

RESEARCH ARTICLE

Dissecting Bayes: Using influence measures to test normative use of probability density information derived from a sample

Keiji Ota^{1,2,3,4*}, Laurence T. Maloney^{1,2}

1 Department of Psychology, New York University, New York, New York, United States, **2** Center for Neural Science, New York University, New York, New York, United States, **3** Institute of Cognitive Neuroscience, University College London, London, United Kingdom, **4** Department of Psychology, School of Biological and Behavioural Sciences, Queen Mary University of London, London, United Kingdom

* keiji.ota@nyu.edu

OPEN ACCESS

Citation: Ota K, Maloney LT (2024) Dissecting Bayes: Using influence measures to test normative use of probability density information derived from a sample. *PLoS Comput Biol* 20(5): e1011999. <https://doi.org/10.1371/journal.pcbi.1011999>

Editor: Stefano Palminteri, Ecole Normale Supérieure, FRANCE

Received: June 5, 2023

Accepted: March 14, 2024

Published: May 1, 2024

Copyright: © 2024 Ota, Maloney. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data frame and the codes used to generate the figures are available at <https://github.com/keijiota/Dissecting-Bayes>.

Funding: This research was supported by a Guggenheim Fellowship awarded to LTM. KO received a salary (post-doctoral research fellowship) by Grant-Aid for JSPS Fellows No. 17J07822. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Bayesian decision theory (BDT) is frequently used to model normative performance in perceptual, motor, and cognitive decision tasks where the possible outcomes of actions are associated with rewards or penalties. The resulting normative models specify how decision makers should encode and combine information about uncertainty and value—step by step—in order to maximize their expected reward. When prior, likelihood, and posterior are probabilities, the Bayesian computation requires only simple arithmetic operations: addition, etc. We focus on visual cognitive tasks where Bayesian computations are carried out not on probabilities but on (1) *probability density functions* and (2) these probability density functions are derived from *samples*. We break the BDT model into a series of computations and test human ability to carry out each of these computations in isolation. We test three necessary properties of normative use of pdf information derived from a sample—*accuracy*, *additivity* and *influence*. Influence measures allow us to assess how much weight *each point* in the sample is assigned in making decisions and allow us to compare normative use (weighting) of samples to actual, point by point. We find that human decision makers violate accuracy and additivity systematically but that the cost of failure in accuracy or additivity would be minor in common decision tasks. However, a comparison of measured influence for each sample point with normative influence measures demonstrates that the individual's use of sample information is markedly different from the predictions of BDT. We will show that the normative BDT model takes into account the geometric symmetries of the pdf while the human decision maker does not. An alternative model basing decisions on a single extreme sample point provided a better account for participants' data than the normative BDT model.

Author summary

Bayesian decision theory (BDT) is used to model human performance in tasks where the decision maker must compensate for uncertainty in order to gain rewards and avoid losses. BDT prescribes how the decision maker can combine available data, prior

knowledge, and value to reach a decision maximizing expected winnings. Do human decision makers actually use BDT in making decisions? Researchers typically compare overall human performance (total winnings or overall percent correct) to the predictions of BDT but we cannot conclude that BDT is an adequate model for human performance based on just overall performance. We break BDT down into elementary operations and test human ability to execute such operations. In two of the tests human performance deviated only slightly (but systematically) from the predictions of BDT. In the third test, we use a novel method to measure the *influence* of each sample point provided to the human decision maker and compare it to the influence predicted by BDT. When we look at what human decision makers do—in detail—we find that they use sensory information very differently from what the normative BDT decision maker does. We advance an alternative non-Bayesian model that better predicts human performance. We propose that influence measures are a more sensitive way to discover discrepancies between human and optimal performance, than comparing overall performance.

Introduction

Bayesian Decision Theory [1–4] is used to model decision and action selection in a wide variety of experimental tasks (perception: [5,6]; visual estimation: [7–9]; movement planning: [10–14]; motor learning: [15]; obstacle avoidance: [16]; eye-hand coordination: [17]; information sampling: [18]; temporal order judgment: [19]). The BDT model allows us to compare actual human performance against normative performance maximizing expected value. The pattern of deviations between human and normative gives us insight into human cognition, perception and motor planning.

