Cognition 263 (2025) 106125

Contents lists available at ScienceDirect

Cognition

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

Full length article

Evaluating the role of mental sampling in probability judgments: Illogical rankings occur in a predictable manner

Xiaotong Liu [a,*], Arndt Bröder [a], Henrik Singmann [b]

[a] University of Mannheim, Mannheim, Germany
[b] University College London, London, UK

ARTICLE INFO

Dataset link: https://osf.io/hw8p9/

Keywords:
Mental sampling
Probability judgments
Fallacies and biases
Cognitive model

ABSTRACT

People's probability judgments often appear to be probabilistically incoherent, as exemplified by the conjunction fallacy. Recently, various sampling-based models have been proposed as an integrative account for different biases and fallacies in probability judgments. In the current study, the novel Event Ranking Task was used to investigate sampling-based models of probability judgments. On each trial of the Event Ranking Task, participants were asked to provide a ranking for an event set consisting of four events, A, not-A, B, and not-B, in terms of their perceived likelihoods. Qualitative predictions were formally derived by assuming direct sampling from a fixed underlying probability distribution. Adding read-out noise in the sampling process – as suggested in the Probability Theory plus Noise model (Costello and Watts, 2014) – did not change the qualitative predictions. Two online experiments, where participants ranked twelve different event sets, yielded results in line with the qualitative predictions, providing evidence for the idea that mental sampling underlies probability judgments.

A widely observed phenomenon in psychology is that people's probability judgments often violate the laws of probability. For example, people tend to commit the conjunction fallacy in the task famously known as the "Linda problem", judging that the conjunction of two events (e.g., "Linda is a bank teller and a feminist".) is more likely than the constituent marginal event (e.g., "Linda is a bank teller".) (Tversky & Kahneman, 1983). There has been an ongoing debate about the explanation for such phenomena. The traditional view suggests that systematic errors in people's probability judgments, such as the conjunction fallacy, result from cognitive processes that are not based on probability theory but rather on alternative approaches such as heuristics (Tversky & Kahneman, 1974) or configural-weighting-and-adding (Juslin et al., 2009; Nilsson et al., 2009). However, recently, the mental sampling framework proposed a view that the brain approximates probabilities (and probabilistic computations) via a process of sampling (for an introduction, see Icard, 2016; Sanborn & Chater, 2016). Various models developed under this theoretical framework demonstrated that cognitive processes in line with probability theory can also lead to a range of apparently irrational effects in probability judgments, including probability matching (Vul et al., 2014), conjunction and disjunction fallacy (Costello & Watts, 2017), and illusory correlation (Bott et al., 2021).

Sampling-based models entail the proposal that people inherently hold coherent subjective probabilities; however, these underlying subjective probabilities cannot be directly accessed. Instead, individuals have to sample instances via either memory or mental simulation to approximate their own underlying subjective probabilities. Such a mental sampling process is analogous to the sampling method used to approximate distributions in statistics, except that mental sampling only utilizes small rather than large samples. According to the resource-rational framework (Griffiths et al., 2015), using small samples allows individuals to optimally allocate time and cognitive resources, as generating samples is presumably effortful and time-consuming. Yet, the vulnerability of small sample sizes also leads individuals' probability judgments to be easily affected by sampling variability (Denison et al., 2013; Vul et al., 2014) and algorithmic properties of the sampling process such as the noise in the sampling processes (Costello & Watts, 2014), correlations of samples (Dasgupta et al., 2017; Lieder, Griffiths, M. Huys, et al., 2018), and decision rules individuals used when drawing samples (Lieder, Griffiths, & Hsu, 2018). Thus, according to sampling-based models, biases and fallacies in probability judgments are natural by-products of the sampling process instead of incoherent underlying representations of probabilities.

* Corresponding author.
E-mail address: liuxiaotong76@outlook.com (X. Liu).

https://doi.org/10.1016/j.cognition.2025.106125
Received 24 June 2024; Received in revised form 6 March 2025; Accepted 16 March 2025
Available online 20 June 2025

By explaining a wide range of biases and fallacies in probability judgments, the mental sampling framework offers an integrative perspective of how people produce probability judgments (Chater et al., 2020). However, so far most evidence for mental sampling comes from a single task – the numerical estimation task (e.g., Costello & Watts, 2014, 2018; Howe & Costello, 2020; Huang et al., 2024; Zhu et al., 2020). This situation limits the scope of the theoretical idea. If mental sampling is a general cognitive process, it should be used not only to produce numerical probability estimates but also in related situations involving probabilistic events.

In the present study, we tested qualitative predictions from the mental sampling framework using the novel *Event Ranking Task*. We focused on a specific class of models, namely, direct sampling models of probability judgments (i.e., Costello & Watts, 2014; Zhu et al., 2020). Based on the core assumptions of these models, we used simulations to derive qualitative predictions that were expected to emerge in participants' responses to the Event Ranking Task. Two online experiments provided support for these predictions, and they offered direct evidence for the idea that people use mental sampling when asked to make judgments about probabilistic events.

This article begins with a brief overview of the direct sampling models of probability judgments. Then, we describe the Event Ranking Task, a model we developed for the task, and a simulation study to derive qualitative predictions from the model. Finally, we present the results of two experiments that tested the predictions and discuss their implications.

1. Direct sampling models of probability judgments

The conceptually simplest method that the mind can use to approximate probabilities is arguably *direct sampling*. According to direct sampling, each time people are asked to estimate the probability of an event, they sample a subset of independent instances via memory and/or a mental simulation to approximate the probability. People's probability judgments of the event are then based on identically and independently distributed (i.i.d.) samples obtained from mental sampling. For instance, suppose people are assessing the probability of event A, "There will be rain on a randomly selected day in Hamburg". People may infer the probability of this event by recalling past days they spent in Hamburg and/or by imagining hypothetical days. Importantly, it is assumed that people cannot exhaust the entire sample space but instead can sample only a subset of random instances, that is, N days. Each sampled day indicates either a day with or without rain. The probability of event A can be estimated by counting the number of rainy days in the N retrieved and/or imagined days.

Formally, such a sampling process can be understood as instantiating a binomial process – that is, N Bernoulli trials (see also Costello & Watts, 2014; Howe & Costello, 2020; Zhu et al., 2020). Each sampled instance indicates either an occurrence of the event being assessed (i.e., rain on a random day in Hamburg) with probability $P(A)$ or a non-occurrence (i.e., NO rain on a random day in Hamburg) with probability $1 - P(A)$. Hereafter, we will refer to $P(A)$ as the underlying probability of event A. The underlying probability of an event governs the mental sampling process, as it is the probability that a randomly sampled instance will indicate an occurrence of the event under evaluation. Mathematically, since the mental sampling process is modeled as a binomial process, the underlying probability is equivalent to the probability of success in a single Bernoulli trial.

The number of instances that indicate the occurrence of event A, hereafter referred to as the number of occurrences of event A, and denoted as O_A, follows the binomial distribution, $O_A \sim \text{Bin}(N, P(A))$. Similarly, if people assess the probability of event A's complementary event, not-A (denoted as $\neg A$), the number of occurrences of this event then follows $O_{\neg A} \sim \text{Bin}(N, P(\neg A))$.

The direct sampling assumption underlies a number of popular sampling-based models of probability judgments, including the probability theory plus noise (PT+N) model (Costello & Watts, 2014, 2016, 2017, 2018; Howe & Costello, 2020) and the Bayesian sampler model (Zhu et al., 2020). The PT+N model and the Bayesian sampler model address how people rely on mental sampling to produce numerical estimations of probabilities. Both models posit that the binomial sampling process is used to evaluate different types of events, including marginal, conjunctive, disjunctive, and conditional events. Additionally, both models assume coherent underlying probabilities of related events, such that the underlying probabilities of two complementary events follow the complement rule ($P(A) + P(\neg A) = 1$).[1] However, the PT+N model and the Bayesian sampler model take different approaches to converting mental samples to probability estimates. Whereas the PT+N model assumes that both mental sampling and response generation are perturbed by noise – captured in a catch-all parameter d, the Bayesian sampler model assumes that people regularize their probability estimates via a prior.

1.1. Evidence from the numerical estimation task

So far, the PT+N model (Costello & Watts, 2014, 2016, 2017; Howe & Costello, 2020) and the Bayesian sampler model (Sundh et al., 2023; Zhu et al., 2022, 2020) have been explored exclusively within the context of the numerical estimation task. In the numerical estimation task, participants are asked to estimate the probability of different types of events, such as marginal and conjunctive events, on a scale from 0% to 100%. Participants' responses consistently exhibit well-documented biases, including conjunction and disjunction fallacies as well as response conservatism. The extent of these biases, however, varies depending on the specific content and other contextual factors employed in the task (Wedell & Moro, 2008).

One piece of evidence for sampling-based models of probability judgments is that they offer a unified explanation for most of the biases and fallacies observed in the numerical estimation task. For example, the (occasional) occurrence of conjunction and disjunction fallacies directly follows from the idea of mental sampling. If participants' probability judgments are based on independent mental samples with relatively small sample sizes, the randomness of the samples is sufficient to expect that sometimes the estimated probability of a single event is larger than the estimated probability of a conjunction including this event. The mechanisms included in the models on top of mental sampling – the noise parameter d in the PT+N model and the prior in the Bayesian sampler model – provide additional explanatory power. For example, both mechanisms provide quantitatively the same explanation for response conservatism by assuming that extreme probability estimates are pushed toward 50%. Furthermore, these additional mechanisms can explain subadditivity, binary complementarity, and varying rates of conjunction and disjunction fallacies in the numerical estimation task (Costello & Watts, 2014, 2017; Zhu et al., 2020).

In addition to being able to explain many of the existing biases and fallacies, the PT+N model and the Bayesian sampler model make specific predictions for people's averaged probability estimates of related events that were confirmed by the data. Costello and colleagues began this line of investigation by combining – adding and subtracting – probability estimates of related events to cancel out or isolate the effects of noise in the sampling process. Using a simple example to illustrate, the expression "$P(A) + P(\neg A)$" has an expected value of 1 according to the PT+N model (i.e., the effects of noise term d cancel out). Costello and colleagues (2014) developed a series of probabilistic expressions based on this idea. In some of these expressions, the effects of noise cancel out, and the expected values are in line with probability

[1] The PT+N model and the Bayesian sampler model conceptualize "underlying probability" differently. According to the Bayesian sampler model, underlying probabilities represent subjective beliefs. However, in the PT + N model, the underlying probabilities are assumed to represent the "objective" relative frequencies of events in memory.

theory. In other expressions, the effects of noise do not cancel out, and the expected values differ from what probability theory predicts. Later, Zhu and colleagues (2020) showed that the Bayesian sampler model makes identical predictions in terms of the expected value of expressions that do not involve conditional probabilities. The predictions for these probabilistic expressions were confirmed in a series of experiments across different pairs of events (Costello & Watts, 2014, 2016, 2018; Zhu et al., 2020).

Taken together, both the PT+N model and the Bayesian sampler model can explain many of the biases typically observed in people's numerical probability judgments, such as the conjunction fallacy and response conservatism. In addition, both models make predictions for the results of a number of probabilistic expressions that were generally confirmed in the aggregated probability estimation data.

2. Testing direct sampling empirically with the Event Ranking Task

The evidence for the mental sampling framework presented above is constrained to the numerical estimation task. However, evidence obtained from one specific experimental paradigm might be tied to the specific features of the paradigm. One feature of the probability estimation task is that it requires a high level of precision in mental representations, as well as a high level of precision in reasoning when responding on a probability scale from 0% to 100%. For example, Sun et al. (2008) discovered people's provided judgments were less coherent when required to reason in a finer-grained manner, and people's probability judgments were also clustered. The potential defects of relying on a single theory-testing paradigm call for converging evidence based on alternative methods (Meiser, 2011). Thus, we introduce the *Event Ranking Task* as an alternative approach for testing sampling-based accounts. In the task, participants need to rank four events according to their perceived probabilities instead of providing numerical estimates of probabilities. This response format relaxes the requirement for precision in probabilistic reasoning.

2.1. Introduction to the Event Ranking Task

In each trial of the Event Ranking Task, participants are presented with an event set consisting of four events, $\{A, \neg A, B, \neg B\}$. More specifically, participants are presented with two event pairs, and each event pair consists of two events that are complementary to each other. An example of an event set is:

- A: There will be rain on a randomly selected day in Hamburg.
- $\neg A$: There will be NO rain on a randomly selected day in Hamburg.
- B: A randomly selected person in Germany lives in a big city.
- $\neg B$: A randomly selected person in Germany does NOT live in a big city.

The event pair $\{A, \neg A\}$ represents an event indicating A occurs (denoted as event A) and a complementary event indicating A does not occur (denoted as event $\neg A$). The same applies to the event pair $\{B, \neg B\}$. Henceforth, we will refer to event A and event B as positive events, and event $\neg A$ and event $\neg B$ as negative events, so that each event pair consists of a positive and a negative event. When describing any negative event in our study, we used the grammatical negative, such as the word "NOT" or "NO" in capital letters. This was done to make it clear that each event pair comprises two mutually exclusive events.

The participants' task is to simultaneously evaluate the probabilities of the four events in a given event set and rank them based on their perceived probabilities. The event(s) that the participant perceives to be the most likely should receive the highest rank (i.e., Rank 1). The event(s) perceived to be second most likely should receive the second highest rank (i.e., Rank 2), and so forth. For instance, a participant might give a ranking such as $\hat{P}(A) > \hat{P}(B) > \hat{P}(\neg B) > \hat{P}(\neg A)$, indicating that event A has the highest perceived probability, event B has the second highest perceived probability, event $\neg B$ has the third highest perceived probability, and event $\neg A$ has the lowest perceived probability.

2.1.1. Logical and illogical rankings

One important feature of the Event Ranking Task is its embedded logical rule. When examining the responses to the task (i.e., rankings of four events, A, $\neg A$, B, $\neg B$, by their perceived probabilities), we can classify them as logical or illogical. A logical ranking conforms to the complement rule; namely, the probabilities of an event and its complement sum to 1, $P(A)+P(\neg A) = P(B)+P(\neg B) = 1$. Consequently, it is illogical to rank both events from the pair $\{A, \neg A\}$ above both events from the pair $\{B, \neg B\}$ because this suggests $P(A) + P(\neg A) > P(B) + P(\neg B)$. If $P(A) \leq P(B)$, then it must follow that $P(\neg A)(= 1 - P(A)) \geq P(\neg B)(= 1 - P(B))$. Conversely, if $P(A) \geq P(B)$, then it must follow that $P(\neg A) \leq P(\neg B)$.

