0.756], respectively, and are significantly different, $F(1,\ 62,972) = 81.93, p < 0.001$. These are consistent with

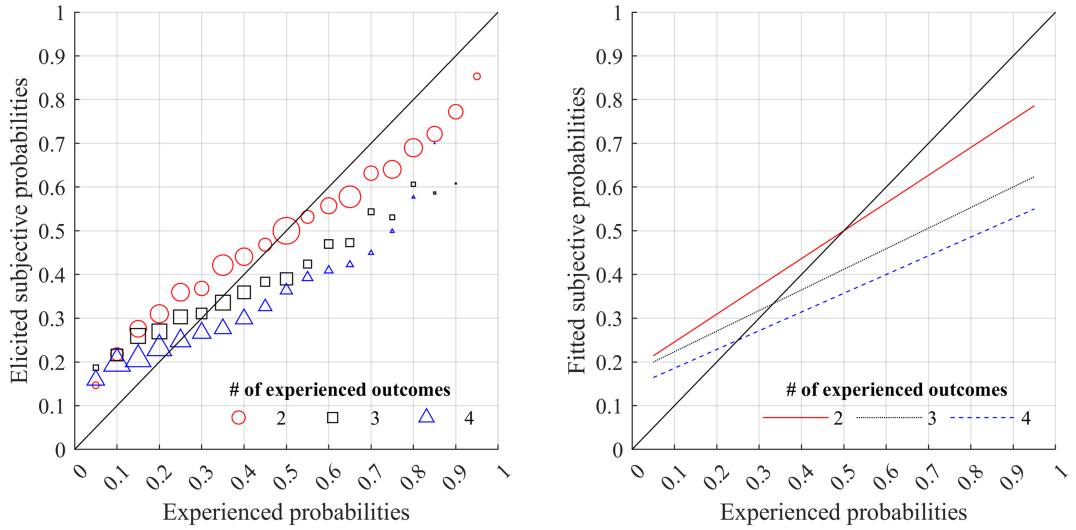

FIGURE 3 | Subjective versus experienced probabilities conditional on the number of experienced outcomes.

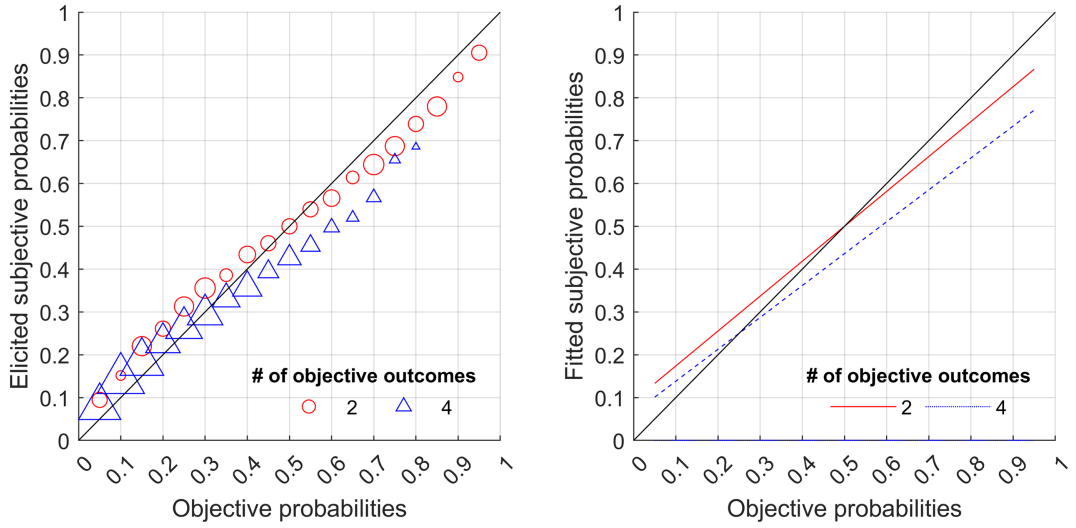

FIGURE 4 | Subjective versus objective probabilities per number of outcomes.

an attenuated (compared to *e*-calibration) inverse-sensitivity effect, as sensitivity falls with the number of events.

A comparison of Figures 3 and 4 reveals that subjective probabilities appear to be closer to the objective probabilities than they are to experienced probabilities. Note, also that sensitivity is now closer to the perfect calibration value of 1, than was the case with *e*-calibration. Define the average degree of miscalibration as the mean of the absolute differences between the mean elicited beliefs and the (binned) experienced or objective probabilities. According to this measure the miscalibration with respect to experienced probabilities is 0.081, 0.127, and 0.13 for N =2, 3, and 4; with respect to objective probabilities, it is 0.045 and 0.063, for N =2 and 4. The dis-aggregated miscalibration results per EP and OP conditional on the number of experienced and objective outcomes (respectively) can be found in Table S1. Bear in mind that these are not exactly comparable as one uses the experienced number of outcomes and the other the objective number of outcomes. Note that the *o*-calibration function is significantly closer to the identity line than the typically estimated probability weighting functions in decisions from description and exhibits less curvature. This implies that the behavior of a Subjective EU decision-maker (using the aggregate of these subjective probabilities) will approximate that of an Expected Value (EV) heuristic, although the degree of similarity to EV behavior in a single task will depend on the specific sampling and error realizations. This evidence suggests that the increase in EV-like behavior in decision from experience is derived from two opposing effects occurring at the different stages of *o*-calibration, the effects of sampling mapping objective probabilities to experienced probabilities, and the *e*-calibration phase mapping experienced to subjective probabilities. The effects of limited sampling, specifically the possibility that some events are not observed at all, serves to bring the *o*-calibration curve closer to perfect calibration. We return to this in more detail in Section 6 discussing the implications for the description-experience gap, and how the conflicting findings of overweighting and underweighting in decisions from experience can be reconciled.

4 | A Cognitive Model of Subjective Probabilities

We propose the similarity-based error diffusion model (SEDM) of ecologically rational probability judgment tempered by noise in the retrieval of the frequentist experienced (or sampled) outcomes. The basic properties of the model in terms of encoding experience from sampling is identical to the PT + N model (Costello and Watts 2014); that is, it is based on the encoding of individual flags representing each observed sample, rather than directly encoding proportions. However, the SEDM differs from PT + N by extending the error mechanism associated with memory retrieval to multiple (more than two) events, by specifying how errors diffuse or are erroneously apportioned to the other events. Furthermore, the search and retrieval mechanism from memory is modified for *simultaneous* estimation of multiple exhaustive events ensuring that the sum of the subjective probabilities is necessarily equal to one. Thus far, these modeling assumptions were derived with knowledge of the empirical regularities presented in Section 3, that is, the 1/N crossover effect and the inverse-sensitivity effect. These assumptions constitute a special case of the SEDM, the error-based model (without similarity). An extension of our original error-based model added a similarity component resulting in the SEDM; note that we had not empirically observed any similarity effect in the data, and therefore, the similarity extension can be considered as making a new prediction rather than an ex post explanation. The decision to add a similarity extension was driven by the fact that similarity-based retrieval (along various dimensions) is found across a wide array of tasks involving memory retrieval and also mediates the degree of interference between memories (McGeoch and McDonald 1931; Conrad 1964; Ratcliff 2022).

