Human Probability Judgments: A Bayesian Sequential Sampler Account

Zeping Li

Department of Psychology, University of Warwick

Target Journal: Psychological Review

Author Note

I have no conflicts of interest to disclose.

I thank Adam Sanborn, Nick Chater, Jianqiao Zhu, Joakim Sundh and Jake Spicer for their guidance.

Correspondence concerning this article should be addressed to Zeping Li, Department of Psychology, University of Warwick, Coventry, CV4 7AL, United Kingdom. Email: Zeping.Li@warwick.ac.uk

Abstract

Sampling-based models have successfully explained many phenomena of human cognition, but how to generate multidimensional continuous responses and when to stop sampling are still underexplored fields. I proposed a new kind of models, Bayesian sequential sampler models for human probability judgments and tested three stopping rules: fixed sample size, fixed time and fixed density in different conditions. The models relatively accurately reproduced judgments and response times in two probability judgment datasets. According to the models, people mainly used the fixed time rule in general situations. They switched to the fixed sample size rule and decreased the strength to generate samples and the non-judgment time under time pressure. When the judgment task became difficult, the fixed time and the fixed sample size rules were equally likely to be used, and the total sample size and the non-judgment time decreased.

Keywords: sequential sampling models, Bayesian sampler models, probability judgments, stopping rules

Human Probability Judgments: A Bayesian Sequential Sampler Account

Living in this highly complex and ever-changing world, people need to make judgments and decisions based on noisy and ambiguous information every day. Sometimes they give quantitative responses, like estimating how likely someone is to get the coronavirus disease if he or she has a continuous cough. Sometimes they make simple choices, like deciding whether to step over or get around a barrier on the road. To deal with the ubiquitous uncertainties in these problems, probabilistic cognition seems a natural solution (Griffiths et al., 2010; Sanborn, 2017). However, a subsequent puzzle is how the brain can conduct exact calculations with probabilities given its physical limits and the fact that some calculations are difficult even for machines. One possible answer is that the brain represents probabilities by samples rather than numeric values (Chater et al., 2006; Chater et al., 2020; Vul et al., 2014). There are two kinds of sampling-based models which have successfully explained many phenomena in human judgments and decisions, sequential sampling models and Bayesian sampler models.

Sequential Sampling Models

Though sequential sampling models (SSMs), or evidence accumulation models were proposed more than four decades ago (e.g., Ratcliff, 1978), they have been widely accepted in the last two decades due to the development of computational techniques and the discovery of similar patterns in the brain (e.g., Shadlen & Newsome, 1996; Wagenmakers et al., 2007). The SSMs are initially designed to model dichotomous perceptual choices like indicating whether a string

presented on the screen is a meaningful or non-sense word (Rubenstein et al., 1970). In relevant tasks, the participants can usually continuously collect sensory information from the environment (e.g., gazing at the presented strings) and indicate their inferences by choosing one of the two responses (e.g., choosing “yes” or “no”). Most SSMs decompose such behaviour into two processes, a decision process for sample accumulation and a non-decision process for all other cognitive activities.

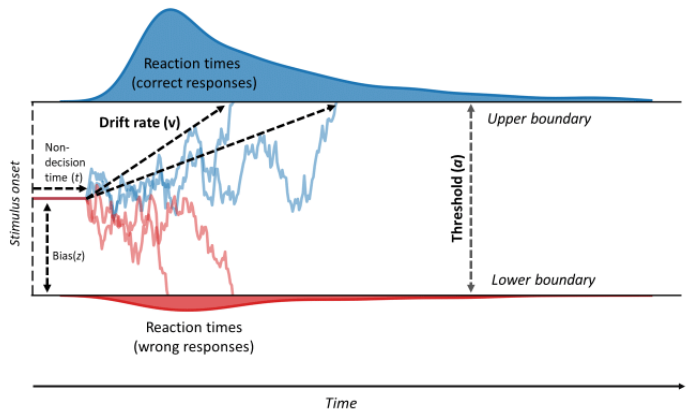
There is a wide range of SSMs and two major families are diffusion decision models (DDMs) and accumulator models (AMs; Ratcliff et al., 2016). The DDMs use a diffusion process to describe the choices. The process starts with a position within two boundaries and each boundary represents a response. The evidence, or samples are constantly drawn from a normal distribution and added to the current position (i.e., the sample value at each moment can be seen as the instantaneous velocity). The mean and the standard deviation of the normal distribution are called drift rate and within-trial variability respectively. The diffusion is terminated once a boundary is reached, and the corresponding response is chosen. The diffusion time and the non-decision time constitute the recorded response time (RT). Figure 1 illustrates the two processes.

The parameters of the DDMs describe different aspects of perceptual choices. The drift rate reflects task difficulty or participant ability. When the upper boundary represents the correct response, a participant will make more and faster correct choices with a higher drift rate. The boundary separation (i.e., the distance between two boundaries) reflects the level of caution. Higher boundary separation leads to

more correct choices but longer RTs. The starting position measures the prior bias. A participant has no bias if the starting position is the midpoint between the two boundaries. But the corresponding choice will be more and faster if the starting position is closer to one boundary. The non-decision time represents the time for any other peripheral activities like stimulus encoding and motor execution. The change of the non-decision time only changes the position but not the shape of the RT distribution. The within-trial variability is a scaling parameter. When it is multiplied by a constant *k*, the model predictions will be unchanged if other parameters relevant to the diffusion process are also multiplied by *k*. Besides, to explain faster and slower error choices and the variability of the .1 quantiles of the RT distribution, DDMs also introduce the inter-trial variability in the starting position, the drift rate and the non-decision time (Ratcliff & McKoon, 2008; Wagenmakers, 2009).

Figure 1

Illustration of a Diffusion Decision Model



Note. From "Volition in Prospective Memory: Evidence Against Differences Between Free and Fixed Target Events," by M. C. Vinding, J. K. Lindeløv, Y. Xiao, R. C. Chan

and T. A. Sørensen, 2021, *Consciousness and Cognition, 94*, p. 3 (https://doi.org/10.1016/j.concog.2021.103175). Copyright 2021 by Elsevier.

The DDMs assume an integrated accumulator collecting relative evidence between two responses, but the AMs assume independent accumulators for different responses and a stopping rule based on absolute evidence. Take the linear ballistic accumulator model (LBAM; Brown & Heathcote, 2008) as an example. It assumes that each response has its own accumulator but shares a common threshold. Each accumulator has a fixed evidence accumulation speed (which is also called drift rate) and a starting position once a decision process starts. The response whose accumulated evidence reaches the threshold first will be chosen. Because the drift rates are fixed during a single decision process, the choice and the RT on each trial are already decided when the drift rates and the starting points are decided. The stochasticity of the choices and the RTs are caused by the inter-trial variability in the drift rate and the starting point for each response. From a sampling perspective, a participant only draws one sample from the distribution of the drift rate for each response and repeatedly uses it to make a final choice. Compared to the DDMs, a natural advantage of the AMs is that they can be applied to tasks with more than two possible responses.

The SSMs have been widely used in different perceptual tasks like lexical decisions (Wagenmakers et al., 2008), visual signal detection (Smith & Ratcliff, 2009) and absolute identification (Brown et al., 2008). And the pattern of sequential sampling has been observed in both low-level activity recordings in animal

experiments (e.g., Roitman & Shadlen, 2002) and high-level region measurements in human experiments (e.g., Summerfield & Koechlin, 2010), which supports the biological feasibility of sampling. Besides, the SSMs have also been successfully extended to value-based choices (Bakkour et al., 2019) and experience-based choices (Fontanesi et al., 2019) recently, which implies a common mechanism of human choices based on sampling.

However, the SSMs are not the sampling models in the general sense. The sampling object is abstract and dimensionless "evidence". And the samples are not used to infer a potential distribution. They are only accumulated until an absorbing state is reached. These properties make the SSMs hard to understand. Besides, though there have been some attempts to apply the SSMs to tasks with continuous responses (Kvam & Turner, 2021; Ratcliff, 2018; Smith, 2016), how to generate responses at a continuous scale from a sequential sampling process is still an underdeveloped field.

Bayesian Sampler Models

Strictly speaking, there are two kinds of Bayesian sampler models (BSMs). The first kind of BSMs focuses on how people update their beliefs about the world based on explicitly given information. They presume that the updating process follows the Bayesian rule and the sampling is used for exploring the posterior distribution. Because of the physical limits of the brain and the heavy computational cost in most Bayesian inference problems, such models assume that people adopt autocorrelated sampling methods like Markov chain Monte Carlo (MCMC; Geman & Geman, 1984) and sample sizes are usually small. The BSMs with these assumptions

have successfully explained many cognitive biases like the unpacking effects (Dasgupta et al., 2017) and the anchoring effect (Lieder et al., 2017).

The second kind of BSMs focuses on how people get the "observed data" and they suggest sampling as a solution. They presume that people can self-generate samples by mental simulation or memory retrieval and make inferences based on them. For similar reasons, they also assume autocorrelated sampling with small sizes. And because the inferences based on limited samples can be highly distorted, they believe that people will adjust them with prior beliefs. Such models have also reconciled many behavioural patterns with rational Bayesian cognition like the conjunction fallacy (Zhu et al., 2020) and the speed-accuracy tradeoff in perceptual choices (Zhu, Sundh, et al., 2021). The models proposed in this article are mainly relevant to this kind of BSMs, more specifically, the BSM for human probability judgments proposed by Zhu et al. (2020).

According to this model, a judgment maker (JM) use a symmetric Beta distribution Beta(β, β) as the prior for probability judgments. When judging the probability of an event A, the JM will self-generate a set of samples containing As and $\neg A$s (not A) based on a true probability $P(A)$ and combine them with the prior. The mean of the posterior is chosen as the judgment. Due to the favourable statistical property of the Beta distribution, the judgment is a linear transformation of the frequency of A. Specifically, suppose the sample size is N and there are N_A samples are A, the judgment is

$$\hat{P}(A) = \frac{N_A}{N+2\beta} + \frac{\beta}{N+2\beta}.$$

If the JM repeatedly judges the probability, the expected value of the judgments is $E(\hat{P}(A)) = \frac{NP(A)}{N+2\beta} + \frac{\beta}{N+2\beta}$.

Let $d = \frac{\beta}{N+2\beta}(0 < d < 0.5)$, the equation can be rewritten as

$$E\left(\hat{P}(A)\right) = (1-2d)P(A) + d = (1-d)P(A) + d\big(1-P(A)\big),$$

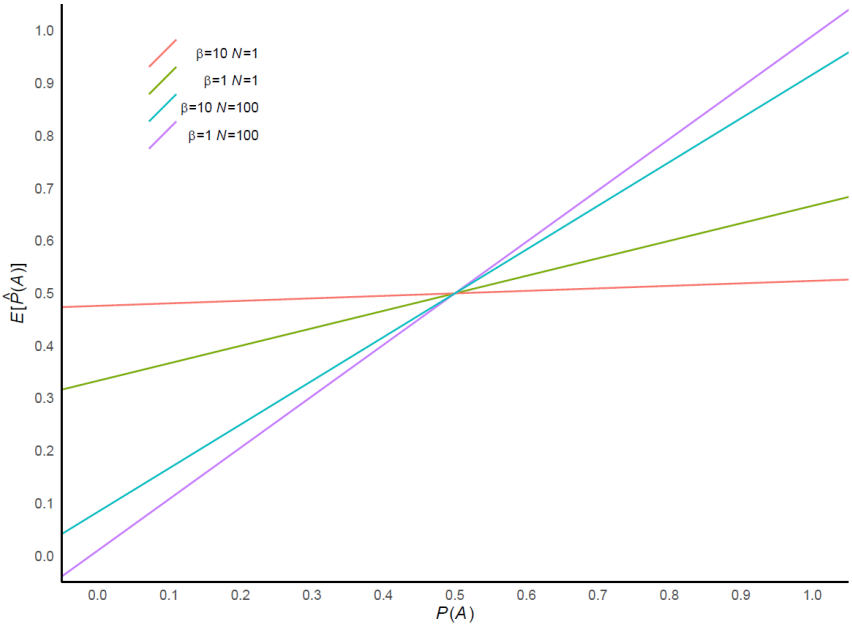
which implies that the expected value of $\hat{P}(A)$ is a weighted mean of $P(A)$ and $P(\neg A)$. Thus, when A is a low probability event, $\neg A$ is a high probability event, which will pull up the expected value of $\hat{P}(A)$. Similarly, the expected value of $\hat{P}(A)$ will be pulled down by $P(\neg A)$ when A is a high probability event. This pattern (i.e., overestimation of small probabilities and underestimation of high probabilities), which is called conservatism, has already been widely observed in human probability judgments (Erev et al., 1994; Hilbert, 2012; Kaufman et al., 1949). And it should be more obvious if N is smaller or β is larger according to the BSM (cf. Figure 2).

The BSM can also capture another cognitive bias, the conjunction fallacy (Tversky & Kahneman, 1983). It means that the mean probability judgment of a conjunction (e.g., a random word selected from a novel ends with "ing") can be higher than that of one of its parent events (e.g., a random word selected from a novel ends with "_n_"). The BSM explains it by assuming that the sampling of conjunctions is more computationally demanding. So the sample sizes are usually smaller than those for the parent events, which enhances the deviation of the mean judgments from the true probabilities. If both the conjunction and one of its parent events occur with a probability lower than 0.5, the adjusted probability judgment of the conjunction can be higher than that of the parent event if the difference in sample

size is large enough.

Figure 2

Illustration of Model Behaviours for the Bayesian Sampler Model



Note. Adapted from "The Bayesian Sampler: Generic Bayesian Inference Causes Incoherence in Human Probability Judgments," by J-Q. Zhu, A. N. Sanborn and N. Chater, 2020, *Psychological Review, 127*(5), p. 726 (https://doi.org/10.1037/rev0000190). Copyright 2020 by the American Psychological Association.

Though the BSM has successfully reproduced many patterns of human probability judgments, it only considers judgments, not RTs. As an important measurement in behavioural science, RTs reflect features of many underlying cognitive processes (e.g., how fast the samples are generated and how many samples are generated). Model fitting using both judgments and RTs should help to give more accurate predictions of human probability judgments. Besides, a default assumption

of the current BSM is that people use a fixed sample size when they give repeated probability judgments. There is no attempt to test whether people will stop sampling based on other rules.

Bayesian Sequential Sampler Models

To predict both judgments and RTs, I propose a new kind of models, Bayesian sequential sampler models (BSSMs).

The BSSMs assume that the continuous RTs can be broken into discrete time steps[1]. Some time steps are used for the non-judgment process and the rest are used for the judgment process. When judging the probability of an event A, the judgment process can be described as follows.

1. At time step t ($t >= 1$), the JM first has a belief about $P(A)$, which follows a Beta distribution Beta(α_t, β_t). When $t = 1$, $\alpha_1 = \beta_1 = \beta$. β ($\beta > 0$) is the prior parameter, which represents how strong the JM's prior belief is.
2. The JM draws a set of samples containing A and $\neg A$. Denote the sample size by N_t. N_t follows a Poisson distribution Pois(λ). λ ($\lambda > 0$) is the strength parameter, which represents the average size of samples the JM can generate at a single time step. Denote the numbers of A in this sample set by N_{t_A}. N_{t_A} follows a binomial

[1] The discretization of RTs is not a new idea in the SSMs (e.g., Audley & Pike, 1965). It decreases the granularity but increases the stability of the predicted RTs, which also makes the model fitting easier. And the representation of RTs will not change much if the time step length is small enough because the actual measurements also have a minimum unit. Besides, if the interval of samples follows an Exponential distribution Exp(λ), the number of samples collected within time length T follows a Poisson distribution Pois(T/λ), and the time to collect N samples follows a Gamma distribution Gamma(N, λ). Thus, using a Poisson distribution to describe the sample size at each time step is equivalent to using a Gamma distribution to describe the continuous distribution of RTs to some extent.

distribution Binom(N_t, p). p ($p > 0$) is the true probability parameter, which represents the actual subjective probability of A.

3. The JM update the belief about $P(A)$ by: $\alpha_{t+1} = \alpha_t + N_{t\_A}$, $\beta_t = \beta_t + N_t - N_{t\_A}$. Then the JM will check whether the accumulated samples meet a specific stopping rule. If the stopping rule is met, the JM will stop sampling and make a judgment. Otherwise, the JM will enter the next time step with the belief Beta(α_{t+1}, β_{t+1}).

Denote the final time step by t_{final}. The JM's judgment should be $\frac{\alpha_{t_{final}+1}}{\alpha_{t_{final}+1}+\beta_{t_{final}+1}}$. And the RT should be the sum of the time step number for the judgment process and that for the non-judgment process (denote it by T_{non}, $T_{non} > 0$), multiplied by the time step length[2].

There are three stopping rules to be tested: fixed sample size, fixed time and fixed density rules. The fixed sample size rule assumes that the JM will check the size of the accumulated samples after each time step and stop once it reaches a threshold d_{size} ($d_{size} > 0$). The fixed time rule assumes that the JM will check how many time steps (for the judgment process) having elapsed and stop once it reaches a threshold d_{time} ($d_{time} > 0$). The fixed density rule assumes that the JM will check the probability density of the posterior mean after each time step and stop once it reaches a threshold $d_{density}$ ($d_{density} > 0$). To explain the variability of the RTs, the time step

[2] The predicted RTs of the BSSMs are inherently discrete. Probably the simplest way to make it continuous is to add a uniformly distributed noise whose mean is 0 and span is the time step length. But I do not add it because the discrete time step numbers are enough to represent RTs and introducing extra random variables can decrease the stability of the model.

number of the non-judgment process is a random variable for the fixed time BSSM[3]. Specifically, the non-judgment time step number is drawn from a shifted Poisson distribution where $P(\text{non-judgment time step number} = T) = \text{Pois}(T - 1|T_{non})$. In other words, parameter T_{non} represents the average non-judgment time step number minus one, rather than a fixed non-judgment time step number for this model (see Appendix A for the reason to choose the shifted Poisson distribution).

Model Features

To explore how predicted judgments and RTs change with different combinations of parameter values, I first conducted a simulation test for each BSSM. The possible values of β, λ, p and T_{non} were the same across three models. Specifically, the possible values of β were 0.05, 0.27 (i.e., empirical prior Zhu et al. (2020) estimated), 1 (i.e., uniform distribution) and 5 (for simplicity, I only showed the results of β = 0.05 and 5 later). The possible values of λ were 2, 5 and 10. The possible values of T_{non} were 5, 15 and 40. The possible values of p were an arithmetic sequence ranging from 0 to 1 with step size 0.1. The possible values of d_{size} were 10, 30 and 50. The possible values of d_{time} were 5, 10 and 20. The possible values of $d_{density}$ were 2, 6 and 10. The possible values of the threshold parameter differed because their meanings were different. But the range of the equivalent average total sample size was similar across three models, which ranged from several to hundreds. Thus, I compared model predictions generated from 1188 combinations of parameter

[3] For the other two BSSMs, I simply assume that the randomness of the sample size at each time step and the samples themselves is the only source of the RT variability. But there is no doubt that the non-judgement time step number should be a random variable.

values for each model, which ensured that I could fully explore the influence of a single parameter or the interaction of parameters on predicted judgments and RTs. For each combination of parameter values, I generated 5000 simulated observations to guarantee the stability of the calculated statistics.

Conservatism and Anti-Conservatism. Mean judgments as a function of true probabilities were illustrated in Figure 3. For the fixed sample size and the fixed time models, the model predictions were similar to those of the original BSM. When $\beta = 0.05$, which represented nearly no prior belief, the mean judgments precisely reproduced the underlying true probabilities given enough observations. And when $\beta = 5$, which represented a strong prior belief, an obvious pattern of conservatism could be found, and the degree of it was mitigated when the JM collected more samples (i.e., a larger threshold). Besides, a larger strength to generate samples could also mitigate conservatism to some extent, especially when the threshold was small. This was because the JM would check the total sample size after each time step rather than after collecting each sample for the fixed sample size model, which led to a larger difference between the average total sample size and the threshold when the strength was larger. And for the fixed time model, the average total sample size was the product of the strength and the threshold, which ensured the same influences of the strength and the threshold. This implied that the change in the strength had a larger effect on the average total sample size for the fixed time model compared to that for the fixed sample size model because the change would be accumulated at each time step. As for why the influence of the strength was smaller for each model when the

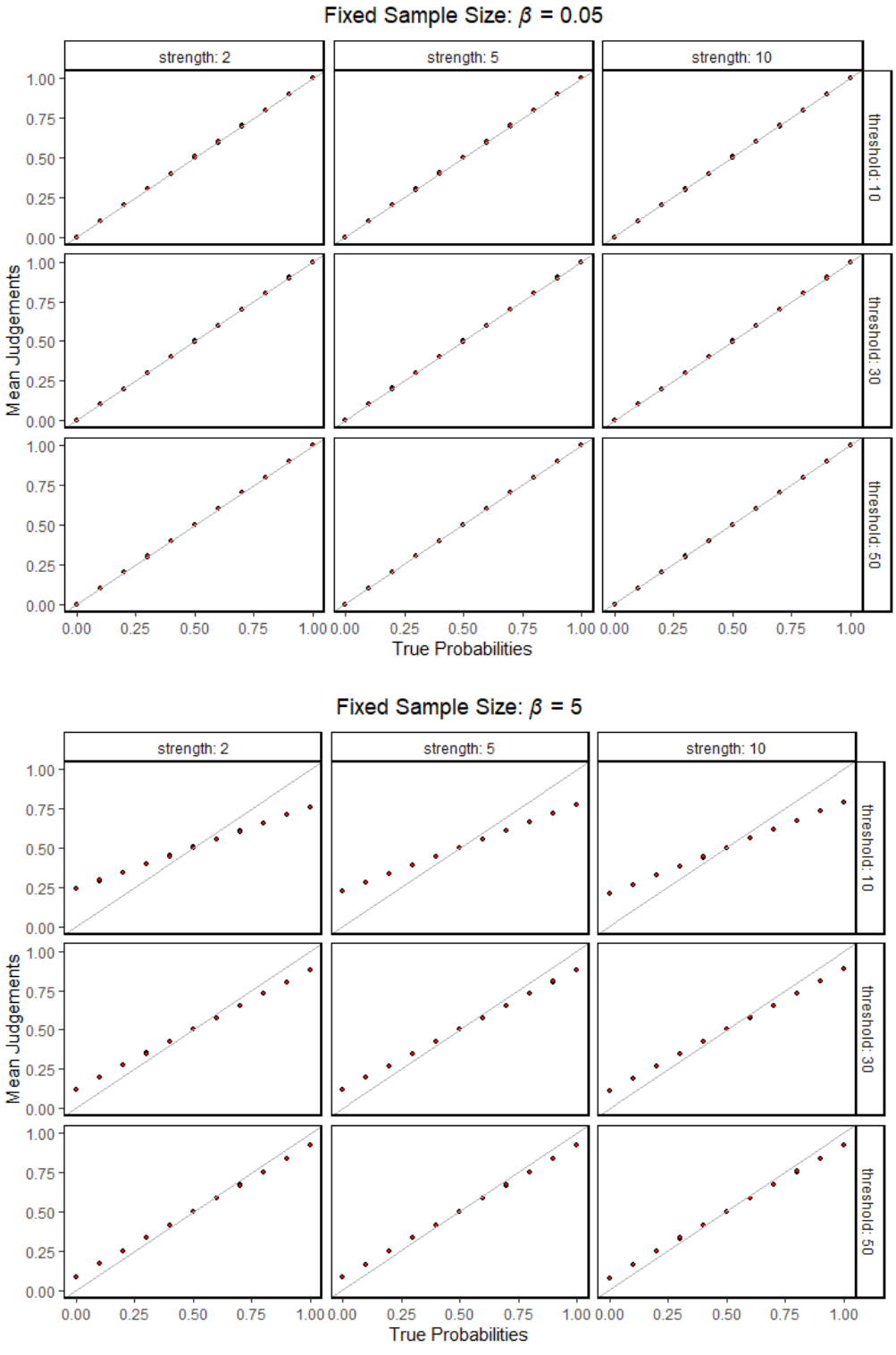
threshold was large. This was because a large threshold guaranteed a large minimum total sample size, which should mitigate conservatism enough.