But demonstrating that *overall* human decision making performance (amount of reward earned) approaches that of a normative BDT decision maker does not prove that human decision makers are carrying out the Bayesian computations in detail. Other decision rules (‘heuristics’) can mimic Bayesian performance arbitrarily closely [20–24]. To fully test the claim that humans are carrying out Bayesian computations, we must also look at performance in detail as we do here.

An advantage of BDT is that it applies to perceptual, motor, and cognitive tasks where uncertainty is captured by the *probability density function* (pdf) of a continuous random variable. A representative task is shown in Fig 1A and 1B. The decision maker may choose to make a speeded reaching movement to any aim point on a display screen. There is an irregular, white target region, T . If the decision maker hits within the region they receive a monetary reward, otherwise nothing. The end point $E = (E^x, E^y)$ where the decision maker touches the screen will differ from the aim point $A = (A^x, A^y)$ because of the motor uncertainty inherent in speeded movements.

Where should the decision maker aim? Because the movement is speeded, the outcome of the movement is distributed as a continuous random variable with *probability density function* (pdf) $\phi(E^x, E^y | A)$. This *population pdf* in reaching tasks is typically close to bivariate Gaussian [10,11,25]. In Fig 1A, a pdf is plotted as a heat map, the aim point as a red diamond, and the target region is in white. The pdf is not itself a probability but it serves to assign a probability that the next reaching movement toward the aim point would end within T :

$$P[T | A] = \iint_T \Phi(E^x, E^y | A) dE^x dE^y \quad [1]$$

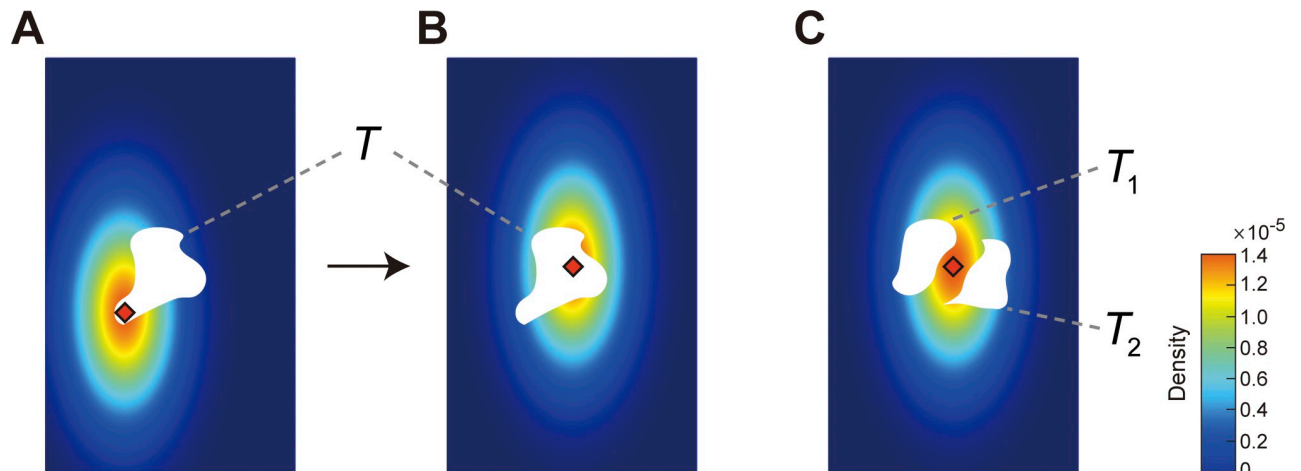


Fig 1. Applications of Bayesian decision theory. **A.** The decision maker is rewarded for making a speeded reaching that terminates in the white target region T . They may choose any aim point (red diamond). Because of motor uncertainty his actual end point $E = (E^x, E^y)$ is distributed as a bivariate Gaussian centred on the aim point, represented here as a heat map. **B.** The decision maker shifts his aim point and the bivariate Gaussian distribution shifts with respect to the aim point. The probability of hitting the target is larger in the panel B than in the panel A. **C.** The target region is divided into two disjoint regions, T_1 and T_2 . Touching either target earns a reward. The decision maker's aim point is shown in red. Bayesian decision theory allows us to calculate the aim point that maximizes the probability of reward [10,11,25].