Consider an exhaustive set of N^* objective outcomes, so that the probabilities of all events sum to one. In our task, these are the possible outcomes associated with the set of events in each prospect. A decision maker sequentially samples outcomes from two prospects (forming a lottery) and encodes these separately—for each prospect—as exemplars or flags in memory. That is, our encoding mechanism does not rely on a strength measure, but assumes that frequencies are represented in the brain independently (Underwood 1969). Define the *experienced* number of outcomes of a prospect as N (specific outcomes are indexed by n) and let $\mathbf{e_{i,t}} = \left(e_{i,t,1}, \ldots, e_{i,t,n}, \ldots, e_{i,t,N}\right)$ be the experienced probabilities and $\mathbf{v_{i,t}} = \left(v_{i,t,1}, \ldots, v_{i,t,n}, \ldots, v_{i,t,N}\right)$ the experienced outcome values for individual i when sampling from a prospect or task t consisting of N^* events with *objective* probabilities $\mathbf{o^*_{i,t}} = \left(o^*_{i,t,1}, \ldots, o^*_{i,t,N^*}\right)$ and event outcomes $\mathbf{v^*_{i,t}} = \left(v^*_{i,t,1}, \ldots, v^*_{i,t,N^*}\right)$. Due to limited sampling, the DM may not experience all the possible outcomes; therefore, we restrict the model to the experienced outcomes. For samples drawn by an individual for each prospect, upon completion of the sampling processes, the memory register will consist of $F_{i,t}$ flags (equivalent to the number of samples), where each flag represents a single experienced sample of a particular outcome n (assuming that there is no encoding error). Consequently, the number of flags in memory for any event n, $f_{i,t,n}$, is equivalent to the product of the experienced probabilities $e_{i,t,n}$ and the total number of encoded flags, $F_{i,t} = \sum_n f_{i,n,t}$. A decision-maker's set of subjective probabilities $\left\{s_{i,t,1}, \ldots, s_{i,t,N}\right\}$ are formed by retrieving each of the flags from memory with error, counting them and dividing by the total number of flags $F_{i,t}$ to convert it to a probabilistic format. That is, upon retrieving a flag, the DM i may attribute it to the wrong outcome with probability $\psi_{i,N}$—the error rate varies by individual and the number of experienced events N in task t. This belongs to the class of enumeration strategies in frequency estimation (Conrad, Brown, and Dashen 2003).

4.1 | The Similarity-Based Error Diffusion Model

According to the SEDM, the expected retrieved evidence for outcome n will accrue from $\left(1-\psi_{i,N}\right)f_{i,t,n}$ correct flags and from incorrect flags derived from the concurrent retrieval of the set of $\neg n$ events. The errors arising from the retrieval of each of $\neg n$ events (denoted by m) are $\psi_{i,N} f_{i,t,m}$. Let these errors diffuse across the events according to the similarity of events along the dimension of the value of their outcomes denoted by $v_{i,t,n}$—see Ratcliff (2022) for empirical evidence that the degree of confusion between numbers decreases exponentially with distance. Define the similarity measure of two events n and m as $e^{-\lambda_{i,N}\left|v_{i,t,n}-v_{i,t,m}\right|}$, where $\lambda_{i,N} > 0$ scales for the importance of similarity and may vary across individuals and the number of

experienced events in a prospect. Define the relative similarity of these events as $\sigma_{i,n,m} = \frac{e^{-\lambda_{i,N}\left|v_{i,t,n}-v_{i,t,m}\right|}}{\sum_{l:l\neq m} e^{-\lambda_{i,N}\left|v_{i,t,m}-v_{i,t,l}\right|}}$, implying that $\sigma_{i,n,m} = \sigma_{i,m,n}$. We assume that the more relatively similar events are, the higher the proportion of the errors that diffuse between them, that is, given that a flag is erroneously recalled, it is more likely to be misattributed to events that are relatively closer in the outcome space than farther. Furthermore, $\psi_{i,N}$ is assumed to be independent of all other variables including $\mathbf{e_{i,t}}$, $\mathbf{v_{i,t}}$ and $\lambda_{i,N}$. Consequently, for each $\neg n$ event the expected number of flags that will accrue erroneously to event n is given by the total number of flags recalled with error for event m,$\psi_{i,N} f_{i,t,m}$, weighted by the relative similarity between events m and n: $\sigma_{i,m,n}\psi_{i,N} f_{i,t,m}$. Note that our definition of the relative similarity can be interpreted as the proportion of errors in the retrieval of flags for event n that are diffused to event m (and by symmetry, vice-versa). Consequently, all errors from the retrieval of flags for any outcome m are attributed to the other $\neg m$ events and therefore the sum of reported subjective probabilities always equals one.

Let the set $\mathbb{T}_N$ contain tasks with N experienced outcomes only. Individual*e*-calibration curves are derived by taking the expectation for an individual i over tasks t with N outcomes and conditioning on the experienced probabilities and outcomes, see Equation (1). The first term in the square brackets is the contribution from the correctly retrieved flags for event n, and the second is the contribution of the accrued flags for event n from the incorrect retrieval of the $\neg n$ events. In the last step, we make use of the fact that the experienced probability of event n is given by $e_{i,t,n} = f_{i,t,n}/F_{i,t}$.

$$E_{t\in\mathbb{T}_N}\left[s_{i,t,n}|e_{i,t,n},N\right] = \frac{1}{F_{i,t}}\left[(1-\psi_{i,N})f_{i,t,n} + \psi_{i,N}\sum_{n\neq m} E_{t\in\mathbb{T}_N}\left(f_{i,t,m}\sigma_{i,m,n}\right)\right] \quad (1)$$

$$= (1-\psi_{i,N})e_{i,t,n} + \psi_{i,N}\sum_{n\neq m} E_{t\in\mathbb{T}_N}\left(e_{i,t,m}\sigma_{i,m,n}\right) \quad (2)$$

Finally, recall that out definition of rational probabilistic judgment is dependent upon the deterministic component of a decision model. If $\psi_{i,N} = 0$ in Equation (1), then subjective probabilities are perfectly calibrated on average as $E_{t\in\mathbb{T}_N}\left[s_{i,t,n}\,|\,e_{i,t,n},N\right] = e_{i,t,n}$. Consequently, we consider the SEDM model to be one of rational judgment. Note, however, that the subjective probability of event n is dependent upon the experienced probabilities not just of event n, but also of each event $\neg n$, that is, on the whole vector $\mathbf{e_{i,t}}$ (and also $\mathbf{v_{i,t}}$, through the similarity measure $\sigma_{i,m,n}$). Consequently, the individual (pooled over tasks) *e*-calibration curve is dependent on the specific tasks employed.