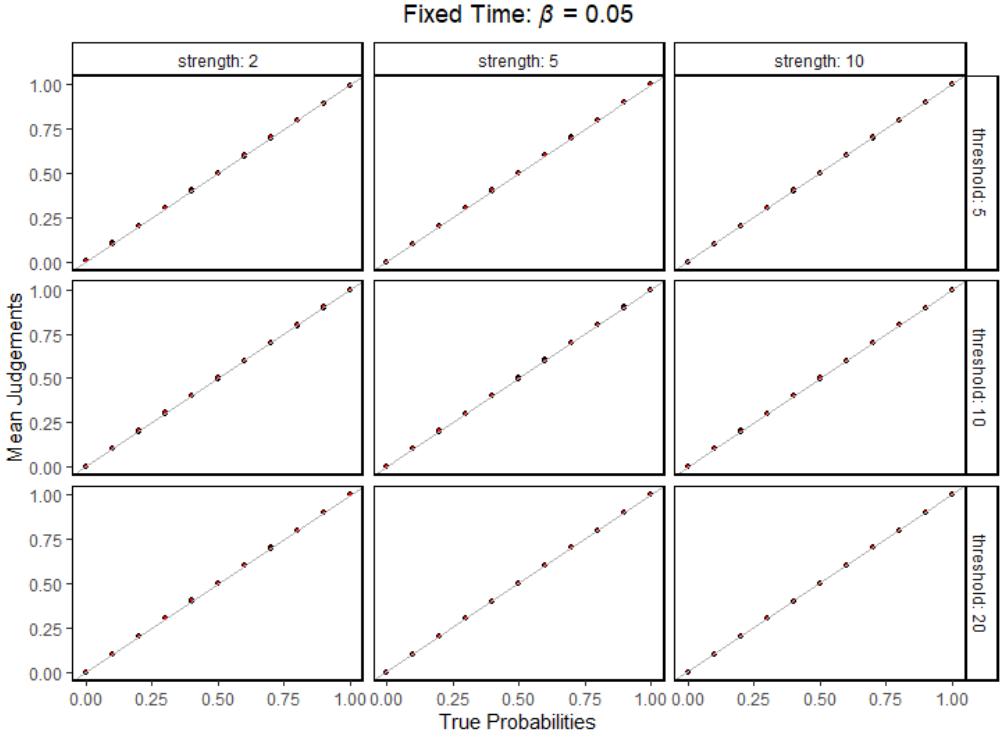
When either the prior, the strength or the threshold was large, the model predictions of the fixed density BSSM were similar to those of the fixed sample size and the fixed time BSSMs. However, when the three parameters were small, a strange pattern was observed, i.e., underestimation of small probabilities and overestimation of high probabilities, which was called anti-conservatism in this article. The pattern was caused by two facts. The first was that the Binomial distribution Binom(N, p) which represented the success times of an event with a probability of p in N observations was highly positively skewed when N and p were both small. For example, suppose a JM drew 10 samples about an event happening with a probability of 0.1. The probability that this JM never seeing this event happening was about 0.35. Thus, the actual relative frequency of a rare event was very likely to be lower than its true probability when the sample size was small. On the contrary, the actual relative frequency of a common event was very likely to be higher than its true probability when the sample size was small. But this was not enough to explain anti-conservatism if enough judgments from the JM had been averaged. Another fact was that a fixed density rule user was easier to stop sampling when the samples were extreme. Specifically, for a Beta distribution Beta(α, β), the probability density of the mean was higher when the absolute difference between α and β was larger if $\alpha + \beta$ was fixed (see Appendix B for a mathematical proof). It meant that a fixed density rule user usually needed a smaller number of samples to

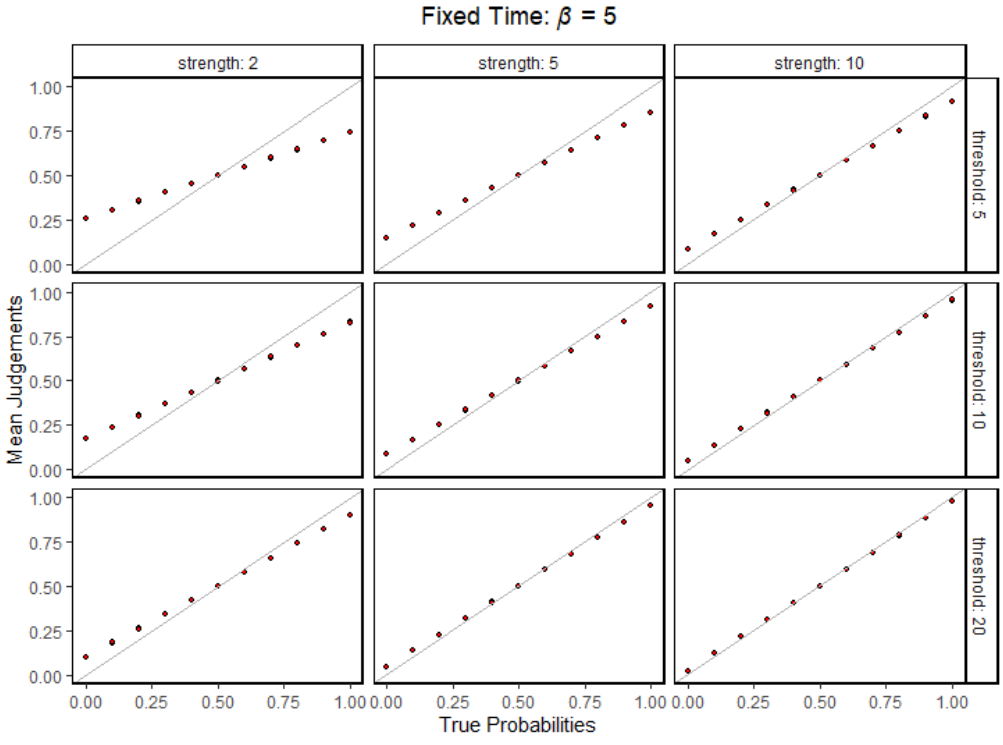
make a judgment when the true probabilities were extreme (which could cause a large difference between α and β easily because the samples mainly consisted of one outcome) compared to when the true probabilities were medium. Thus, a JM who could have seen some rare outcomes (i.e., the occurrence of a rare event or the nonoccurrence of a common event) might stop too early before seeing them, which made the influence of the high skewness of success times based on limited samples not be compensated with enough judgments. And the pattern of anti-conservatism was observed. It was obvious that the occurrence of anti-conservatism needed two conditions. One was that the judgments should be mainly based on the samples, which required a small prior. The other was that the total sample size used for the judgments should be small, which required a small strength and threshold. This explained why anti-conservatism could only be observed when all three parameters were small.

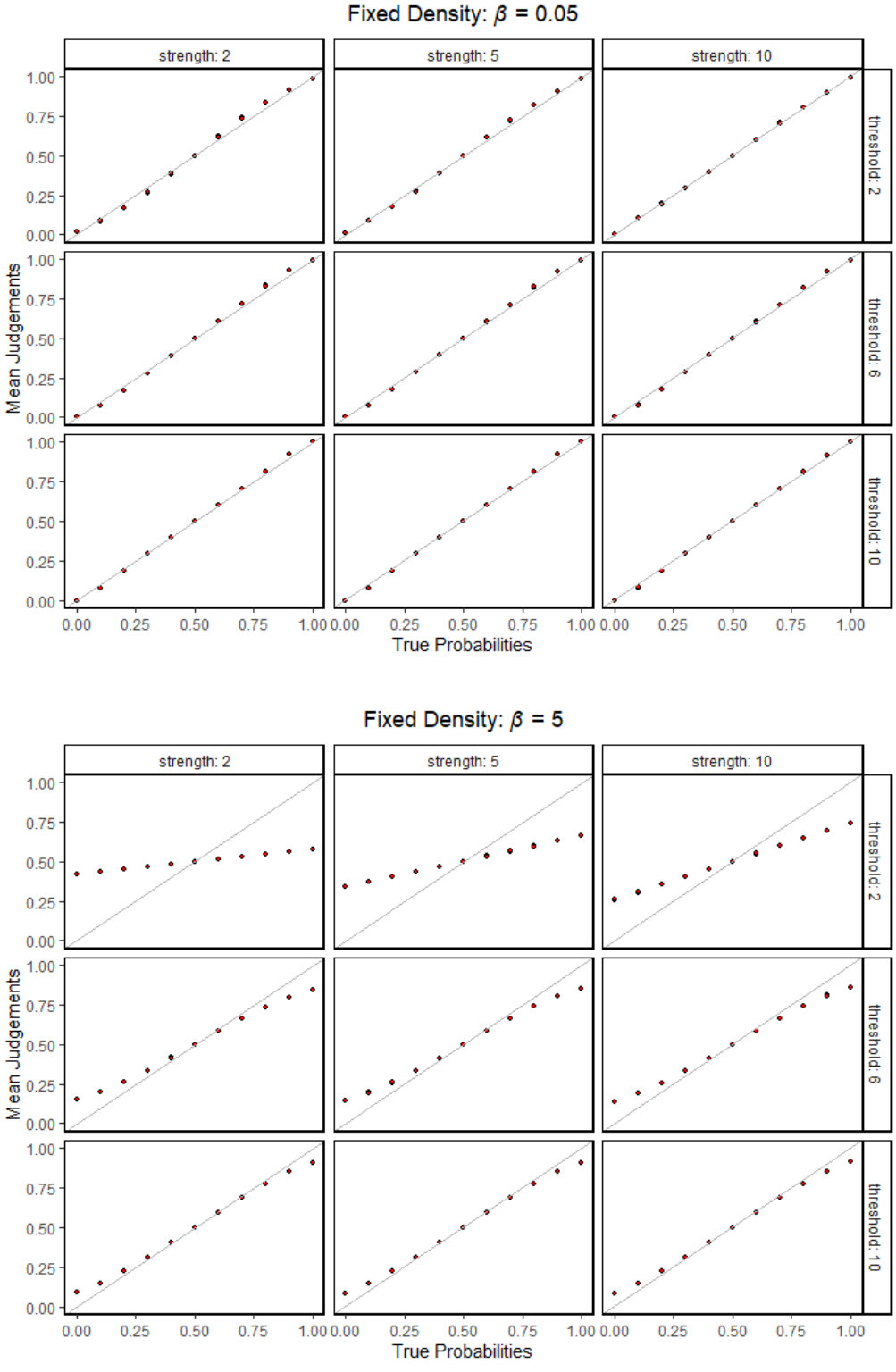
Figure 3

Mean Judgments as a function of True probabilities for Different Models Under Different Combinations of Parameter Values




Fixed Time: β = 0.05
strength: 2
strength: 5
strength: 10
threshold: 5
threshold: 10
threshold: 20
Mean Judgements
True Probabilities
0.00
0.25
0.50
0.75
1.00


Fixed Time: β = 5
strength: 2
strength: 5
strength: 10
threshold: 5
threshold: 10
threshold: 20
Mean Judgements
True Probabilities
0.00
0.25
0.50
0.75
1.00



Note. The grey lines are identity lines.

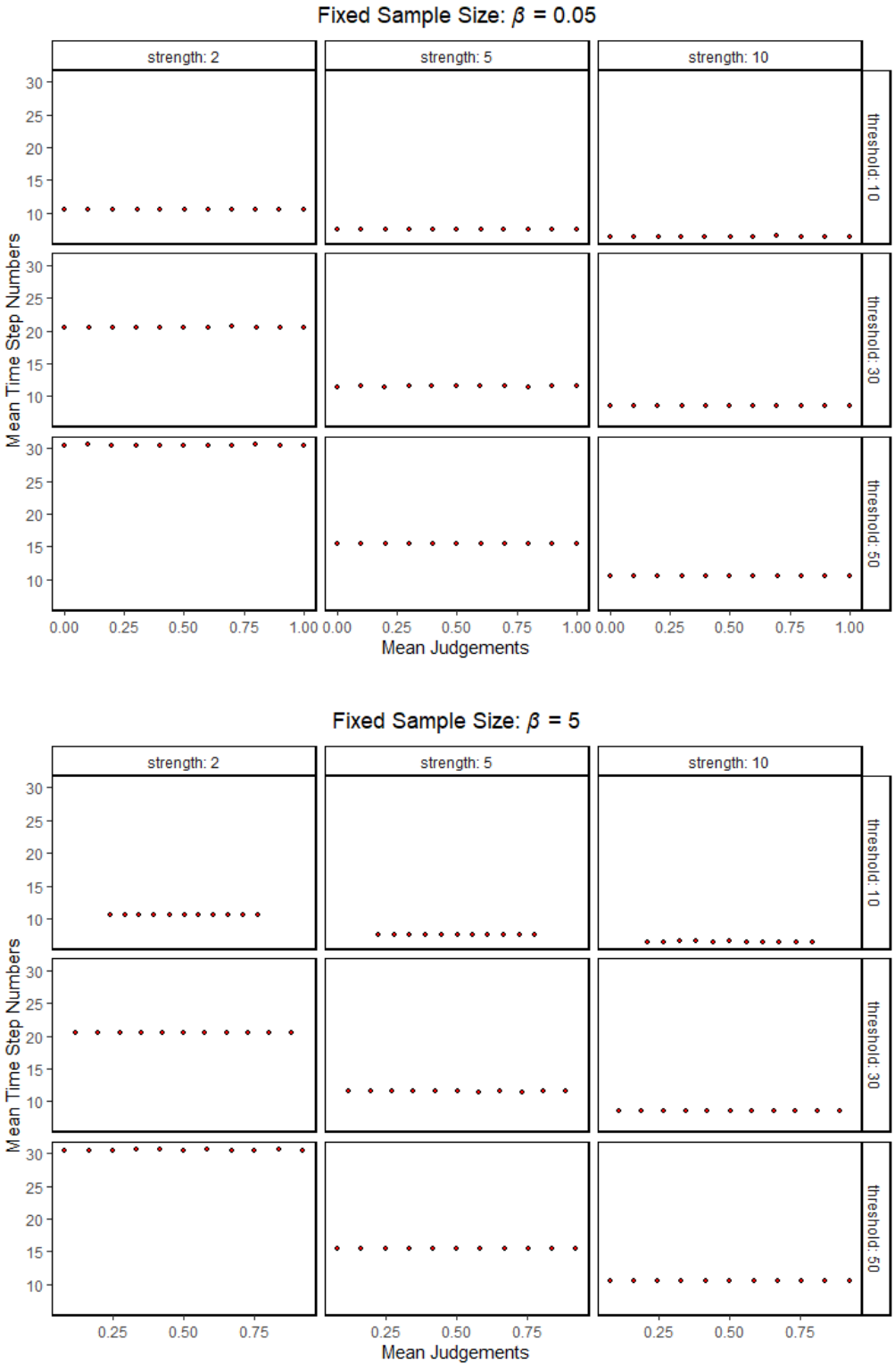
Though conservatism was a prevalent phenomenon at the group level, the existence of anti-conservatism at the individual level was also widely reported (Khaw

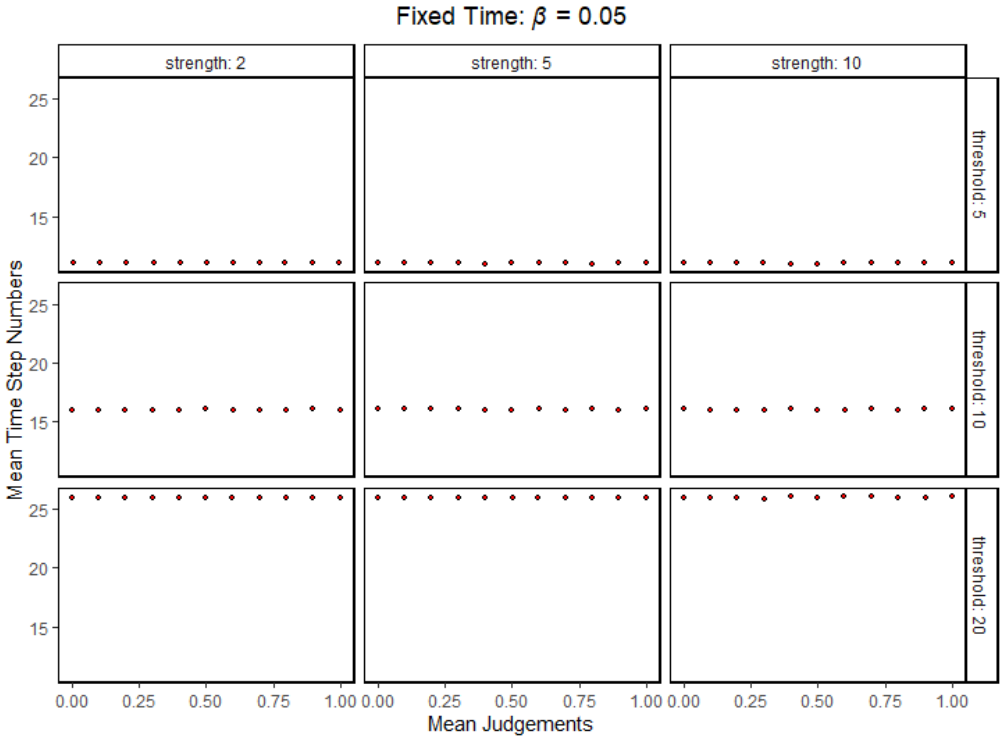
et al., 2021; Zhang et al., 2020). It seemed that whether a JM showed anti-conservatism could help to distinguish the stopping rule. But unfortunately, it was also easy to find the pattern of anti-conservatism or some mixed patterns if there were not enough judgments gathered from the JM even if the fixed sample size or the fixed time rule was used (see Appendix B). An easier way to distinguish the rule a JM used was to check the relation between the judgments and the RTs.

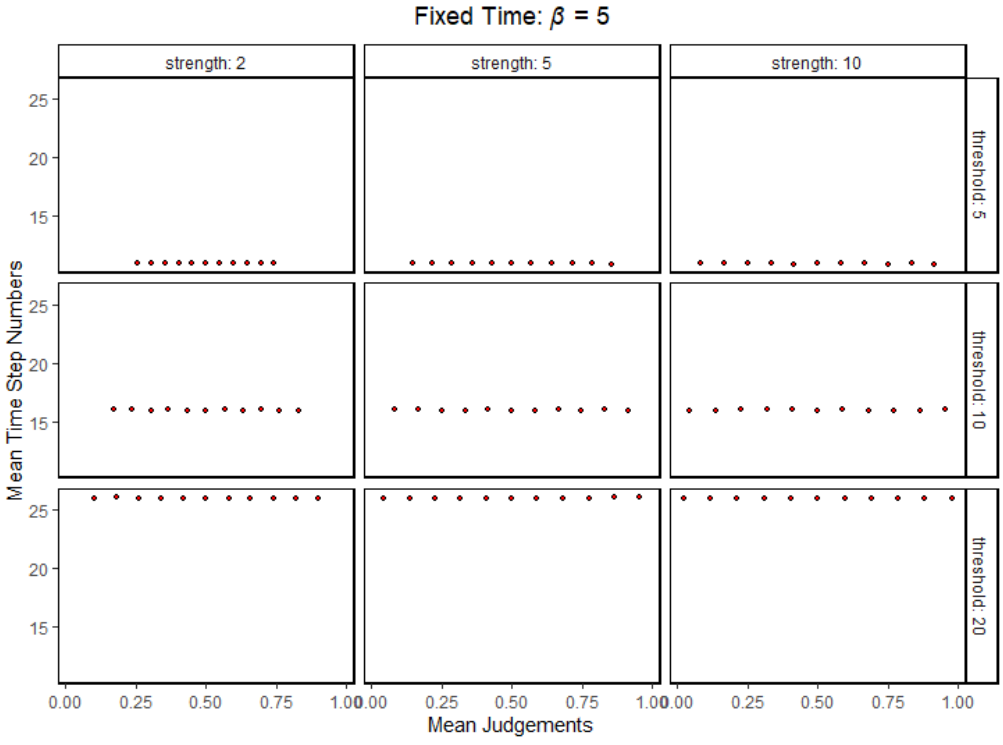
Relation Between the Judgments and the Response Times. Mean total time step numbers (T_{non} was fixed to 5 because it was irrelevant) as a function of mean judgments were illustrated in Figure 4. For the fixed sample size and the fixed time models, there was no obvious relation between the mean total time step numbers (i.e., RTs) and the mean judgments, which could be regarded as a linear transformation of the true probabilities. In other words, the JM used similar times to give judgments regardless of the true probabilities. Meanwhile, for the fixed sample size model, the mean RTs were influenced by the strength and the threshold. A larger threshold and a smaller strength could both make the mean RTs increase. For the fixed time model, the mean RTs were only influenced by the threshold. A larger threshold could increase the mean RTs. This was because the prior and the true probability, no matter how extreme they were, would not influence when to stop according to these two rules.