<https://doi.org/10.1371/journal.pcbi.1011999.g001>

In Fig 1B, we plot the target and pdf for an alternative aim point, also marked in red. Small shifts of the aim point are equivalent to rigidly translating the pdf. We assume that small shifts of aim point do not alter the covariance of the Gaussian distribution of end points. See, for example, Trommershäuser et al. [10,11,25].

The probability that the decision maker hits the target is plausibly higher in Fig 1B than in Fig 1A. In maximizing the probability of hitting the target (Eq 1) the decision maker must in effect choose not just between these two aim points but among all possible aim points. The normative BDT decision maker—maximizing expected value—selects the aim point that maximizes Eq (1) for any choice of target region [1,10,11,25].

Accuracy. We refer to the decision maker's ability to correctly evaluate Eq (1) for any target as **accuracy**. The decision maker's estimates of probability can be accurate for all possible targets only to the extent that they have an accurate estimate of the Gaussian pdf [20] or, equivalently the parameters, that determine the Gaussian pdf, the population mean $\mu = (\mu_x, \mu_y)$ and the population covariance $\Sigma = (\sigma_x^2, \sigma_y^2, \rho)$. Intuitively, the population mean, μ , specifies the location of the pdf on the plane and the population covariance, Σ , its size and "shape". The probabilities for any target all together completely determine the pdf (Riesz-Fischer Theorem). We can in principle reconstruct the pdf from the target probabilities [20].

Additivity. A second task is illustrated in Fig 1C. The target region is now divided into two disjoint regions T_1 and T_2 (i.e. $T_1 \cap T_2 = \emptyset$). A reaching movement ending in *either* region earns the same, fixed reward and, as the regions are disjoint, the normative BDT decision maker should seek to maximize

$$P[T|A] = P[T_1|A] + P[T_2|A] \quad [2]$$

where $T = T_1 \cup T_2$. We refer to the decision maker's ability to correctly evaluate Eq (2) as **additivity**. In the first part of this article we compare human estimation with pdfs in simple tasks to that of normative BDT.