4.2 | The Error Diffusion Model (Without Similarity)

A special case of the SEDM occurs if similarity is irrelevant $(\lambda_{i,N} = 0)$, henceforth referred to simply as the error diffusion model (EDM). For $\lambda_{i,N} = 0$, the individual *e*-calibration function in Equation (1) collapses to (see the derivation in Appendix S4, Equation S4):

$$E_{t\in\mathbb{T}_N}\left[s_{i,t,n}|e_{i,t,n},N\right] = \underbrace{\frac{\psi_{i,N}}{N-1}}_{\text{constant}} + \underbrace{\left[1-\left(\frac{N}{N-1}\right)\psi_{i,N}\right]}_{\textit{gradient}} e_{i,t,n} \quad (3)$$

Similarly, the EDM is also a rational model of probability judgment as letting $\psi_{i,N} = 0$ leads to perfect calibration, as was also the case with the SEDM. The aggregate *e*-calibration function can be derived by further taking the expectation over all individuals, where $E_i\left[\psi_{i,N}\right]$ is denoted by $\overline{\psi}_N$:

$$E_i\left[E_{t\in\mathbb{T}_N}\left[s_{i,t,n}|e_{i,t,n},N\right]\right] = \underbrace{\frac{\overline{\psi}_N}{N-1}}_{\text{constant}} + \underbrace{\left[1-\left(\frac{N}{N-1}\right)\overline{\psi}_N\right]}_{\textit{gradient}} e_{i,t,n} \quad (4)$$

For both the individual (Equation 3) and aggregate (Equation 4) *e*-calibration functions, it is clear that there is a linear relationship between the expectation of subjective probabilities and experienced probabilities, conditional on the error rate and the number of experienced outcomes. If significant nonlinearities need to be modeled for observations very close to the endpoints, a similar extension to that made to the PT + N model by Howe and Costello (2020) might be appropriate.[9] The crossover point occurs when $E_{t\in\mathbb{T}_N}\left[s_{i,t,n}|e_{i,t,n},N\right] = e_{i,t,n}$, substituting this into Equation (3) reveals this as $1/N$ as we verified empirically earlier—the same result holds in the aggregate. Notably, the crossover point is independent of the level of the error parameter. However, the sensitivity or gradient of the linear EDM $1-\psi_{i,N}N/(N-1)$ is dependent upon both $\psi_{i,N}$ and the number of outcomes N. An increase in the error rate or the number of outcomes leads to a decline in sensitivity, that is, a flatter slope. The gradient is necessarily less than 1 and greater than zero as long as $\psi_{i,N} < (N-1)/N$—this is always true for any N, as long as $\psi_{i,N} < 0.5$. Furthermore, as the EP tends to zero, the SP tends to $\psi_{i,N}/(N-1)$, that is, the constant in a linear regression of SP on EP.[10] The same relationships can be shown to hold for the aggregate *e*-calibration function (with $\psi_{i,N}$ replaced by its aggregate mean $\overline{\psi}_N$) after performing the same steps as above on Equation (4).

5 | Modeling Results

5.1 | *e*-Calibration

We estimate the model in Equation (1) per participant and number of outcomes, and examine the estimated error rates and similarity measure—see Tables S2, S3, and S4 for the detailed regression results and parameter estimates.

Table 1 presents the median estimates of the error and similarity parameters and the proportion of the individual estimates of the latter that are greater than zero, as implied by the similarity hypothesis. The medians of the estimated error rates are increasing with N, or complexity as defined by the number of outcomes to track, 0.18, 0.36, and 0.46 for N =2, 3, 4 respectively. The complexity hypothesis can also be tested

TABLE 1 | Summary statistics of individually estimated parameters conditional on *N*.

	$\widehat{\psi}_{i,N}$	$\widehat{\lambda}_{i,N}\,(\times 10^{-2})$	
# of outcomes	**Median**	**Median**	$p\left(\widehat{\lambda}_{i,N}\right) > 0$
2	0.180	—	—
3	0.358	0.29	0.76
4	0.462	0.36	0.79

more rigorously within-participant. The total possible number of rank orderings of the three error rates of each participant is $3! = 6$; therefore, the expected percentage if there is no association between complexity and error rates is 16.7%. The rank ordering is consistent with the complexity hypothesis for 72% of participants, and the probability of at least such a high result by chance is practically zero (to at least 6 decimal places, signed-rank test, two-sided binomial).

Turning to the similarity-based error diffusion process, Table 1 presents the median of the similarity parameter estimates $\widehat{\lambda}_{i,N}$ for N =3 and 4 and the probability that individual estimates per participant are greater than zero. Evidence for the similarity hypothesis is confirmed at the individual level as 76% (for $N = 3$) and 79% (for $N = 4$) of the participants' estimates are greater than zero. Examining whether $\widehat{\lambda}_{i,N}$ estimates were significantly different from zero using one-sided 95% (bootstrapped) confidence intervals revealed that $\lambda_{i,3}$ and $\lambda_{i,4}$ were significantly greater than zero for 30.2% and 40.4% of the participants, respectively—these values are significantly different (binomial test, $p < 0.0001$) from what would be expected due to the multiple comparisons (on average 5% due to chance).[11] Recall that for $\lambda_{i,N} = 0$, the SEDM collapses to the simplified linear EDM. Consequently, we conclude that 30.2%–40.4% of participants' behavior was significantly more accurately modeled by the SEDM, while the remaining percentage is represented by the EDM.

Our categorization above of participants into two types (SEDM and EDM) was based on statistical significance. Another interesting question is how different are the two models conditional on our task in terms of behavioral predictions? Specifically, what is the improvement in estimated errors moving from the EDM to the SEDM (with $\lambda_{i,N}$ as a free parameter)? Consider a comparison in the RMSE for participants for whom $\widehat{\lambda}_{i,N}$ was identified as significantly greater than zero above: For N =3 and 4, the EDM's RMSEs are 0.132 and 0.099, and the SEDM's are 0.128 and 0.096, respectively. The distribution of the differences between the two models for all participants for N =3 and 4 is presented in Appendix S5 (Figure S1). In general, the differences between the two models are small (in terms of magnitude), despite the finding that $\widehat{\lambda}_{i,N} > 0$ for many participants. Consequently, the improvement of allowing $\lambda_{i,N}$ to vary, rather than fixing it to zero (the EDM), are minimal for our particular set of prospects—this need not be the case for a different set that may include prospects for which the two models diverge more. In our tasks, for behavioral analyses, we could assume the linear EDM arising from $\lambda_{i,N} = 0$ with a negligible practical loss in performance. This will allow us to easily test the inverse-sensitivity effect; as for the EDM, the sensitivity is constant across the whole range of experienced probabilities and equal to the gradient $\left[1 - \left(\frac{N}{N-1}\right)\psi_{i,N}\right]$, recall Equation (3). The median estimated sensitivity of individuals verifies the inverse-sensitivity effect previously validated at the aggregate level: 0.641, 0.467, and 0.385 for N =2, 3, and 4, respectively. Examining the inverse-sensitivity effect within-participant, we find that the expected rank-order relationship holds for 54% of participants. When $\widehat{\lambda}_{i,N} > 0$ in the SEDM, then whether subjective probabilities satisfy the 1/N crossover effect depends on each tasks' specific outcome values. This is evident in Figure 5, a graphical representation of the individual-level predicted subjective beliefs of the EDM and SEDM. However, there is still an aggregate 1/N crossover effect when averaging over all our tasks, even though divergence can be seen for individual tasks—see the black line in the SEDM subgraph, which is a linear fit to the SEDM predictions, whose crossover points are virtually identical to 1/N.