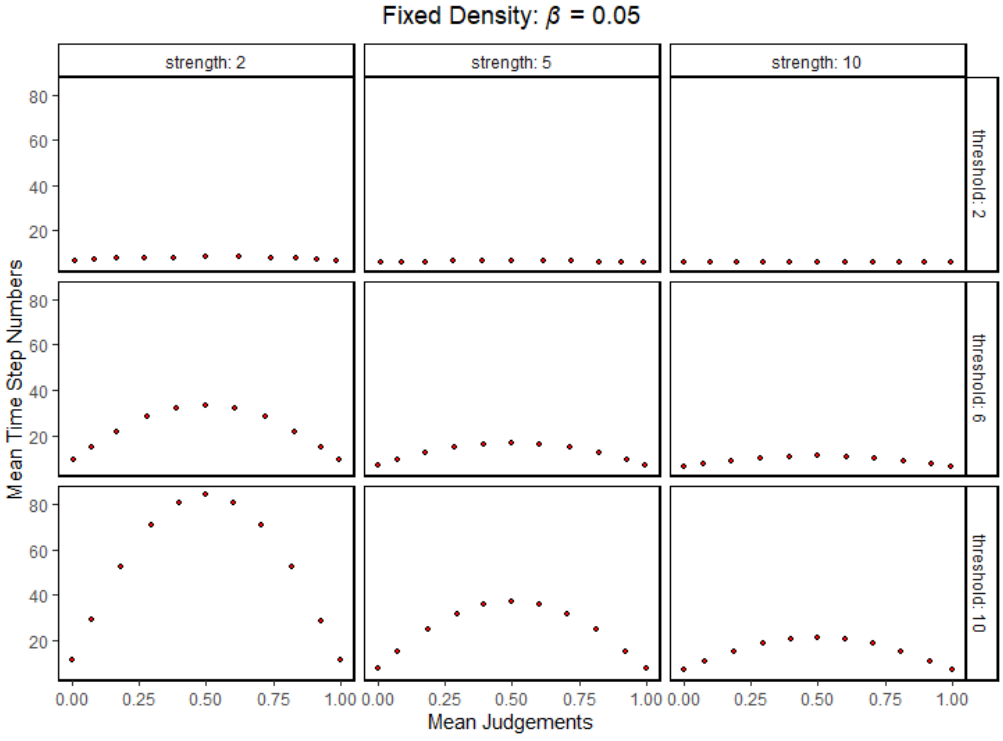
Figure 4

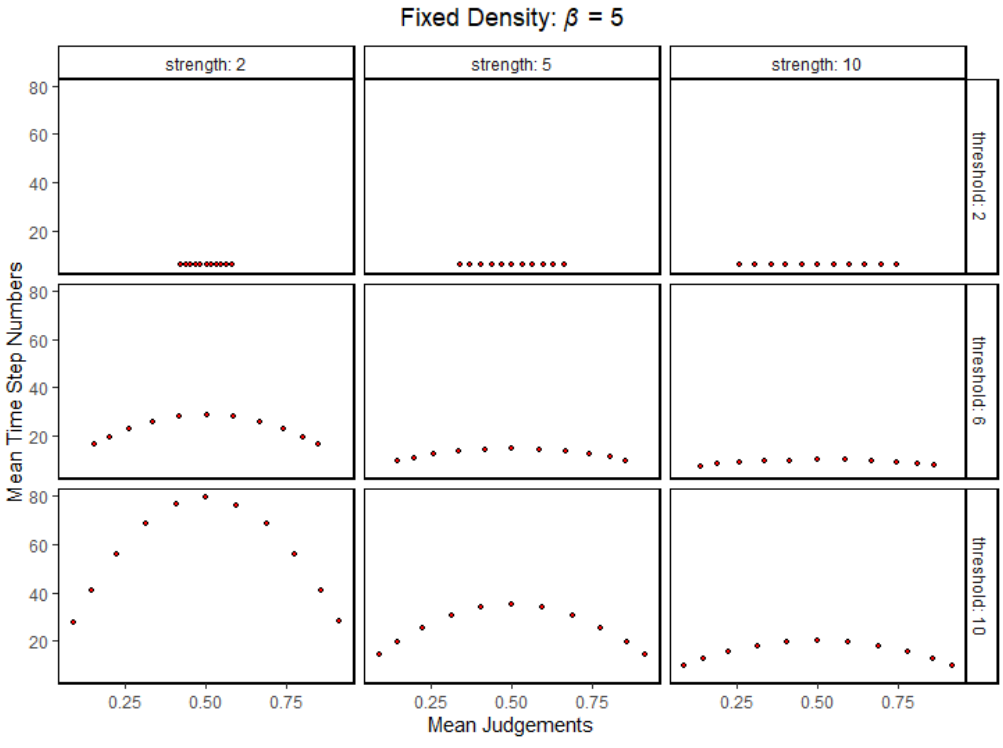
Mean Time Step Numbers as a function of Mean Judgments for Different Models Under Different Combinations of Parameter Values




Fixed Time: β = 0.05
strength: 2
strength: 5
strength: 10
threshold: 5
threshold: 10
threshold: 20
Mean Time Step Numbers
25
20
15
0.00
0.25
0.50
0.75
1.00
Mean Judgements


Fixed Time: β = 5
strength: 2
strength: 5
strength: 10
threshold: 5
threshold: 10
threshold: 20
Mean Time Step Numbers
25
20
15
0.00
0.25
0.50
0.75
1.00
Mean Judgements





For the fixed density model, the model predictions were similar to those of the fixed time model when the threshold was small. But when the threshold increased, a bell curve seemed to illustrate the relation between the two variables properly, and

its curvature seemed to be adjusted by the strength. This was because the JM needed fewer samples to make an extreme judgment based on the fixed density stopping rule, as previously mentioned. When the threshold was large, the difference in the total sample size needed to make an extreme and a medium judgment was amplified and a clear bell curve could be observed. Meanwhile, the time step number for the judgment process was in inverse proportion to the strength when other parameters were fixed. Thus, a smaller strength could increase the curvature of the curve. When the threshold was small, even the density of the prior mean (i.e., 0.5) could be larger than the threshold, which made that drawing the samples once was enough to give a judgment regardless of the content of the samples This caused straight lines in small threshold conditions. Besides, the curve was gentler with a larger prior. This was because a larger prior meant that the JM behaved as if having seen more symmetric samples where the relative frequency of the event was 0.5 before the sampling process. Therefore, a set of highly skewed samples only caused a smaller shift from 0.5 and the change in the density of the posterior mean was just a little higher than that induced by a set of symmetric samples when the prior was large. Coupled with the fact that the density of the prior mean was high enough in this situation, the curve became gentler.

Because the RT distributions were notoriously positively skewed (Luce, 1986; McCormack & Wright, 1964), I also recorded the skewness of the simulated total time step number distribution for each combination of parameter values. The mean skewness for the fixed sample size, the fixed time and the fixed density models was

0.355(0.004)[4], 0.288(0.004) and 0.846(0.088) respectively. The value for the fixed density model was extreme because when the threshold was small but the strength was large, the skewness could be very high (because usually sampling once was enough).

Perception-Based Probability Judgments

Perception-based probability judgments refer to the probability judgments about features of perceptual stimuli (Balci et al., 2009; Howe & Costello, 2017; Peterson & Beach, 1967). In relevant tasks, the participants will be presented with a set of stimuli sequentially or simultaneously. The stimuli differ in one or multiple dimensions (like color and shape). And the participants should judge the proportion of the stimuli with specific features in the set. I conducted an experiment to get the data of such judgments. In this experiment, the participants only needed to judge the probability of a simple event, i.e., the percentage of red or blue circles in a set of colored circles.

Methods

Participants

89 participants were recruited through the Prolific participant pool. They were asked to complete a 30-minute online experiment for a monetary reward (£4). The experiment was conducted in Pavlovia.

Procedures

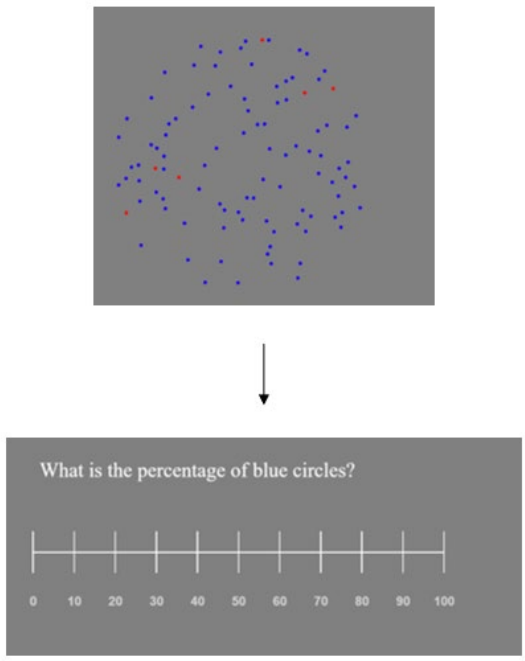
The experiment stimuli were displays of red and blue circles. Each display

[4] The values in the parentheses indicated standard errors.

containing 100 circles and the positions of the circles were randomly decided on each trial. After a display disappeared, the participants would see a linear ruler ranging from 0 to 100 on the screen (cf. Figure 5). They should click the corresponding position on the ruler to indicate their judgment about the percentage of a specific kind of colored circles on the display. The judgment and the RT from the ruler onset to the click would be recorded. The inter-trial interval was 0.4s. The experiment was written by PsychoPy (Peirce et al., 2019).

Figure 5

Illustration of the Experimental Stimuli



There were three conditions in the experiment: accuracy, speed and difficulty. In the accuracy condition, the participants were asked to give judgments as accurately as possible. In the speed condition, the participants were asked to give judgments as fast as possible but still trying to be accurate. They would see a warning message if

they did not respond within a given time and the corresponding trial was labelled as a missing trial. In the difficulty condition, the presentation time of the displays was shorter and the participants were still asked to give judgments as accurately as possible. Before the formal experiment, the participants should finish a practice stage. At that stage, the participants' judgments would be displayed for 1s after the clicks to help them understand the task better. There was no feedback about the judgments in the formal experiment. The presentation time of the displays was 0.4s in the difficulty condition and 0.8s in the other two conditions. The time constraint in the speed condition was the product of 0.9 and the median RT at the practice stage[5]. There were no time constraints in the other two conditions. The experimental parameter settings at the practice stage were the same as those in the accuracy condition. The type of the colored circles whose percentage should be judged was randomly decided at the beginning of the experiment and was consistent across the practice stage and three conditions. The participants would complete the three conditions in a random order after the practice stage. The practice stage contained 12 trials and each condition contained 100 trials.

Before each condition and the practice stage, four values would be drawn from the uniform distribution U(0.05, 0.15), U(0.3, 0.45), U(0.55, 0.7) and U(0.85, 0.95) respectively. These values would be rounded to two decimal places and used as the possible percentages to be judged for the next 100 or 12 trials. Each percentage

[5] The time constraint was individual-specific to reduce the influence of individual differences in response speed. Besides, I used the median RT to reduce the influence of too fast or too slow responses at the practice stage. The final choices of the time constraint and the display presentation time were decided based on a pilot study.

would be judged equal times in a random order. For example, if a participant would finish a formal condition containing 100 trials and the four percentages generated before that condition were 0.1, 0.4, 0.6 and 0.9. The participant would see displays containing 10, 40, 60 and 90 target colored circles, 25 times for each in the following condition. And the order of them was random. This manipulation ensured that I could get judgments of low, medium-low, medium-high and high probabilities from each participant, which could help to identify conservatism or anti-conservatism. It also ensured that the participants would not feel bored because the percentage to be judged could change trial by trial.

Results

1 participant was excluded because of finishing the experiment too fast (less than 10 mins). For the rest 88 participants, too fast or too slow responses (i.e., the responses whose RTs deviated more than 3 standard deviations from the mean) for each true percentage were deleted. Then I calculated the number of abnormal responses for each participant. An abnormal response meant giving a judgment larger than 0.5 for a low probability (i.e., $p <= 0.15$) or giving a judgment smaller than 0.5 for a high probability (i.e., $p >= 0.85$). Though these responses were possible to appear from a sampling perspective, they were more likely to be caused by some random errors like a trembling hand. 15 participants were excluded because of giving more than 3 abnormal responses during the whole experiment.

Mean judgments as a function of true percentages for the rest 73 participants were illustrated in Figure 6. Each participant provided 4 points in each sub-figure

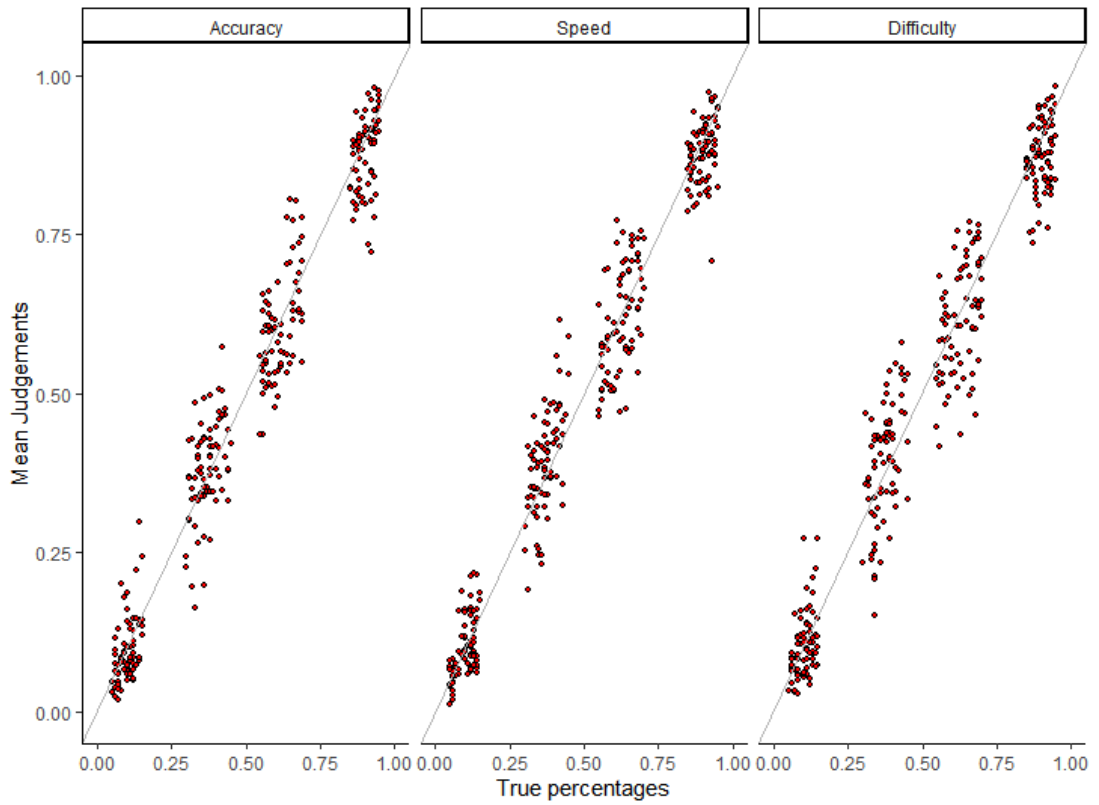
because there were 4 possible percentages to be judged in each condition. There was no clear pattern of conservatism or anti-conservatism in each condition. The points fluctuated around the identity line across different magnitudes of the true percentages. To identify conservatism or anti-conservatism at the group level quantitatively, I fitted the weighting function used by Tversky and Kahneman (1992) to the data. The function was

$$w(p) = \frac{p^{\gamma}}{(p^{\gamma}+(1-p)^{\gamma})^{1/\gamma}},$$

where p was a true percentage and $w(p)$ was the mean judgment of this percentage according to my dataset. γ ($\gamma > 0$) was the parameter indicating conservatism or anti-conservatism. When it was smaller than 1, the mean judgments showed the pattern of conservatism. When it was equal to 1, the mean judgments matched the true percentages. When it was larger than 1, the mean judgments showed the pattern of anti-conservatism. There were two advantages of this function. The first was that the degree (or existence) of conservatism or anti-conservatism could be controlled by one parameter. The second was that it could produce asymmetric conservatism or anti-conservatism (i.e., the crossover between the mean judgments-true percentages curve and the identity line was not necessary to be 0.5).

Figure 6

Mean Judgments as a Function of True Percentages in Different Conditions



Note. The grey lines are identity lines.

The value of γ that minimized the mean squared error between the predicted mean judgments and the actual mean judgments was searched using the PORT routines (Gay, 1990) with 10 random starting points, and a 95% Bias Corrected and accelerated (BCa) confidence interval (CI; DiCiccio & Efron, 1996) was constructed through a bootstrap procedure with 2000 replicates. For the accuracy condition, the estimate of γ was 0.935 and its 95% BCa CI was (0.900, 0.975). For the speed condition, the estimate was 0.941 and its 95% BCa CI was (0.911, 0.975). For the difficulty condition, the estimate was 0.921 and its 95% BCa CI was (0.883, 0.956). Thus, there was a slight pattern of conservatism across three conditions at the group

level.

Because each participant only judged four true percentages in each condition, fitting the weighting function at the individual level was unstable. I used a rough criterion to identify conservatism or anti-conservatism as a solution. That was, for each participant, I checked the mean judgments of the low percentage and the high percentage in each condition[6]. A participant was thought to show conservatism/anti-conservatism if he or she overestimated/underestimated the small percentage and underestimated/overestimated the high percentage. The results showed that there were 23 conservative participants and 18 anti-conservative participants in the accuracy condition, 28 conservative participants and 19 anti-conservative participants in the speed condition, and 28 conservative participants and 21 anti-conservative participants in the difficulty condition respectively. Though the mean judgments were not very credible based on limited observations (see Appendix B), the results to some extent indicated that there was a substantial portion of participants whose judgments of low and high percentages deviated from the true percentages in opposite directions. The number of conservative participants was slightly larger than that of anti-conservative participants, which might be the cause of the slight conservatism at the group level.

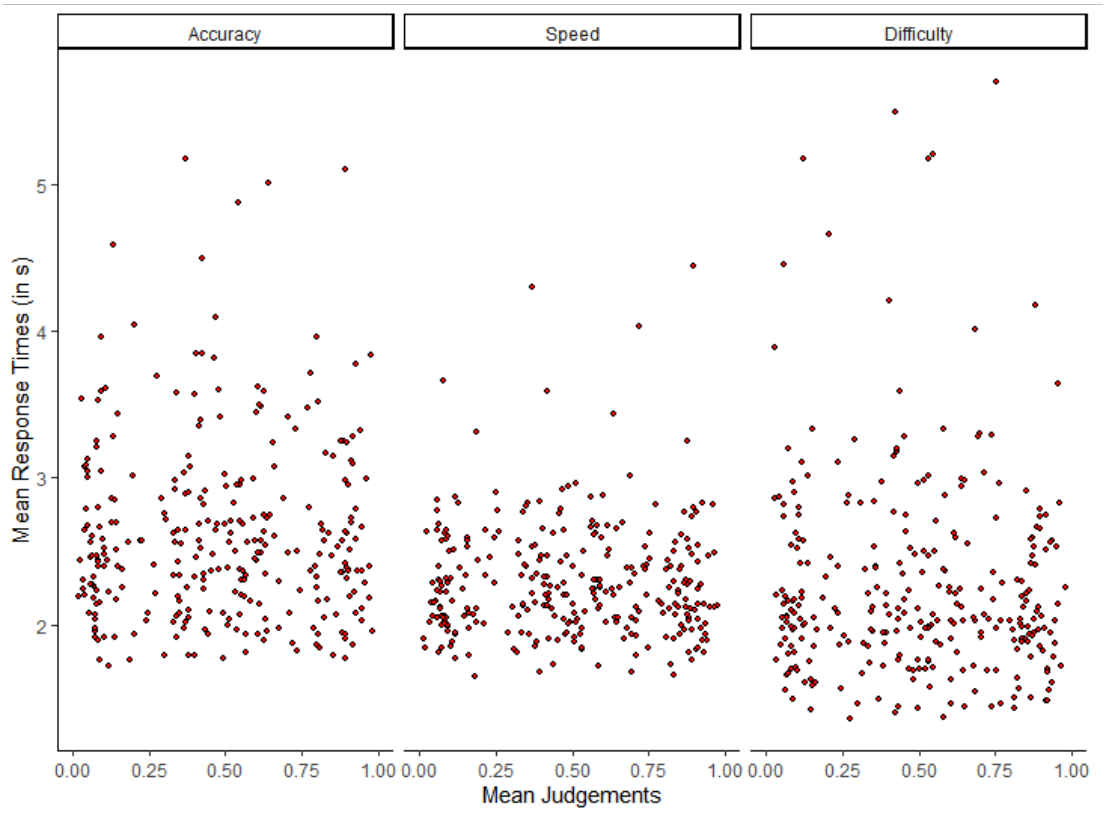
Mean RTs as a function of mean judgments were illustrated in Figure 7. There was no clear relation between the mean RTs and the mean judgments in each

[6] The judgements of the medium-low and medium-high percentages were not used finally because the samples of them were noisier, which meant that they were less credible.

condition. To identify the relation quantitatively, I fitted a quadratic regression model with dependent variable the mean RTs and independent variable the mean judgments. The axis of symmetry of the regression line was fixed to mean judgment = 0.5 and the coefficient of interest was the quadratic coefficient. If this coefficient was larger than 0, the mean RTs were shorter when the mean judgments were closer to 0.5. If this coefficient was smaller than 0, the mean RTs were longer when the mean judgments were closer to 0.5.

Figure 7

Mean Response Times as a Function of Mean Judgments in Different Conditions



Similarly, the values of the coefficients that minimized the mean squared error between the predicted and the actual mean RTs were searched using the PORT routines with 10 random starting points, and a 95% BCa CI for the quadratic

coefficient was constructed through a bootstrap procedure with 2000 replicates. For the accuracy condition, the estimate of the quadratic coefficient was -0.438 and its 95% BCa CI was (-1.361, 0.426). For the speed condition, the estimate was -0.333 and its 95% BCa CI was (-0.932, 0.266). For the difficulty condition, the estimate was -0.373 and its 95% BCa CI was (-1.444, 0.715). Thus, there was no significant relation between the mean RTs and the mean judgments across three conditions at the group level.