To illustrate the above statements with an example, suppose a person is asked to rank the probabilities of four events: "rain in Hamburg" (A), "NO rain in Hamburg" ($\neg A$), "a person lives in a big city" (B), and "a person does NOT live in a big city" ($\neg B$), and they rank the event "rain in Hamburg" (A) as the most probable among the four events. Then, they must rank the event "NO rain in Hamburg" ($\neg A$) to be the least probable event for the complement rule to hold. The rank order of the two remaining events, a person lives/does NOT live in a big city, does not matter as long as they are ranked to be more probable than "No rain" and less probable than "rain". Namely, when A is the most probable event, rankings that follow the complement rule should be one of the following three: $\hat{P}(A) > \hat{P}(B) > \hat{P}(\neg B) > \hat{P}(\neg A)$ or $\hat{P}(A) > \hat{P}(\neg B) > \hat{P}(B) > \hat{P}(\neg A)$ or $\hat{P}(A) > \hat{P}(\neg B) = \hat{P}(B) > \hat{P}(\neg A)$. In contrast, suppose the person ranks the event "rain in Hamburg" (event A) as more likely than the event "a person lives in a big city" (event B) and consider these two events to be the most probable, while ranking "NO rain in Hamburg" (event $\neg A$) to be more likely than the event "a person does NOT live in a big city" (event $\neg B$) and considering these two events to be the least probable. This ranking $\hat{P}(A) > \hat{P}(B) > \hat{P}(\neg A) > \hat{P}(\neg B)$ is illogical, as it simultaneously suggests $\hat{P}(A) > \hat{P}(B)$ and $\hat{P}(\neg A) > \hat{P}(\neg B)$. Such illogical rankings frequently occurred in our data.

The occurrence of such illogical rankings cannot be explained from a strictly normative perspective. However, it can be anticipated from the perspective of mental sampling if one assumes that people draw independent samples for different events. Consider a hypothetical scenario where a person draws two independent samples via memory/mental simulation for events A and $\neg A$, each with a sample size of five and underlying probabilities of 0.8 and 0.2, respectively. The samples that perfectly match A and $\neg A$'s underlying probabilities are both $\{A, A, A, A, \neg A\}$. Yet, in reality, samples rarely perfectly match their underlying probabilities due to sampling variation. Over- and under-representation happen often. For example, a person might sample $\{A, A, A, A, A\}$ for event A and $\{A, A, A, \neg A, \neg A\}$ for event $\neg A$ by chance, both over-representing the true underlying probabilities of A and $\neg A$. Similarly, for events B and $\neg B$, each with a sample size of five and underlying probabilities of 0.6 and 0.4, respectively, the person might sample $\{B, B, B, \neg B, \neg B\}$ (perfect representation) and $\{B, B, B, B, \neg B\}$ (under-representation). Such a sampling result would lead the person to erroneously rank $\hat{P}(A) > \hat{P}(B) > \hat{P}(\neg A) > \hat{P}(\neg B)$ in the example above.

It is important to stress that a sampling-based model can predict illogical rankings in the Event Ranking Task *only* if it assumes that a ranking is based on independent samples for each event. More specifically, we need to assume that even for two complementary events (e.g., A and $\neg A$), people draw independent samples and do not reuse one sample. While this assumption first appears questionable from a

Table 1
Full response space and mapping to ranking category.

	Category	Ranking			
		Rank 1	Rank 2	Rank 3	Rank 4
1	logical ranking	A	B	$\neg B$	$\neg A$
2		A	$\neg B$	B	$\neg A$
3		$\neg A$	B	$\neg B$	A
4		$\neg A$	$\neg B$	B	A
5		B	A	$\neg A$	$\neg B$
6		B	$\neg A$	A	$\neg B$
7		$\neg B$	A	$\neg A$	B
8		$\neg B$	$\neg A$	A	B
9	stacked-illogical ranking	A	$\neg A$	B	$\neg B$
10		A	$\neg A$	$\neg B$	B
11		$\neg A$	A	B	$\neg B$
12		$\neg A$	A	$\neg B$	B
13		B	$\neg B$	A	$\neg A$
14		B	$\neg B$	$\neg A$	A
15		$\neg B$	B	A	$\neg A$
16		$\neg B$	B	$\neg A$	A
17	interlaced-illogical ranking	A	B	$\neg A$	$\neg B$
18		A	$\neg B$	$\neg A$	B
19		$\neg A$	B	A	$\neg B$
20		$\neg A$	$\neg B$	A	B
21		B	A	$\neg B$	$\neg A$
22		B	$\neg A$	$\neg B$	A
23		$\neg B$	A	B	$\neg A$
24		$\neg B$	$\neg A$	B	A

Note. The table presents all possible responses in the ties-not-allowed condition of the Event Ranking Task, along with their corresponding ranking categories. An event assigned Rank 1 is perceived to be the most likely, Rank 2 as the second most likely, Rank 3 as the third most likely, and Rank 4 as the least likely.

resource rational perspective, assuming otherwise that people reuse the same sample for evaluating A and $\neg A$ would suggest when A is over-represented, $\neg A$ will always be under-represented. This means that the scenario above where both A and $\neg A$ are over-represented, which is pivotal for the illogical ranking above to occur, would never happen. The same holds for the event B and event $\neg B$. Reusing any pairs of samples above (e.g., using two samples $\{A, A, A, A, A\}$ and $\{B, B, B, \neg B, \neg B\}$ for the four events or $\{A, A, A, \neg A, \neg A\}$ and $\{B, B, B, B, \neg B\}$) would lead to rankings that follow the complement rule. Consequently, no illogical ranking can occur, which contradicts our data. In the General Discussion section, we will discuss this point in more detail to explain why assuming four independent samples for the four events in the event set is essential and provides a parsimonious account of the data presented in this paper. It remains to be seen whether our explanation can eventually be incorporated into a resource rational account.

2.1.2. Handling rankings with ties

Another crucial aspect of the Event Ranking Task is whether to allow participants to provide rankings with ties. When considering all possible rankings of four events (A, $\neg A$, B, $\neg B$), there are 75 potential rankings. Of these, 24 are full orders without ties, where each event is assigned a unique rank (e.g., $\hat{P}(A) > \hat{P}(B) > \hat{P}(\neg B) > \hat{P}(\neg A)$), while the remaining 51 are partial orders that include ties, where some events share the same rank (e.g., $\hat{P}(A) > \hat{P}(B) = \hat{P}(\neg B) > \hat{P}(\neg A)$).

In our experiments, we included both possibilities in a between-subjects design. The ties-not-allowed condition was only allowed to produce rankings without ties, while the ties-allowed condition was permitted to produce rankings with ties. Predictions from the sampling-based model for the Event Ranking Task can be derived separately for both conditions and were qualitatively very similar. Because rankings with ties were rare in the ties-allowed condition and results showed the same patterns, we only focus on the simpler ties-not-allowed condition in this paper. Full details regarding the ties-allowed condition – including the task, model, simulation, and empirical investigations of model predictions – are provided in Supplementary Material S2.

2.1.3. Ranking categories

To facilitate a nuanced investigation of the sampling-based models using the Event Ranking Task, we have classified responses into three categories: logical rankings and two distinct types of illogical rankings.

Table 1 provides a comprehensive enumeration of the possible responses, along with the categorization of each response. The enumeration and categorization of responses presented here apply exclusively to the ties-not-allowed condition. For the enumeration and categorization of the responses under the ties-allowed conditions, see Supplementary Materials S2.1, where the same principles are applied with adjustments for additional partial orders.

- *Logical Rankings*: Logical rankings must obey the rule that one event pair in a given event set occupies Ranks 1 and 4 while the other pair occupies Ranks 2 and 3. Only these constellations are compatible with the complement rule.
- *Stacked-Illogical Rankings*: Both events in one event pair are simultaneously ranked higher than both events in the other event pair. In other words, in a stacked-illogical ranking, Rank 1 and Rank 2 are assigned exclusively to a pair of complementary events.
- *Interlaced-Illogical Rankings*: One pair occupies Ranks 1 and 3, the other pair Ranks 2 and 4.

2.2. The Ranking Model

In this section, we present a cognitive model for the Event Ranking Task based on direct sampling, the *Ranking Model*. We begin by outlining the basic model, the *Basic Ranking Model*, and then introduce an extended version of it, the *Ranking Model with Read-Out Noise*, which incorporates an additional assumption of read-out noise that affects the mental sampling process. In the subsequent section, both the basic model and the extended model will be used to derive qualitative predictions, which will then be tested empirically.

To provide a preliminary overview, the Ranking Model calculates the probability of each ranking given the comprehensive set of all possible rankings for a given event set. In the basic model, the probability of each possible ranking is determined by three parameters that characterize the mental sampling processes: $P(A)$, $P(B)$, and N. The parameters $P(A)$ and $P(B)$ represent the underlying probabilities that govern the sampling process. Because $P(A) + P(\neg A) = 1$ and $P(B) + P(\neg B) = 1$, specifying only one parameter per event pair is enough.

In the extended model, an additional assumption is introduced: the sampling process is affected by read-out noise, denoted as d. Specifically, there is a fixed probability d that a participant mistakenly interprets an instance of an event as an instance of the event's complement (i.e., reading an instance of event A as an instance of event $\neg A$, or vice versa).

In this paper, we investigate participants' responses only at the level of the ranking categories (i.e., logical rankings, stacked-illogical rankings, and interlaced-illogical rankings) rather than at the level of individual rankings. Accordingly, we focus on the model predictions at the ranking category level. We aggregate the model predictions for the rankings that belong to the same ranking category to derive model predictions.

Notably, the Ranking Model presented here is tailored to the ties-not-allowed condition of the Event Ranking Task. While partial orders are included in the ties-allowed condition – leading to differences in how rankings are enumerated and probabilities calculated – the Ranking Models for the ties-allowed and ties-not-allowed conditions share core assumptions and produce very similar qualitative predictions. Supplementary Materials S1.1 provides the equations for the Ranking Model specific to the ties-allowed condition, S2.2 outlines the assumptions, and S2.3 presents the derived predictions.

2.2.1. Basic Ranking Model

During a trial of the Event Ranking Task, participants rank the four events, A, $\neg A$, B, and $\neg B$, by their perceived probabilities. The Ranking Model assumes that participants begin by drawing independent samples for each of these four events. Imagine that in a trial, participants are presented with the previous example event set {A: rain in Hamburg, $\neg A$: NO rain in Hamburg, B: a person lives in a big city, $\neg B$: a person does NOT live in a big city}. To evaluate event A, participants recall/simulate a sample of N random days in Hamburg and count the number of days with rain, denoted as O_A. Similarly, independent samples are drawn for the remaining three events, resulting in $O_{\neg A}$ (for event $\neg A$), O_B (for event B) and $O_{\neg B}$ (for event $\neg B$).

Consistent with previous literature (Costello & Watts, 2014; Howe & Costello, 2020; Zhu et al., 2020), the sampling outcomes O are distributed according to a binomial distribution, $O \sim \text{Bin}(N, p)$, where N represents the sample size used for the event under evaluation and p is the underlying probability of the event under evaluation. Thus, the sampling outcomes for A are modeled as $O_A \sim \text{Bin}(N_A, P(A))$. Because the Ranking Model adheres to the complement rule, which states that the probabilities of complementary events sum to 1, we can model $O_{\neg A}$ as $O_{\neg A} \sim \text{Bin}(N_{\neg A}, 1-P(A))$. The same reasoning applies to deriving the binomial distributions for the sampling outcomes of event B and its complement, $\neg B$. The probability mass function (PMF) of the binomial distribution, $f(i, N, p)$, is used to compute the probability of observing $O = i$ occurrences of the event in a sample of N instances:

$$P(O = i) = f(i, N, p) = \binom{N}{i} p^i (1-p)^{N-i}. \tag{1}$$

To derive the sampling outcome distribution for any of the four events A, $\neg A$, B, and $\neg B$, we only need to substitute the parameters (i.e., the underlying probability p and the sample size of mental sampling N). For example, $P(O_A = i) = f(i, N_A, P(A)) = \binom{N_A}{i} P(A)^i (1-P(A))^{N_A - i}$.

Importantly, the Ranking Model posits that, during a single trial of the Event Ranking Task, participants use a constant sample size, N, to evaluate all four events $(A, \neg A, B, \neg B)$ from the same event set. This means that in any single trial, $N_A = N_B = N_{\neg A} = N_{\neg B} = N$. After obtaining mental samples for each event, the Ranking Model assumes that participants derive a ranking by directly comparing the numbers of occurrences (i.e., O_A, $O_{\neg A}$, O_B, $O_{\neg B}$). Events with the largest number of occurrences in the obtained samples are ranked highest, followed by those with the second-largest number of occurrences, and so on. Both the assumption of independent samples for each event and the use of a constant sample size are central assumptions of the Ranking Model. Without the constant sample size assumption, participants would be unable to provide a ranking based solely on the binomial sampling outcomes. Instead, additional assumptions, such as converting sampling results into interim metrics (e.g., relative frequencies) before ranking, would be required.

As an example to illustrate the proposed process, imagine a scenario where a participant uses a sample size of 10 to assess the example event set introduced above and obtains the following sampling results: for event A, 5 out of 10 instances are rainy days (i.e., $O_A = 5$); for event $\neg A$, 7 out of 10 instances are non-rainy days (i.e., $O_{\neg A} = 7$); for event B, 2 out of 10 people live in a big city (i.e., $O_B = 2$); and for event $\neg B$, 4 out of 10 people do not live in a big city (i.e., $O_{\neg B} = 4$). Based on these sampling outcomes, $O_{\neg A} > O_A > O_{\neg B} > O_B$, participants would derive the following ranking: $\hat{P}(\neg A) > \hat{P}(A) > \hat{P}(\neg B) > \hat{P}(B)$. This ranking falls into the category of stacked-illogical rankings; although the underlying probabilities adhere to the complement rule, the sampling outcomes do not due to sampling variability.