5.2 | Relationship to Other Theories

In this section, we discuss the relevance of the following alternative models of probability judgment to our experiential task, in some cases with necessary modifications to extend them to multiple events: the Probability Theory + Noise (PT + N) model (Costello and Watts 2014), the Bayesian sampler (BS) model (Zhu, Sanborn, and Chater 2020; Zhu et al. 2022), and partition-dependent support theory (ST) (Fox and Rottenstreich 2003; Fox and Clemen 2005).

First, the PT + N model can accommodate the elicitation of multiple subjective beliefs over a set of mutually exclusive events by introducing an additional normalization stage after using the standard model of sequential and separate estimation of each event's likelihood.[12] Note that the EDM (without similarity) collapses to such a normalized version of the PT + N model—see Appendix S3 for details. The standard PT + N model was originally introduced for tasks involving judgments rather than outright probability learning. Uncertainty about probabilities in the tasks of Costello and Watts (2014) is a product of lack of knowledge (epistemic uncertainty), whereas in our case it arises from the sampling and the experiential process of a truly stochastic environment (aleatory uncertainty). There are other important differences with their model, as we require it to *simultaneously* predict the likelihoods of multiple outcomes, not binary outcomes (or multiple individual outcomes *sequentially*).[13]

Proofs of the following claims and a more detailed discussion can be found in Appendix S3. The BS model can be extended to multiple outcomes by substituting the Beta distribution prior with a Dirichlet prior. The ST model does not require any modification as it already captures multiple partitions by construct. All three modified models correctly predict the 1/N crossover effect. The PT + N model does this through the error mechanism, which is similar to that of the EDM. By contrast, the ST and BS models

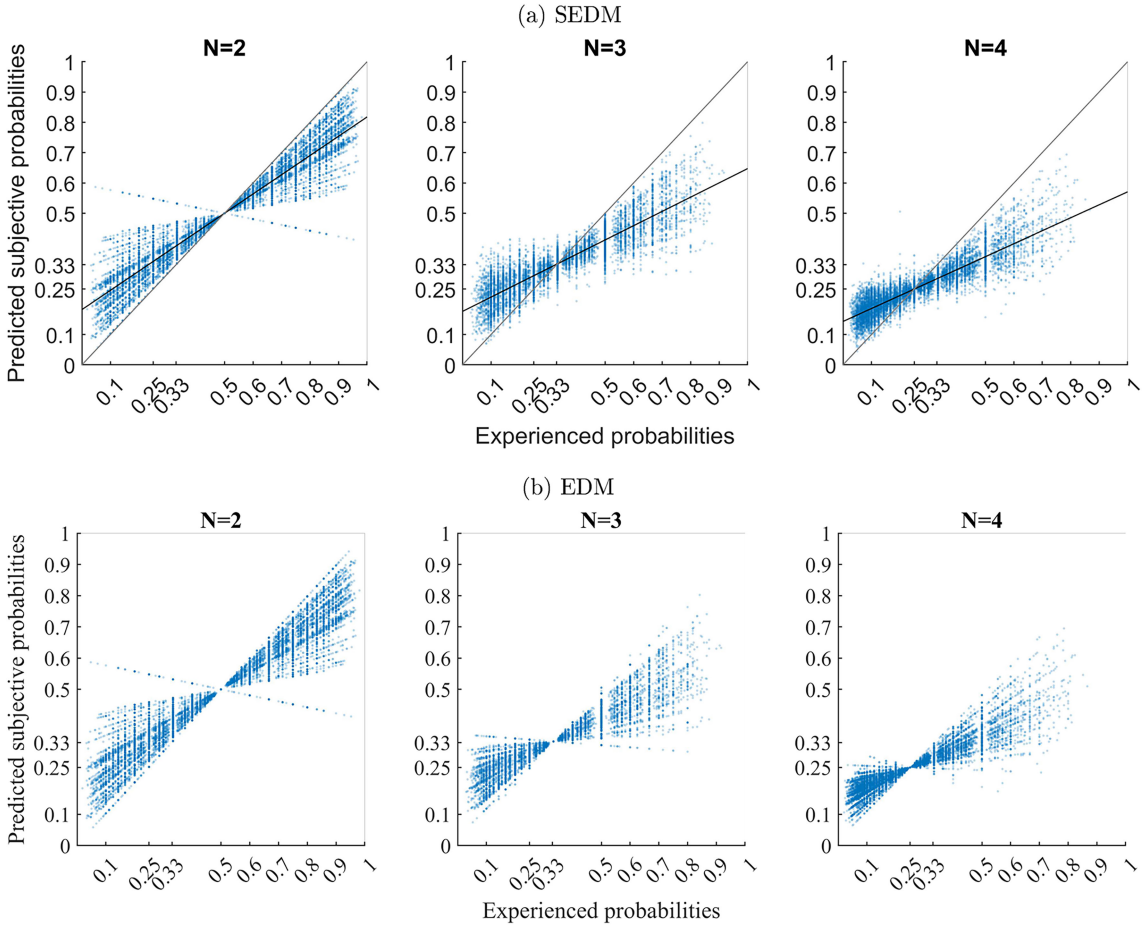

FIGURE 5 | Predicted subjective probabilities conditional on *N* for SEDM and EDM.

do so by assuming that subjective probabilities are influenced by uniform priors over the set of events. It is not clear how priors would play a role in our specific task (although they are relevant for the original tasks these models were applied to). A further implication of models using flat priors is that it is reasonable to expect that the relative weight attached to priors would be related to how much participants sampled. That is, the more they sampled, the greater the sensitivity to experienced probabilities as relatively less weight would be attached to the flat priors. We did not discover such a systematic relationship in our data, see Appendix S3. Note that it is trivial to extend the SEDM and EDM for other tasks where decision-makers may have priors over the likelihood of events (perhaps even unobserved events).[14]

Furthermore, we show the following results in Appendix S3. The normalized PT + N model and the BS model with Dirichlet priors are isomorphic to our EDM model—that is, they would make the same predictions as the EDM model, but not the SEDM. Consequently, our results that showed that the SEDM model better accounted for 30%–40% of participants than the EDM model, implies that the SEDM is also more appropriate than the PT + N and BS models for these participants. This is because the ST, PT + N, and BS models cannot capture the dependence on outcome values that the SEDM does. Also, the ST model is not isomorphic to the EDM as it is a nonlinear function in the experienced probabilities, whereas the EDM, PT + N, and BS models are.