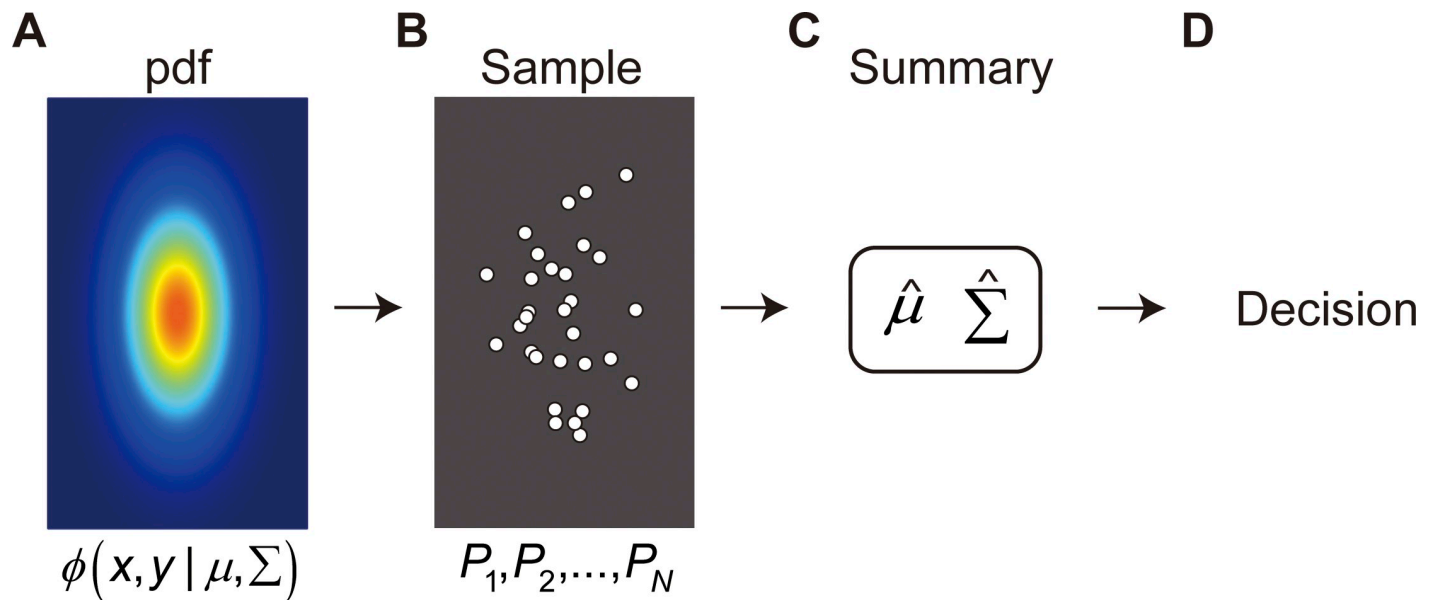


Fig 2. Parametric decision making based on a sample. A. Bivariate Gaussian PDF (referred to as the "population"). The population PDF is not known to the decision maker. B. The decision maker is given only a sample P_1, \dots, P_N of size N drawn from the Gaussian. C. The Gaussian parametric decision maker reduces a large number of sample values to the values of a small number of parameters. For the bivariate Gaussian, the sample is often reduced to five parameters that are estimates of population parameters $\hat{\mu}, \hat{\Sigma}$ (referred to as "statistics"). D. The normative decision maker then makes decisions based only on these statistics, ignoring any "accidental" structure in the sample not captured by the parameters. For convenience in presentation we assume throughout that the Gaussian pdf is elongated (anisotropic) so that there are exactly two orthogonal axes of symmetry. The excluded possibility is that the Gaussian is isotropic (circularly symmetric). All the Gaussian pdfs used in the experiment were anisotropic, vertically elongated.

<https://doi.org/10.1371/journal.pcbi.1011999.g002>

Decisions based on samples

In many tasks the decision maker does not know the exact population pdf (Fig 2A). The decision maker is instead given a random sample from the population pdf (P_1, P_2, \dots, P_N) as shown in Fig 2B. The sample may be based on prior experience, often provided in an explicit training phase [1]. The first question we address is, what information from the sample is retained and used by the decision maker in making any decision? How are the different points in the sample combined to reach a decision? We refer to this information as the **summary** (Fig 2C). The summary is the information about the pdf available to the decision maker and any decision is based solely on this summary (Fig 2D).

Non-parametric approaches. At one extreme the decision maker could retain the entire sample in every detail. The summary is then the sample. If the sample size were large enough then the decision maker could use normative non-parametric approaches such as resampling [26] that assume nothing about the population pdf. For example, they could estimate Eq (1) by simply calculating the proportion of sample points that fall within the region T . As sample size increases, this value converges in probability to the correct answer as a consequence of the weak law of large numbers [27]. Here, though, we will work only with small sample sizes (5 or 30 points) where approaches based on resampling would lead to implausible predictions of human behavior. With a sample of size five and a target that is small compared to the inter-point spacings in the sample, for example, the only possible probabilities would be 0.0, 0.2, . . . , 1.0 and most would be 0.0 (no point within the target) or 0.1 (one point within the target). See S2 Fig. We consider instead parametric approaches [28].

Parametric approaches. In parametric estimation we restrict the choice of pdf's to a family of pdf's summarized by a small number of parameter values corresponding to the selected

pdf. Intuitively, the parametric decision maker “knows” more than just the sample. In Fig 2A and 2B they know that the pdf is unimodal and symmetric about two axes. For the bivariate Gaussian, in particular, the decision maker could replace the sample of size N shown in Fig 2B by the *summary* in Fig 2C comprising the sample mean (\bar{x}, \bar{y}) , the sample covariance (s_x^2, s_y^2, s_{xy}) and the sample size N . When the underlying pdf is Gaussian, these sample statistics are the corrected maximum likelihood estimates of the population mean and covariance $\hat{\mu}, \hat{\Sigma}$. The estimates $\hat{\mu}, \hat{\Sigma}$ are *jointly sufficient statistics* [4,29] that capture all of the data relevant to estimating the parameters of population pdf. Not every pdf family has jointly sufficient statistics [4,29].