To calculate the probability of obtaining the sampling outcomes with the order $O_{\neg A} > O_A > O_{\neg B} > O_B$, we need to enumerate all possible combinations of O_A, $O_{\neg A}$, O_B, and $O_{\neg B}$ that have this order. Continuing with the previous hypothetical scenario, we consider all possible sampling outcomes resulting from a sample size of 10. What combinations of O_A, $O_{\neg A}$, O_B, and $O_{\neg B}$ have the aforementioned order? O_B, as the smallest value, must be smaller than the other three numbers of occurrences. Thus, O_B can take values ranging from 0 to 7 $(= 10-3)$. $O_{\neg B}$ must be larger than O_B, but its value cannot exceed the other two numbers of occurrences. Thus, $O_{\neg B}$ can take values ranging from O_B+1 to 8 $(= 10-2)$. Using the same reasoning, O_A can take values ranging from $O_{\neg B}+1$ to 9 $(= 10-1)$, and $O_{\neg A}$ can take values ranging from O_A+1 to 10. To generalize to sampling outcomes obtained using any sample size, the probability of obtaining the sampling outcomes with the aforementioned order is

$$\begin{aligned} &P(O_{\neg A} > O_A > O_{\neg B} > O_B) \\ &= \sum_{O_B=0}^{N-3} f(O_B, N, P(B)) \sum_{O_{\neg B}=O_B+1}^{N-2} f(O_{\neg B}, N, P(\neg B)) \\ &\sum_{O_A=O_{\neg B}+1}^{N-1} f(O_A, N, P(A)) \sum_{O_{\neg A}=O_A+1}^{N} f(O_{\neg A}, N, P(\neg A)), \end{aligned} \tag{2}$$

where the function $f(i, N, p)$ calculates the probability of obtaining i occurrences in a mental sample of size N, as given by Eq. (1). Four summations are calculated for O_A, $O_{\neg A}$, O_B, and $O_{\neg B}$, respectively (e.g., $P(O_{\neg A} \in [O_A+1, N]) = \sum_{O_{\neg A}=O_A+1}^{N} f(O_{\neg A}, N, P(A))$) to give us the probability of them falling into a range of values that allows the desired order (e.g., $O_{\neg A} > O_A$). Finally, using the product rule, we enumerate all possible combinations of O_A, $O_{\neg A}$, O_B, and $O_{\neg B}$ that result in the desired order.

One possible result, directly based on the sampling outcomes, is that the numbers of occurrences for two or more events are equal (e.g., $O_A = O_{\neg A} > O_{\neg B} > O_B$). In the main text of the paper, we focus on the ties-not-allowed condition of the Event Ranking Task, where participants cannot provide a ranking with ties directly based on such sampling results, producing a ranking such as $\hat{P}(A) = \hat{P}(\neg A) > \hat{P}(\neg B) > \hat{P}(B)$. We assume that participants would instead randomly assign an order to the tied events while maintaining the linear order suggested by the sample outcomes. Specifically, participants would follow the orders $O_A > O_{\neg B} > O_B$ and $O_{\neg A} > O_{\neg B} > O_B$, but would randomly assign an order for A and $\neg A$, with each event having a probability of 0.5 of being ranked first. Consequently, participants would either produce the ranking $\hat{P}(\neg A) > \hat{P}(A) > \hat{P}(\neg B) > \hat{P}(B)$ or $\hat{P}(A) > \hat{P}(\neg A) > \hat{P}(\neg B) > \hat{P}(B)$, with equal probability. The probability of arriving at either of these two rankings is the probability of generating the original partial ordering, $\hat{P}(A) = \hat{P}(\neg A) > \hat{P}(\neg B) > \hat{P}(B)$ (given by Equation A27 in Supplementary Material S1.1.1), multiplied by 0.5.[2]

As a consequence of ties being randomly assigned to generate a ranking without ties, the ranking $\hat{P}(\neg A) > \hat{P}(A) > \hat{P}(\neg B) > \hat{P}(B)$ can arise from a range of sampling outcomes. One possibility is that the sampling outcomes follow the exact order without ties: $O_{\neg A} > O_A > O_{\neg B} > O_B$, with the probability of obtaining such outcomes given by Equation A21 in Supplementary Material S1.1.1. Other possibilities involve sampling outcomes with ties, which may follow the ordering $O_{\neg A} > O_A > O_{\neg B} > O_B$ once ties are resolved. Ties can occur between any two ranks, and there can be any number of ties. Equation B12 in Supplementary Material S1.2.1 calculates the probability of participants providing the ranking $\hat{P}(\neg A) > \hat{P}(A) > \hat{P}(\neg B) > \hat{P}(B)$ considering all possible sampling outcomes with and without ties. Supplementary Materials S1.2.1 provides equations that calculate the probabilities of all 24 possible rankings.

[2] If there are three (or four) ties in the sampling result, participants would need to randomly decide among six (or twenty-four) potential full orders after ordering the ties. The probability of reporting any of the six (or twenty-four) orders would be the probability of obtaining the original sampling result with ties, multiplied by $\frac{1}{6}$ (or $\frac{1}{24}$).

2.2.2. Ranking Model with Read-out Noise

In an extension to the Basic Ranking Model, we introduce read-out noise, denoted by parameter d, into the sampling processes (i.e., the model now has four parameters, $P(A)$, $P(B)$, N, and d). This read-out noise, first proposed in the PT+N model (Costello & Watts, 2014), regresses the underlying probabilities governing mental sampling toward 0.5. The rationale for this extension is twofold: first, to explore whether the model's predictions remain intact after considering the effect of regressed underlying probabilities on sampling outcomes, and second, to align the model more closely with established literature, where the noise (or, alternatively, Bayesian belief updating) plays a central role in explaining findings such as the probabilistic expressions in people's probability estimations of related events (Costello & Watts, 2014; Zhu et al., 2020), and the varying rates of the occurrence of the conjunction fallacy (Costello & Watts, 2017).

We modeled the "read-out noise" using the same approach as the PT+N model. Under the influence of the read-out noise, each sampled instance might be mistakenly read as its complement (A be identified as $\neg A$ or vice versa) at a constant rate. Each sampled instance for event A has a probability $P_{reg}(A)$ (instead of $P(A)$) of indicating the occurrence of A, and a probability $1 - P_{reg}(A)$ of indicating the occurrence of $\neg A$. The probability of obtaining a number of i occurrences in a sample of N instances is given by the PMF of the binomial distribution (Eq. (1)) where p is replaced with the regressed underlying probability p_{reg}, with $p_{reg} = (1 - 2d)p + d$. Furthermore, by replacing p with p_{reg} in Eq. (2) (e.g., replacing $P(A)$ with $P_{reg}(A) = (1 - 2d)P(A) + d$), we obtain the equation for the sampling result $O_{\neg A} > O_A > O_{\neg B} > O_B$ according to the Ranking Model with Read-Out Noise.

Although we followed the same modeling approach as the PT+N model, the noise parameter d in the Ranking Model differs conceptually from that in the PT+N model. In the PT+N model, d is treated as a catch-all for multiple sources of errors in a continuous response task, whereas in our model, d plays the more specific role of internal read-out noise only, which disrupts the sampling process. Consequently, one might expect the value of d in the Event Ranking Task to be smaller compared to the value of d in numerical estimation tasks. However, exploring this issue in detail falls outside the scope of the current paper.

3. Simulation study

In this section, we present a simulation study that generates qualitative predictions about the occurrence of different ranking categories across various event sets in the Event Ranking Task. These response patterns were examined using both the Basic Ranking Model and the Ranking Model with Read-Out Noise. The same parameter settings were applied to both models, except for the additional noise parameter (d), which was only relevant to the Ranking Model with Read-Out Noise. The simulation results show that the same qualitative pattern is predicted to appear in the data, regardless of whether noise in the sampling process is assumed. The simulation was implemented in the R environment (R Core Team, 2024). The simulation code and the results are available at the Open Science Framework (https://osf.io/hw8p9/).

The simulation study presented here was tailored to the ties-not-allowed condition of the Event Ranking Task. Supplementary Materials S2.3 describes a comparable simulation study tailored to the ties-allowed condition, using the same simulation settings as in the ties-not-allowed condition. Both simulations yielded the same predictions for ranking categories shared by both conditions.

3.1. Procedure

The Basic Ranking Model has three parameters: $P(A)$, $P(B)$, and N. The Ranking Model with Read-Out Noise includes an additional parameter, d, alongside these three parameters. $P(A)$ and $P(B)$ (and their complementary probabilities, $1 - P(A)$ and $1 - P(B)$) represent the underlying probabilities of the four events that constitute an event set presented in a single trial of the Event Ranking Task. N represents the sample size of mental sampling used for all four events in a given trial. d represents a constant read-out noise that affects the mental sampling process across all trials of the Event Ranking Task.

First, we systematically varied all parameters in the Basic Ranking Model to assess their impact on the occurrence of ranking categories. Our main focus in the simulation was on the underlying probabilities of the events, for which we created three different types of event sets with distinct underlying probabilities of the constituent events. We derived qualitative predictions about the occurrence of different ranking categories across different types of event sets, irrespective of the sample size of mental sampling, N.

Second, we investigated whether the qualitative predictions would hold when considering read-out noise in the sampling process. To this end, we ran the simulation of the Ranking Model with Read-Out Noise. We applied the same parameter values for N and the underlying probabilities, $P(A)$ and $P(B)$. Additionally, we incorporated the read-out noise parameter, d, setting it to a value near what we considered its maximum plausible limit to explore whether the predictions of the Basic Ranking Model would be distorted.

3.1.1. Basic Ranking Model

We varied $P(A)$ and $P(B)$ together to generate three different types of event sets: edge-event sets, mid-event sets, and mixed sets. Recall that each event set consists of two pairs of events, each consisting of two complementary events. An edge-event set comprises two event pairs whose constituent events have underlying probabilities close to 1 and 0, respectively (we term such pairs as edge-event pairs). For each edge-event pair, the underlying probability of one constituent event is determined by drawing a random value from a Beta(1, 10) distribution (Fig. 1, left panel), and the underlying probability of the remaining event is obtained by subtracting this value from 1. A mid-event set comprises two event pairs whose constituent events have underlying probabilities close to 0.5 (we term such pairs as mid-event pairs). For each mid-event pair, the underlying probability of one constituent event is determined by drawing a random value from a Beta(10, 10) distribution (Fig. 1, right panel), and the underlying probability of the remaining event is obtained by subtracting this value from 1. A mixed set consists of one edge-event pair and one mid-event pair. The realizations of different types of event sets are summarized in Table 2.

We varied the sample size N from 1 to 50 in a total of 26 levels. For values of N from 1 to 20 we increased the sample size in steps of 1 and for values of N from 20 to 50 in steps of 5. The reason for choosing different step sizes across the range of N was that we expected more changes in the qualitative pattern of predictions for small values of N (i.e., $N < 20$) compared to the changes expected for larger values of N. Ultimately, we wanted qualitative predictions that generalize across a wide range of Ns, so we did not have to make questionable assumptions about N in our experiments.

Taken together, the simulation of the Basic Ranking Model varied two factors: event-set type (three levels) and sample size N (26 levels). For each combination of factor levels, we performed 10,000 simulation runs, which returned the predicted probabilities for three ranking categories as the output. Specifically, in each run, we generated an event set and determined the constituent events' underlying probabilities by drawing from corresponding Beta distributions (see Table 2). Plugging the underlying probabilities and sample sizes into the Basic Ranking Model, we calculated the predicted probabilities of all 24 possible rankings for each event set. Finally, we classified the rankings into three categories as introduced in Section *Ranking Categories* and summed the probabilities of the rankings within each ranking category.

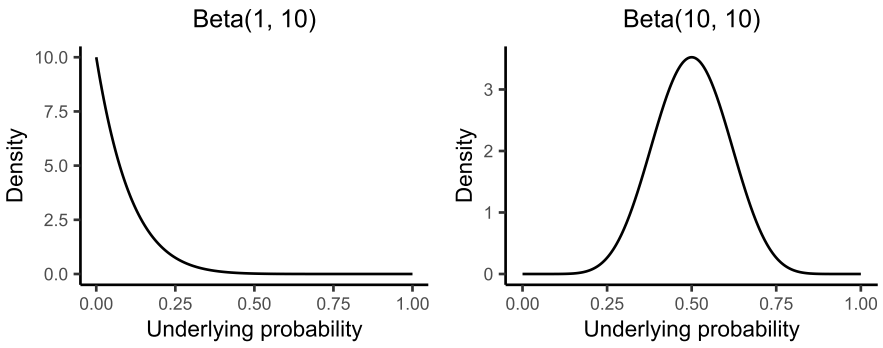

Fig. 1. Beta distributions used in the simulation study to generate the underlying probabilities.
The left panel shows the Beta(1, 10) distribution used to generate underlying probabilities for events in the edge-event pairs. The right panel shows the Beta(10, 10) distribution to generate underlying probabilities for events in the mid-event pairs.

Table 2
Summary of realizations of event sets in the simulation.

Event-set type	Definition	Distributions for random value generation
Edge-event set $\{A, \neg A, B, \neg B\}$	A set consisting of two edge-event pairs with constituent events having underlying probabilities near 1 and 0, respectively	$P(A) \sim \text{Beta}(1, 10),\ P(\neg A) = 1 - P(A)$ $P(B) \sim \text{Beta}(1, 10),\ P(\neg B) = 1 - P(B)$
Mid-event set $\{A, \neg A, B, \neg B\}$	A set consisting of two mid-event pairs with constituent events having underlying probabilities close to 0.5	$P(A) \sim \text{Beta}(10, 10),\ P(\neg A) = 1 - P(A)$ $P(B) \sim \text{Beta}(10, 10),\ P(\neg B) = 1 - P(B)$
Mixed set $\{A, \neg A, B, \neg B\}$	A set consisting of one edge-event pair and one mid-event pair	$P(A) \sim \text{Beta}(1, 10),\ P(\neg A) = 1 - P(A)$ $P(B) \sim \text{Beta}(10, 10),\ P(\neg B) = 1 - P(B)$

3.1.2. Ranking Model with Read-out Noise

In addition to the simulation of the Basic Ranking Model, we explored the effect of read-out noise parameter d on the predicted probabilities of different ranking categories by simulating the Ranking Model with Read-out Noise. For the simulation of the Ranking Model with Read-out Noise, we used the same parameter settings for $P(A)$, $P(B)$ (across the same three levels), and N (across the same 26 levels) as described above. Same as in the simulation of the Basic Ranking Model, for each combination of event-set type and sample size, we performed 10,000 simulation runs and generated 10,000 sets of predicted responses. In all simulation runs, the additional parameter, read-out noise d, was set at a fixed value of 0.3, which we consider to be a reasonable approximation of its upper bound.

We determined the maximum value for the noise d based on its empirically estimated and theoretical maximum values. For its empirically estimated values, Costello and Watts (2016) reported that the average rate of mistaking a sampled instance (as its complement) was 0.24.[3] The maximum theoretical value that the noise d can take is 0.5, which regresses the probability that governs the sampling process to 0.5 regardless of the original underlying probability. Therefore, setting d to a value of 0.3 allows us to investigate the effect of read-out noise on the occurrence of ranking categories to a sufficient extent.