6 | *o*-Calibration and Underweighting Decisions From Experience

In decisions from experience, limited sampling means that the experienced probabilities of low probability outcomes are more often lower than the objective probabilities than they are higher, although the expectation is the same[15] (Hertwig et al. 2004). This effect can contribute to an as-if underweighting of probabilities in choices; estimated probability weighting functions in decisions from experience have mixed forms, in some cases inverse-S shaped and in other cases S-shaped (Wulff, Mergenthaler-Canseco, and Hertwig 2018).[16] How do these findings about *o*-calibration relate to our model and the interaction between sampling and *e*-calibration? The tendency for underweighting due to limited sampling is only one of two opposing effects. As we showed above participants' subjective probabilities of low probability events typically overestimate the experienced probabilities, contributing to an as-if overweighting of probabilities. As the two effects are in opposite directions, the total effect (captured by *o*-calibration) is dependent on the relative magnitude of these effects. These magnitudes will depend on how much an individual samples (more samples decrease the magnitude of the sampling effect), the objective probabilities as the sampling effect and the degree of overestimation is not constant across the probability range, and the error rates of individuals during the subjective probability stage (which determine the degree of overestimation

captured by *e*-calibration). Whether underweighting or overweighting occurs on average will depend on the relative magnitude of these two effects.

Recall from Section 3 that *o*-calibration curves were better calibrated than the *e*-calibration curves—see Figure 4. This occurs due to the cases where the experienced probabilities are zero for events, whose associated objective probabilities are non-zero. Overestimation of low probabilities occurs in *e*-calibration, but only events that are *observed* with non-zero probability are considered by decision-makers. Unobserved events during sampling are an exception and are underestimated as they are implicitly assigned a subjective probability of zero. The key difference is that in *o*-calibration, all events are considered when mapping objective to subjective probabilities, but subjective probability formation is captured by *e*-calibration of experienced events only. Since the probability of not sampling an outcome that occurs with positive probability is higher the rarer the outcome is, we should expect low probability events to be better *o*-calibrated, which is what we showed earlier.

7 | Discussion

Our manuscript sought to systematically investigate within-participant calibration functions relating subjective probabilities to experienced probabilities arising from the sampling of multiple-outcome event distributions. We empirically determined the existence of two important regularities conditional on the number of events whose likelihoods must be estimated, the 1/N crossover effect and the inverse-sensitivity effect. These two regularities lead to changes in both the elevation and curvature of the calibration curves relating subjective probabilities both to experienced probabilities from sampling and the objective probabilities of the true distributions, impacting the degree of underextremity that subjective probabilities typically exhibit. This is important in considering the external validity of the existing literature, which has focused primarily on modeling probability judgments of two-outcome distributions. It is of direct relevance to the decisions from experience literature where decision makers learn about multiple probabilities by sampling from the environment. Consequently, our findings speak strongly to real-world decision making, which often involves multiple stochastic events, whose likelihoods are rarely described rather than experienced. At the same time, probabilistic reasoning and judgment are a core element of psychological theorizing and economic behavior—understanding the intricate relationships between ecological or objective probabilities, experienced and subjective probabilities is of paramount importance.

We have also shown the importance of considering process models of cognition when considering the rationality (or lack thereof) of human behavior. At a superficial level, our empirical findings seem strongly suggestive of biases in human reasoning, as not only were subjective probabilities not perfectly calibrated, but also sensitive to the number of events under observation. Despite the significant empirical evidence we presented of deviations from perfect calibration, we show that these can arise from a (noisy) rational model of probability judgment. The proposed similarity-based error diffusion model of subjective probability formation effectively captures within-participant variation in calibration curves conditional on the number of events to be estimated, and between-participant variation driven primarily from differences in error rates and the importance of similarity in the retrieval mechanism. That is, assuming no errors in memory encoding or retrieval, the deterministic component of our model collapses to perfect calibration, where the expectation of the subjective probabilities equals the experienced probabilities. A special case of our proposed model (without similarity) still predicts the 1/N crossover and inverse-sensitivity effects, and leads only to a small degradation in performance compared to the full model.

Another key empirical finding of our study is that subjective probabilities are better calibrated with respect to the ecological (or objective) probabilities than to the experienced (or sampled) probabilities. Our cognitive model provides a succinct explanation for this intriguing finding. Noisy retrieval of the sampled events generally leads to overestimation of low probabilities. However, limited sampling produces a counteracting effect, leading to the underestimation of low probabilities, on average. The result is that limited sampling actual serves to improve calibration with respect to the ecological probabilities, and consequently can be considered an ecologically rational strategy in the face of the irreducible nature of noisy memory. This is a novel justification for the robustness of limited sampling, which is pervasive in experimental and field studies.

Based on the evidence we have presented, we believe that the most important effects of the number of events on the calibration of subjective probabilities seem to be captured very well by our model; however, this is not to say that more complex models of exemplar encoding and retrieval are irrelevant, and would not capture other effects that our model may be missing. One example is the possibility of significant nonlinearities at extreme probabilities that are not generally predicted by error-based models including our own, although an extension similar to that in Howe and Costello (2020) could also be applied to our model. Another example, in decisions from experience many participants may learn the conditional (rather than just unconditional) probabilities of events, for example, the probability of observing a pattern of outcomes (Plonsky, Teodorescu, and Erev 2015; Plonsky and Erev 2017). Similarly, conditional probabilities or pattern learning has been found in strategic interactions, or games, including cognitive operations such as forgetting, similarity-matching between patterns during retrieval (Spiliopoulos 2012; 2013a; 2013b). On the other hand, the impact of patterns may be reduced as the number of outcomes increases, as limited sampling in decisions from experience will drastically reduce the frequency of occurrence (and repetition) of patterns in the sample, undermining the learning of conditional probabilities due to increased uncertainty. In contrast to such sequential pattern-based similarity, our similarity measure was defined over the dimension of outcome values. Subjective probabilities may depend on the magnitude of outcomes via other mechanisms as well. For example, outcomes large in magnitude, whether negative or positive, may mediate the attention given to these experiences, thereby possibly impacting the encoding and retrieval of exemplars from memory.

Our goal was to model the formation of subjective probabilities when decision-makers observe a sequence of outcomes by sampling from the true distribution (decisions from experience) and must concurrently estimate likelihoods of a set of collectively exhaustive events, whose probabilities necessarily sum up to one. Future work should be directed at examining our model's predictive ability in tasks where only a subset of experienced events are to be estimated, not just simultaneously, but also sequentially.

Acknowledgments

The authors would like to thank David Budescu, Fintan Costello, Tomás Lejarraga, and anonymous referees for comments and feedback. Leonidas Spiliopoulos acknowledges support from the Alexander von Humboldt-Stiftung in the form of a renewed research stay. Open Access funding enabled and organized by Projekt DEAL.

Data Availability Statement

The data that support the findings of this study are openly available in OSF at https://osf.io/ywkvn/.