Though each participant still only judged four true percentages in each condition, the judgments of a true percentage could vary because of the randomness of the samples and the sample sizes. And the relation between the RTs and the judgments was directly caused by the samples according to the fixed density stopping rule. For example, if the true percentage was 0.1 but the actual relative frequency of the samples was 0.5, a fixed density rule user still took a long time to give a medium judgment. In other words, the relation between the mean RTs and the mean judgments was only a simplified representation of that between the RTs and the judgments (because extreme true probabilities were easy to produce extreme samples) and directly fitting the regression model to the original RTs and judgments was reasonable. Thus, for each participant, I fitted the quadratic model whose axis of symmetry was judgment = 0.5 and constructed the 95% BCa CI of the quadratic coefficient. The results showed that there were 11 participants whose upper limit of the 95% BCa CI was smaller than 0 and 3 participants whose lower limit of the 95% BCa CI was larger than 0 in the accuracy condition. The numbers were 12 and 1 in

the speed condition, and 6 and 2 in the difficulty condition. Thus, there were not many participants whose RTs and judgments showed the pattern predicted by the fixed density model when the threshold was large.

Besides, 34 of the rest 73 participants had seen the warning message at least once (i.e., missing at least one trial) in the speed condition. The maximum number of missing trials was 4. For each participant, I also calculated the mean RTs in the three conditions and compared the mean of the mean RTs across conditions. The results showed that the mean of the mean RTs in the speed condition was significantly smaller than that in the accuracy condition (difference = -0.320, SE = 0.049), $t(72) = -6.47$, d (Cohen's d) = -0.763, $p < .001$, but there was no significant difference between it and that in the difficulty condition (difference = 0.026, SE = 0.058), $t(72) = 0.45$, $d = 0.053$, $p = .655$. The mean of the mean RTs in the difficulty condition was significantly smaller than that in the accuracy condition (difference = -0.346, SE = 0.065), $t(72) = -5.33$, $d = -0.628$, $p < .001$ (all the p-values above were adjusted by the Bonferroni-Holm correction). The results suggested that the time pressure manipulation was effective, but the participants gave judgments even faster in the difficulty condition. The causes of this phenomenon were uncovered by computational modelling.

Model Fitting and Comparison

Because of the computational cost of the model fitting at the individual level. I only fitted the data of the first 42 participants. All of them had given at least 22 judgments for each true percentage in each condition. I divided the RTs by 0.1s and

rounded the results to integers to transform them to time step numbers[7]. And to make the range of the judgments and that of the time step numbers similar, I divided the time step numbers by 100 to generate the scaled time step numbers (STSNs). This transformation was conducted to ensure that the weights of the two variables were roughly equal in model fitting and comparison.

Parameter Estimation

It was a notorious problem for the SSMs to get the analytical likelihood function (Wagenmakers, 2009), and this problem also existed for the BSSMs, especially the fixed density BSSM. Thus, I estimated the parameters using a simulation-based method. Specifically, for the data of a certain participant in a certain condition, the parameter estimation of a BSSM was conducted as follows.

1. Divide the data into four groups. Each group contained the judgments and the STSNs for a specific true percentage. Calculate the mean and the standard deviation of the judgments and the STSNs respectively for each group.
2. For a specific combination of parameter values, generate 500 pairs of simulated judgment and STSN for each group. The true probabilities in the simulations matched the true percentages, so the parameters to be estimated were β, λ, d and T_{non}. Calculate the mean and standard deviation of the simulated judgments and STSNs respectively for each group.

[7] The time step length was arbitrary like the within-trial variability in the SSMs to some extent. When it was multiplied by k, the model predictions of the mean estimates and mean RTs would be roughly unchanged if the strength was multiplied by k and the non-judgement time was divided by k for the three BSSMs without considering the problem of rounding.

3. Calculate the differences (absolute differences for the fixed sample size and the fixed time BSSMs, and squared differences for the fixed density BSSM) between the simulated results and the actual data in the four statistics for each group. Multiply the differences in the mean and the standard deviation of the judgments and multiply the differences in the mean and the standard deviation of the STSNs. Add the two products together and average the sums across groups to get a one-shot distance.
4. Repeat step 2 and 3 ten times and average all one-shot distances to get the integrated distance. This procedure aimed to reduce the influence of the randomness of the simulation results.
5. Use the particle swarm optimization method (Parsopoulos & Vrahatis, 2002) to search for the combination of parameter values that minimize the integrated distance with 5 random starting points. To reduce the computational cost, the maximum times of iteration was restricted to 80. The ranges of the β, λ and T_{non} were (e^{-7}, $e^{2.5}$), (1, $e^{2.5}$) and (1, e^{4}) respectively. The ranges of d_{size} and d_{time} were both (1, e^{5}). The range of $d_{density}$ was (1, $e^{2.5}$)[8]. Some parameters like d_{size} should be an integer. Though the search of them was still conducted in a continuous space, they were rounded when generating the simulation results and the final estimates of them would also be rounded.

Appendix C showed the reasons why I used the statistics-based metric to quantify the difference between the simulation results and the actual data. Compared

[8] The limits of the ranges were represented in the form of e^{x} because the actual parameter space searched was a logarithmic space.

to the likelihood-based parameter estimation, the simulation-based estimation was more unstable because it introduced the randomness in the simulations. But it was more robust when there were some not-deleted extreme responses like too long or short RTs (Ratcliff & Tuerlinckx, 2002). The parameter estimation results were shown in Table 1. The parameter estimates of a single BSSM were meaningless to some extent because I found that the participants tended to use different strategies in different conditions.

Table 1

Means and Standard Errors of Parameter Estimates for Different Models in Different Conditions

Model	Mean estimate			
	Prior	Strength	Threshold	Non-judgment time
Accuracy				
Fixed sample size	0.810(0.221)	1.247(0.073)	21.476(1.671)	7.476(0.861)
Fixed time	0.719(0.162)	5.319(0.619)	6.738(0.690)	18.414(1.351)
Fixed density	0.923(0.158)	4.531(0.589)	3.647(0.238)	17.667(0.975)
Speed				
Fixed sample size	0.763(0.171)	1.783(0.173)	22.786(3.209)	9.524(0.775)
Fixed time	0.741(0.169)	2.71(0.394)	10.310(0.594)	11.734(0.858)
Fixed density	0.858(0.148)	3.876(0.495)	3.593(0.230)	16.095(0.912)
Difficulty				
Fixed sample size	0.843(0.258)	1.307(0.078)	17.786(1.479)	7.833(0.907)
Fixed time	0.81(0.202)	4.695(0.503)	5.333(0.591)	16.85(1.502)
Fixed density	1.015(0.199)	3.977(0.586)	3.251(0.208)	15.619(1.182)

Note. Standard errors are presented in parentheses.

Model Comparison

Because the integrated distance was based on the absolute differences for the fixed sample size and the fixed time BSSMs but squared differences for the fixed density BSSM, I used a distribution-based metric to compare the model predictions

and the actual data. The metric was the 2-dimensional (2d) wasserstein distance (Rüschendorf, 1985), which was the best metric among the distribution-based metrics in the parameter recovery test (see Appendix C) that could directly compare the distance between two joint distributions of judgments and STSNs. Specifically, for the data of a certain participant in a certain condition, the model comparison was conducted as follows.

1. Divide the data into four groups. Each group contained the judgments and the STSNs for a specific true percentage. Duplicate each observation 200 times within each group to generate the enlarged groups. This was because the 2d wasserstein distance could only compare two sets of 2d samples whose sizes were equal and the size of the simulated results should be large enough to reduce the influence of the randomness in the simulations.
2. For a specific BSSM with its parameter estimates, generate a set of simulated judgments and STSNs whose size equals the enlarged group size for each group. Calculate the wasserstein distance of order 1 or 2 between the simulation results and the enlarged actual data and average the results across groups to get the integrated wasserstein distance.
3. Choose the BSSM which produced the smallest integrated wasserstein distance as the best model to describe this participant in this condition.

Numbers of participants best described by each model in different conditions were shown in Table 2. Combining the results based on the waseerstein distance of order 1 (WD1) and those based on the wasserstein distance of order 2 (WD2), the

participants mainly used the fixed time rule in the accuracy condition and the fixed sample size rule in the speed condition. In the difficulty condition, the two rules seemed equally likely to be used. Besides, there were always some participants using the fixed density rule in each condition, but the number was smaller in the speed condition.

Table 2

Numbers of Participants Best Described by Each Model in Different Conditions

Model	Participant number		
	Accuracy	Speed	Difficulty
Wasserstein (order 1)			
Fixed sample size	11	23	16
Fixed time	23	16	19
Fixed density	8	3	7
Wasserstein (order 2)			
Fixed sample size	15	27	21
Fixed time	20	11	13
Fixed density	7	4	8

Parameter Change

To explore how different manipulations changed the cognitive processes underlying the probability judgments, I should compare the parameter values across three conditions. But a problem was that the participants seemed to have a strategy switch when facing different conditions. Thus, for the three common parameters (i.e., β, λ and T_{non}), I used the estimates from the best model for each participant in each condition to illustrate the corresponding cognitive processes.

When the best models were decided based on the WD1, I analyzed the estimates using a Greenhouse-Geisser corrected ANOVA with condition (accuracy,

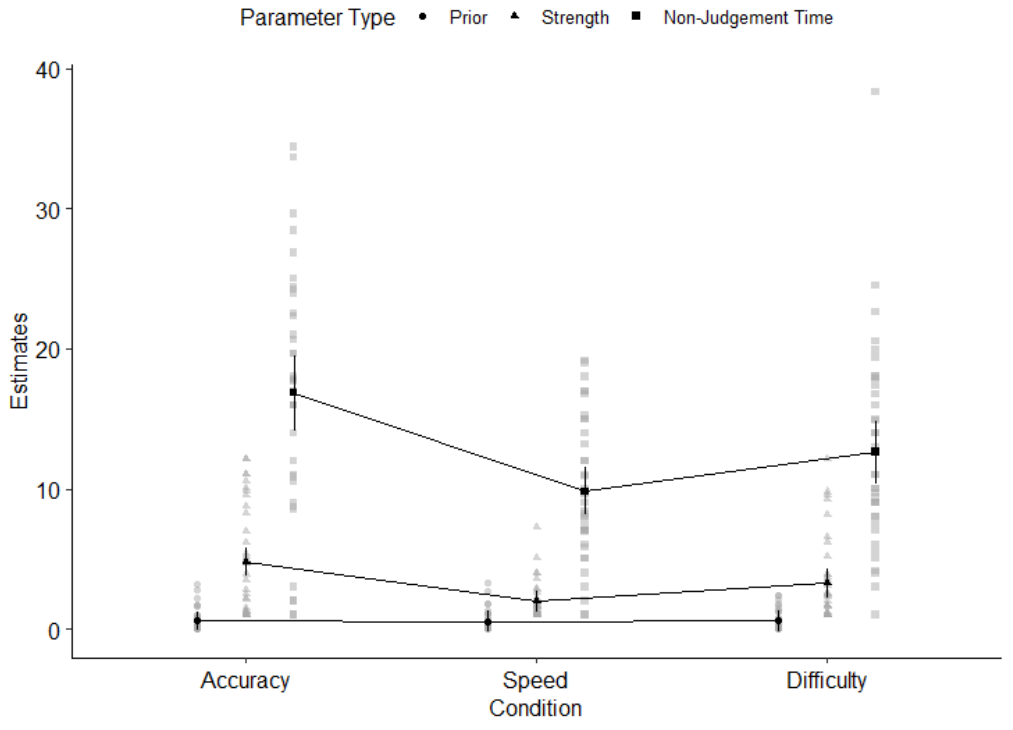
speed and difficulty) and parameter type (β, λ and T_{non}) as repeated measures. The results showed that the mean effect of condition was significant, $F(1.81, 74.11) = 7.75$, η^2(generalized eta-squared) = .055, $p = .003$. The main effect of parameter type was significant, $F(1.31, 53.84) = 208.89$, $\eta^2 = .531$, $p < .001$. The interaction of condition and parameter type was significant, $F(2.27, 93.16) = 5.51$, $\eta^2 = .041$, $p = .004$. I also ran a Kenward-Roger corrected mixed linear model to analyze the estimates with condition and parameter type as fixed effects and a random intercept for each participant after deleting the extreme estimates (i.e., deviating more than three standard deviations from the mean) of each parameter within each condition, which produced similar results, ps < .001.

Then I tested the simple effects of condition. The results showed that there was no significant difference in the mean β across three conditions (|difference| < 0.095), $|t(41)| < 0.41$, d (an approximation of Cohen's d) < 0.063, $p = 1.000$. For the mean λ, there was no significant difference between the accuracy condition and the difficulty condition (difference = 1.012, SE = 0.819), $d = 0.193$, $t(41) = 1.24$, $p = .224$. But the mean λ in the speed condition was significantly lower than that in the accuracy condition (difference = -2.573, SE = 0.575), $d = -0.699$, $t(41) = -4.48$, $p < .001$, as well as that in the difficulty condition (difference = -1.561, SE = 0.662), $d = -0.368$, $t(41) = -2.36$, $p = .047$. For the mean T_{non}, there was no significant difference between the accuracy and the difficulty condition (difference = 2.959, SE = 1.871), $d = 0.247$, $t(41) = 1.58$, $p = .122$, or between the speed and the difficulty condition (difference = -3.242, SE = 1.648), $d = -0.307$, $t(41) = -1.97$, $p = .112$. But

the mean T_{non} in the speed condition was significantly lower than that in the accuracy condition (difference = -6.200, SE = 1.574), d = -0.615, $t(41)$ = -3.94, p = .001 (all the p-values above, including those for the F tests in the two-way ANOVA, were adjusted by the Bonferroni-Holm correction). When the best models were decided based on the WD2 or when analyzing the estimates of each parameter separately using Greenhouse-Geisser corrected ANOVAs with condition as a repeated measure (or their equivalent Kenward-Roger corrected mixed linear models), the basic pattern kept unchanged (cf. Figure 8 and 9).

Figure 8

Parameter Estimates From Best Models Based on the Wasserstein Distance of Order 1 in Different Conditions

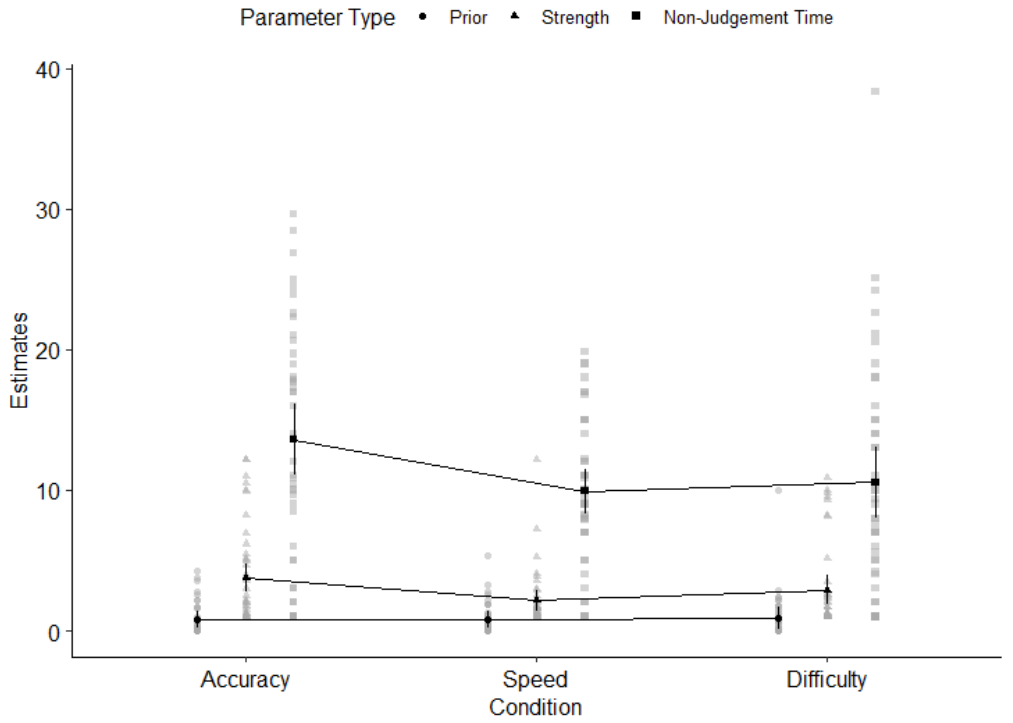


Note. Overlapping points are shown with lower transparency. Bold points in the foreground show the means. Error bars show 95% within-subjects confidence

intervals.

Figure 9

Parameter Estimates From Best Models Based on the Wasserstein Distance of Order 2 in Different Conditions



Note. Overlapping points are shown with lower transparency. Bold points in the foreground show the means. Error bars show 95% within-subjects confidence intervals.

Because the meaning of the threshold was different for different BSSMs, I changed it to the equivalent average sample size (EASS). Specifically, the EASS equalled the threshold parameter for the fixed sample size BSSM[9]. For the fixed time BSSM, the EASS equalled the product of the strength parameter and the threshold

[9] Strictly speaking, the EASS should be a little higher than the threshold for the fixed sample size BSSM, but simply treating them as the same thing would not influence the results much because the strength was typically small.

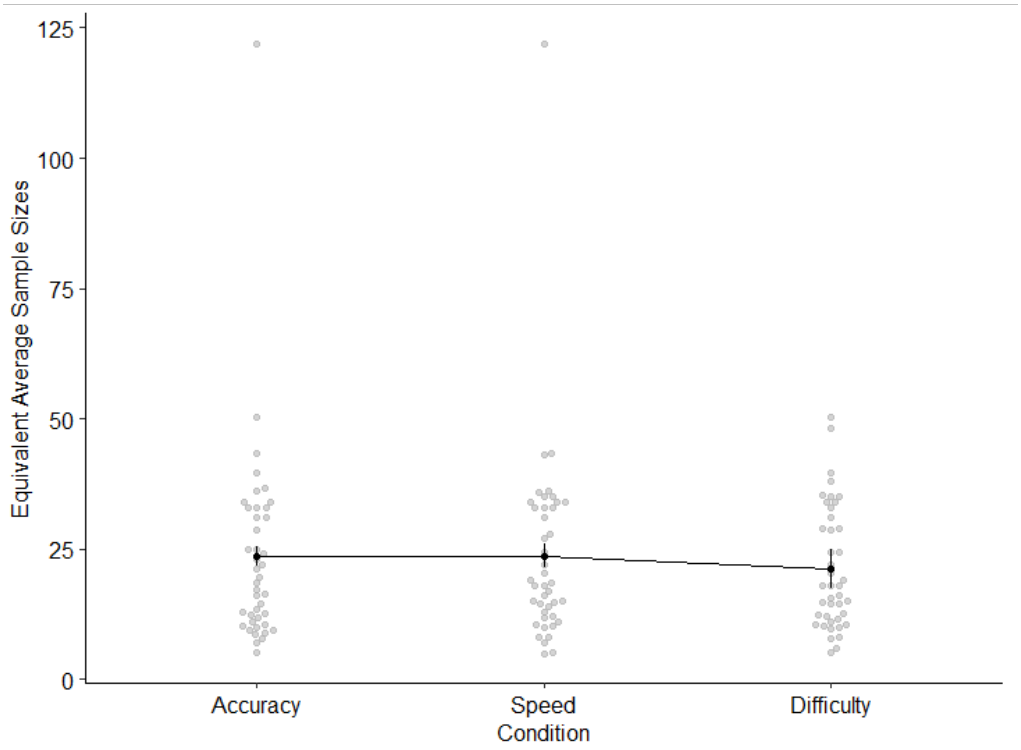
parameter. For the fixed density BSSM, the EASS could not be calculated simply because it was also relevant to the true probabilities. Thus, for each participant in each condition, I generated 5000 simulations for each true percentage he or she had judged using the fixed density BSSM with the corresponding parameter estimates and recorded the total sample size of each simulation. And I averaged the 20000 sample sizes to get an approximation of the EASS.

Similarly, I used the EASS from the best model for each participant in each condition to represent the actual average total sample size. And I analyzed the EASSs using a Greenhouse-Geisser corrected ANOVA with condition (accuracy, speed and difficulty) as a repeated measure when the best models were decided based on the WD1. The results showed that the effect of condition was not significant, $F(1.13, 46.31) = 1.03$, $\eta^2 = .004$, $p = .326$. A Kenward-Roger corrected mixed linear model with condition as the fixed effect and a random intercept for each participant after deleting the extreme EASSs in each condition produced similar results, $p = .951$. However, when the best models were decided based on the WD2, the results showed that the effect of condition was significant, $F(1.78, 73.07) = 3.24$, $\eta^2 = .026$, $p = .050$. The post hoc test showed that there was no significant difference between the accuracy and the speed condition (difference = 0.342, SE = 2.138), $d = 0.050$, $t(41) = 0.16$, $p = .874$, or between the accuracy and the difficulty condition (difference = 6.016, SE = 2.834), $d = 0.663$, $t(41) = 2.12$, $p = .119$, or between the speed and the difficulty condition (difference = 5.674, SE = 2.914), $d = 0.608$, $t(41) = 1.95$, $p = .119$. Though the differences were all non-significant, there was still an obvious pattern

that the mean EASS was smaller in the difficulty condition (cf. Figure 10 and 11). And the equivalent Kenward-Roger corrected mixed linear model showed marginal significant differences between the difficulty condition and the two other conditions, *p*s = .061 (all the *p*-values above were adjusted by the Bonferroni-Holm correction).