3.2. Results

Fig. 2 shows the mean predicted probabilities of participants' responses falling into each ranking category as a function of the event-set type, sample size, and whether read-out noise is present.[4] A visual inspection of Fig. 2 reveals qualitative differences among event sets (columns) in the probability distribution of ranking categories. For instance, the predicted probability of logical rankings is smallest for the mid-event sets, regardless of sample size N. Additionally, the predicted probability of stacked-illogical rankings is largest for the mid-event sets, also irrespective of sample size N. Notably, these qualitative differences across event-set types remain consistent in the predictions derived from both the Basic Ranking Model and the Ranking Model with Read-Out Noise.

Since the predicted probabilities of providing different ranking categories are not independent and must sum to one, directly deriving testable predictions from the full distribution shown in Fig. 2 is challenging. To circumvent this problem, we decomposed the distribution of ranking categories into (conditional) probabilities of ranking categories, considering that the occurrence of different ranking categories is mutually exclusive (this decomposition procedure is also known as "nested dichotomies" in statistics; e.g., J. Fox, 2015). As a first step, we examined the probability of participants giving a logical ranking versus an illogical ranking. As a second step, we examined the conditional probability of giving a stacked-illogical ranking versus an interlaced-illogical ranking, given that the ranking was illogical.

Fig. 3 shows the decomposed probabilities of different ranking categories. Note that the composition of the figure is different from that of Fig. 2: While the two rows still correspond to the two models without and with read-out noise and the x-axis still represents the sample size, the columns now refer to the decomposed ranking categories and the lines represent different event-set types. Fig. 3 now clearly shows the effects of event-set types on the predicted (conditional) probabilities of the ranking categories. Furthermore, the observed qualitative patterns are very similar across both rows, indicating that the effect of read-out noise d is negligible. Under mild conditions – specifically, assuming a sample size greater than 3 for the Basic Ranking Model and greater than 10 for the Ranking Model with Read-Out Noise (where the noise term d is set to its maximum value) – the following qualitative patterns consistently emerge from predictions derived from both models, regardless of the sample size N:

- The probability of giving a logical ranking is highest for the mixed sets, second highest for the edge-event sets, and lowest for the mid-event sets;

[3] Costello and Watts (2016) estimated the average value of the read-out noise, d, by using participants' probability estimations about related events to form probabilistic expressions.

[4] To calculate the mean predictions, we averaged the simulation outputs from 10,000 runs for each combination of the event-set type and sample size, resulting in 3×26 $(= 78)$ predicted probabilities. This averaging process was done for the simulation of the Basic Ranking Model and the simulation of the Ranking Model with Read-out noise, respectively.

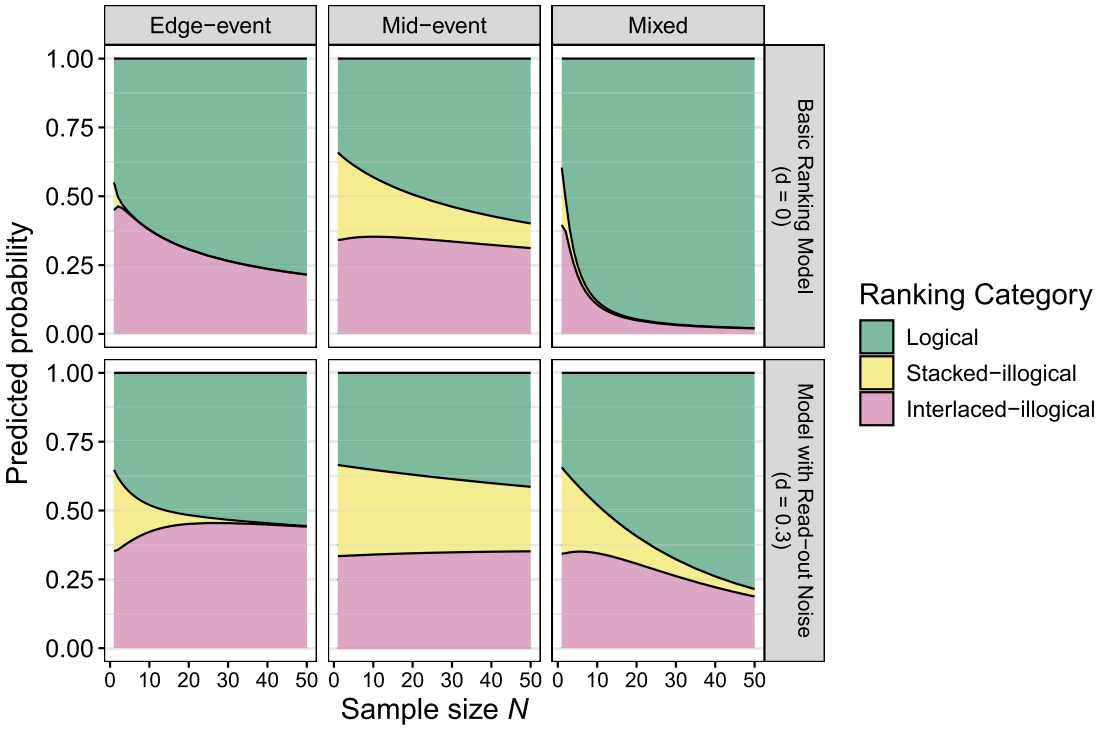

Fig. 2. Simulation results before decomposing the distribution of ranking categories.
Predicted response probabilities of different ranking categories as a function of event-set type and sample size N, derived from the Basic Ranking Model (first row) and the Ranking Model with Read-Out Noise (second row). The probabilities of different ranking categories always sum to one for a given sample size and event set. The x-axis represents the sample size N, columns refer to different types of event sets, and colors represent different ranking categories.

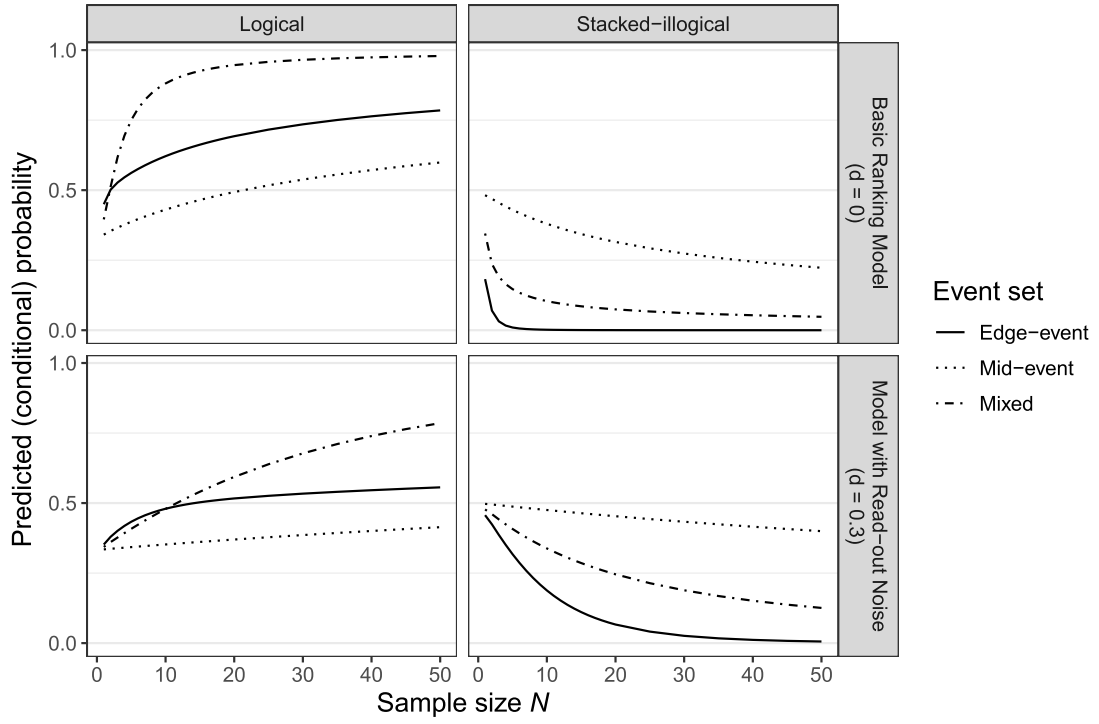

Fig. 3. Simulation results after decomposing the distribution of ranking categories.
Predicted (conditional) probabilities of different ranking categories as a function of event-set type (lines) and sample size N (x-axis), derived from the Basic Ranking Model (top row) and the Ranking Model with Read-Out Noise (bottom row). The predicted (conditional) probabilities for each plot are as follows. "Logical": probability of giving logical rankings versus all other rankings. "Stacked-illogical": conditional probability of giving stacked-illogical rankings versus interlaced-illogical rankings.

Table 3
Examples of event pairs selected in the pilot study.

Event pair	Ways of construction	Constituent events	Example
Edge-event	$P(A) \approx 1$	Positive event	A randomly selected person in Germany was born in a hospital.
	$P(\neg A) \approx 0$	Negative event	A randomly selected person in Germany was NOT born in a hospital.
	$P(A) \approx 0$	Positive event	In a randomly selected year, it will snow in Germany in June.
	$P(\neg A) \approx 1$	Negative event	In a randomly selected year, it will NOT snow in Germany in June.
Mid-event	$P(A) \approx 0.5$	Positive event	On a randomly selected day in Hamburg, there will be rain.
	$P(\neg A) \approx 0.5$	Negative event	On a randomly selected day in Hamburg, there will be NO rain.

- The conditional probability of giving a stacked-illogical ranking, given that participants do not give a logical ranking, is highest for the mid-event sets, second highest for the mixed sets, and lowest for the edge-event sets.

4. Overview of experiments

In the following, we aim to test the qualitative predictions derived from the Ranking Model empirically. As a first step, we conducted a pilot study to obtain normed event pairs for which people share common beliefs about their potential range of probabilities. The normed event pairs were then used in Experiment 1 and Experiment 2 to construct different types of event sets. Experiment 1 tested the qualitative predictions we derived for the mid-event and edge-event sets. Experiment 2 tested the qualitative predictions we derived for the mid-event, edge-event, and mixed sets.

In both Experiments 1 and 2, we investigated two conditions of the Event Ranking Task, namely, the ties-allowed condition and the ties-not-allowed condition, using a between-subjects design. To streamline the presentation, we report the ties-not-allowed condition in the main text and defer the more complex ties-allowed condition, which produced similar results, to the Supplementary Materials S2.

5. Pilot study

The goal of the pilot study was to generate normed edge-event and mid-event pairs. Full details of the pilot study can be found in Appendix A. Briefly, we first generated 200 event pairs, each of which consisted of two events: a positive event (e.g., "On a randomly selected day in Hamburg, there will be rain".) and a complementary negative event (e.g., "On a randomly selected day in Hamburg, there will be NO rain".). All of the event pairs were related to Germany, and all participants in all experiments were located in Germany. Two hundred event pairs were divided into four groups, each with 50 pairs. Each participant was randomly assigned to one group and estimated the probabilities of both positive and negative events for each pair on a probability scale from 0% to 100% in steps of one percentage point.

To obtain a set of normed event pairs, we inspected the density plots of participants' probability estimates for each event pair. Based on these density plots, we selected a set of 24 event pairs whose density distributions visually resembled the shapes of the Beta distributions used in the simulation and whose constituent events appeared to be complementary to each other. The density plots for the selected event pairs are shown in Fig. 4, examples are presented in Table 3, and the full list of selected event pairs is given in Table A.1 in Appendix A.

6. Experiment 1

Experiment 1 tested the qualitative predictions derived from the Ranking Model for the mid-event and edge-event sets. Specifically, according to the simulation results, we expected the probability of logical rankings versus illogical rankings to be larger for edge-event sets compared to mid-event sets. Furthermore, we expected the conditional probability of stacked-illogical rankings versus interlaced-illogical rankings to be larger for mid-event sets compared to edge-event sets.

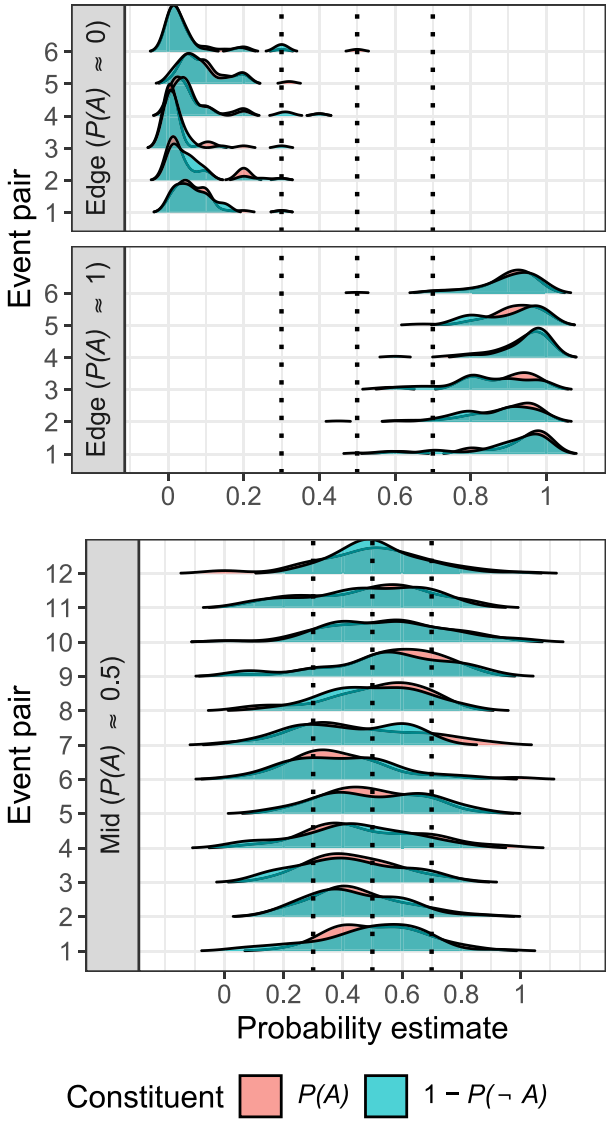

Fig. 4. Density Plot of the probability estimates of the positive event (in red) and the complementary estimates of the negative event (in blue) for each selected event pair. The density plot of the complementary estimates of the negative event is constructed in two steps. In the first step, the estimates provided by participants for a given negative event were subtracted from 100%. In the second step, the density plots were drawn using these complementary probability estimates. The upper panel shows six selected edge-event pairs in which the constituent positive events have probabilities close to 0. The middle panel shows six selected edge-event pairs in which the constituent positive events have probabilities close to 1. The bottom panel shows twelve selected mid-event pairs.