Endnotes

[1] For example, see the sequence learning literature (Remillard and Clark 2001; Sun and Giles 2001; Clegg, DiGirolamo, and Keele 1998; Nissen and Bullemer 1987), which examines whether conditional probabilities can be learned through experience and the decisions from experience literature, where outcomes and their relative frequencies must be learned through sampling (Barron and Erev 2003; Hertwig et al. 2004; Hertwig and Erev 2009; Erev and Roth 2014).

[2] A similar distinction is made by Erev, Wallsten, and Budescu (1994), referring to uncertainty that is internal to the decision maker (e.g., due to lack of knowledge) or external to the decision maker (e.g., it is inherent to the environment).

[3] Uncertainty about the latter is composed of irreducible aleatory uncertainty related to the true likelihood of events occurring, but also of epistemic uncertainty related to the learning process of these likelihoods. Thus, the more the decision-maker samples, the closer the experienced samples will be to the true likelihood, reducing the epistemic uncertainty of subjective estimates.

[4] There is a subtle distinction between the subjective beliefs of the objective probabilities and those of the experienced probabilities. The former must still be based on the experienced probabilities—unless a strong prior exists—but would require sophisticated corrections for the small sample properties of estimators of outcome likelihoods. We are unaware of empirical evidence of such bias corrections in peoples' estimates; therefore, we believe that attempting to elicit either of these two quantities would likely result in very similar reported beliefs. Note that participants must also first learn the outcome space from sampling before inferring the likelihood of outcomes within that space. This further encourages the convergence of the subjective beliefs of objective and experienced probabilities, as unobserved outcomes will not be reported in either case.

[5] We use the term *objective* probabilities in keeping with the terminology in the decisions from experience literature. Alternatively, they may be referred to as the *generating* probabilities of the realized samples drawn from a prospect or the *true* probabilities of a prospect.

[6] Upon observing k occurrences of an event from a sample of size N, the optimal inference of the objective probability of said event is not k/N, but rather $k+1/N+2$. This can be extended to multiple events E to $k+1/N+E$, which however assumes that E is known a priori, which will not be the case in our tasks.

[7] Our experimental setup necessarily implied that the elicited beliefs in the former were zero (as only experienced outcomes were presented in the table that participants had to fill out), and presenting only a single outcome (for an experienced probability of one) clearly implied a subjective probability of one. We will return later to these special cases of experienced probabilities (EP) equal to zero or one and their importance for the transformation of objective probabilities (OP) to subjective probabilities (SP).

[8] Conducting two-sided signed-rank tests with the null hypothesis that the median of the distribution of the subjective probabilities is equal to 1/N: $N=3$ $(z=-0.889, p=0.374)$ and $N=4$ $(z=-2.275, p=0.023)$. While for $N=4$, the hypothesis of no difference is not accepted at the 5% level (due to a large number of observations); for all intents and purposes, they are identical.

[9] This extension accommodates an inverse-S shape because proportions exhibit less variation at the endpoints of the probability scale, thereby diminishing the regression effect.

[10] We examine the parametric form of the model for $0<e_n<1$, as the endpoints exhibit discontinuities at $e_n=0$ or 1. In the case of $e_n=0$, an outcome that has not been encoded in memory has no chance of being retrieved either, therefore $s_n=0$. If $e_n=1$, since only a single outcome has been observed, there are no other outcomes to be erroneously retrieved; therefore, $s_n=1$.

[11] We consider the one-sided test to be more relevant as the similarity hypothesis clearly implies that λ is greater than zero. For the sake of completeness, we note that using instead a 95% two-sided confidence interval we conclude that 24% and 30.9% were significantly different from zero for $N=$3 and 4, respectively.

[12] We are unaware of any published implementation of such a normalization stage, and we thank an anonymous reviewer for suggesting this.

[13] Their model predicts empirical regularities from the judgment literature, for example, conservatism, subadditivity, the conjunction, and disjunction fallacies.

[14] Consider that said priors, however they may arise, are simply encoded and retrieved as flags just as the sample outcomes are. The error-retrieval based mechanism could then be applied to them as well; that is, flags represent both prior and experienced relative likelihoods.

[15] The majority of samples of size k will not include a rare event with probability q if $(1-q)^k>0.5$, or equivalently if $k<\log(0.5)/\log(1-q)$. For $q=0.1$ this will be the case for $k<6.57$. Therefore, if k is less than or equal to six, most samples will not include the rare event (Teoderescu, Amir, and Erev 2013). Note that contrary to prior assertions of consistent underestimation for probabilities below 0.5, Shteingart and Loewenstein (2015) find the existence of a zigzag or alternating pattern of underestimation and overestimation.

[16] Regarding the as-if underweighting of rare probabilities as determined from differences in choice proportions across description and experience, in this case, what matters is not necessarily the *expectation* of the subjective probabilities, but the *likelihood* that individual estimates underestimate or overestimate the objective probabilities. As is well known, the OP to EP stage is more prone to underestimation of rare events than overestimation. The EP to SP stage is more likely to overestimate the (low) experienced probabilities; the degree of which depends on the error rate. Consequently, whether subjective probabilities are more or less likely to underestimate the objective probabilities will depend on the relative magnitude of these opposite effects: the first arising from the OP-EP stage and the second from the erroneous retrieval in the EP-SP stage.

References

Attneave, F. 1953. "Psychological Probability as a Function of Experienced Frequency." *Journal of Experimental Psychology* 46, no. 2: 81–86. https://doi.org/10.1037/h0057955.

Barron, G., and I. Erev. 2003. "Small Feedback-Based Decisions and Their Limited Correspondence to Description-Based Decisions." *Journal of Behavioral Decision Making* 16, no. 3: 215–233. https://doi.org/10.1002/bdm.443.

Barron, G., and E. Yechiam. 2009. "The Coexistence of Overestimation and Underweighting of Rare Events and the Contingent Recency Effect." *Judgment and Decision Making* 4: 447–460.

Betsch, T., M. Glauer, F. Renkewitz, I. Winkler, and P. Sedlmeier. 2010. "Encoding, Storage and Judgment of Experienced Frequency and Duration." *Judgment and Decision Making* 5, no. 5: 347–364.

Brunswik, E. 1943. "Organismic Achievement and Environmental Probability." *Psychological Review* 50, no. 3: 255–272. https://doi.org/10.1037/h0060889.

Clegg, B. A., G. J. DiGirolamo, and S. W. Keele. 1998. "Sequence learning." *Trends in Cognitive Sciences* 2, no. 8: 275–281. https://doi.org/10.1016/s1364-6613(98)01202-9.

Conrad, R. 1964. "Acoustic Confusions in Immediate Memory." *British Journal of Psychology* 55, no. 1: 75–84. https://doi.org/10.1111/j.2044-8295.1964.tb00899.x.

Conrad, F. G., N. R. Brown, and M. Dashen. 2003. "Estimating the Frequency of Events From Unnatural Categories." *Memory & Cognition* 31, no. 4: 552–562. https://doi.org/10.3758/bf03196096.