Figure 10

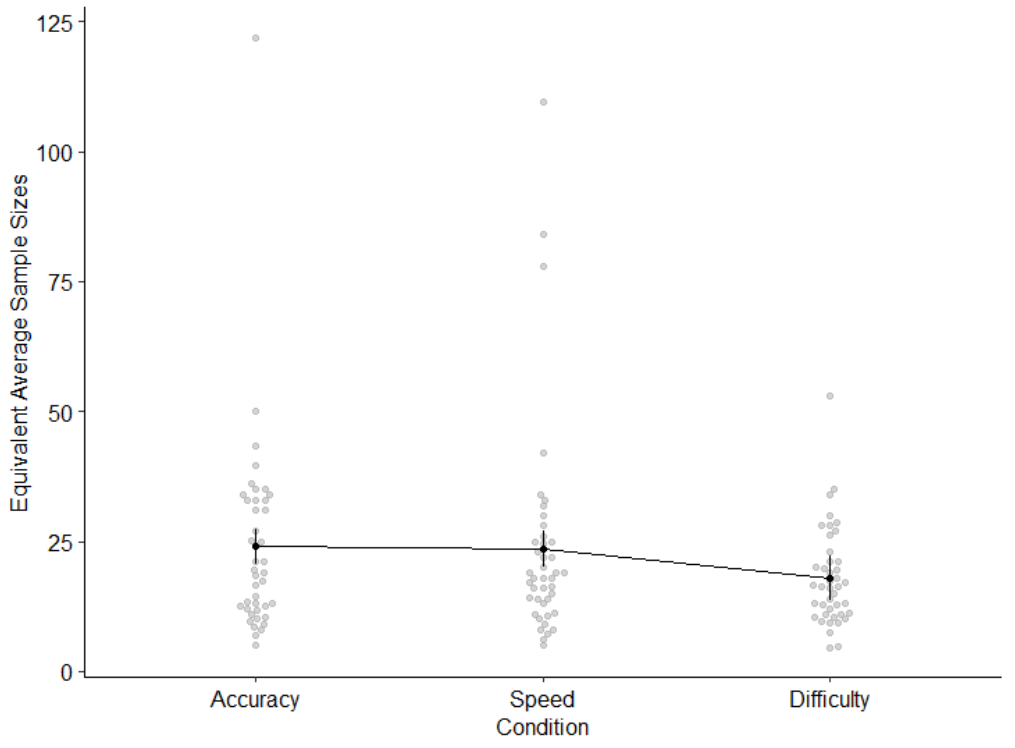
Equivalent Average Sample Sizes From Best Models Based on the Wasserstein Distance of Order 1 in Different Conditions



Note. Points are jittered to prevent overlapping. Bold points in the foreground show the means. Error bars show 95% within-subjects confidence intervals.

Figure 11

Equivalent Average Sample Sizes From Best Models Based on the Wasserstein Distance of Order 2 in Different Conditions



Note. Points are jittered to prevent overlapping. Bold points in the foreground show the means. Error bars show 95% within-subjects confidence intervals.

Combining the results based on the WD1 and those based on the WD2. The participants' abilities to generate samples decreased in the speed condition. They also spent shorter time on the non-judgment process to give faster responses. In the difficulty condition, they used fewer samples to make judgments and the non-judgment time also slightly decreased.

Model Predictions

To explore how well the BSSMs could describe the real data, I compared the predicted and the actual mean judgments and RTs. Specifically, for each participant in

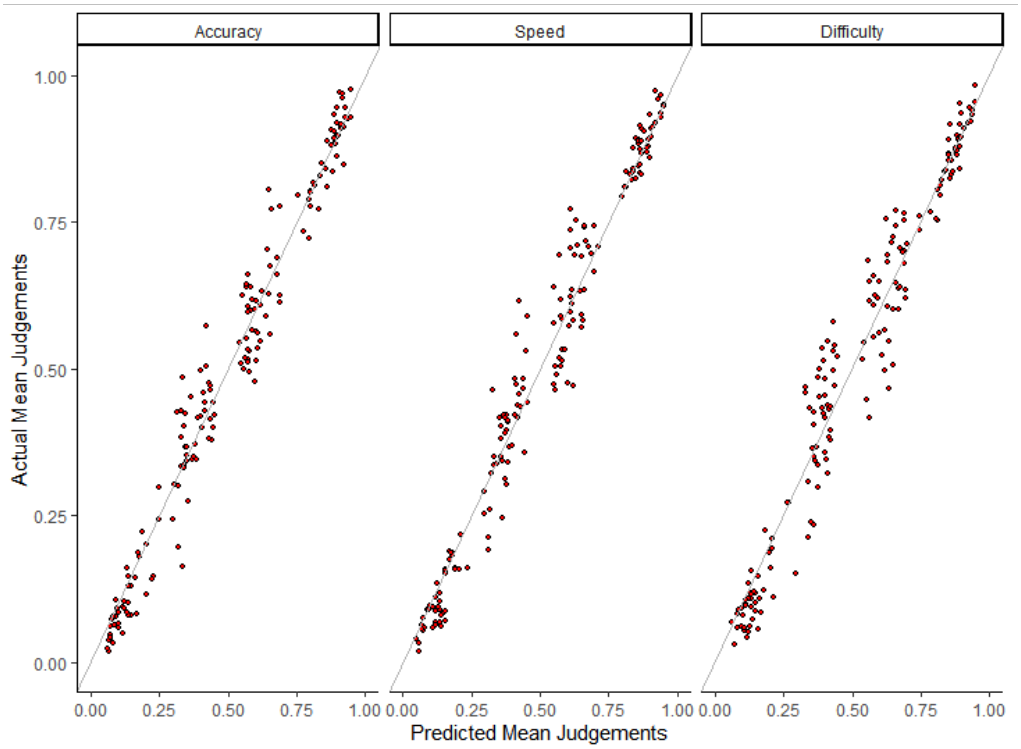
each condition, I generated 5000 simulations for each true percentage he or she had judged using the best BSSM for him or her and calculated the mean judgments and RTs (the time step numbers were transformed to RTs by being multiplied by 0.1s).

When the best models were decided based on the WD1, actual mean judgments as a function of predicted mean judgments were illustrated in Figure 12. The BSSMs slightly overestimated the mean judgments for low probabilities but not underestimated those for high probabilities, which suggested a pattern of asymmetric conservatism. To evaluate the performances quantitatively, I ran a simple linear regression model with independent variable the predicted mean judgments and dependent variable the actual mean judgments and compared the regression line with the identity line. The results showed that there was no significant difference between the slope and 1 (difference = 0.021, SE = 0.014), $d = 0.114$, $t(166) = 1.47$, $p = .142$, or between the intercept and 0 (difference = -0.016, SE = 0.008), $d = 0.151$, $t(166) = -1.95$, $p = .106$ in the accuracy condition. There was no significant difference between the slope and 1 (difference = 0.026, SE = 0.017), $d = 0.115$, $t(166) = 1.49$, $p = .278$, or between the intercept and 0 (difference = -0.015, SE = 0.010), $d = -0.115$, $t(166) = -1.49$, $p = .278$ in the difficulty condition. However, in the speed condition, the slope was significantly larger than 1 (difference = 0.036, SE = 0.015), $d = 0.184$, $t(166) = 2.37$, $p = .037$, and the intercept was significantly smaller than 0 (difference = -0.019, SE = 0.009), $d = -0.172$, $t(166) = -2.22$, $p = .037$. When the best models were decided based on the WD2, the regression line was not different from the identity line even in the speed condition, ps = .062 (all the p-values above were

adjusted by the Bonferroni-Holm correction).

Figure 12

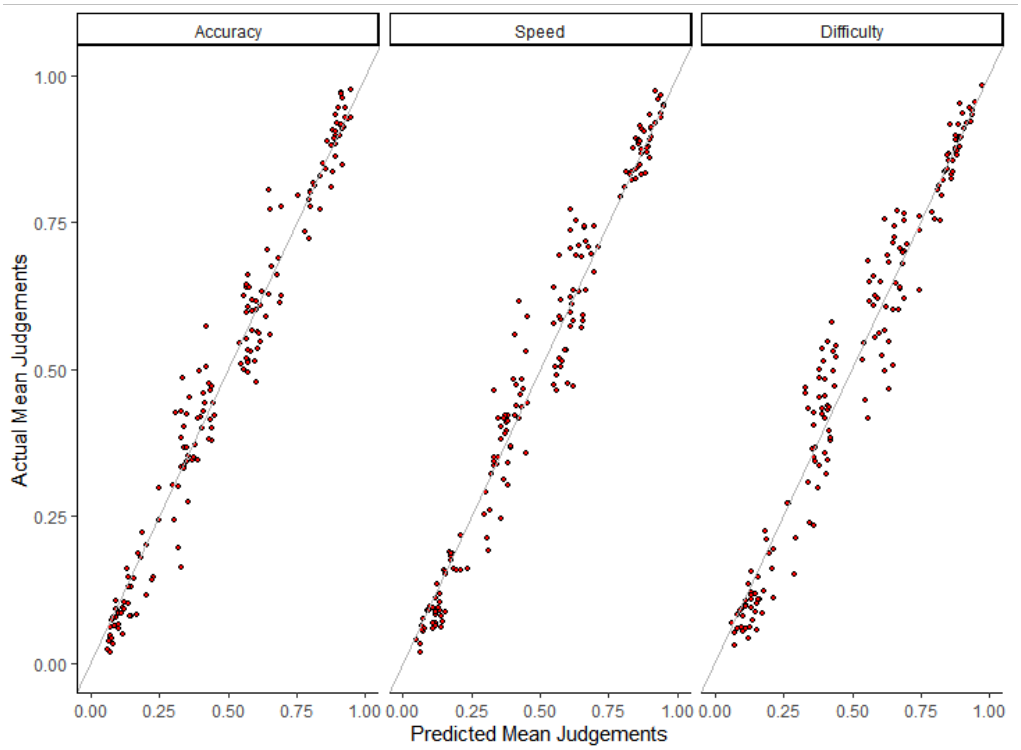
Actual Mean Judgments as a function of Predicted Mean Judgments based on the Wasserstein Distance of Order 1 in Different Conditions



Note. The grey lines are identity lines.

Figure 13

Actual Mean Judgments as a function of Predicted Mean Judgments based on the Wasserstein Distance of Order 2 in Different Conditions



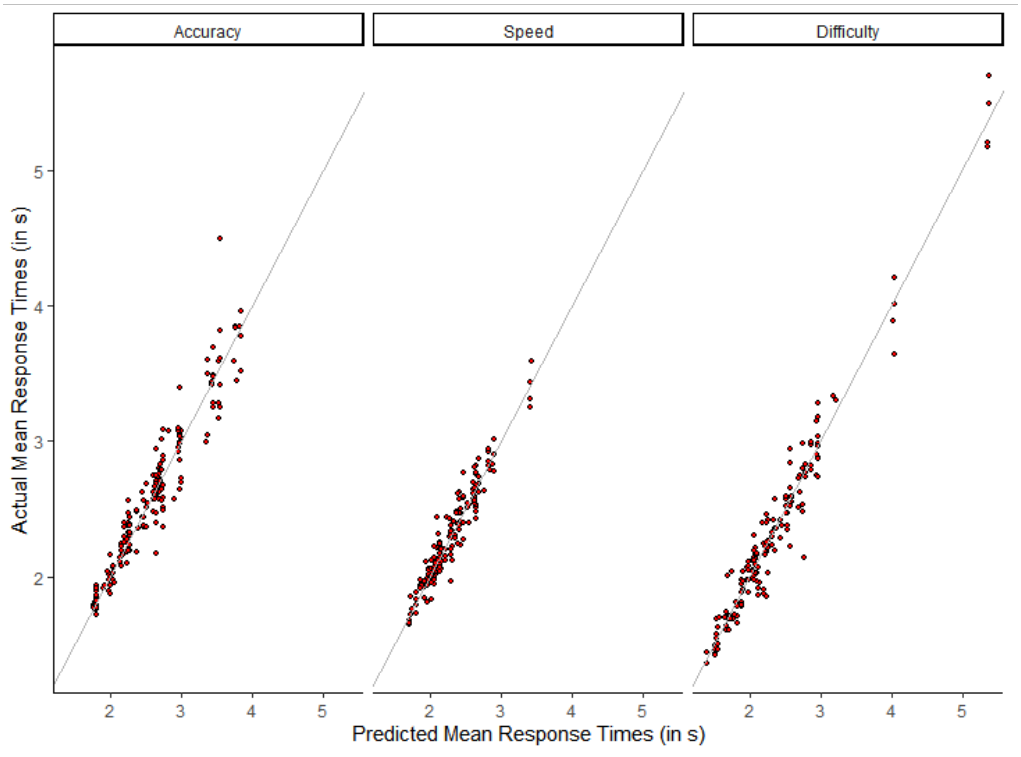
Note. The grey lines are identity lines.

When the best models were decided based on the WD1, actual mean RTs as a function of predicted mean RTs were illustrated in Figure 14. The BSSMs relatively accurately predicted the mean RTs expect for one outlier in the accuracy condition. After deleting that outlier, I ran a simple linear regression model with independent variable the predicted mean RTs and dependent variable the actual mean RTs and compared the regression line with the identity line. The results showed that in the accuracy condition, the slope was significantly smaller than 1 (difference = -0.058, SE = 0.021), d = -0.218, $t(165)$ = -2.80, p = .012, and the intercept was significantly larger than 0 (difference = 0.140, SE = 0.055), d = 0.198, $t(165)$ = 2.54, p = .012.

There was no significant difference between the slope and 1 (difference = -0.020, SE = 0.022), d = -0.071, t(166) = -0.91, p = .655, or between the intercept and 0 (difference = 0.051, SE = 0.051), d = 0.076, t(166) = 0.98, p = .655 in the speed condition. There was no significant difference between the slope and 1 (difference = -0.007, SE = 0.015), d = -0.035, t(166) = -0.46, p = 1.000, or between the intercept and 0 (difference = 0.011, SE = 0.038), d = 0.023, t(166) = 0.30, p = 1.000 in the difficulty condition (all the p-values above were adjusted by the Bonferroni-Holm correction). The results did not change when the best models were decided based on the WD2.

Figure 14

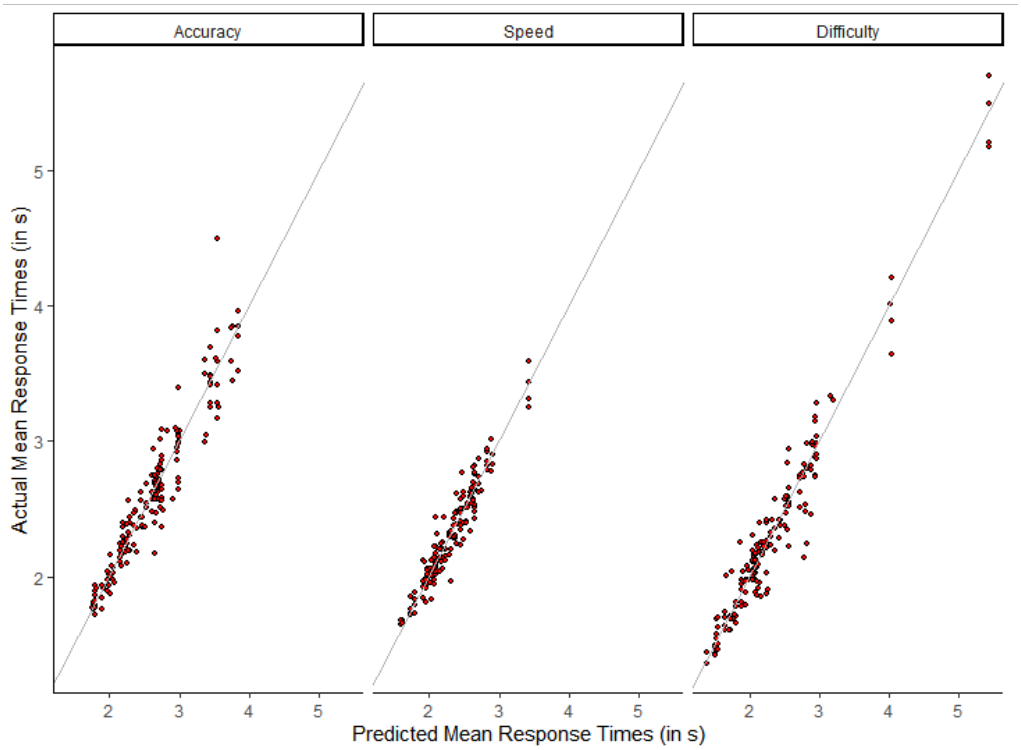
Actual Mean Response Times as a function of Predicted Mean Response Times based on Wasserstein Distance of Order 1 in Different Conditions



Note. The grey lines are identity lines.

Figure 15

Actual Mean Response Times as a function of Predicted Mean Response Times based on Wasserstein Distance of Order 2 in Different Conditions



Note. The grey lines are identity lines.

Combining the results based on the WD1 and those based on the WD2, the BSSMs properly predicted the mean judgments and RTs in general, though they underperformed when predicting the mean judgments in the speed condition and predicting the mean RTs in the accuracy condition.

Knowledge-Based Probability Judgments

Knowledge-based probability judgments or probability reasoning refer to the probability judgments about descriptive events (Costello, 2009; Tversky & Kahneman, 1983; Zhu et al., 2020). In relevant tasks, the participants will be directly asked to judge the probabilities of some everyday events like weather or infer the

probabilities of some events (e.g., the flipping results of a coin) with descriptive information (e.g., the flipping history of that coin). The objective probabilities of the events can sometimes be ambiguous or even unavailable, but those are known for the dataset I used. The experimental data of such judgments came from an unpublished study (Zhu, Newall, et al., 2021). In the experiment, the participants should judge the occurrence probabilities of some combination of poker cards.

Methods

Participants

906 participants were recruited from an online professional poker player community (twoplustwo.com). Because small monetary incentives could not motivate professional players, they received feedback about their performances in the experiment as rewards.

Procedures

The participants' main task was to judge the occurrence probabilities of different 3-card combinations from a standard deck of 52 poker cards. They judged the probabilities in a frequency format (Gigerenzer & Hoffrage, 1995). That was, they judged the number of occurrences of the combinations within 1000 repetitions. The participants should give judgments within 60s to avoid explicit calculation or searching on the internet, which should not be seen as time pressure if they indeed just gave their subjective judgments.

There were nine combinations whose probabilities were judged: 3 cards with the same rank (0.002), 2 cards with the same rank (0.169), no cards with the same

rank (0.829), 3 cards with the same suit (0.052), 2 cards with the same suit (0.550), no cards with the same suit (0.398), 3 cards to a straight (0.035), 2 cards to a straight (0.370), no cards to a straight (0.595). The values in the parentheses indicated the true probabilities. Each participant judged the probability of each combination only once.

Before the formal experiment, the participants were asked about the number of suits (4) and the number of cards per suit (13) in a standard deck of poker cards to ensure that they understood what the term meant in the questions. The values in the parentheses indicated the correct answers.

Results

There were 223 participants left after excluding the participants who did not finish the whole experiment or gave wrong answers to the questions about suits. Because each participant only gave one judgment to each probability, the data was analyzed at the group level, which meant the judgments and RTs would be treated as coming from one participant who gave repeated probability judgments.

After deleting too fast or too slow responses for each "true percentage"[10] and abnormal responses[11], mean judgments as a function of true percentages were illustrated in Figure 16. There was an obvious pattern of asymmetric conservatism. The participants overestimated small probabilities and underestimated high, even medium probabilities. To verify this pattern quantitatively, I fitted the weighting

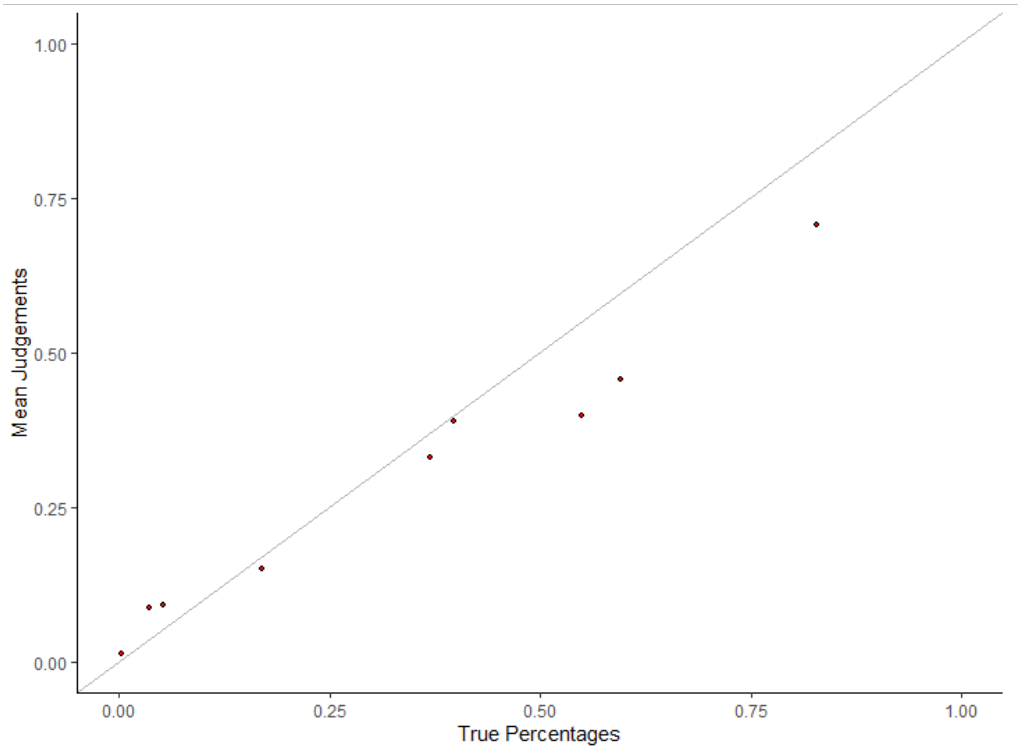
[10] To distinguish the "true probability" parameter in the BSSMs and the "true probabilities" of the events to be judged, I still used the term "true percentages" like what was used in the perception-based probability judgement experiment.

[11] Because most probabilities judged in this experiment were small probabilities, the "abnormal response" was defined as give a judgement larger than 0.8 for a true percentage smaller than 0.2 or giving a judgement smaller than 0.2 for a true percentage larger than 0.8.

function to the original judgments (the model would be too unstable if I fitted the function to the mean judgments because there were only 9 observations). The estimate of γ was 0.642 and its 95% BCa CI was (0.617, 0.669), which supported the existence of conservatism.