Any invertible transformation of the estimates $\hat{\mu}, \hat{\Sigma}$ are also *jointly sufficient statistics* [4,29] and replacing $\hat{\mu}, \hat{\Sigma}$ by the transformed parameters would lead to the same conclusions as those we reach. We refer to the model decision maker in Fig 2 applied to any judgment or estimation as the *normative BDT* decision maker for that task (the “normative decision maker” or “normative model” when context permits). We emphasize that this normative model is representative of a broad class of equally normative models based on jointly sufficient statistics.

Summarizing the sample by $\hat{\mu}, \hat{\Sigma}$ reduces a potentially large number of sample values to just five numbers, a remarkable degree of data compression, made possible by the parametric assumption that the sample is drawn from a bivariate Gaussian distribution. The decision maker can base his decision on this summary with no loss.

There are four advantages to basing all decisions in any task on a small number of parameters. The first, most obvious, is *parsimony*. The full complexity of the sample is replaced by a handful of estimated parameters. Second, the calculation of parameters (Fig 2C) does not depend on the task (Fig 2D). We use the same summary for all tasks. The third advantage is that the summary ignores *accidental structure* in the sample that provides no useful information about the underlying population. In Fig 2B, for example, the cluster of five points at the bottom of the sample is an accident of sampling. If we sampled again, we are unlikely to encounter a similar cluster. The cluster is visually salient but it would be a mistake to give more weight (or less weight) to the points in these clusters simply because they form an apparent cluster. In particular, the Gaussian pdf is symmetric about its mean in the vertical direction and this symmetry implies that the illusory 5-point cluster would be as likely to appear near the top of the sample—reflected about a horizontal line—as at the bottom where it is.

The last advantage is that knowledge of the form of the pdf can potentially improve any non-parametric decision or estimate based solely on a sample. The parametric decision maker knows more than just the sample. Knowing, for example, that a pdf is bimodal even without knowing the locations of the modes is one example. Knowledge of the evident vertical and horizontal symmetry of the bivariate Gaussian pdf in Fig 2B is potentially useful in making decisions.

We will compare human performance to the normative Gaussian BDT decision maker sketched in Fig 2A–2D. To compare human performance to normative BDT in detail, we measure the *influence* of each point P in the sample on the decision maker’s choices and on the choices of the normative model. Historically, influence measures were extensively used in the theory of robust estimation in mathematical statistics [30,31] and in research concerning depth cue combination [32,33]. We sketch here how we measure influence experimentally. The details are to be found in the Methods section.

Influence measures. In Fig 3A, a bivariate Gaussian sample is shown in white circles. The underlying population pdf is drawn as a heat map and lightly sketched as elliptical contours of equal probability density. The decision maker’s hypothetical task is to estimate the probability