Costello, F. 2009. "How Probability Theory Explains the Conjunction Fallacy." *Journal of Behavioral Decision Making* 22, no. 3: 213–234. https://doi.org/10.1002/bdm.618.

Costello, F., and P. Watts. 2014. "Surprisingly Rational: Probability Theory Plus Noise Explains Biases in Judgment." *Psychological Review* 121, no. 3: 463–480. https://doi.org/10.1037/a0037010.

Costello, F., and P. Watts. 2019. "The Rationality of Illusory Correlation." *Psychological Review* 126, no. 3: 437–450. https://doi.org/10.1037/rev0000130.

Edwards, W. 1968. "Conservatism in Human Information Processing." In *Formal Representations of Human Judgment*, edited by B. Kleinmuntz, 17–52. Wiley.

Erev, I., and A. E. Roth. 2014. "Maximization, Learning, and Economic Behavior." *Proceedings of the National Academy of Sciences* 111, no. 3: 10818–10825. https://doi.org/10.1073/pnas.1402846111.

Erev, I., T. S. Wallsten, and D. V. Budescu. 1994. "Simultaneous Over- and Underconfidence: The Role of Error in Judgment Processes." *Psychological Review* 101, no. 3: 519–527.

Ert, E., and S. T. Trautmann. 2014. "Sampling Experience Reverses Preferences for Ambiguity." *Journal of Risk and Uncertainty* 49, no. 1: 31–42. https://doi.org/10.1007/s11166-014-9197-9.

Fiedler, K., and C. Unkelbach. 2014. "Regressive Judgment: Implications of a Universal Property of the Empirical World." *Current Directions in Psychological Science* 23, no. 5: 361–367. https://doi.org/10.1177/0963721414546330.

Fox, C. R., and R. T. Clemen. 2005. "Subjective Probability Assessment in Decision Analysis: Partition Dependence and Bias Toward the Ignorance Prior." *Management Science* 51, no. 9: 1417–1432. https://doi.org/10.1287/mnsc.1050.0409.

Fox, C. R., and L. Hadar. 2006. ""Decisions From Experience" = Sampling Error+ Prospect Theory: Reconsidering Hertwig, Barron, Weber & Erev (2004)." *Judgment and Decision Making* 1, no. 2: 159–161.

Fox, C. R., and Y. Rottenstreich. 2003. "Partition Priming in Judgment Under Uncertainty." *Psychological Science* 14, no. 3: 195–200. https://doi.org/10.1111/1467-9280.02431.

Gigerenzer, G. 1996. "Why Do Frequency Formats Improve Bayesian Reasoning? Cognitive Algorithms Work on Information, Which Needs Representation." *Behavioral and Brain Sciences* 19, no. 1: 23–24. https://doi.org/10.1017/s0140525x00041248.

Gigerenzer, G., and U. Hoffrage. 1995. "How to Improve Bayesian Reasoning Without Instruction: Frequency Formats." *Psychological Review* 102, no. 4: 684–704. https://doi.org/10.1037/0033-295x.102.4.684.

Griffin, D., and L. Brenner. 2004. "Perspectives on Probability Judgment Calibration." *Blackwell Handbook of Judgment and Decision Making* 199: 158–177.

Hertwig, R., G. Barron, E. U. Weber, and I. Erev. 2004. "Decisions From Experience and the Effect of Rare Events in Risky Choice." *Psychological Science* 15, no. 8: 534–539. https://doi.org/10.1111/j.0956-7976.2004.00715.x.

Hertwig, R., and I. Erev. 2009. "The Description-Experience Gap in Risky Choice." *Trends in Cognitive Sciences* 13, no. 12: 517–523. https://doi.org/10.1016/j.tics.2009.09.004.

Hertwig, R., and G. Gigerenzer. 1999. "The 'Conjunction Fallacy' Revisited: How Intelligent Inferences Look Like Reasoning Errors." *Journal of Behavioral Decision Making* 12, no. 4: 275–305. https://doi.org/10.1002/(sici)1099-0771(199912)12:4<275::aid-bdm323>3.0.co;2-m.

Hintzman, D. L. 1976. Repetition and Memory." *Psychology of Learning and Motivation* 10: 47–91. https://doi.org/10.1016/s0079-7421(08)60464-8.

Howe, R., and F. Costello. 2020. "Random Variation and Systematic Biases in Probability Estimation." *Cognitive Psychology* 123: 101306. https://doi.org/10.1016/j.cogpsych.2020.101306.

Kahneman, D., and A. Tversky. 1972. "Subjective Probability: A Judgment of Representativeness." *Cognitive Psychology* 3, no. 3: 430–454. https://doi.org/10.1016/0010-0285(72)90016-3.

Kelly, M. H., and S. Martin. 1994. "Domain-General Abilities Applied to Domain-Specific Tasks: Sensitivity to Probabilities in Perception, Cognition, and Language." *Lingua* 92: 105–140. https://doi.org/10.1016/0024-3841(94)90339-5.

Lejarraga, T., J. K. Woike, and R. Hertwig. 2016. "Description and Experience: How Experimental Investors Learn About Booms and Busts Affects Their Financial Risk Taking." *Cognition* 157: 365–383. https://doi.org/10.1016/j.cognition.2016.10.001.

Malmendier, U., and S. Nagel. 2011. "Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?." *Quarterly Journal of Economics* 126, no. 1: 373–416. https://doi.org/10.1093/qje/qjq004.

Marchiori, D., S. D. Guida, and I. Erev. 2015. "Noisy Retrieval Models of Over- and Undersensitivity to Rare Events." *Decision* 2, no. 2: 82–106. https://doi.org/10.1037/dec0000023.

McGeoch, J. A., and W. T. McDonald. 1931. "Meaningful Relation and Retroactive Inhibition." *American Journal of Psychology* 43, no. 4: 579. https://doi.org/10.2307/1415159.

Nissen, M. J., and P. Bullemer. 1987. "Attentional Requirements of Learning: Evidence From Performance Measures." *Cognitive Psychology* 19, no. 1: 1–32. https://doi.org/10.1016/0010-0285(87)90002-8.

Peterson, C. R., and L. R. Beach. 1967. "Man as an Intuitive Statistician." *Psychological Bulletin* 68, no. 1: 29. https://doi.org/10.1037/h0024722.

Plonsky, O., and I. Erev. 2017. "Learning in Settings With Partial Feedback and the Wavy Recency Effect of Rare Events." *Cognitive Psychology* 93: 18–43. https://doi.org/10.1016/j.cogpsych.2017.01.002.

Plonsky, O., K. Teodorescu, and I. Erev. 2015. "Reliance on Small Samples, the Wavy Recency Effect, and Similarity-Based Learning."

Psychological Review 122, no. 4: 621–647. https://doi.org/10.1037/a0039413.