Table 4
Event sets constructed for Experiment 1.

Event-set type	Ways of creating the event set	No.
Edge-event sets $\{A, \neg A, B, \neg B\}$	$P(A) \approx 0,\ P(B) \approx 0$	2
	$P(A) \approx 1,\ P(B) \approx 1$	2
	$P(A) \approx 0,\ P(B) \approx 1$	2
Mid-event sets $\{A, \neg A, B, \neg B\}$	$P(A) \approx 0.5,\ P(B) \approx 0.5$	6

Note. Every time we constructed an event set, we randomly selected two event pairs from the list of event pairs obtained in the pilot study that had not been previously used. No. = number of sets created this way for each participant.

6.1. Methods

6.1.1. Design

The experiment implemented a 2 × 2 mixed design with factors event-set type and possibility of giving ties. Event-set type was a within-subjects factor with two levels, mid-event sets and edge-event sets. Each participant was asked to provide rankings for thirteen event sets: six mid-event sets, six edge-event sets, and one set which was used as a comprehension check item. For the comprehension check, participants were asked to rank an event set with a clear ranking, as the events in the set were widely recognized and their probabilities widely accepted.[5]

We manipulated between subjects if participants were allowed to give ties in their provided rankings of events. In the ties-allowed condition, participants were allowed to give ties (i.e., assign the same rank to more than one event). In the ties-not-allowed condition, participants were not allowed to give ties. Here, we only report the results from the ties-not-allowed condition (See Supplementary Materials S2.5 for the methods, results, and discussion of the ties-allowed condition).

6.1.2. Participants

186 participants located in Germany were recruited via Prolific (www.prolific.co), among whom 8 were excluded because they did not pass the comprehension check, and 1 was excluded because they indicated that they were not proficient in German. Of the remaining participants, 86 participants were assigned to the ties-not-allowed condition (43 females, 41 males, and 2 others) with a mean age of 26.20 (SD = 8.26) years. 91 participants were randomly assigned to the ties-allowed condition. Participants were compensated with £2 for their participation.

6.1.3. Materials

We constructed two types of event sets: six edge-event sets and six mid-event sets. Each event set was constructed by randomly selecting, for each participant anew, two event pairs without replacement from the 24 event pairs selected in the pilot study so that every event pair was used and used only once. The different ways of constructing the event sets are summarized in Table 4.

6.1.4. Procedure

The experiment was programmed in `lab.js` (Henninger et al., 2022). It consisted of thirteen ranking trials. In each trial, participants were presented with one event set consisting of two pairs of complementary events, A, $\neg A$, B, $\neg B$. Participants were asked to create a ranking of these events based on the perceived probabilities. The order of the twelve event sets (i.e., six edge-event and six mid-event sets) was randomized for each participant. After ranking six event sets, on the seventh trial of the experiment, participants were presented with a comprehension check item. Afterward, participants ranked the remaining six event sets. At the end of the experiment, participants were asked for their demographic information.

Fig. 5 shows one example trial of the Event Ranking Task (translated to English). The four events from one event set were presented simultaneously on the right side of the screen, and the participants were asked to create a ranking by dragging and dropping the events to the left side of the screen. Participants in the condition presented here were not allowed to give ties (see Supplementary Material S2.5.1.4 for an example trial under the ties-allowed condition, in which participants received instructions explicitly permitting ties).

6.1.5. Analysis

For the data analysis, we used a multinomial processing tree (MPT) model (Riefer & Batchelder, 1988). Typically, MPT models are used as cognitive measurement models that relate probabilities underlying observed response frequencies to latent cognitive processes (Batchelder & Riefer, 1999; Erdfelder et al., 2009; Schmidt et al., 2023; Singmann et al., 2024). However, in the present study, we used an MPT model purely as a statistical tool to map out and decompose the underlying multinomial distribution of observed responses, in line with the decomposition used in the simulation. The benefit of using this decomposition also for the analysis is that it permits testing each prediction derived from the simulation in a statistically independent manner.

We constructed the MPT model following the decomposition introduced in the simulation section, as illustrated in Fig. 6. The first step is to assess the logicality of the response; with probability l a logical ranking is produced and with probability $1 - l$ an illogical ranking is produced. If a ranking is illogical, we take the second step to assess what category of illogical rankings the ranking belongs to. Given that the ranking is illogical, the conditional probability of it belonging to the stacked- versus interlaced-illogical rankings is represented by the parameter s. The MPT model is fully saturated and perfectly describes any data pattern that can emerge in the Event Ranking Task. Furthermore, the model is globally identifiable, with each branch terminating in a distinct ranking category.

We fitted the corresponding MPT model to the data from different event-set conditions using a hierarchical-Bayesian approach (Klauer, 2010; Singmann et al., 2024). Two sets of group-level parameters θ were estimated for two event-set conditions, with θ_e for the edge-event sets and θ_m for the mid-event sets. The model fitting was implemented via the R package `TreeBUGS` (Heck et al., 2018).

The relations between MPT parameters estimated for edge-event and mid-event set conditions (l_e vs. l_m and s_e vs. s_m) should align with the predictions of the Ranking Model:

1. The probability of logical rankings is greater when ranking edge-event sets compared to when ranking mid-event sets: $l_e > l_m$.
2. The conditional probability of stacked-illogical rankings (versus interlaced-illogical rankings) is greater when ranking mid-event sets compared to when ranking edge-event sets: $s_m > s_e$.

To assess the difference between parameters estimated for different event-set conditions, we calculated the posterior difference distributions for parameters l and s. For ease of interpretation, we always subtracted the distribution of the expected smaller parameter estimate from the distribution of the expected larger parameter estimate.[6] Thus, results are in line with the predictions of the Ranking Model if the difference distributions are positive. We considered there to be a statistically meaningful difference between the group-level parameter estimates obtained from two event-set conditions if more than 95% of the probability mass of the difference distribution was above 0.

[5] Details about the comprehension check question can be found in the OSF repository (https://osf.io/hw8p9/).

[6] Specifically, for parameter l, we calculated l_e - l_m. For parameter s, we calculated $s_m - s_e$.

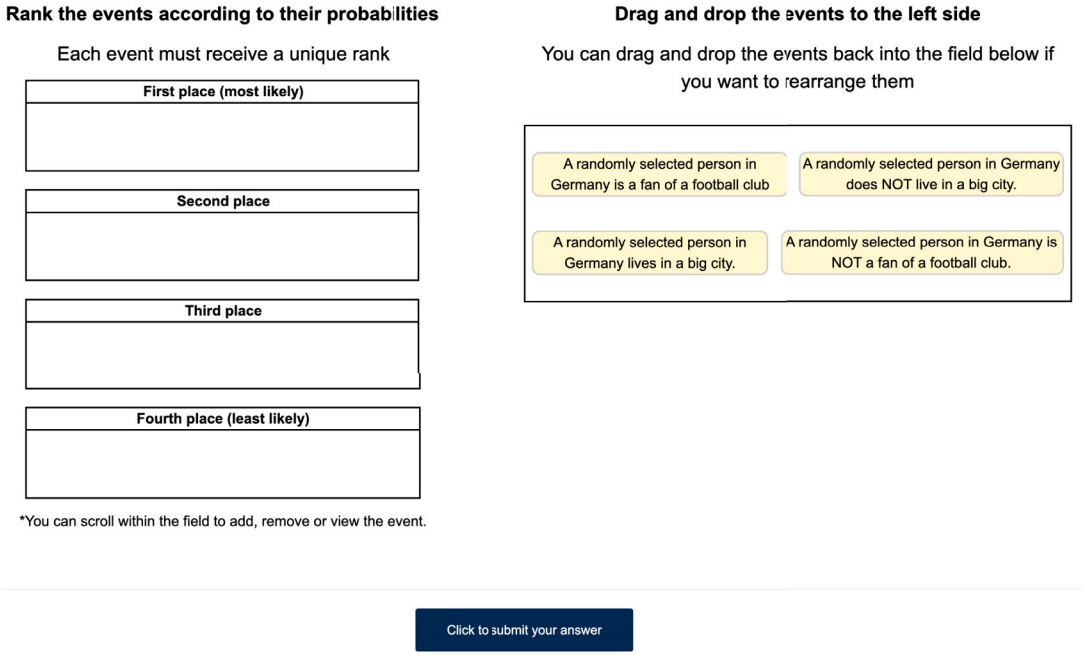

Fig. 5. Example trial of the Event Ranking Task.
The original experiment was conducted in German. We translated the instructions and the materials shown in this screenshot from German to English.

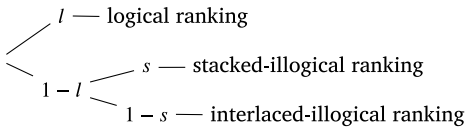

Fig. 6. MPT model for decomposing the rankings.
Multinomial processing tree (MPT) model that decomposes the ternary ranking category into two independent and nested binomials. l = probability of giving a logical ranking; s = conditional probability of giving a stacked-illogical ranking given that participants did not give a logical ranking.

Table 5
Parameter estimates of the MPT model in Experiment 1.

Parameter	Edge-event sets	Mid-event sets
l	.79 [.74; .84]	.52 [.46; 0.58]
s	.05 [.01; .12]	.34 [.24; 0.44]

Note. MPT parameter estimates for the edge-event set and mid-event set conditions in Experiment 1. l = probability of giving a logical ranking; s = conditional probability of giving a stacked-illogical ranking given that participants did not give a logical ranking. The brackets indicate the 95% credibility intervals.

6.2. Results

The proportion of logical, stacked-illogical, and interlaced-illogical rankings in participants' responses can be found in Fig. 7. This figure clearly shows that participants frequently provided illogical rankings. Furthermore, the proportions with which participants provided different ranking categories differed across event-set types. In line with the prediction, the edge-event set condition produced more logical rankings than the mid-event set condition, and the mid-event set condition produced more stacked-illogical rankings than the edge-event set condition. To statistically substantiate these results patterns, we performed the MPT analysis in the following.

6.2.1. Model-based results

Table 5 provides the group-level estimates of MPT model parameters. Fig. 8 shows the posterior difference distributions comparing the group-level estimates across event-set types.

In line with the first prediction from the Ranking Model, parameter l, representing the probability of providing a logical ranking, was meaningfully larger for edge-event sets than for mid-event sets. In line with the second prediction, parameter s, the conditional probability of providing a stacked-illogical ranking versus an interlaced-illogical ranking was meaningfully larger for mid-event sets than for edge-event sets.

6.3. Discussion

When asked to rank two pairs of complementary events, participants frequently produced illogical rankings, around 25% for edge-event sets and 50% for mid-event sets. The Ranking Model, a sampling-based model for ranking tasks, can not only explain the occurrence of illogical rankings but also correctly predict the qualitative pattern of ranking categories across event-set types.

Just as the simulation predicted, participants exhibited a greater tendency to provide illogical rankings for mid-event sets. Additionally, we observed that the conditional probability of participants providing stacked-illogical rankings (versus interlaced-illogical rankings) was higher when ranking mid-event sets compared to edge-event sets. These behavioral results provide evidence for the idea that mental sampling underlies probability judgments.

7. Experiment 2

Experiment 2 aimed to offer a more stringent test of the Ranking Model by assessing its predictions for mixed sets in addition to those for mid-event and edge-event sets. Additionally, we intended to replicate the findings from Experiment 1. We preregistered Experiment 2, including its sample size, hypotheses, design, and analysis. The preregistration can be found in the Open Science Framework Registries https://osf.io/hw8p9/.

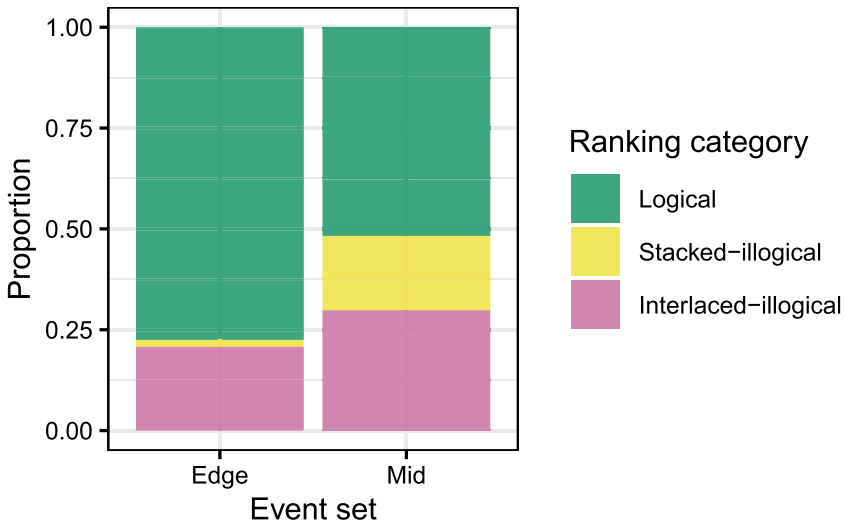

Fig. 7. Proportion of three ranking categories in Experiment 1.
Colors indicate the ranking categories. For edge-event sets, the proportions of logical, stacked-illogical, and interlaced-illogical are 0.78, 0.02, and 0.21, respectively. For mid-event sets, the proportions are 0.52, 0.18, and 0.30, respectively.

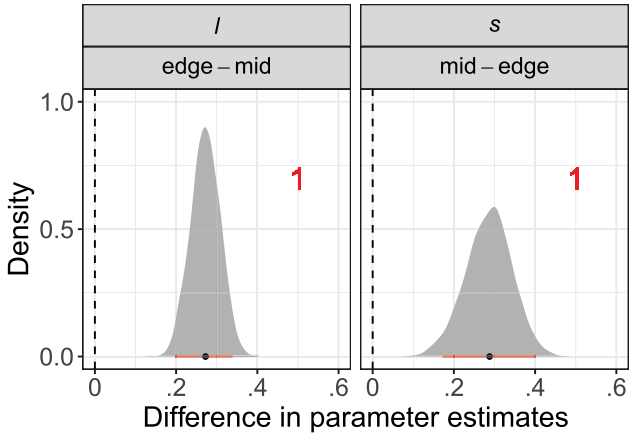

Fig. 8. Posterior difference distributions comparing the MPT parameter estimates across event-set types calculated for Experiment 1.
l and *s* are the specific MPT parameters being compared (see Table 5 for their estimated values and interpretations). "edge – mid" and "mid – edge" specify how the event sets were compared: "edge – mid" means the estimate for mid-event set was subtracted from the estimate for the edge-event set, and "mid – edge" means the opposite. All posterior difference distributions were expected to be positive, as the predicted smaller estimate was subtracted from the predicted larger estimate when making comparisons. A posterior with 95% of its probability mass greater than 0 indicates a credible difference between event sets in line with the predictions of the Ranking Model. The gray area shows the full posterior difference distribution. The black dot shows the median. The red line shows the 95% credibility interval. The red number indicates the proportion of probability mass of the difference distribution larger than 0.