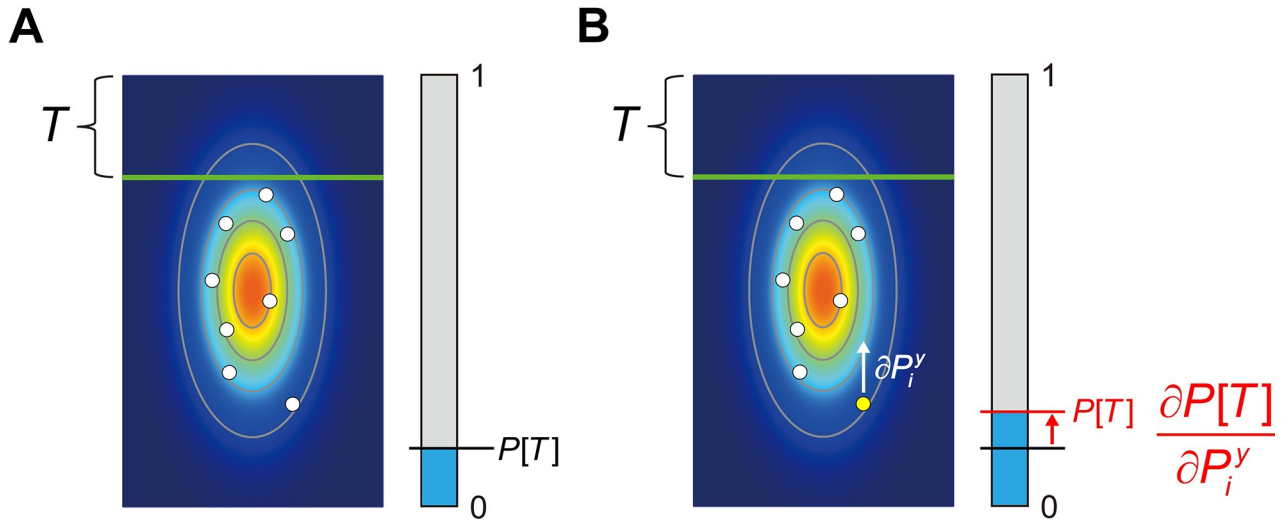


Fig 3. Measuring influence. **A.** A hypothetical experiment. A sample is drawn from a bivariate Gaussian pdf marked by a heat map and contours of equal probability density. The blue bar represents the decision maker’s estimate of the probability that an additional point drawn from the same underlying pdf will be in the region above the green line T . The precise task is not important. **B.** Measuring influence. The (vertical) influence of one point in the sample can in principle be measured by perturbing it slightly in the vertical direction and measuring the effect of the perturbation on the decision maker’s estimate $P[T]$. The ratio of the change in estimate to the magnitude of perturbation is the *influence* of the point on the setting. We do not use this method (single point perturbation) but instead use a method based on linear regression. See *Methods*. The influence measures allow us to characterize how each point in the sample affects decision-making.

<https://doi.org/10.1371/journal.pcbi.1011999.g003>

$P[T]$ that an additional point drawn from the same pdf will be above the green line, a region marked by T . We wish to evaluate how each point in the sample enters into this estimate.

The decision maker’s estimate of the probability $P[T]$ is shown in a blue/grey scale in Fig 3A. Now suppose we alter the sample by shifting a point P slightly upward. In Fig 3B, the change is shown as a vertical white arrow and it is exaggerated in size to make it visible. The small change in P may result in a shift of the estimate and we are in effect computing a numerical estimate of the gradient of partial derivatives of probability estimate with respect to the vertical direction $\partial P[T]/\partial P_i^y$, the influence in the vertical direction of the point. We could similarly estimate the influence of the point in the horizontal direction but for the remainder of this article we focus on influence in the vertical direction only; for simplicity we will leave out the y -superscript in all following equations.

Intuitively, we could imagine grasping each point in turn and “wiggling” it up and down to see how $P[T]$ changes. If the influence is 0, for example, then we would conclude that the point P is not used by the visual system in computing $P[T]$. We will estimate the decision maker’s **empirical influence** $\hat{I}(P) = \partial \hat{P}[T]/\partial P$ by regression analysis (see *Results*).

For any normative model, we can also compute the **normative influence** $I(P) = \partial P[T]/\partial P$ for any task we choose analytically, by numerical differentiation or by Monte Carlo simulation. Influence is a signed measure and to compare empirical influence and normative influence we form the ratio

$$\Phi(P) = \frac{\hat{I}(P)}{I(P)} \tag{3}$$

the **influence ratio** for any given point P . If the decision makes use of the points as the normative model does, the influence ratios will all be 1. The influence ratio gives an indication of which points have too great or too little influence in absolute magnitude compared to normative. If the influence ratio is negative then the influence measure for the human decision

maker is of the opposite sign to that of the normative BDT decision maker. We could measure it for each of the five points in the cluster we discussed in Fig 2.

To summarize, in this article, we test three properties of normative use of pdf's derived from samples: **accuracy** (Eq 1), **additivity** (Eq 2), and **influence** (Eq 3). We anticipate that participants will fail to match the normative BDT model perfectly. Our primary interest is in *patterned* deviations from normative. Such patterns might indicate what participants are actually doing in carrying out the task. Influence measures will allow us to seek such patterns. The influence task is potentially a powerful tool that could be combined with many other tasks including our first estimation task, accuracy and additivity. We chose to test the three properties in separate experiments to improve the readability of the resulting article.