Figure 16

Mean Judgments as a Function of True Percentages



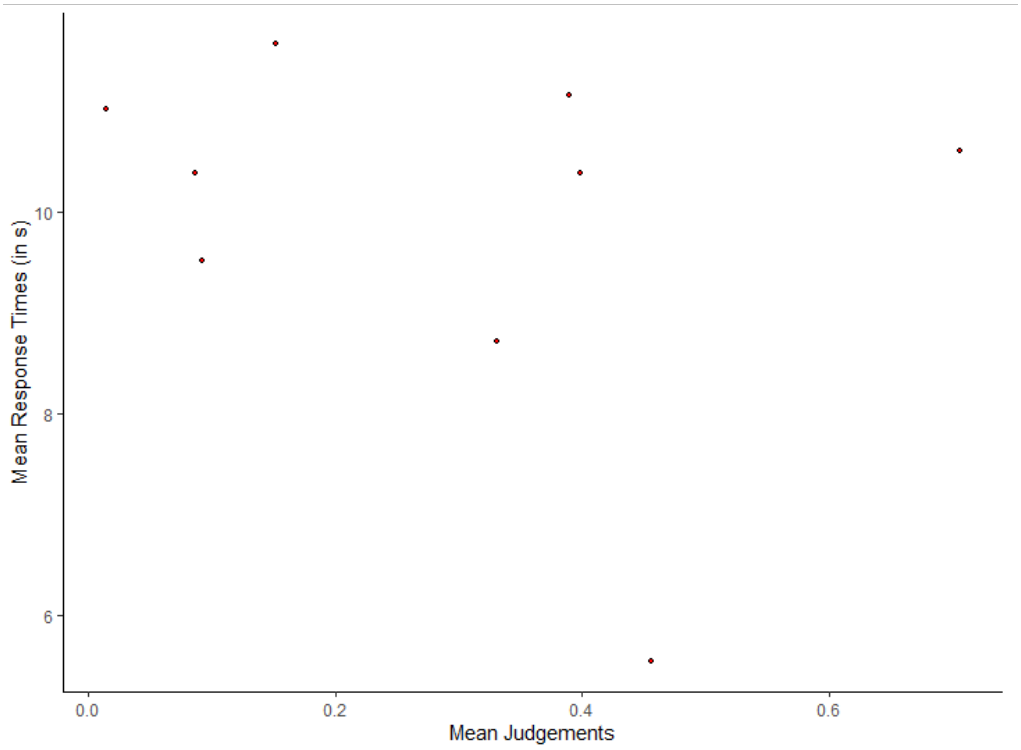
Note. The grey line is identity line.

Mean RTs as a function of mean judgments were illustrated in Figure 17. There was no clear pattern between the mean RTs and the mean judgments. To identify the relation quantitatively, I fitted the quadratic model whose axis of symmetry was 0.5 to the original judgments and RTs. The estimate of the quadratic coefficient was 0.155 and its 95% BCa CI was (0.044, 0.266), which suggested that there was even a slight U shape relation between the RTs and the judgments. Though

the results were not very trustworthy because most true percentages were smaller than 0.5, it at least showed that the fixed density rule was not the prevalent stopping rule.

Figure 17

Mean Response Times as a function of Mean Judgments



Model Results

The RTs were divided by 0.5s and 100 to be transformed to STSNs (because the RTs were typically larger than those in the perception-based probability judgment experiment). The parameter estimation procedures were the same as those in the perception-based probability judgment experiment, and the estimate of each parameter of each BSSM at the group level was shown in Table 3.

Table 3

Parameter Estimates for Different Models

Model	Estimate			
	Prior	Strength	Threshold	Non-judgment time
Fixed sample size	0.195	1.125	5	16
Fixed time	0.229	6.114	1	19.014
Fixed density	0.273	5.426	1.236	19

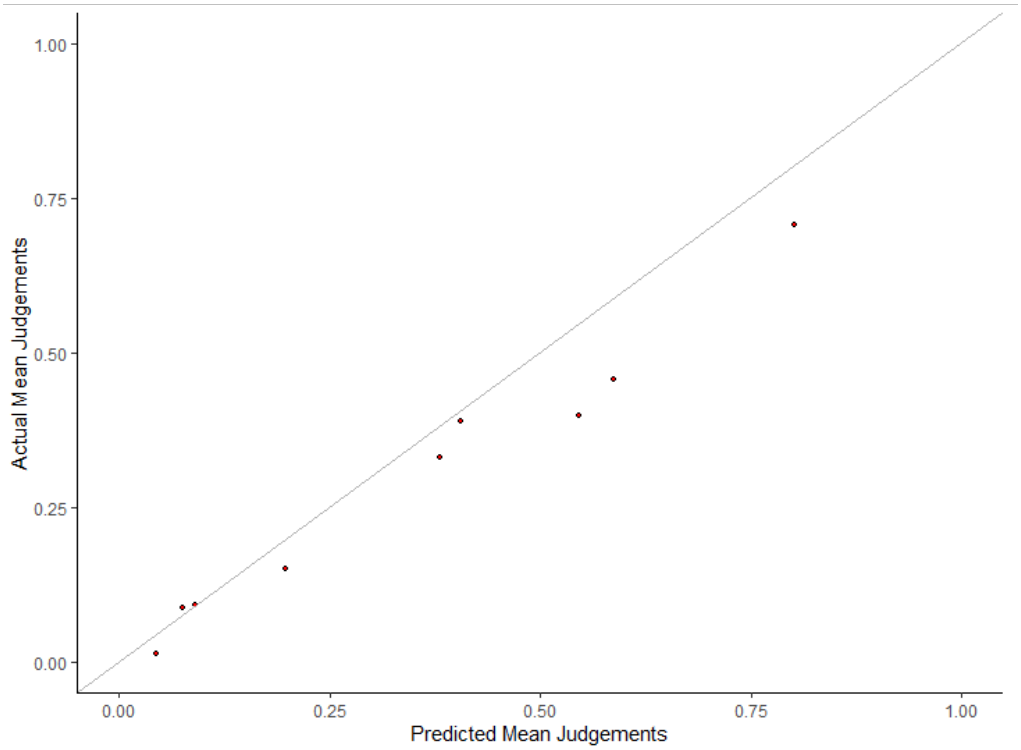
I still used the integrated wasserstein distances to compare BSSMs. The integrated wasserstein distances of order 1 for the three BSSMs were 0.174, 0.160 and 0.181 respectively. The integrated wasserstein distances of order 2 for the three BSSMs were 0.211, 0.193 and 0.224 respectively. Thus, the fixed time BSSM was always the best model.

Because the fixed time BSSM was the best model, I only presented the model predictions based on it. The predicted mean judgments and mean RTs were generated in the same way as that in the perception-based probability judgment experiment.

Actual mean judgments as a function of predicted mean judgments were shown in Figure 18. It was clear that the BSSM overestimated the mean judgments for the medium and high probability events. This was because the BSSM with a symmetric prior only predicted symmetric conservatism but the actual data showed a pattern of asymmetric conservatism, where even medium probabilities were underestimated.

Figure 18

Actual Mean Judgments as a function of Predicted Mean Judgments

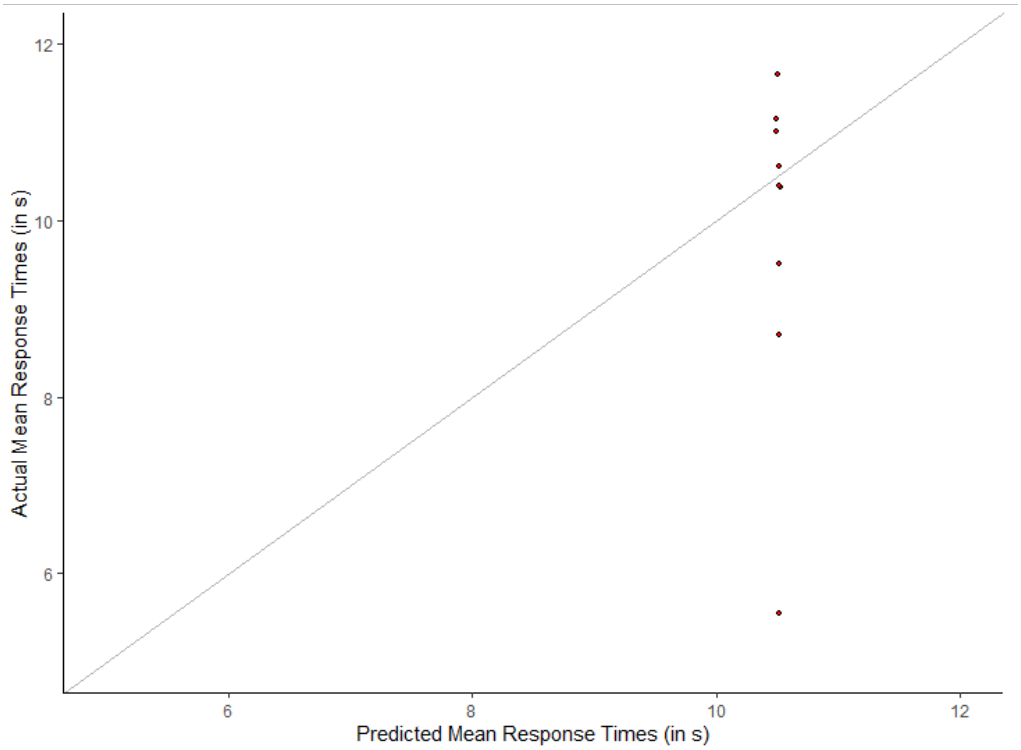


Note. The grey line is identity line.

Actual mean RTs as a function of predicted mean RTs were shown in Figure 19. It was not strange that the model predictions were always the same because it was the fixed time BBSM. The actual mean RTs had a relatively large fluctuation, but they were shown to have nearly no relation with the true percentages. Considering the only predicted RT of the BSSM was roughly at the middle of the actual mean RTs, the BSSM still gave an acceptable prediction in terms of the RTs.

Figure 19

Actual Mean Response Times as a function of Predicted Mean Response Times



Note. The grey line is identity line.

I did not run simple linear models and compare them with the identity line in this section because the observations of the mean judgments and RTs were too few.

Discussion

Changes in the Stopping Rule and the Cognitive Processes

Combining the model results of both perception-based and knowledge-based probability judgment datasets, the fixed time stopping rule seems a general rule when people make probability judgments. Timing is an important and frequently-used skill for survival and goal achieving for humans and animals (Richelle & Lejeune, 1980). And as a simple heuristic, the fixed time rule has been widely tested in fields like animal foraging (Green, 1984), information retrieval (Maxwell et al., 2015) and

perceptual choices (Swensson & Thomas, 1974). This rule can usually perform as well as the optional stopping rules if the environment is not very complicated and it is equivalent to the fixed sample size rule if the speed of searching or sampling is constant. But in the framework of the BSSMs, this rule is different from the fixed sample size rule because the sample size at each time step is a random variable. Among the three stopping rules considered in this article, the fixed time rule should be the most cognitively economical one without considering the difficulty in timing. If checking whether to stop is a separate process after updating the belief at each time step, the fixed time rule users just need to add 1 to the time having spent. But the fixed sample size rule users need to count how many samples collected at this time step and the fixed density rule users even need to calculate the probability density of the posterior mean. And if d_{time} is large enough, simply sampling for a fixed time still ensures that the judgments are based on enough samples, which makes them close to the true probabilities. In the two experiments where the participants should repeatedly give probability judgments for the same or different probabilities, a fixed time rule seems ecologically rational.

However, a counterintuitive fact is that the participants tended to switch to the fixed sample size rule rather than remain to use the fixed time rule with a lower d_{time} under time pressure, when a simple solution to avoid timing out is to finish the sampling within a fixed time. One explanation is that the participants do not only care about speed but also accuracy in this condition (the instructions required them to give judgments as fast as possible but still trying to be accurate). There is no objective

criterion for the participants to evaluate their accuracy because the true percentages are unknown, but they can be aware of their consistency, i.e., how consistent they are when judging the events seeming to have similar probabilities, which is a necessary condition of accuracy. When the EASS of the fixed time BSSM and the fixed sample size BSSM is the same, the fluctuation of the total sample size is smaller for the fixed sample size BSSM. This is because the randomness in the total sample size for the fixed time BSSM is decided by the randomness in the sample size at each time step. But that for the fixed sample size BSSM is decided only by the randomness in the sample size at the last time step and the randomness in the total sample size one step before (which is limited because that size must be smaller than d_{size}). In other words, though a fixed time rule user may get a lot of samples to give a judgment, he or she may also get limited or even no samples after a fixed time. On the contrary, a fixed sample rule user may have some variability in the judgment time, he or she always makes judgments based on samples whose size is larger than a threshold, which ensures higher consistency[12]. Because the time limit in the experiment is tolerant in order to get enough judgments from the participants (more than half participants even never saw the warning message though they still sped up), the variability in the judgment time may be affordable in order to get higher consistency. And some previous experiments also showed that people tended to switch to strategies with

[12] From this perspective, the fixed time rule can be seen as a risky rule and the fixed sample size rule can be seen as a safe rule. Time pressure, to some extent, changes the strategy selection to a risky choice problem and people switch to the fixed sample size rule because of risk aversion. Besides, time pressure may not only change the problem structure but also increase risk aversion (Ben Zur & Breznitz, 1981; Busemeyer, 1993).

higher consistency under time pressure (Olschewski & Rieskamp, 2021; Rice & Trafimow, 2012), though the overall consistency might not decrease because the parameters values of different strategies might also change.

Another explanation is that time pressure hinders the process of timing. Moon and Anderson (2013) found that time pressure in multitask situations caused a too-early bias in timing. And according to Zakay's (1993) attention-based model, judgments under time pressure is essentially a kind of dual-task. The attention assigned to judgment (or sampling) will decrease the performance of timing. If timing becomes more inaccurate under time pressure, the advantages of the fixed time rule are naturally weakened.

Admittedly, the dominance of the fixed sample size BBSM may only be an artifact, because the fixed sample size and the fixed density BSSMs do not consider the variability of the non-judgment time. Thus, the three BSSMs actually differ not only in the stopping rule but also in the representation of the non-judgment time. Considering that the fixed sample size and the fixed time BSSMs are similar in many aspects, some participants using the fixed time rule may be wrongly identified to use the fixed sample size rule.

Compared to the former two conditions, the strategies are more heterogeneous in the difficulty condition, probably because the shorter presentation time also encourages the use of a more robust stopping rule or disturbs the process of timing, though the degree of which is not as strong as that induced by an explicit time limit. It is also possible that an ambiguous environment encourages the diversity of

strategies to increase the probability of success at the group level.

Across three conditions, the fixed density rule is always the most infrequent rule, especially in the speed condition. But in the meantime, this rule is the only optional stopping rule with high statistical efficiency. And this rule is the most "Bayesian" rule because it considers the relative strength of the target hypothesis (i.e., the posterior mean). Why is this rule the most unpopular? As shown in Appendix B, the probability density of the posterior mean is a complex function, which should be much more computationally costly than generating samples. If people use sampling to reduce the computational cost of representing the exact probability distributions, there is no reason to believe that they will conduct complicated calculations of probability densities after getting the samples[13]. This explanation can also account for the change in the number of the fixed density rule users. In the accuracy and the difficulty conditions, some people still use this rule probably because they have enough time and the statistical efficiency of the rule overwhelms the computational cost. But in the speed condition where fast responses are required, the cognitive resources devoted to the judgment process are restricted and thus fewer people insist on this rule.

But similarly, the weakness of the fixed density BSSM may also be an artifact. The parameter values of the fixed sample size and the fixed time BSSMs are estimated by minimizing the absolute differences but those of the fixed density

[13] A subsequent question is how people can get the value of the posterior mean without an integral. The answer can be that people simply "know" the results of the Bayesian updating without explicit calculation. In other words, the judgement process is Bayesian-like, not actually Bayesian.

BSSMs are estimated by minimizing the squared differences. The wasserstein distance may be influenced more by the absolute differences, thus prefers the former two BSSMs. But this view is not very trustworthy because there are not many participants showing the pattern predicted by the fixed density BSSM (i.e., longer RTs for medium judgments).

The manipulations of time pressure and difficulty also change the underlying cognitive processes. In the speed condition, the strength to generate samples and the non-judgment time decrease. The decrease in the strength is within expectation because time pressure typically impairs various cognitive performances like creative cognitive processing (Amabile et al., 2002), working memory capacity (Tohill & Holyoak, 2000) and response inhibition (Endres et al., 2020). The BSSMs assume that the samples are generated by memory retrieval or mental simulations. The decrease in general cognitive abilities should naturally decrease the speed to generate samples. And some SSM-based studies (Heathcote & Love, 2012; Starns et al., 2012) also found that the drift rate (which is like the strength in the BSSMs) decreased under time pressure. As for the decrease in the non-judgment time, though the SSMs typically explain the acceleration under time pressure by the decrease of the threshold, Dambacher and Hübner (2014) showed that time pressure could also decrease the non-decision time through a two-stage SSM. And in the experiment, the participants give judgments by clicking rather than simply pressing buttons. The decrease in motor execution should be captured more easily by the models if the participants indeed accelerate this process under time pressure. The non-change of the ESSA in

the speed condition can be explained as a result of adaptation. Because the time limit is tolerant and accuracy is still stressed, the participants may try to find an optimal ESSA which ensures that they can give judgments with enough samples before the time limit (they actually spend more time sampling because their ability to generate samples decrease).

In the difficulty condition, the participants slightly decrease the non-judgment time and the ESSA. The manipulation of difficulty is realized by decreasing the presentation time of stimuli, but the shorter presentation time may also work as a priming (Lashley, 1951). That is, a shorter presentation time may motivate the participants to give faster responses internally. Ans because it is not an external time limit, it only decreases the non-judgment time but not the strength to generate samples. The participants use fewer samples probably because they do not trust the samples as much as what they do in the other two conditions even if they still have learned the true percentages unconsciously (like what the BSSMs assume). After all, the learning is more difficult in a shorter time. As a result, the influence of the prior is relatively stronger in this condition, which is also reflected in a higher degree of conservatism. In fact, an unpublished study (Yuan, 2020) showed that when people judged the probabilities of complex perceptual events (i.e., the percentage of figures in a specific color and configuration), which was another way to increase difficulty, the judgments for high and low probability events were very similar (both were close to 0.5). This supported that people will use fewer samples to make judgments when the task is difficult. Finally, the smaller ESSA and non-judgment time naturally

explain why the RTs are even shorter in the difficulty condition than those in the speed condition.

Asymmetric Priors and Learning in Repeated Judgments

For knowledge-based probability judgments, an obvious pattern is asymmetric conservatism. Even the medium probabilities which are smaller than 0.5 are underestimated. In the meantime, numerous studies of risky decisions have shown that the crossover of the weighting function and the identity line is lower than 0.5 (e.g., Wu & Gonzalez, 1996; Tversky & Kahneman, 1992). So how can the BSSMs accommodate this pattern? One method is to use asymmetric priors.

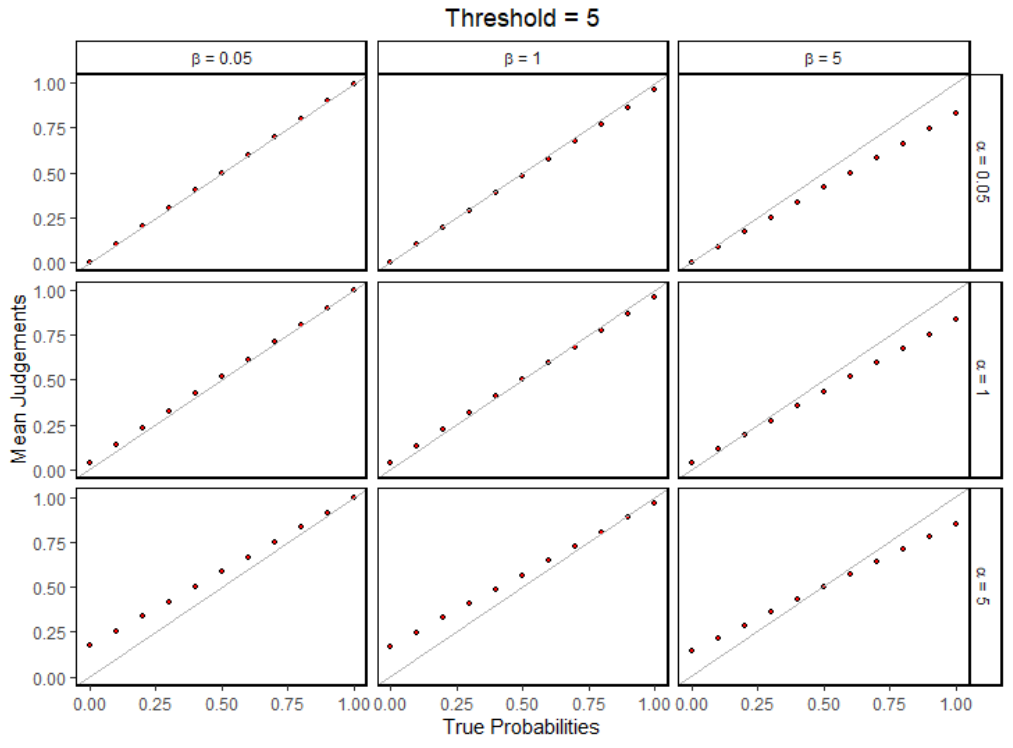
The reason why the original BSM uses a symmetric prior is to mimic a similar model, the probability theory plus noise model (Costello & Watts, 2014). There are some advantages of using a symmetric prior. Firstly, a symmetric prior is "non-informative" in a way when β is small. There is indeed no reason to believe an unfamiliar event will happen with a probability higher or lower than 0.5. Secondly, it ensures that the sum of the mean judgments of two complementary events (i.e., A and $\neg A$) is 1, which is a good statistical property. However, these advantages are relative. When β increases, the prior becomes more and more "informative", which represents a stronger belief that the event happens with a probability of 0.5. And the property of complementarity does not hold for three or more events (e.g., three mutually exclusive events whose probabilities are all 1/3).