7.1. Methods

The experiment used the same methods as Experiment 1, with the sole difference that we included mixed sets in addition to mid-event and edge-event sets.

7.1.1. Design

The experiment implemented a 3 × 2 mixed design with factors event-set type and possibility of giving ties. Event-set type was a within-subjects factor with three levels, mid-event sets, edge-event sets, and mixed sets. As in Experiment 1, participants were asked to provide rankings for thirteen event sets, four mid-event sets, four edge-event sets, four mixed sets, and one set which was the same comprehension check used in Experiment 1. We manipulated between subjects if participants were allowed or not allowed to give ties in rankings. Here, we only report the results obtained from the ties-not-allowed condition (See Supplementary Materials S2.6 for the methods, results, and discussion of the ties-allowed condition).

Table 6
Event sets constructed for Experiment 2.

Event-set type	Ways of creating the event set	No.
Edge-event sets $\{A, \neg A, B, \neg B\}$	$P(A) \approx 0,\ P(B) \approx 0$	1
	$P(A) \approx 1,\ P(B) \approx 1$	1
	$P(A) \approx 0,\ P(B) \approx 1$	2
Mid-event sets$\{A, \neg A, B, \neg B\}$	$P(A) \approx 0.5,\ P(B) \approx 0.5$	4
Mixed sets $\{A, \neg A, B, \neg B\}$	$P(A) \approx 0,\ P(B) \approx 0.5$	2
	$P(A) \approx 1,\ P(B) \approx 0.5$	2

Note. Every time we constructed an event set, we randomly selected two event pairs from the list of normed event pairs that had not been used. No. = number of sets created this way for each participant.

7.1.2. Participants

The target sample size was chosen in order to have roughly the same number of observations from each event-set condition as we had in Experiment 1. Therefore, 310 participants located in Germany were recruited from Prolific (www.prolific.co), among whom 9 were excluded because they did not pass the comprehension check. Of the remaining participants, 151 participants were assigned to the ties-not-allowed condition (69 females, 80 males, and 2 others) with a mean age of 29.20 (SD = 10.25) years. 150 participants were randomly assigned to the ties-allowed condition. Participants were compensated with £2 for their participation.

7.1.3. Materials

The materials were prepared in the same manner as those in Experiment 1 based on the event sets selected in the Pilot Study. The mixed set was constructed by randomly selecting one normed edge-event pair and one normed mid-event pair. The edge-event sets and the mid-event sets were constructed using the same method as in Experiment 1. Table 6 summarizes the construction of the event sets, which was done randomly for each participant.

7.1.4. Procedure

The procedure was the same as in Experiment 1.

Table 7
Parameter estimates of the MPT models in Experiment 2.

Parameter	Edge-event sets	Mid-event sets	Mixed sets
l	.72 [.68; .77]	.45 [.40; .50]	.86 [.82; .90]
s	.06 [.01; .12]	.41 [.33; .48]	.17 [.03; .31]

Note. MPT parameter estimates for the edge-event set, mid-event set and mixed set conditions in Experiment 2. l = probability of giving a logical ranking; s = conditional probability of giving a stacked-illogical ranking given that participants did not give a logical ranking. The brackets indicate the 95% credibility intervals.

7.1.5. Analysis

We followed the same analysis steps as in Experiment 1, employing the MPT model shown in Fig. 6. We estimated the MPT model jointly for three different event-set conditions. Separate sets of group-level parameters were estimated for three event-set conditions: one set θ_e for the edge-event set condition, one set θ_m for the mid-event set condition, and one set θ_x for the mixed set condition.

The ordinal relationships of MPT parameters estimated for different event-set conditions should align with the qualitative predictions – derived from the simulation – about the occurrence of different ranking categories across different event sets as follows:

1. The probability of giving a logically possible ranking is highest for mixed sets, second highest for the edge-event sets, and lowest for the mid-event sets: $l_x > l_e > l_m$.
2. The conditional probability of giving a stacked-illogical ranking (versus an interlaced-illogical ranking) is highest for mid-event sets, second highest for the mixed sets, and lowest for the edge-event sets: $s_m > s_x > s_e$.

To compare the MPT parameter estimates for different event sets, we calculated the posterior difference distributions for parameters l and s. For each parameter, we made three pairwise comparisons and computed three posterior difference distributions to fully test the predicted ordinal relationships (e.g., for parameter l, $l_x > l_e > l_m$). Specifically, we first compared the expected largest estimate and the expected second largest estimate (e.g., l_x and l_e), then compared the expected second largest with the expected smallest (e.g., l_e and l_m). Lastly, we compared the expected largest with the expected smallest parameter (e.g., l_x and l_m). As in Experiment 1, we always subtracted the distribution of the expected smaller parameter estimates from the distribution of the expected larger parameter estimates. Hence, the Ranking Model predicts a positive posterior difference distribution for every MPT parameter comparison.

7.2. Results

The proportion of logical, stacked-illogical, and interlaced-illogical rankings in participants' responses can be found in Fig. 9. As in Experiment 1, illogical rankings were produced frequently. Furthermore, there were clear differences across event sets. In line with the predictions of the Ranking Model, logical rankings were most common in the mixed set condition, followed by the edge-event set condition, and least common in the mid-event set condition. For the illogical rankings, the pattern also appeared to be in line with the predictions; stacked-illogical rankings were most common in the mid-event set condition, followed by the mixed and edge-event set conditions. To statistically substantiate these results patterns, we performed the MPT analysis.

7.2.1. Model-based results

Table 7 provides the group-level estimates of MPT model parameters. Fig. 10 shows the posterior distributions of the differences between the parameters estimated for different event-set conditions.

For parameter l, the probability of providing a logical ranking, the predictions regarding the ordinal relationships among all three types of event sets (i.e., mixed, mid-event, and edge-event sets) were strongly supported: l was largest for the mixed sets, second largest for the mid-event sets, and smallest for the edge-event sets. This also implies that we replicated the results for mid-event and edge-event sets from Experiment 1.

The predictions for the parameter s, the conditional probability of providing a stacked-illogical ranking versus an interlaced-illogical ranking, were mostly supported by the MPT model. The predictions are that $s_m > s_x > s_e$. While the qualitative pattern supported the predictions – all posterior medians and more than 90% of posterior mass were positive – the 95% credibility interval for one posterior difference distribution included 0. Specifically, we replicated the ordinal pattern for mid-event sets and edge events already observed in Experiment 1. We also found that the conditional probability of providing a stacked-illogical ranking versus an interlaced-illogical ranking was meaningfully larger for mid-event sets than for mixed sets ($s_m > s_x$ was supported). However, the comparison between mixed sets and edge-event sets did not reach our inference criterion ($s_x > s_e$ was not supported). Still, more than 90% of posterior mass in the distribution for comparing the mixed and edge-event sets were positive, suggesting that there is more evidence for the predicted ordinal relationship holding than not.

7.3. Discussion

In Experiment 2, participants were asked to provide rankings for mixed sets alongside mid-event and edge-event sets. As in Experiment 1, we again showed that participants frequently produced illogical rankings. Furthermore, the frequency with which participants produced illogical rankings across the three different types of event sets again matched the predictions of the Ranking Model. Illogical rankings were least likely for mixed sets (around 20%), more likely for edge-event sets (around 30%), and most likely for mid-event sets (around 55%).

The Ranking Model also made specific predictions regarding the frequency with which different types of illogical rankings should occur across the three event sets. In total, we tested three predicted ordinal relationships. Two of these ordinal relationships met our inference criterion. However, the predicted ordinal relationship comparing the edge-event and mixed sets, in terms of the conditional probability providing stacked- versus interlaced-illogical rankings, did not meet our pre-specified inference criterion. This appears to be a power issue due to the low frequency of stacked-illogical rankings occurring under both the edge-event set condition and the mixed set condition (around 3% of responses for edge-event sets and around 4% for mixed sets). Notably, the Ranking Model predicts that the probability of participants providing stacked-illogical rankings for mid-event sets should be relatively large, while the predicted probabilities for the edge-event sets and for the mixed sets are small except in cases where readout noise is large and the mental sampling size is small (see Fig. 2). Additionally, the model predicts a relatively larger difference in the conditional probability of providing stacked- versus interlaced-illogical rankings when comparing edge-event and mid-event sets, as well as when comparing mid-event and mixed sets. In contrast, the model predicts a smaller difference for the comparison between mixed and edge-event sets (see Fig. 3). Thus, the results for stacked-illogical rankings seem to be entirely in line with the predictions from the Ranking Model. Comparing the edge-event and mixed sets with a larger sample size in order to detect a small difference in the conditional probability of providing stacked- versus interlaced-illogical rankings appears a promising direction for future research. Taken together, these results provide further evidence for the idea that mental sampling underlies probability judgments.

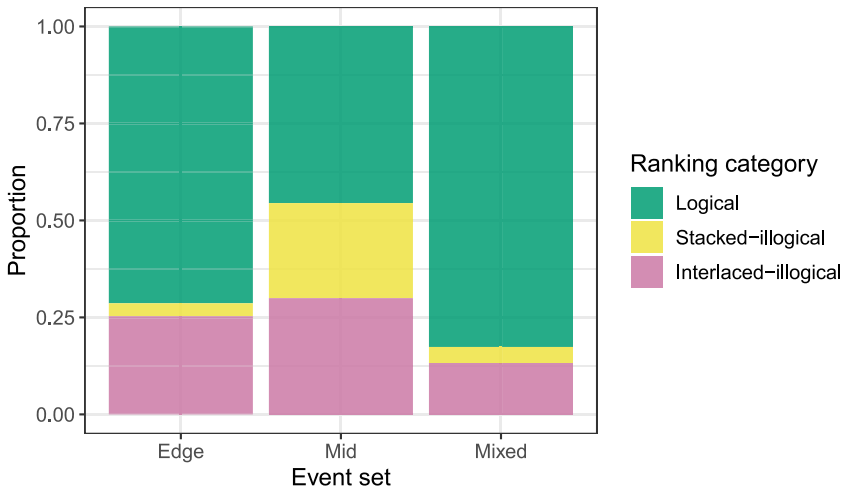

Fig. 9. Proportion of three ranking categories in Experiment 2.
Colors indicate the ranking categories. For edge-event sets, the proportions of logical, stacked-illogical, and interlaced-illogical are 0.71, 0.03, and 0.25, respectively. For mid-event sets, the proportions are 0.46, 0.25, and 0.30, respectively. For mixed sets, the proportions are 0.83, 0.04, and 0.13, respectively.

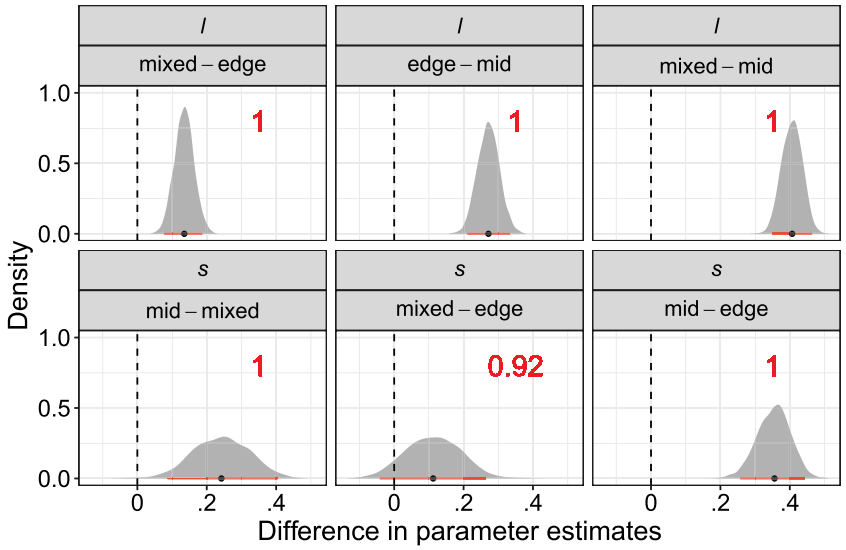

Fig. 10. Posterior Difference distributions comparing the MPT parameter estimates across event-set types calculated for Experiment 2.
l and *s* are the specific MPT parameters being compared (see Table 7 for their estimated values and interpretations). "mixed – edge", "edge – mid", "mixed – mid", "mid – mixed", "mixed – edge " and "mid – edge" specify how the event sets were compared: the estimate for the latter event set was subtracted from the estimate for the former event set. For example, "mixed – edge" means the estimate for the edge-event set was subtracted from the estimate for the mixed set. Three pairwise comparisons were conducted to test the ordinal relationship mixed > edge > mid for the parameter *l*. Similarly, three pairwise comparisons were conducted to test the ordinal relationship mid > mixed > edge for the parameter *s*. All posterior difference distributions were expected to be positive, as the predicted smaller estimate was subtracted from the predicted larger estimate when making comparisons. A posterior with 95% of its probability mass greater than 0 indicates a credible difference between event sets in line with the predictions of the Ranking Model. The gray area shows the full posterior difference distribution. The black dot shows the median. The red line shows the 95% credibility interval. The red number indicates the proportion of probability mass of the difference distribution larger than 0.

8. General discussion

Models developed under the mental sampling framework have been successful in explaining observed effects in people's probability judgments and probabilistic reasoning. However, less effort has been devoted to empirical investigations of the fundamental ideas underlying the mental sampling framework. In this study, we set out to derive testable predictions from the mental sampling framework using the novel Event Ranking Task. We began our theory testing effort by developing a formal model tailored to the Event Ranking Task – the Ranking Model – based on existing direct sampling models of probability estimation (Costello & Watts, 2014; Zhu et al., 2020). The Ranking Model makes novel predictions that when creating the rankings of probabilities, people will violate the complement rule in a predictable manner: the probability of providing illogical rankings, as well as the occurrence of different types of illogical rankings, depend on the underlying probabilities that govern the sampling processes. Such predictions were derived from a simulation study where we varied the range of the underlying probability parameters in different event-set conditions. In two online experiments, we tested the qualitative predictions derived from the simulation using experimental manipulations that map the

setting in the simulation. The predicted pattern that the occurrence of different ranking categories changes across event-set conditions was closely confirmed in these two online experiments (in both the ties-not-allowed condition reported here as well as the ties-allowed condition reported in Supplementary Material S2).