Results

Testing accuracy and additivity

Participants first completed interval estimation tasks designed to test both accuracy and additivity. The participants were shown a sample of 5 or 30 points from an anisotropic bivariate Gaussian distribution (Fig 4A). The population means of the bivariate Gaussian distribution was fixed on a center of the screen, whereas the population covariance (the "size" and "shape" of the Gaussian pdf) randomly changed with each trial.

After a fixed interval (1 sec), two green horizontal lines across the distribution of the white points appeared (Fig 4B). The participants were asked to judge the probability that a new point drawn from the same population mean and population covariance would fall within the region delimited by the two lines. Accurate performance in this interval estimation task was equivalent to integrating the probability density of the Gaussian distribution within the specified region (Eq 1). We compared participants' estimates to the correct estimates (a test of *accuracy*).

The participants recorded their probability from 0% to 100% by moving a digitized pen horizontally (Fig 4B). After their judgment, 10 new, yellow points drawn from the same distribution that generated the sample were presented for 2.5 sec. The new points fell inside or outside the horizontal lines giving participants performance feedback concerning their probability estimates.

On every trial, we selected a new Gaussian distribution with a randomly chosen variance. We generated samples from these population distributions. The participants, during the estimation tasks, became familiar with the range of possible variances of the population distributions from which samples were drawn. Therefore, they had considerable opportunity to observe typical Gaussian samples.

The horizontal lines appeared in any of three configurations with respect to the center of the screen. The lines covered a symmetric interval S , the upper half of the symmetric interval SU , or the lower half of the symmetric interval SL (Fig 4C). The two half regions are non-overlapping and together form a symmetric region. These triples allowed a test of *additivity*: $P[S] = P[SU] + P[SL]$. We varied the interval width to make nine probability conditions equally spaced from 0.1 to 0.9 for the symmetric interval $P[S]$ and from 0.05 to 0.45 for the asymmetric intervals $P[SU]$ and $P[SL]$. The participants saw all three types of intervals on different trials. They repeated each condition 5 times, resulting in 270 (2 sample sizes \times 3 configurations \times 9 probabilities \times 5 times) trials in total.

Test of accuracy. We tested whether the participant's estimates of all intervals S, SU, SL were accurate. Before doing so, we confirmed that their estimates did not change across trials (S1 Fig). We then averaged the data over the five trials for each probability condition. We also