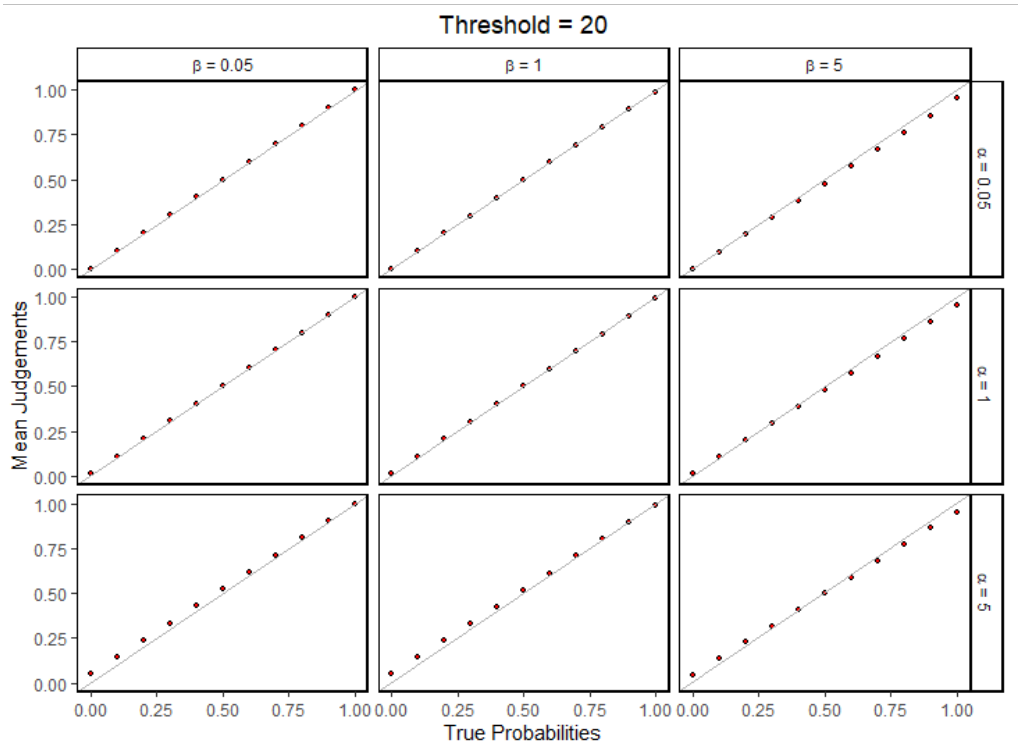
Here I show some simulation results of a fixed time BSSM based on an asymmetric prior Beta(α, β). Suppose $\lambda = 5$ and $T_{non} = 5$, Figure 20 show how the

mean judgments change with α, β and d_{time}. It is clear that the mean judgments are still linear transformations of the true probabilities. But when $\alpha > \beta$, the crossover between the mean judgments-true probabilities line and the identity line is higher than 0.5, and it increases when α was larger. When $\alpha < \beta$, the crossover is lower than 0.5, and it decreases when α is smaller, which reflects the pattern observed in the knowledge-based probability judgment experiment. Besides, the deviations of the mean judgments are still adjusted by the threshold. A higher threshold can make the mean judgments closer to the true probabilities (i.e., a slope closer to 1).

Figure 20

Mean Judgments as a function of True Probabilities for the Fixed Time Model Under Different Combinations of Parameter Values





Note. The grey lines are identity lines.

Thus, asymmetric priors can flexibly reproduce different patterns of conservatism. The results of the fixed sample size BSSM should be similar to those of the fixed time BSSM because the contents of the samples will not influence when to stop based on this rule either. The results for the fixed density model may be more complicated, but the basic patterns should keep unchanged.

But how can asymmetric priors arise? One possibility is that they are the results of adaptive learning. For Bayesian inference, a good property is that the posterior of one test can be the prior of the next test, which means that people can gradually adjust their beliefs with sequential samples, and it is also the mechanism of the BSSMs. Such adjustments may not only happen within one judgment, but also across judgments. In other words, after giving a small probability judgment, the prior mean for the next judgment may also slightly decrease. This view explains why the

medium-low probabilities are also underestimated in the knowledge-based judgments. This is because most probabilities judged in the experiment are low probabilities, which shifts the prior to the 0 end, i.e., decreasing the value of α in the above simulations. Actually, Yuan (2020) found that the probability judgments after judging a series of high probabilities were significantly larger than those after judging a series of low probabilities, which supported the adaptive learning in repeated probability judgments.

Extensions of the Sampling Algorithm and the Stopping Rule

The BSSMs in this article assume that the generation of each sample is independent. But as previously mentioned, the BSMs usually adopt autocorrelated sampling methods like MCMC. In other words, the generation of one sample depends not only on the true probability, but also on one or more samples just generated. Such autocorrelation is widely observed in tasks relevant to mental simulation or memory retrieval (Gilden et al., 1995; Troyer et al., 1997). And Zhu, Sundh, et al. (2021) successfully explained many empirical results of perceptual choices in an autocorrelated BSM framework. Besides, the autocorrelated sampling is even considered in the SSMs. Recall the details of the LBAM. This model assumes that a sample is repeatedly used for each response in a choice, which is the most extreme situation of autocorrelation.

Another possible extension in sampling is amortized sampling, which means reusing samples across judgments. This is an effective way to reduce cognitive loads for computation when the judgments are related and such patterns have already been

observed in both perception-based and knowledge-based judgments (Gershman & Goodman, 2014; Zhu et al., 2019). However, the difficulty of applying it in the BSSMs is that the influence of it may be confounded with that of adaptive learning. After all, the judgments are decided by both prior and samples.

The stopping rules can also be improved. One direction is to add some common components to all three rules. Currently, all rules assume a fixed strength and threshold, but some neurophysiological experiments (Bowman et al., 2012; Thura et al., 2012) suggested that less evidence was needed to make a choice when time passed. This phenomenon can be explained by an urgency signal (i.e., an increasing strength to generate samples) or a collapsing threshold (i.e., a decreasing threshold). The BSSMs with these components may have a new account for the decrease in the RTs in the speed and the difficulty conditions.

Another direction is to modify the rules individually. For the fixed density rule, checking the density of the posterior mean is not a natural Bayesian stopping rule. Firstly, the mean and the mode do not overlap when the Beta distribution is asymmetric, i.e., $\alpha \neq \beta$. Even if the JM uses the mean as an integrated point estimate of the posterior distribution, the probability density of the mode (i.e., the height of the peak) of the posterior is still the most straightforward metric of how many samples having been collected or how strong the posterior is. Secondly, the probability density is not a reliable metric which is not comparable across distributions. From a Bayesian perspective, checking the probability that the belief falls in a specific interval (e.g., (mode - 0.05, mode + 0.05)) is more appropriate

though an extra free parameter representing how wide the interval is seems inevitable. For the fixed sample size and the fixed time rules, they just ignore the influence of the prior on the total sample size. As a measure of previous experience or stereotypes, a stronger prior may decrease the total sample size needed. To accommodate such possible effects, one way is to assume the JM checks the "equivalent total sample size" of the posterior belief (i.e., $\alpha + \beta$ for a posterior belief Beta(α, β)) rather than the actual total sample size, or simply assume d_{size} and d_{time} are decreasing functions of β.

Finally, I make a simplified assumption that people can learn the true probabilities of the events (even unconsciously) when fitting the BSSMs, which is not very realistic, especially for the knowledge-based probability judgments. Future work can let the true probability also be a free parameter if enough judgments are available. It can help to explore whether people can indeed learn the actual states of the world. Besides, current BSSMs use the variability of the total sample size and the samples to explain the variability of judgments for a specific true probability, but a more Bayesian way is to assume that the JM just draws a sample from the posterior belief as the judgment. In other words, sampling is used not only to generate judgments, but also to give judgments.

Conclusion

Bayesian methods provide a reasonable solution for sampling-based models to predict responses of different scales and multiple dimensions. In this article, I proposed the BSSMs for human probability judgments and tested different stopping rules in different conditions. The BSSMs gave relatively accurate predictions of both

judgments and RTs overall. And according to the BSSMs, people tended to use the fixed time rule in general situations but the fixed sample size rule under time pressure. They seemed equally likely to use the two rules when the task became difficult. People's ability to generate samples and the non-judgment time decreased under time pressure. And they used fewer samples and shorter non-judgment time when the task became difficult. The BSSMs reconcile the Bayesian rationality and the cognitive biases in human probability judgments in a sequential sampling framework and provide a new starting point for the sampler account of human cognition.

References

Amabile, T., Mueller, J. S., Simpson, W. B., Hadley, C. N., Kramer, S. J., & Fleming, L. (2002). *Time pressure and creativity in organizations: A longitudinal field study* (No. 02-073). Harvard Business School Working Paper. https://www.hbs.edu/faculty/Pages/item.aspx?num=11879

Audley, R. J., & Pike, A. R. (1965). Some alternative stochastic models of choice. *British Journal of Mathematical and Statistical Psychology*, *18*(2), 207–225. https://doi.org/10.1111/j.2044-8317.1965.tb00342.x

Bakkour, A., Palombo, D. J., Zylberberg, A., Kang, Y. H., Reid, A., Verfaellie, M., Shadlen, M. N., & Shohamy, D. (2019). The hippocampus supports deliberation during value-based decisions. *eLife*, *8*. https://doi.org/10.7554/elife.46080

Balci, F., Freestone, D., & Gallistel, C. R. (2009). Risk assessment in man and mouse. *Proceedings of the National Academy of Sciences*, *106*(7), 2459-2463. https://doi.org/10.1073/pnas.0812709106

Ben Zur, H., & Breznitz, S. J. (1981). The effect of time pressure on risky choice behavior. *Acta Psychologica*, *47*(2), 89-104. https://doi.org/10.1016/0001-6918(81)90001-9

Bowman, N., Kording, K., & Gottfried, J. (2012). Temporal integration of olfactory perceptual evidence in human Orbitofrontal cortex. *Neuron*, *75*(5), 916-927. https://doi.org/10.1016/j.neuron.2012.06.035

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153-178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Brown, S. D., Marley, A. A., Donkin, C., & Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. *Psychological Review*, *115*(2), 396-425. https://doi.org/10.1037/0033-295x.115.2.396

Buhusi, C. V., & Meck, W. H. (2005). What makes us tick? Functional and neural mechanisms of interval timing. *Nature Reviews Neuroscience*, *6*(10), 755-765. https://doi.org/10.1038/nrn1764

Busemeyer, J. R. (1993). Violations of the speed–accuracy tradeoff relation: Decreases in decision accuracy with increases in decision time. In O. Svenson & A. J. Maule (Eds.), *Time pressure and stress in human judgment and decision making* (pp. 181-193). Plenum Press. https://doi.org/10.1007/978-1-4757-6846-6_13

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287-291. https://doi.org/10.1016/j.tics.2006.05.007

Chater, N., Zhu, J., 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. https://doi.org/10.1177/0963721420954801

Costello, F. J. (2009). Fallacies in probability judgments for conjunctions and disjunctions of everyday events. *Journal of Behavioral Decision Making*, *22*(3), 235-251. https://doi.org/10.1002/bdm.623

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463-480. https://doi.org/10.1037/a0037010

Dambacher, M., & Hübner, R. (2014). Time pressure affects the efficiency of perceptual processing in decisions under conflict. *Psychological Research*, *79*(1), 83-94. https://doi.org/10.1007/s00426-014-0542-z

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

Deza, M. M., & Deza, E. (2013). *Encyclopedia of Distances* (2nd ed.). Springer. https://doi.org/10.1007/978-3-642-30958-8

DiCiccio, T., & Efron, B. (1996). Bootstrap Confidence Intervals. *Statistical Science*, *11*(3), 189-212. http://www.jstor.org/stable/2246110

Endres, D. N., Byrne, K. A., Anaraky, R. G., Adesegun, N., Six, S. G., & Tibbett, T. P. (2020). Stop the clock because I can't stop: Time pressure, but not monitoring pressure, impairs response inhibition performance. *Journal of Cognitive Psychology*, *32*(7), 627-644. https://doi.org/10.1080/20445911.2020.1810692

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*(3), 519-527. https://doi.org/10.1037/0033-295x.101.3.519

Fontanesi, L., Gluth, S., Spektor, M. S., & Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin & Review*, *26*(4), 1099-1121. https://doi.org/10.3758/s13423-018-1554-2

Gay, D. M. (1990). *Usage summary for selected optimization routines*. Computing science technical report, AT&T Bell Laboratories. http://netlib.bell-labs.com/cm/cs/cstr/153.pdf

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-6*(6), 721-741. https://doi.org/10.1109/tpami.1984.4767596

Gershman, S. J., & Goodman, N. (2014). Amortized inference in probabilistic reasoning. In M. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual meeting of the cognitive science society* (pp. 517-522). Austin TX: Cognitive Science Society. https://escholarship.org/uc/item/34j1h7k5

Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. *Psychological Review*, *84*(3), 279-325. https://doi.org/10.1037/0033-295x.84.3.279

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684-704. https://doi.org/10.1037/0033-295x.102.4.684

Gilden, D., Thornton, T., & Mallon, M. (1995). 1/F noise in human cognition. *Science*, *267*(5205), 1837-1839. https://doi.org/10.1126/science.7892611

Green, R. F. (1984). Stopping rules for optimal foragers. *The American Naturalist*, *123*(1), 30-43. https://doi.org/10.1086/284184

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357-364. https://doi.org/10.1016/j.tics.2010.05.004

Heathcote, A., & Love, J. (2012). Linear deterministic accumulator models of simple choice. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00292

Heathcote, A., Popiel, S. J., & Mewhort, D. J. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin*, *109*(2), 340-347. https://doi.org/10.1037/0033-2909.109.2.340

Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin*, *138*(2), 211-237. https://doi.org/10.1037/a0025940

Howe, R., & Costello, F. (2017). Probability judgment from samples: accurate estimates and the conjunction fallacy. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th annual conference*

of the cognitive science society (pp. 2224-2229). Austin, TX: Cognitive Science Society. https://cogsci.mindmodeling.org/2017/papers/0424/paper0424.pdf

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American Journal of Psychology*, *62*(4), 498-525. https://doi.org/10.2307/1418556

Khaw, M. W., Stevens, L., & Woodford, M. (2021). Individual differences in the perception of probability. *PLOS Computational Biology*, *17*(4), e1008871. https://doi.org/10.1371/journal.pcbi.1008871

Kvam, P. D., & Turner, B. M. (2021). Reconciling similarity across models of continuous selections. *Psychological Review*, *128*(4), 766-786. https://doi.org/10.1037/rev0000296

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior; the Hixon Symposium* (pp. 112-146). Wiley.

Lieder, F., Griffiths, T. L., M. Huys, Q. J., & Goodman, N. D. (2017). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, *25*(1), 322-349. https://doi.org/10.3758/s13423-017-1286-8

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195070019.001.0001

Maxwell, D., Azzopardi, L., Järvelin, K., & Keskustalo, H. (2015). Searching and stopping: An analysis of stopping rules and strategies. In *Proceedings of the 24th ACM international on conference on information and knowledge management* (pp. 313-322). Association for Computing Machinery. https://doi.org/10.1145/2806416.2806476

McCormack, P. D., & Wright, N. M. (1964). The positive skew observed in reaction time distributions. *Canadian Journal of Psychology*, *18*(1), 43-51. https://doi.org/10.1037/h0083285

Moon, J., & Anderson, J. R. (2013). Timing in multitasking: Memory contamination and time pressure bias. *Cognitive Psychology*, *67*(1-2), 26-54. https://doi.org/10.1016/j.cogpsych.2013.06.001

Olschewski, S., & Rieskamp, J. (2021). Distinguishing three effects of time pressure on risk taking: Choice consistency, risk preference, and strategy selection. *Journal of Behavioral Decision Making*. https://doi.org/10.1002/bdm.2228

Oprisan, S. A., & Buhusi, C. V. (2014). What is all the noise about in interval timing? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1637), 20120459. https://doi.org/10.1098/rstb.2012.0459

Palmer, E. M., Horowitz, T. S., Torralba, A., & Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 58-71. https://doi.org/10.1037/a0020747

Parsopoulos, K. E., & Vrahatis, M. N. (2002). Recent approaches to global optimization problems through particle swarm optimization. *Natural computing*, *1*(2), 235-306. https://doi.org/10.1023/a:1016568309421

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195-203. https://doi.org/10.3758/s13428-018-01193-y

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*(1), 29-46. https://doi.org/10.1037/h0024722

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59-108. https://doi.org/10.1037/0033-295x.85.2.59

Ratcliff, R. (2018). Decision making on spatially continuous scales. *Psychological Review*, *125*(6), 888-935. https://doi.org/10.1037/rev0000117

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873-922. https://doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260-281. https://doi.org/10.1016/j.tics.2016.01.007

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter

variability. *Psychonomic Bulletin & Review*, *9*(3), 438-481. https://doi.org/10.3758/bf03196302

Rice, S., & Trafimow, D. (2012). Time pressure heuristics can improve performance due to increased consistency. *The Journal of General Psychology*, *139*(4), 273-288. https://doi.org/10.1080/00221309.2012.705187

Richelle, M., & Lejeune, H. (1980). *Time in animal behaviour*. Elsevier. https://doi.org/10.1016/C2009-0-00224-4

Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *The Journal of Neuroscience*, *22*(21), 9475-9489. https://doi.org/10.1523/jneurosci.22-21-09475.2002

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 487-494. https://doi.org/10.1016/s0022-5371(70)80091-3

Rüschendorf, L. (1985). The Wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, *70*(1), 117-129. https://doi.org/10.1007/bf00532240

Sanborn, A. N. (2017). Types of approximation for probabilistic cognition: Sampling and variational. *Brain and Cognition*, *112*, 98-101. https://doi.org/10.1016/j.bandc.2015.06.008

Shadlen, M. N., & Newsome, W. T. (1996). Motion perception: Seeing and deciding. *Proceedings of the National Academy of Sciences*, *93*(2), 628-633. https://doi.org/10.1073/pnas.93.2.628

Smith, P. L. (2016). Diffusion theory of decision making in continuous report. *Psychological Review*, *123*(4), 425-451. https://doi.org/10.1037/rev0000023

Smith, P. L., & Ratcliff, R. (2009). An integrated theory of attention and decision making in visual signal detection. *Psychological Review*, *116*(2), 283-317. https://doi.org/10.1037/a0015156

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of zROC slopes with response time data and the diffusion model. *Cognitive Psychology*, *64*(1-2), 1-34. https://doi.org/10.1016/j.cogpsych.2011.10.002

Summerfield, C., & Koechlin, E. (2010). Economic value biases uncertain perceptual choices in the parietal and prefrontal cortices. *Frontiers in Human Neuroscience*, *4*. https://doi.org/10.3389/fnhum.2010.00208

Swensson, R. G., & Thomas, R. (1974). Fixed and optional stopping models for two-choice discrimination times. *Journal of Mathematical Psychology*, *11*(3), 213-236. https://doi.org/10.1016/0022-2496(74)90019-4

Thura, D., Beauregard-Racine, J., Fradet, C., & Cisek, P. (2012). Decision making by urgency gating: Theory and experimental support. *Journal of Neurophysiology*, *108*(11), 2912-2930. https://doi.org/10.1152/jn.01071.2011

Tohill, J. M., & Holyoak, K. J. (2000). The impact of anxiety on analogical reasoning. *Thinking & Reasoning*, *6*(1), 27-40. https://doi.org/10.1080/135467800393911

Troyer, A. K., Moscovitch, M., & Winocur, G. (1997). Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. *Neuropsychology*, *11*(1), 138-146. https://doi.org/10.1037/0894-4105.11.1.138

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293-315. https://doi.org/10.1037/0033-295x.90.4.293

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297-323. https://doi.org/10.1007/bf00122574

Vinding, M. C., Lindeløv, J. K., Xiao, Y., Chan, R. C., & Sørensen, T. A. (2021). Volition in prospective memory: Evidence against differences between free and fixed target events. *Consciousness and Cognition*, *94*, 103175. https://doi.org/10.1016/j.concog.2021.103175

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. https://doi.org/10.1111/cogs.12101

Wagenmakers, E. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of*

Cognitive Psychology, *21*(5), 641-671. https://doi.org/10.1080/09541440802205067

Wagenmakers, E., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*(1), 140-159. https://doi.org/10.1016/j.jml.2007.04.006

Wagenmakers, E., Van Der Maas, H. L., & Grasman, R. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*(1), 3-22. https://doi.org/10.3758/bf03194023

Wu, G., & Gonzalez, R. (1996). Curvature of the probability weighting function. *Management Science*, *42*(12), 1676-1690. https://doi.org/10.1287/mnsc.42.12.1676

Yuan, S. (2020). *Probability judgment and fallacies* [Unpublished master's thesis]. University of Warwick.

Zakay, D. (1993). The impact of time perception processes on decision making under time stress. In O. Svenson & A. J. Maule (Eds.), *Time pressure and stress in human judgment and decision making* (pp. 59-72). Plenum Press. https://doi.org/10.1007/978-1-4757-6846-6_4

Zhang, H., Ren, X., & Maloney, L. T. (2020). The bounded rationality of probability distortion. *Proceedings of the National Academy of Sciences*, *117*(36), 22024-22034. https://doi.org/10.1073/pnas.1922401117

Zhu, J-Q., Leon-Villagra, P., Chater, N., & Sanborn, A. N. (2021, June 25). Understanding the structure of cognitive noise. https://doi.org/10.31234/osf.io/qfyzw

Zhu, J-Q., Newall, P. W., Sundh, J., Chater, N., & Sanborn, A. N. (2021). Clarifying the relationship between coherence and accuracy in probability judgments [Under review].

Zhu, J-Q., Sanborn, A. N., & Chater, N. (2019). Why decisions bias perception: An amortised sequential sampling account. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st annual conference of the cognitive science society* (pp. 3220-3226). Cognitive Science Society. https://cogsci.mindmodeling.org/2019/papers/0540/0540.pdf

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. https://doi.org/10.1037/rev0000190

Zhu, J-Q., Sundh, J., Chater, N., & Sanborn, A. N. (2021). The Autocorrelated Bayesian sampler for estimation, choice, confidence, and response times. https://doi.org/10.31234/osf.io/3qxf7

Appendix A

Choice of the Non-Judgment Time Distribution

For the fixed time BSSM, the time step number for the judgment process is fixed, which equals d_{time}. Thus, a positively skewed discrete distribution for the non-judgment time step number is necessary to explain the variability of the RTs. Besides, to keep the number of parameters the same across three models, this distribution should be controlled by one parameter.

I used the experimental data of an unpublished study (Zhu, Leon-Villagra, et al., 2021) to decide on the distribution. In this between-subjects experiment, the participants' main task was to give a tap when they felt that a fixed period of time had passed after the beginning of each trial and the RTs were recorded. There were three conditions only differing in the duration to be reproduced, which were 1/3s, 1s and 3s respectively. It was natural to decompose the RTs in this task into two parts. One for the inner timing process (i.e., a fixed time process) and the other for the non-timing process. Similarly, the non-timing time step number should follow a positively skewed discrete distribution after discretizing the RTs to explain the RT variability.