Rapoport, A., and T. S. Wallsten. 1972. "Individual Decision Behavior." *Annual Review of Psychology* 23, no. 1: 131–176. https://doi.org/10.1146/annurev.ps.23.020172.001023.

Ratcliff, R. 2022. "Integrated Diffusion Models for Distance Effects in Number Memory." *Cognitive Psychology* 138: 101516. https://doi.org/10.1016/j.cogpsych.2022.101516.

Remillard, G., and J. M. Clark. 2001. "Implicit Learning of First-, Second-, and Third-Order Transition Probabilities." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27, no. 2: 483–498.

P. Sedlmeier, and T. Betsch, eds. 2002. *Etc. Frequency Processing and Cognition*. Oxford, UK: Oxford University Press.

See, K. E., C. R. Fox, and Y. S. Rottenstreich. 2006. "Between Ignorance and Truth: Partition Dependence and Learning in Judgment Under Uncertainty." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32, no. 6: 1385–1402. https://doi.org/10.1037/0278-7393.32.6.1385.

Shteingart, H., and Y. Loewenstein. 2015. "The Effect of Sample Size and Cognitive Strategy on Probability Estimation Bias." *Decision* 2, no. 2: 107–117. https://doi.org/10.1037/dec0000027.

Spiliopoulos, L. 2012. "Pattern Recognition and Subjective Belief Learning in a Repeated Constant-Sum Game." *Games and Economic Behavior* 75, no. 2: 921–935. https://doi.org/10.1016/j.geb.2012.01.005.

Spiliopoulos, L. 2013. "Beyond Fictitious Play Beliefs: Incorporating Pattern Recognition and Similarity Matching." *Games and Economic Behavior* 81: 69–85. https://doi.org/10.1016/j.geb.2013.04.005.

Spiliopoulos, L. 2013. "Strategic Adaptation of Humans Playing Computer Algorithms in a Repeated Constant-Sum Game." *Autonomous Agents and Multi-Agent Systems* 27, no. 1: 131–160. https://doi.org/10.1007/s10458-012-9203-z.

Spiliopoulos, L., and R. Hertwig. 2023. "Variance, Skewness and Multiple Outcomes in Described and Experienced Prospects: Can One Descriptive Model Capture It All?." *Journal of Experimental Psychology: General* 152, no. 4: 1188–1222. https://doi.org/10.1037/xge0001323.

Stolarz-Fantino, S., E. Fantino, D. J. Zizzo, and J. Wen. 2003. "The Conjunction Effect: New Evidence for Robustness." *American Journal of Psychology* 116, no. 1: 15. https://doi.org/10.2307/1423333.

Sun, R., and C. L. Giles. 2001. "Sequence Learning: Paradigms, Algorithms, and Applications."

Szollosi, A., G. Liang, E. Konstantinidis, C. Donkin, and B. R. Newell. 2019. "Simultaneous Underweighting and Overestimation of Rare Events: Unpacking a Paradox." *Journal of Experimental Psychology. General* 148, no. 12: 2207–2217. https://doi.org/10.1037/xge0000603.

Tannenbaum, D., C. R. Fox, and G. Ülkümen. 2017. "Judgment Extremity and Accuracy Under Epistemic vs. Aleatory Uncertainty." *Management Science* 63, no. 2: 497–518. https://doi.org/10.1287/mnsc.2015.2344.

Teoderescu, K., M. Amir, and I. Erev. 2013. "Chapter 6 The Experience–Description Gap and the Role of the Inter Decision Interval." *Progress in Brain Research* 202: 99–115. https://doi.org/10.1016/b978-0-444-62604-2.00006-x.

Tversky, A., and D. Kahneman. 1973. "Availability: A Heuristic for Judging Frequency and Probability." *Cognitive Psychology* 5, no. 2: 207–232.

Tversky, A., and D. Kahneman. 1974. "Judgment Under Uncertainty: Heuristics and Biases." *Science* 185, no. 4157: 1124–1131. https://doi.org/10.1126/science.185.4157.1124.

Underwood, B. J. 1969. "Attributes of Memory." *Psychological Review* 76, no. 6: 559–573. https://doi.org/10.1037/h0028143.

Ungemach, C., N. Chater, and N. Stewart. 2008. "Are Probabilities Overweighted or Underweighted When Rare Outcomes Are Experienced (Rarely)?." *Psychological Science* 20, no. 4: 473–479. https://doi.org/10.1111/j.1467-9280.2009.02319.x.

Varey, C. A., B. A. Mellers, and M. H. Birnbaum. 1990. "Judgments of Proportions." *Journal of Experimental Psychology: Human Perception and Performance* 16, no. 3: 613–625. https://doi.org/10.1037/0096-1523.16.3.613.

Vlek, C. A. J. 1970. "Learning Probabilities of Events." *Acta Psychologica* 34: 160–171. https://doi.org/10.1016/0001-6918(70)90014-4.

Wagenaar, W. A., and G. B. Keren. 1985. "Calibration of Probability Assessments by Professional Blackjack Dealers, Statistical Experts, and Lay People." *Organizational Behavior and Human Decision Processes* 36, no. 3: 406–416. https://doi.org/10.1016/0749-5978(85)90008-1.

Wallsten, T. S., and D. V. Budescu. 1983. "State of the Art–Encoding Subjective Probabilities: A Psychological and Psychometric Review." *Management Science* 29, no. 2: 151–173. https://doi.org/10.1287/mnsc.29.2.151.

Wulff, D. U., M. Mergenthaler-Canseco, and R. Hertwig. 2018. "A Meta-Analytic Review of Two Modes of Learning and the Description-Experience Gap." *Psychological Bulletin* 144, no. 2: 140–176. https://doi.org/10.1037/bul0000115.

Zacks, R. T., and L. Hasher. 2002. "Frequency Processing: A Twenty-Five Year Perspective." In *Etc Frequency Processing and Cognition*, edited by P. Sedlmeier, and T. Betsch, 21–36. New York: Oxford University Press.

Zhang, H., and L. T. Maloney. 2012. "Ubiquitous Log Odds: A Common Representation of Probability and Frequency Distortion in Perception, Action, and Cognition." *Frontiers in Neuroscience* 6: 1. https://doi.org/10.3389/fnins.2012.00001.

Zhang, H., X. Ren, and L. T. Maloney. 2020. "The Bounded Rationality of Probability Distortion." *Proceedings of the National Academy of Sciences* 117, no. 36: 22024–22034. https://doi.org/10.1073/pnas.1922401117.

Zhu, J.-Q., P. W. S. Newall, J. Sundh, N. Chater, and A. N. Sanborn. 2022. "Clarifying the Relationship Between Coherence and Accuracy in Probability Judgments." *Cognition* 223: 105022. https://doi.org/10.1016/j.cognition.2022.105022.

Zhu, J.-Q., A. N. Sanborn, and N. Chater. 2020. "The Bayesian Sampler: Generic Bayesian Inference Causes Incoherence in Human Probability Judgments." *Psychological Review* 127, no. 5: 719–748. https://doi.org/10.1037/rev0000190.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.