8.1. Complementarity in aggregated versus trial-level probability judgments

Our results appeared to conflict with *binary complementarity* (Tversky & Koehler, 1994), which refers to the phenomenon that the averaged estimates (or medians as shown by C. R. Fox, 1999) across trials and participants for two complementary events roughly sum up to 1, in line with the complement rule. However, the current study showed that people violated the complement rule, providing illogical rankings in a significant proportion of trials. This is because we were focusing on a different level than previous studies: whereas previous studies focused on probability estimations at an aggregated level, the present study focused on the probability rankings at the trial level.

Mental sampling makes different predictions regarding the agreement with the complement rule at the aggregated and the trial levels. According to mental sampling, the underlying probabilities of two complementary events sum up to one. Even though individual probability estimates of an event based on a single sample are subject to sampling variation, they are assumed to be distributed around the event's underlying probability. Along this line of reasoning, the expected sum of probabilities of two complementary events is distributed around 1.[7] At the trial level, however, mental sampling posits that people draw only a small sample to evaluate each event's probability. For instance, in the Event Ranking Task, the Ranking Model assumes that at each trial, participants draw four independent samples to evaluate the four events A, $\neg A$, B, and $\neg B$ – one sample for each event. Because each sample is prone to random fluctuation, these one-time draws may not reflect the true relationships among the events' underlying probabilities. For example, even if the underlying probability of event A is higher than that of B, a single set of small samples might suggest the opposite ordering. Furthermore, when different samples are drawn for two complementary events, apparent violations of the complement rule can arise at the trial level. The judgments at the trial level (e.g., individual rankings) are not derived from expected values but rather from the specific outcomes of individual sampling processes. Hence, the observed individual rankings can deviate from what might be expected when the means of the estimates are considered.

8.2. Comparing the Ranking Model and the direct sampling models of probability estimations

Compared with the two most popular direct sampling models for the probability estimation task (Costello & Watts, 2014; Zhu et al., 2020), the Ranking Model has the advantage of specifying how participants engage in the experimental task. Especially, the Ranking Model allows us to derive predictions for the probability of every possible ranking for any given set of parameters. The models for the probability estimation task, however, cannot make predictions that cover the full response space (e.g., the probability scale from 0 to 1). With a given value of the sample size of mental sampling, the Bayesian sampler and the PT+N model predict only a limited number of possible point estimates on the probability scale. For instance, if the sample size of mental sampling is 3, then the PT+N model can only predict fractions with a denominator of 3 as possible estimates, because probability estimations are based on the relative frequency of instances that support or do not support the event under evaluation in the obtained sample. Therefore, the predicted estimates can only be $\frac{0}{3}$, $\frac{1}{3}$, $\frac{2}{3}$, $\frac{3}{3}$. However, infinite decimals like $\frac{1}{3}$ or $\frac{2}{3}$ are impossible to give on the probability scale by participants. Additionally, in the raw data of Costello and Watts (2014), prime numbers such as 0.01 or 0.71 appeared frequently. If the model does not introduce additional assumptions about how participants round their estimates, estimates like 0.01 or 0.71 can only occur with a sample size ≥ 100. Using a sample size greater than 100 for a query appears unlikely from a resource rational perspective (Griffiths et al., 2015). It, therefore, appears more likely that participants incorporate a secondary rounding process when responding on a probability scale from 0% to 100%. However, neither the PT+N model nor the Bayesian sampler model specified the rounding process for probability estimation. The Ranking Model can be treated as an experimental model (Kellen, 2019), which removes the ambiguity regarding how participants would use the response scale.

8.3. Necessity of model assumptions and possible alternatives

The Ranking Model is based on the binomial sampling process and involves several important assumptions. Firstly, it assumes that people draw independent samples to evaluate different events. Secondly, it assumes that people use a fixed sample size to evaluate a given event set and convert samples to a ranking based on the counts of events that occur in the samples. In this section, we address possible critiques or alternatives of these assumptions and argue that these assumptions are supported by and provide a parsimonious account of the data.

First, as mentioned in the Introduction (the *Logical and Illogical Rankings* section), the assumption that people draw independent samples for different events, even if they are complementary to each other, seems implausible from a resource-rational perspective. However, we argue that the alternative assumption – people reusing the same sample for complementary events – fails to account for the data in this study. To examine how this alternative assumption impacts the model predictions, we modify the Ranking Model by considering two scenarios.

In the first scenario, we assume people evaluate a pair of complementary events based on a single sample and maintain the assumption that people use a consistent sample size N for evaluating different event pairs. Specifically, a sample of size N including x instances of A and $N-x$ instances of $\neg A$ is drawn for the event pair $\{A, \neg A\}$. A sample of size N including y instances of B and $N-y$ instances of $\neg B$ is drawn for the event pair $\{B, \neg B\}$. According to these sampling results, when x (the number of occurrences of A, or O_A) $\geq y$ (O_B), it must follow that $N-x$ ($O_{\neg A}$) $\leq N-y$ ($O_{\neg B}$). The ranking, based on these sampling results, would then always follow the complement rule and be logical, which contradicts the data in which illogical rankings were common.

One might argue that illogical rankings might still occur if we further relax the assumption that people use a constant sample size. Thus, in the second scenario, we assume that people draw different sample sizes for the two pairs of complementary events. This assumption also cannot predict illogical rankings. The proof is as follows. Imagine that people draw a sample of N instances to evaluate the event A, including x instances of A and $N-x$ instances of $\neg A$. People draw another sample of M instances, including y instances of B and $M-y$ instances of $\neg B$. Since the sample sizes used for evaluating two pairs of complementary events differ, we additionally assume that participants first convert the instances into a relative frequency and then derive a ranking based on relative frequencies, with $\frac{x}{N}$ instances indicating the proportion of A, $\frac{N-x}{N}$ indicating the proportion of $\neg A$, $\frac{y}{M}$ indicating the proportion of B, and $\frac{M-y}{M}$ indicating the proportion of $\neg B$. Mathematically, we can prove that when $O_A \geq O_B$; namely, $\frac{x}{N} \geq \frac{y}{M}$, then $1-\frac{x}{N} \leq 1-\frac{y}{M}$. This means that $\frac{N-x}{N} \leq \frac{M-y}{M}$. Recall that $O_{\neg A} = \frac{N-x}{N}$ and $O_{\neg B} = \frac{M-y}{M}$, which means that $O_{\neg A} \leq O_{\neg B}$ follows $O_A \geq O_B$ directly.

[7] Alternatively, according to the PT+N model, individual estimates are assumed to be distributed around an expected value that depends on underlying probabilities as well as noise in the sampling process. When combining/adding up the expected values of the two complementary events together, the "noise" will be canceled out, only leaving out the sum of the two complementary events' underlying probabilities.

With the two scenarios above, we show that the illogical rankings in our data cannot be predicted without assuming that participants use independent samples for complementary events. Next, we move on to discuss the necessity of the constant sample size assumption and the assumption that people derive rankings directly by counting the occurrence for each event. We do not think relaxing these assumptions offers any benefits compared to the models presented here, as the Ranking Model can already fully predict the observed patterns. Relaxing the constant sample size assumption would require three more parameters; calculating relative frequencies for making comparisons would require one more intermediary step compared to directly comparing counts. Surely, such more complex models could still predict the data pattern that matches the observed data. However, whether other model variants make the same predictions seems immaterial, given that the current predictions already match the observed data. In sum, the current model provides a parsimonious description. Given that there is no empirical necessity for assuming that people do not directly compare observed counts, the principle of parsimony suggests that it is unnecessary at this time to explore more complicated model variants with additional assumptions. If future work using the Event Ranking Task finds empirical patterns that are not predicted by the Ranking Model described here, it might be worthwhile to revisit these assumptions.

8.4. Limitations and future directions

So far, one aspect that has been neglected is the individual differences between participants. We observed that around 7% participants in the ties-not-allowed condition of Experiment 1 ($N = 6$) and around 5% participants in the ties-not-allowed condition of Experiment 2 ($N = 8$) produced no illogical rankings at all. Because participants in our study had to produce 12 rankings in total, such an outcome is unlikely to be solely the result of random sampling alone.[8] From the standpoint of the sampling model, producing no illogical rankings consistently is only possible when sampling an infinite number (or at least a very large number) of instances. Instead of assuming that people use an infinite/a very large number of instances, we hypothesize that these individuals consistently applied logical rules (i.e., the complement rule) when creating rankings. It suggests that there might be differences between participants who rely on mental sampling alone and others who also use logical rules. This raises important considerations for the development of more comprehensive models. If logical-rule-based reasoning indeed plays a part in the probability judgments for a group of individuals, future iterations of sampling models might benefit from incorporating this possibility explicitly. Such a model could account for a mixture of rule-based and sampling-based reasoning across participants, potentially offering a more complete understanding of the cognitive mechanisms at play.

One might wonder why we did not fit the Ranking Model to investigate individual differences in parameter estimates and instead relied solely on its qualitative predictions. Fitting the model at the individual level is challenging. The Basic Ranking Model consists of 75 equations with three free parameters. For the ties-not-allowed condition presented in the main text, these 75 equations are combined to generate predictions for 24 full orders. A single set of parameters corresponds to a probability distribution over these 24 possible responses. However, for each set of parameters (characterizing a single trial of the Event Ranking Task), we observe only one response out of the 24 possibilities. Even when aggregating data across trials for each participant, we obtain only 12 observed responses, which are insufficient to reliably estimate the three parameters (N, $P(A)$, and $P(B)$) for individual trials or participants. The sparsity of data also prevents us from deriving predictions for each of the 24 possible rankings. To address this limitation, we categorized the rankings to generate predictions at the level of ranking categories. Future research could explore alternative categorizations to test new predictions.

In the current study, we focus on investigating the influence of underlying probabilities on the occurrence of different ranking categories. However, the Ranking Model also predicts how other model terms should influence the occurrence of different ranking categories. For example, the model predicts that employing a larger N for drawing mental samples would lead to a higher rate of logical rankings, holding all other model terms constant. The model also predicts that having a higher read-out noise d in the sampling processes would lead to a lower rate of providing logical rankings, holding all other model terms constant. These predictions warrant empirical investigation in future research. One concrete approach could involve manipulating the sample size N through task difficulty and thinking time, as demonstrated by Hamrick et al. (2015). As for the noise parameter d, it could be measured in the probability estimation task, using the approach pursued by Costello and Watts (2018). The noise levels measured by the probability estimation task can then be used to predict the task performance in the Event Ranking Task. Additionally, there have been limited efforts to ground the important sampling terms in a psychological context (but see Lloyd et al., 2019). Thus, another intriguing line of research would be to see how individual differences in cognitive abilities, such as fluid intelligence and working memory, correlate with the model terms, especially the parameters N and d.

Finally, future research can extend the current Ranking Model to evaluate if the sampling-based ranking process proposed in this paper can explain how people rank not only marginal events, but also more complex events, such as conjunctions and disjunctions. In fact, the Linda problem introduced by Tversky and Kahneman (1983) is essentially a ranking task. The Linda problem was presented as a two-alternative forced-choice question in which participants ranked the probabilities of the marginal event A and the conjunctive event $A \wedge B$ (where $\wedge$ represents "and"). Variants of the Linda problem that involve ranking complex events can provide materials for further empirical investigations of the sampling-based ranking model.

Another important research question is whether the sampling-based ranking model can replicate the observed patterns of conjunction fallacies as effectively as or more effectively than sampling-based probability estimation models. For example, studies have shown that in some cases, such as the Linda problem, participants exhibited a high conjunction fallacy rate (around 80%) (Tversky & Kahneman, 1983). Costello and Watts (2017) demonstrated that this high rate can be explained by assuming greater noise for conjunctive events and lower noise for marginal events in the sampling process. Without this assumption, the original PT+N model predicts a ceiling rate of 50%. An important question is whether the ranking model, based on counts in mental samples, also predicts a ceiling rate of 50% without assuming different noise levels or if the ranking model can predict a high rate of conjunction fallacy without additional assumptions. Costello and Watts (2017) further illustrated that the sampling-based model of probability estimation effectively explains variations in conjunction fallacy rates as influenced by alterations in the underlying probabilities of individual events ($P(A)$, $P(B)$, and $P(A \wedge B)$). It is worthwhile to investigate whether the ranking model can reproduce the same pattern.

In addition to explaining previously observed effects in conjunction fallacy experiments, future research can generate new predictions for situations that involve both marginal events and complex events. When ranking the probabilities of the four events – two marginal events, their conjunction, and their disjunction ($P(A), P(B), P(A \wedge B), P(A \vee B)$) – there are only two full orders and one partial order that do not violate the conjunction and disjunction rules across all possible rankings of

[8] For example, in Experiment 1, according to the simulation study (Fig. 2), the smallest probability of producing illogical rankings for edge-events is .2, and for mid-events, it is .4. With these probabilities, the likelihood of producing at least one illogical ranking across 12 trials (6 edge-event and 6 mid-event) is $1 - ((1 - 0.2)^6 \times (1 - 0.4)^6) = .988$.

these four events. These three logical rankings are: $\hat{P}(A \vee B) > \hat{P}(A) > \hat{P}(B) > \hat{P}(A \wedge B)$, $\hat{P}(A \vee B) > \hat{P}(B) > \hat{P}(A) > \hat{P}(A \wedge B)$ or $\hat{P}(A \vee B) > \hat{P}(A) = \hat{P}(B) > \hat{P}(A \wedge B)$. Future research can extend the current Ranking Model and see whether the sampling-based ranking model makes testable predictions regarding the occurrence of these logical rankings.

CRediT authorship contribution statement

Xiaotong Liu: Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Arndt Bröder:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition. **Henrik Singmann:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Funding

This research was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG, GRK 2277) to the Research Training Group "Statistical Modeling in Psychology" (SMiP).

Code availability

See OSF repository mentioned above.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Pilot study

A.1. Methods

A.1.1. Participants

Two hundred participants located in Germany were recruited via Prolific (www.prolific.co). Two participants were excluded for indicating they did not take the experiment seriously, and eight were excluded for reporting inadequate German proficiency, as the experiment was conducted in the German language. In order to avoid grossly careless responding, we further excluded 33 participants based on their responses according to the following criterion: Participants were excluded if the sum of the probability estimates they provided for two constituent complementary events exceeded 125% in more than 1/4 of the event pairs they evaluated. Participants received £4.5 for their participation. The final sample included 157 participants (54 females, 102 males, and 1 other) with a mean age of 30.54 years (SD = 8.84).