I used the RT data of 30 participants who have finished at least 500 trials, 10 participants in each condition, to test different distributions. Before I discretized the RTs, I found that the variability of the RTs increased when the duration to be reproduced increased (the within-condition standard deviations of the RTs were 0.083s, 0.465s and 1.248s respectively). Actually, this phenomenon that the timing noise was proportional to the target time was widely observed across different species

and tasks (e.g., Buhusi & Meck, 2005; Gibbon, 1977; Oprisan & Buhusi, 2014). Thus, it was improper to use a common time step length for three conditions. Because the mean and the standard deviation of a discrete distribution controlled by a single parameter were usually positively correlated like the Poisson distribution. A common time step length would indicate that the JM used more time steps for the non-timing process when the duration to be reproduced was longer, which was unrealistic. Thus, I used time step lengths proportional to the durations to be reproduced. Specifically, the time step lengths for the three conditions were 1/60s, 1/20s and 3/20s respectively.

After discretizing the RTs, I fitted a simple model to the data at the individual level. The model assumed that the total time step number was the sum of the timing process time step number and the non-timing process time step number. The former time step number was fixed and represented by an integer parameter ranging from 1 to the minimum observed total time step number minus 1. The latter time step number followed a discrete distribution whose possible values were all positive integers.

Considering that many previous RT models used multi-parameter continuous distributions to describe the RTs (e.g., Heathcote et al., 1991; Luce, 1986; Palmer et al., 2011) and I only found two positively skewed discrete distributions properly controlled by one parameter (i.e., the Poisson distribution and the Geometric distribution), I used some discretized continuous distributions where only one parameter was free as some of my candidate distributions. The discretization method

was simple: For a continuous distribution whose cumulative distribution function was $F(x)$ ($x > 0$), the discretized version of it specified that P(non-timing time step number = T) = $F(T) - F(T - 1)$ (T was an positive integer). As for the two inherently discrete distributions, I used the shifted versions of them because their original versions had some probability mass at 0 (I assumed that there should be at least 1 time step for the non-timing process). That was, if the probability mass function of a discrete distribution was $f(x)$ (x was a non-negative integer), the shifted version of it specified that P(non-timing time step number = T) = $f(T - 1)$ (T was a positive integer). 9 discrete or discretized distributions were finally tested, which were the shifted Poisson distribution, the shifted Geometric distribution, the discretized Exponential distribution, the discretized Chi-squared distribution, the discretized Rayleigh distribution (i.e., the Weibull distribution with shape parameter = 2), the discretized Log-norm distribution with log standard deviation = 1, the discretized Log-logistic distribution with shape parameter = 1, the discretized Gamma distribution with scale parameter = 1 and the discretized Inverse gaussian distribution with dispersion parameter = 1.

The value of the parameter that maximized the log-likelihood function was searched using the PORT routines with 10 random starting points. For each participant, I compared the models assuming different non-timing time step number distributions and chose the distribution whose corresponding model had the largest maximum log-likelihood value as the best distribution.

For 26 participants, the best distribution was the shifted Poisson distribution.

For 3 participants, the best distribution was the discretized Rayleigh distribution. For the rest 1 participant, the best distribution was the discretized Exponential distribution or the shifted Geometric distribution (the two distributions were equivalent because they were both memoryless). Thus, the shifted Poisson distribution showed an obvious advantage in describing the time step number of the non-timing process.

Appendix B

Anti-Conservatism Produced by Bayesian Sequential Sampler Models

For the fixed density BSSM, anti-conservatism is a stable phenomenon even with enough observations when the prior, the threshold and the strength are all low. This is caused by an important characteristic of the Beta distribution. That is, the probability density of the mean is higher when the absolute difference between the two parameters is larger while the sum of them is fixed.

Specifically, for a Beta distribution Beta(α, β), the mean of it is $\frac{\alpha}{\alpha+\beta}$, and its probability density is

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} * \left(\frac{\alpha}{\alpha+\beta}\right)^{\alpha-1} * \left(\frac{\beta}{\alpha+\beta}\right)^{\beta-1} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} * \frac{\alpha^{\alpha-1} * \beta^{\beta-1}}{(\alpha+\beta)^{\alpha+\beta-2}},$$

Where Γ is the gamma function. Suppose $\alpha + \beta = x$, where x is a fixed value. The probability density can be regarded as a function of α, where $\beta = x - \alpha$. Denote this function by $f(\alpha)$. Then

$$\frac{d}{d\alpha} ln(f(\alpha)) = \frac{\alpha-1}{\alpha} + ln(\alpha) - \psi(\alpha) - [\frac{\beta-1}{\beta} + ln(\beta) - \psi(\beta)],$$

where ψ is the digamma function. Because the monotonicity of $f(\alpha)$ and that of $ln(f(\alpha))$ is the same, I only need to prove the above function is smaller than 0 when $0 < \alpha < \frac{x}{2}$, but larger than 0 when $\frac{x}{2} < \alpha < x$ in order to prove that the probability density will decrease when the absolute difference between α and β decreases. Suppose $g(\alpha) = \frac{\alpha-1}{\alpha} + ln(\alpha) - \psi(\alpha)$, the above function can be rewritten as $g(\alpha) - g(x - \alpha)$. Thus, the problem is transformed to proving that $g(\alpha)$ is an increasing function when $0 < \alpha < x$.

In the meantime, the weierstrass definition of the gamma function is

$\Gamma(y) = \frac{e^{-\gamma y}}{y}\prod_{n=1}^{\infty}(1+\frac{y}{n})^{-1}e^{y/n}$,

where γ is the Euler-Mascheroni constant. Then

$\psi(y) = \frac{d}{dy}\ ln\big(\Gamma(y)\big) = -\gamma - \frac{1}{y} + \sum_{n=1}^{\infty}\left(\frac{1}{n} - \frac{1}{y+n}\right) = -\gamma + \sum_{n=1}^{\infty}\left(\frac{1}{n} - \frac{1}{y+n-1}\right)$.

Thus,

$\frac{d}{d\alpha}g(\alpha) = \frac{1}{\alpha} + \frac{1}{\alpha^2} - \sum_{n=1}^{\infty}\frac{1}{(n+\alpha-1)^2} = \frac{1}{\alpha} - \sum_{n=1}^{\infty}\frac{1}{(n+\alpha)^2} > \frac{1}{\alpha} - \int_0^{\infty}\frac{1}{(t+\alpha)^2}dt = 0$,

which supports that $g(\alpha)$ is an increasing function when $0 < \alpha < x$. And the characteristic is proven.

However, when the observations are limited, even the fixed sample size and the fixed time BSSMs can reproduce anti-conservatism when the prior, the threshold and the strength are all low. I will show this with two examples.

For the fixed sample size BSSM, suppose β = 0.05, λ = 2, d_{size} = 10 and T_{non} = 5 (T_{non} is an irrelevant parameter with an arbitrary value). I ran 100000 sets of simulations based on the parameter values for different true probabilities. Each set contained 25 simulated judgments and time step numbers, which was the maximum number of the judgments a participant could provide for a specific true percentage in the perception-based probability judgment experiment. For each simulation set, I calculated the mean of the judgments. If it was smaller than the true probability and the true probability was smaller than 0.5, or it was larger than the true probability and the true probability was larger than 0.5, the set was labelled as an anti-conservatism set. Relative frequencies of the anti-conservatism sets under different true probabilities were shown in Table 4. The relative frequencies were close to 0.5 under nearly all true probabilities. And the relative frequency of the anti-conservatism sets

slightly increased when the true probability was closer to 0.5. This was because the samples were noisier based on the Bernoulli distribution when the true probability was closer to 0.5.

Table 4

Relative Frequencies of the Anti-Conservatism sets for the Fixed Sample Size Model Under Different True Probabilities

True probability	Relative frequency(%)
0.1	43.234
0.2	46.033
0.3	47.505
0.4	48.803
0.6	48.994
0.7	47.767
0.8	46.169
0.9	43.41

For the fixed time BSSM, suppose $\beta = 0.05$, $\lambda = 2$, $d_{time} = 5$ and $T_{non} = 5$. The simulation procedures were the same. Relative frequencies of the anti-conservatism sets under different true probabilities were shown in Table 5. The basic patterns kept unchanged, which meant that the occurrences of conservatism and anti-conservatism were nearly equally likely in such situations.

Table 5

Relative Frequencies of the Anti-Conservatism sets for the Fixed Time Model Under Different True Probabilities

True probability	Relative frequency(%)
0.1	42.361
0.2	45.917
0.3	47.244
0.4	48.840
0.6	48.737
0.7	47.247
0.8	46.037
0.9	42.481

Thus, the pattern of anti-conservatism was not diagnostic to distinguish different BSSMs with limited observations.

Appendix C

Parameter Recovery Test

To conduct the simulation-based parameter estimation, the most important thing is to find a metric which can properly measure the differences between the simulated results and the actual data. In other words, the parameter estimation based on this metric should accurately recover the parameter values from simulated data of the model.

I compared two families of metrics commonly used in the simulation-based parameter estimation. The first family were distribution-based metrics (Deza & Deza, 2013), which could directly compare two distributions. There were two kinds of distribution-based metrics. The first were 2-dimensional (2d) distribution-based metrics. Such metrics could directly measure the similarity between two joint distributions of judgments and STSNs. There were six 2d distribution-based metrics tested: 2d wasserstein distance of order 1, 2d wasserstein distance of order 2, 2d hellinger distance, 2d symmetric chi-squared distance, 2d L^1 distance and 2d L^2 distance. The second were 1-dimensional (1d) distribution-based metrics. Such metrics could only measure the similarity between two univariate distributions of judgments or STSNs. Thus, I needed a method to combine the 1d distance between the judgment distributions and that between the STSN distributions to get an integrated distance. I tried two possible methods. One was adding them together. The other was multiplying them together. There were nine 1d distribution-based metrics tested: 1d wasserstein distance of order 1, 1d wasserstein distance of order 2, 1d

hellinger distance, 1d symmetric chi-squared distance, 1d L^1 distance, 1d L^2 distance, 1d minkowski distance of order 1, 1d minkowski distance of order 2 and 1d intersection distance.

The second family were statistics-based metrics. Such metrics used the differences in the statistics between two distributions to represent their similarity. There were two kinds of statistics commonly used to describe a distribution: statistics of central tendency and statistics of dispersed tendency. For the first kind, the members tested were the mean (arithmetic, geometric, root and harmonic) and the median. For the second kind, the members tested were the standard deviation and the interquantile range. To reduce the computational cost, I first chose the proper statistics of central tendency and dispersed tendency for each variable and tested the methods to combine the differences based on them afterwards.

If the statistics could help to identify the parameter values behind the data, it should be both robust and unique. The robustness meant that for different sets of simulations generated from the same combination of parameter values, the calculated statistics should be stable across simulation sets. The uniqueness meant that for different sets of simulations generated from different combinations of parameter values, the calculated statistics should be "diverse enough" across simulation sets. To get an integrated indicator which combined the robustness and the uniqueness of the statistics, I conducted the following procedures.

1. For a specific combination of parameter values, generate 25 simulations (which was the maximum number of the judgments a participant can provide for a

specific true percentage in the perception-based probability judgment experiment). Calculate the target statistics for the target distribution (judgments or STSNs). Repeat the above operations 100 times and calculate the standard deviation of the target statistics.

2. Repeat step 1 for all possible combinations of parameter values and calculate the mean of the standard deviations. This was the robustness indicator.
3. For a specific combination of parameter values, generate 5000 simulations. Calculate the target statistics for the target distribution.
4. Repeat step 2 for all possible combinations of parameter values and calculate the standard deviation of the target statistics. This was the uniqueness indicator.
5. Divide the uniqueness indicator by the robustness indicator to get an integrated indicator. The statistics performed better when the indicator was larger.

The possible combinations of parameter values were the same as those in the Model Features section in the main text. The results showed that for different BSSMs, different means could describe the central tendency for both judgment and STSN distributions equally well, which were better than the median. Similarly, the standard deviation could better describe the dispersed tendency for both judgment and STSN distributions than the interquantile range. Considering the arithmetic mean and the standard deviation had a favourable mathematical relation, they were finally used to generate the statistics-based metrics.

After deciding the statistics, the next step was to decide how to measure the differences in the statistics between the simulated results and the actual data. I tested

two differences: absolute difference and squared difference. The final problem was how to combine the differences within judgments and STSNs and how to combine them together. Because there was no reason to assume that the differences were combined in one way within judgments but another way within STSNs, I tested four combination methods: adding the two internal differences and adding the sums together (sum + sum), adding the two internal differences and multiplying the sums together (sum + product), multiplying the two internal differences and adding the products together (product + sum) and multiplying the internal two differences and multiplying the products together (product + product). There were 8 statistics-based metrics tested: absolute + sum + sum, absolute + sum + product, absolute + product + sum, absolute + product + product, squared + sum + sum, squared + sum + product, squared + product + sum and squared + product + product.

To conduct the parameter recovery test, I first generated some simulated data as the “actual data”. specifically, for each combination of parameter values (not including the true probability here because it was not estimated in the main text), I generate 25 simulated pairs of judgment and STSN based on a low, medium-low, medium-high and high true probability respectively (the classification of the true probabilities was the same as that in the main text). And the parameter recovery performance was calculated as follows.

1. For each dataset, estimate the parameter values based on a specific metric according to the procedures in the Parameter Estimation section. To reduce the computational cost, the optimization was not realized by the particle swarm

method, but by randomly choosing 100 points in the parameter space and using the point which produced the smallest difference between the simulation results and the "actual data". Calculate the absolute difference between the estimate and the actual value and divide it by the actual value for each parameter (this was because the meanings of different parameters were different). Average the results to get the integrated relative error.

2. Repeat step 1 for each data set. After deleting the extreme values (this was because the rough optimization method could sometimes produce incredibly high integrated relative errors), average the integrated relative errors to get the average relative error. The metric performed better when this error was smaller.

Relative errors based on different metrics were illustrated in Table 6. 2d wasserstein distance with order 1 or 2 was the best distribution-based metric overall. The statistics-based metrics were typically better than the distribution-based metrics. For the fixed sample size and the fixed time BSSMs, the best metric should be absolute + product + sum. For the fixed density BSSM, the best metric should be squared + product + sum.

Table 6

Average Relative Errors Based on Different Metrics for Different Models

Metric	Relative error (%)		
	Fixed sample size	Fixed time	Fixed density
distribution-based			
2d wasserstein^1	79.394	85.264	65.121

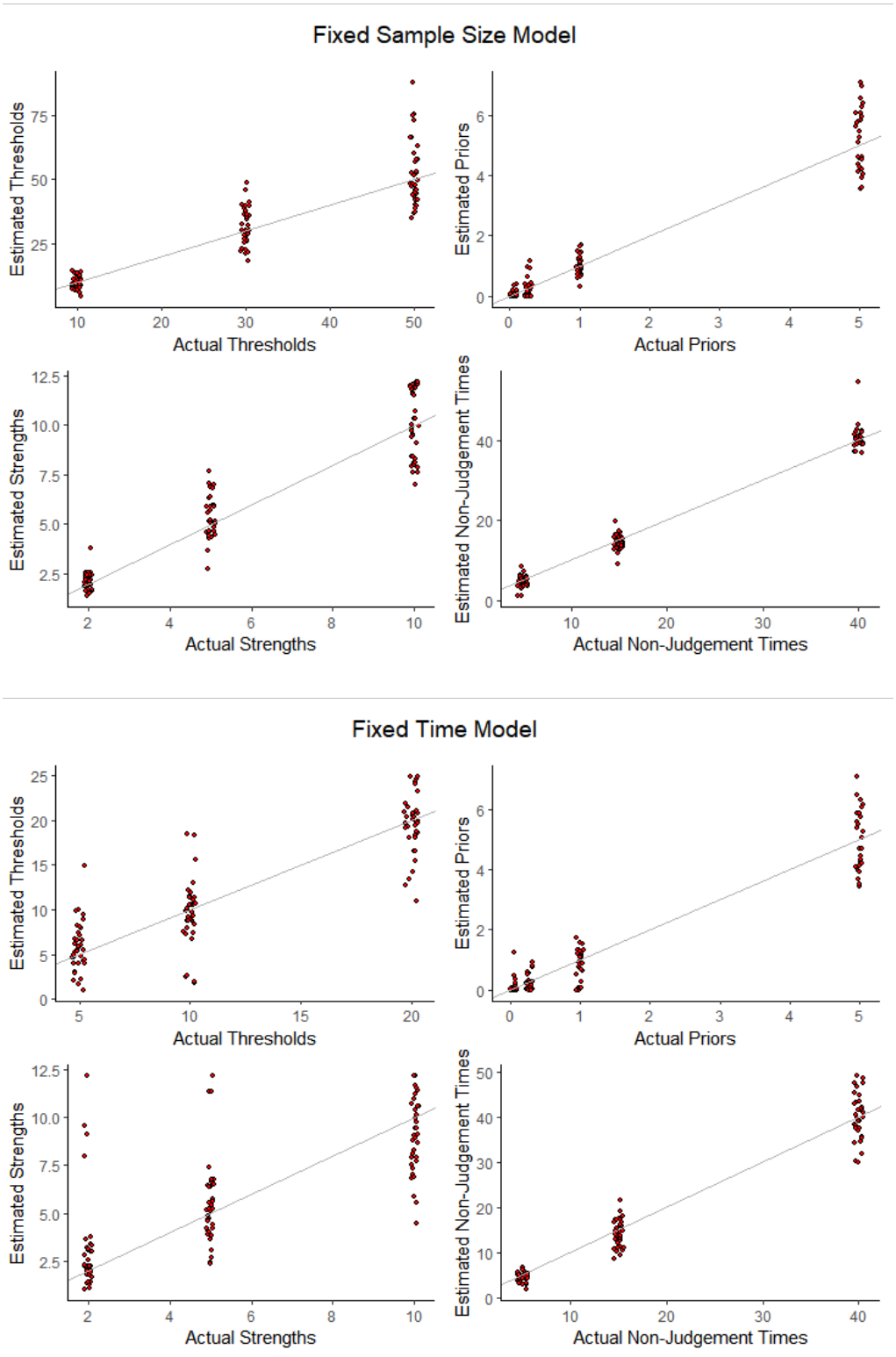
2d wasserstein^2	90.651	67.777	76.352
2d L^1	154.884	141.367	257.965
2d L^2	217.201	74.759	383.496
2d hellinger	150.148	94.413	152.534
2d chi-squared	209.696	107.897	144.410
1d wasserstien^1 (sum)	426.730	364.259	309.806
1d wasserstien^2 (product)	154.242	237.718	252.104
1d L^1 (sum)	106.381	65.032	96.224
1d L^1 (product)	114.776	75.227	83.958
1d L^2 (sum)	186.060	101.370	123.590
1d L^2 (product)	139.390	86.322	105.671
1d hellinger (sum)	107.546	63.278	76.865
1d hellinger (product)	87.118	102.182	75.248
1d chi-squared (sum)	93.648	75.513	76.159
1d chi-squared (product)	104.105	119.157	74.370
1d minkowski^1 (sum)	97.328	78.489	99.502
1d minkowski^1 (product)	125.283	80.293	93.778
1d minkowski^2 (sum)	113.673	73.600	92.657
1d minkowski^2 (product)	150.691	78.412	73.375
1d intersection (sum)	114.708	79.494	81.404
1d intersection (product)	111.327	81.505	94.527

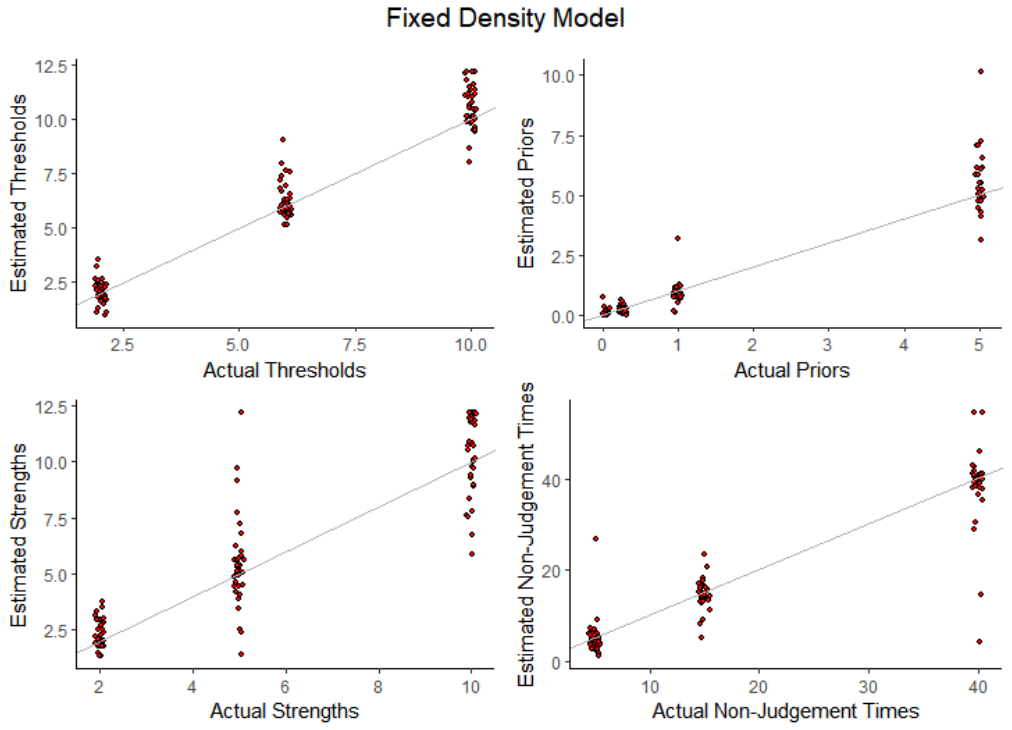
statistics-based			
absolute + sum + sum	82.416	74.855	64.285
absolute + sum + product	62.077	79.374	56.412
absolute + product + sum	53.381	66.526	52.469
absolute + product + product	73.206	68.051	54.943
squared + sum + sum	60.807	82.834	59.309
squared + sum + product	96.299	78.215	51.494
squared + product + sum	54.477	83.593	47.407
squared + product + product	64.979	86.635	51.512

To better explore the parameter recovery performance, I estimated the parameters based on the best metric for each model using the particle swarm method with 3 random starting points and maximum iteration times of 30 (This was due to the computational cost. The optimization in the main text should be more accurate with more starting points and larger maximum iteration times). Parameter estimates as a function of actual parameter values were illustrated in Figure 21. The parameter estimation relatively accurately recovered the actual parameter values, though some distortions were inevitable (e.g., $\beta = 0.01$ or 0.001 were nearly the same for each BSSM because the minimum unit of samples was one, which was already much larger than the prior).

Figure 21

Parameter Estimates as a Function of Actual Values for Different Models





Note. Points are jittered to prevent overlapping. The grey lines are identity lines.