A.1.2. Materials

The experiment was programmed in lab.js (Henninger et al., 2022). Using our intuitions about the probabilities of everyday events in Germany, we generated two hundred event pairs for selection, including 100 presumably mid-event and 100 presumably edge-event pairs. A complete list of event pairs can be found in the OSF repository (https://osf.io/hw8p9/). We used two different ways to construct the edge-event pairs: we generated 50 edge-event pairs with the probability of the constituent positive event close to 0, and 50 edge-event pairs with the probability of the constituent positive event close to 1. We took the frequentist probability approach to define the probability of the event with a clearly defined reference class to make our queries of probabilities as unambiguous as possible. Specifically, we queried about the occurrence and non-occurrence of the events that are regularly observed in everyday life in Germany, such as weather events. Table 3 shows examples of two edge-event pairs constructed in two different ways and a mid-event pair.

A.1.3. Design

Among the 200 event pairs, participants were asked to rate 50 of them, including a mix of edge-event and mid-event pairs. This enabled us to avoid long experiments and encourage participants to use the full probability scale. To this end, the generated event pairs were divided into four groups: The mid-event and edge-event pairs were divided equally (25 pairs per group). Moreover, we counter-balanced different ways of constructing the edge-event pairs. Among the four groups of event pairs, two groups have thirteen edge-event pairs with the probability of the constituent positive event A close to 1 and twelve edge-event pairs with the probability of the constituent positive event A close to 0. The other two have twelve edge-event pairs with the probability of the constituent positive event A close to 1 and thirteen edge-event pairs with the probability of the constituent positive event A close to 0.[9]

A.1.4. Procedure

Participants were randomly assigned to evaluate one of the four groups of events, each consisting of 50 event pairs. The experiment had two blocks. 25 of the 50 pairs had their positive events shown in block one and negative events shown in block two, and the other 25 pairs had their positive events shown in block two and negative events shown in block one. Therefore, participants were presented with an equal number of positive and negative events (i.e., 25 positive events and 25 negative events) in each block. The rationale for presenting event pairs in separate blocks is to prevent events from the same pair from being shown consecutively and to ensure independent judgments for each event. For example, this approach prevents participants from calculating the probability of one event based on its complementary event. The order of blocks, as well as the order of events within a block, was randomly determined for each participant.

A.2. Results

We created density plots of participants' probability estimates for the constituent positive and negative event, respectively, for each event pair using R package `ggridges` (Wilke, 2024). To allow the comparison of the probabilities of positive and negative events of the same pair, we subtracted the estimates provided by participants from 100% for the negative events. The density plots for all event pairs can be found in the OSF repository (https://osf.io/hw8p9/).

Fig. 4 shows the density plots of the event pairs that were finally selected. To match the realizations of the edge-event and mid-event pairs in the simulation, we selected the event pairs according to the following criteria:

1. In the two density plots for an event pair, most of their probability mass should fall between 0% and 30% or between 70% and 100% for edge-event pairs and between 30% and 70% for mid-event pairs.
2. The two density plots should be peaked and centered around a value close to 0% (or 100%) for the edge-event pairs and 50% for the mid-event pairs.
3. The two density plots should show a large overlap.

Criteria 1 and 2 were adopted to approximate the shapes of the distributions we used in the simulation for modeling people's underlying probabilities of events in two types of event pairs (i.e., Beta$(1, 10)$ and Beta$(10, 10)$ in edge-event and mid-event pairs respectively). Criterion 3 was adopted to identify event pairs in which the complementary relationship is evident to participants. Table A.1 provides the complete list of selected event pairs that meet these three criteria.

[9] Due to programming errors, we collected probability estimates only for the positive event for one event pair, thus, had to drop the data for this event pair. Additionally, four event pairs were mistakenly presented twice, leading to the collection of probability estimates for these event pairs from around 100 (instead of 50) participants.

Table A.1
Event pairs used for constructing event sets in Experiment 1 and Experiment 2.

Event pair type	Positive event	Negative event
Mid-event pair	A randomly selected person over the age of 30 in Germany is married.	A randomly selected person over the age of 30 in Germany is NOT married.
	A randomly selected person aged between 20 and 25 in Germany is studying at a university or college.	A randomly selected person aged between 20 and 25 in Germany is NOT studying at a university or college.
	A randomly selected person in Germany lives in Bavaria, Baden-Württemberg, or North Rhine-Westphalia.	A randomly selected person in Germany does NOT live in Bavaria, Baden-Württemberg, or North Rhine-Westphalia.
	On a randomly selected day in Hamburg, there will be rain.	On a randomly selected day in Hamburg, there will NOT be rain.
	A randomly selected person in Germany will eventually die of cardiovascular disease.	A randomly selected person in Germany will NOT eventually die of cardiovascular disease.
	A randomly selected person over the age of 18 in Germany has an office job.	A randomly selected person over the age of 18 in Germany does NOT have an office job.
	A randomly selected person in Germany lives in a big city.	A randomly selected person in Germany does NOT live in a big city.
	A randomly selected tree in Germany is a deciduous tree.	A randomly selected tree in Germany is NOT a deciduous tree.
	A randomly selected person in Germany is a fan of a football club.	A randomly selected person in Germany is NOT a fan of a football club.
	A randomly selected car on the road in Germany was manufactured in Germany.	A randomly selected car on the road in Germany was NOT manufactured in Germany.
	On a randomly selected day of the year, the temperature in Germany will be above 15 °C.	On a randomly selected day of the year, the temperature in Germany will NOT be above 15 °C.
	A randomly selected German is a member of the Christian church.	A randomly selected German is NOT a member of the Christian church.
Edge-event pair (in which positive event has a probability close to 1)	A randomly selected student speaks English.	A randomly selected student does NOT speak English.
	A randomly selected German adult can ride a bicycle.	A randomly selected German adult can NOT ride a bicycle.
	In a randomly selected German household, at least one washing machine can be found.	In a randomly selected German household, NO washing machines can be found.
	A randomly selected person in Germany walks more than 100 steps a day.	A randomly selected person in Germany does NOT walk more than 100 steps a day.
	A randomly selected person in Germany owns at least one device that can connect to the Internet.	A randomly selected person in Germany does NOT own a device that can connect to the Internet.
	A randomly selected person in Germany was born in a hospital.	A randomly selected person in Germany was NOT born in a hospital.
Edge-event pair (in which positive event has a probability close to 0)	A randomly selected person in Germany has more than five siblings.	A randomly selected person in Germany does NOT have more than five siblings.
	A randomly selected person in Germany plays volleyball every day.	A randomly selected person in Germany does NOT play volleyball every day.
	In a randomly selected year, it will snow in Germany in June.	In a randomly selected year, it will NOT snow in Germany in June.
	A randomly selected person in Germany can speak more than four languages.	A randomly selected person in Germany can NOT speak more than four languages.
	A randomly selected person in Germany lives in Saarland.	A randomly selected person in Germany does NOT live in Saarland.
	A randomly selected person in Germany will contract malaria in the course of their life.	A randomly selected person in Germany will NOT contract malaria in the course of their life.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cognition.2025.106125.

Availability of data and material

Data and scripts can be found on Open Science Framework (OSF): https://osf.io/hw8p9/.

References

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*(1), 57–86. http://dx.doi.org/10.3758/BF03210812.

Bott, F. M., Kellen, D., & Klauer, K. C. (2021). Normative accounts of illusory correlations. *Psychological Review*, *128*(5), 856–878. http://dx.doi.org/10.1037/rev0000273.

Chater, N., Zhu, J.-Q., Spicer, J., Sundh, J., León-Villagrá, P., & Sanborn, A. (2020). Probabilistic biases meet the Bayesian brain. *Current Directions in Psychological Science*, *29*(5), 506–512. http://dx.doi.org/10.1177/0963721420954801.

Costello, F., & Watts, D. (2014). Surprisingly rational: probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463–480. http://dx.doi.org/10.1037/a0037010.

Costello, F., & Watts, D. (2016). People's conditional probability judgments follow probability theory (plus noise). *Cognitive Psychology*, *89*, 106–133. http://dx.doi.org/10.1016/j.cogpsych.2016.06.006.

Costello, F., & Watts, D. (2017). Explaining high conjunction fallacy rates: the probability theory plus noise account. *Journal of Behavioral Decision Making*, *30*(2), 304–321. http://dx.doi.org/10.1002/bdm.1936.

Costello, F., & Watts, P. (2018). Invariants in probabilistic reasoning. *Cognitive Psychology*, *100*, 1–16. http://dx.doi.org/10.1016/j.cogpsych.2017.11.003.

Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive Psychology*, *96*, 1–25. http://dx.doi.org/10.1016/j.cogpsych.2017.05.001.

Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. L. (2013). Rational variability in children's causal inferences: The sampling hypothesis. *Cognition*, *126*(2), 285–300. http://dx.doi.org/10.1016/j.cognition.2012.10.010.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift Für Psychologie*, (Journal of Psychology) *217*(3), 108–124. http://dx.doi.org/10.1027/0044-3409.217.3.108.

Fox, C. R. (1999). Strength of evidence, judged probability, and choice under uncertainty. *Cognitive Psychology*, *38*(1), 167–189. http://dx.doi.org/10.1006/cogp.1998.0711.

Fox, J. (2015). *Applied Regression Analysis and Generalized Linear Models*. Sage.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229. http://dx.doi.org/10.1111/tops.12142.

Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? The amount of mental simulation tracks uncertainty in the outcome. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 37).

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An r package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*, *50*(1), 264–284. http://dx.doi.org/10.3758/s13428-017-0869-7.

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2022). Lab. js: A free, open, online study builder. *Behavior Research Methods*, *54*(2), 556–573. http://dx.doi.org/10.3758/s13428-019-01283-5.

Howe, R., & Costello, F. (2020). Random variation and systematic biases in probability estimation. *Cognitive Psychology*, *123*, Article 101306. http://dx.doi.org/10.1016/j.cogpsych.2020.101306.

Huang, J., Busemeyer, J., Ebelt, Z., & Pothos, E. (2024). Bridging the gap between subjective probability and probability judgments: the quantum sequential sampler. *Psychological Review*, http://dx.doi.org/10.1037/rev0000489.

Icard, T. (2016). Subjective probability as sampling propensity. *Review of Philosophy and Psychology*, *7*, 863–903. http://dx.doi.org/10.1007/s13164-015-0283-y.

Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review*, *116*(4), 856–874. http://dx.doi.org/10.1037/a0016979.

Kellen, D. (2019). A model hierarchy for psychological science. *Computational Brain & Behavior*, *2*, 160–165. http://dx.doi.org/10.1007/s42113-019-00037-y.

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, *75*(1), 70–98. http://dx.doi.org/10.1007/s11336-009-9141-0.

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. http://dx.doi.org/10.1037/rev0000074.

Lieder, F., Griffiths, T. L., M. Huys, Q. J., & Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, *25*, 322–349. http://dx.doi.org/10.3758/s13423-017-1286-8.

Lloyd, K., Sanborn, A., Leslie, D., & Lewandowsky, S. (2019). Why higher working memory capacity may help you learn: sampling, search, and degrees of approximation. *Cognitive Science*, *43*(12), Article e12805. http://dx.doi.org/10.1111/cogs.12805.

Meiser, T. (2011). Much pain, little gain? Paradigm-specific models and methods in experimental psychology. *Perspectives on Psychological Science*, *6*(2), 183–191. http://dx.doi.org/10.1177/1745691611400241.

Nilsson, H., Winman, A., Juslin, P., & Hansson, G. (2009). Linda is not a bearded lady: Configural weighting and adding as the cause of extension errors. *Journal of Experimental Psychology: General*, *138*(4), 517–534. http://dx.doi.org/10.1037/a0017351.

R Core Team (2024). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, URL https://www.R-project.org/.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*(3), 318–339. http://dx.doi.org/10.1037/0033-295X.95.3.318.

Sanborn, A., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893. http://dx.doi.org/10.1016/j.tics.2016.10.003.

Schmidt, O., Erdfelder, E., & Heck, D. W. (2023). How to develop, test, and extend multinomial processing tree models: A tutorial. *Psychological Methods*, http://dx.doi.org/10.1037/met0000561.

Singmann, H., Heck, D. W., Barth, M., Erdfelder, E., Arnold, N. R., Aust, F., Calanchini, J., Gümüsdagli, F. E., Horn, S. S., Kellen, D., Klauer, K. C., Matzke, D., Meissner, F., Michalkiewicz, M., Schaper, M. L., Stahl, C., Kuhlmann, B. G., & Groß, J. (2024). Evaluating the robustness of parameter estimates in cognitive models: A meta-analytic review of multinomial processing tree models across the multiverse of estimation methods. *Psychological Bulletin*, *150*(8), 965. http://dx.doi.org/10.1037/bul0000434.

Sun, Y., Wang, H., Zhang, J., & Smith, J. W. (2008). Probabilistic judgment on a coarser scale. *Cognitive Systems Research*, *9*(3), 161–172. http://dx.doi.org/10.1016/j.cogsys.2007.03.001.

Sundh, J., Zhu, J.-Q., Chater, N., & Sanborn, A. (2023). A unified explanation of variability and bias in human probability judgments: How computational noise explains the mean–variance signature. *Journal of Experimental Psychology: General*, *152*(10), 2842–2860. http://dx.doi.org/10.1037/xge0001414.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131. http://dx.doi.org/10.1126/science.185.4157.1124.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293. http://dx.doi.org/10.1037/0033-295X.90.4.293.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547. http://dx.doi.org/10.1037/0033-295X.101.4.547.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637. http://dx.doi.org/10.1111/cogs.12101.

Wedell, D. H., & Moro, R. (2008). Testing boundary conditions for the conjunction fallacy: Effects of response mode, conceptual focus, and problem type. *Cognition*, *107*(1), 105–136. http://dx.doi.org/10.1016/j.cognition.2007.08.003.

Wilke, C. O. (2024). ggridges: ridgeline plots in 'ggplot2'. R Package Version 0.5.6 URL https://CRAN.R-project.org/package=ggridges.

Zhu, J.-Q., Newall, P. W., Sundh, J., Chater, N., & Sanborn, A. N. (2022). Clarifying the relationship between coherence and accuracy in probability judgments. *Cognition*, *223*, Article 105022. http://dx.doi.org/10.1016/j.cognition.2022.105022.

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The Bayesian sampler: Generic Bayesian inference causes incoherence in human probability judgments. *Psychological Review*, *127*(5), 719–748. http://dx.doi.org/10.1037/rev0000190.