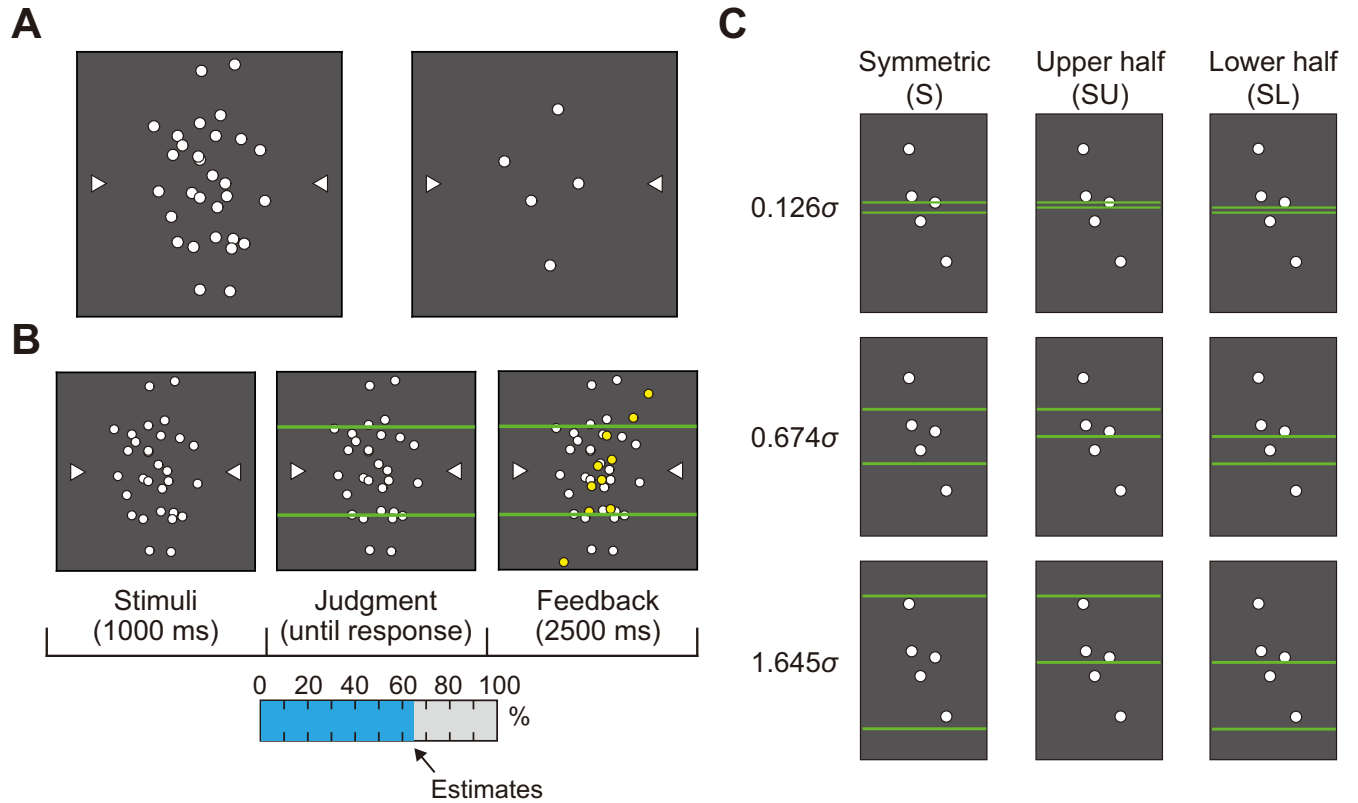


Fig 4. Design of the interval estimation tasks. **A.** Examples of 30-point and 5-point samples. The horizontal and vertical coordinates of the points are independent random variables drawn from a bivariate Gaussian distribution. **B.** The trial sequence of the interval estimation tasks. The sample appears and then two horizontal lines. Participants judged the probability that an additional sample from the same distribution would fall into the region delimited by the horizontal lines. **C.** Three configurations of interval estimation with respect to the center of the screen. The upper half and lower half intervals are non-overlapping; their set-theoretic union is the symmetric interval. The vertical interval distances are expressed as in proportions of σ , the standard deviation of the population pdf in the vertical direction. The vertical distances 0.126σ , 0.674σ , and 1.645σ correspond with 10%, 50%, and 90% probabilities in the symmetric interval, respectively and 5%, 25%, and 45% probabilities in the upper half and the lower half intervals, respectively.

<https://doi.org/10.1371/journal.pcbi.1011999.g004>

confirmed that the participants did not take a non-parametric approach which counts how many points in the sample fall within the interval marked by horizontal lines (S2 Fig).

Fig 5A illustrates the mean estimates for a symmetric interval $\hat{P}[S]$ (averaged across participants and number of trials) against the correct probability $P[S]$ of 30-point and 5-point, respectively. The observed deviation of the estimates for the symmetric interval is similar to the patterns of distortion in decision under risk [34–36]: participants systematically overestimated the probability induced by small and medium intervals.

There are several models used to model distortions in the estimates of probability [34–36]. Zhang and Maloney [37] used linear transformations of the log odds of probability. The linear in log odds model (LLO) is defined by the implicit equation

$$Lo(\hat{P}[S]) = \gamma Lo(P[S]) + (1 - \gamma) Lo(p_0) \tag{4}$$

where $Lo(P[S]) = \log \frac{P[S]}{1-P[S]}$ is the log odds [38] or logit function [39]. We write the equation in terms of the symmetric interval S for convenience. It applies equally to SU and SL as well. Eq (4) has two free parameters. The parameter γ is the slope of linear transformation of the log

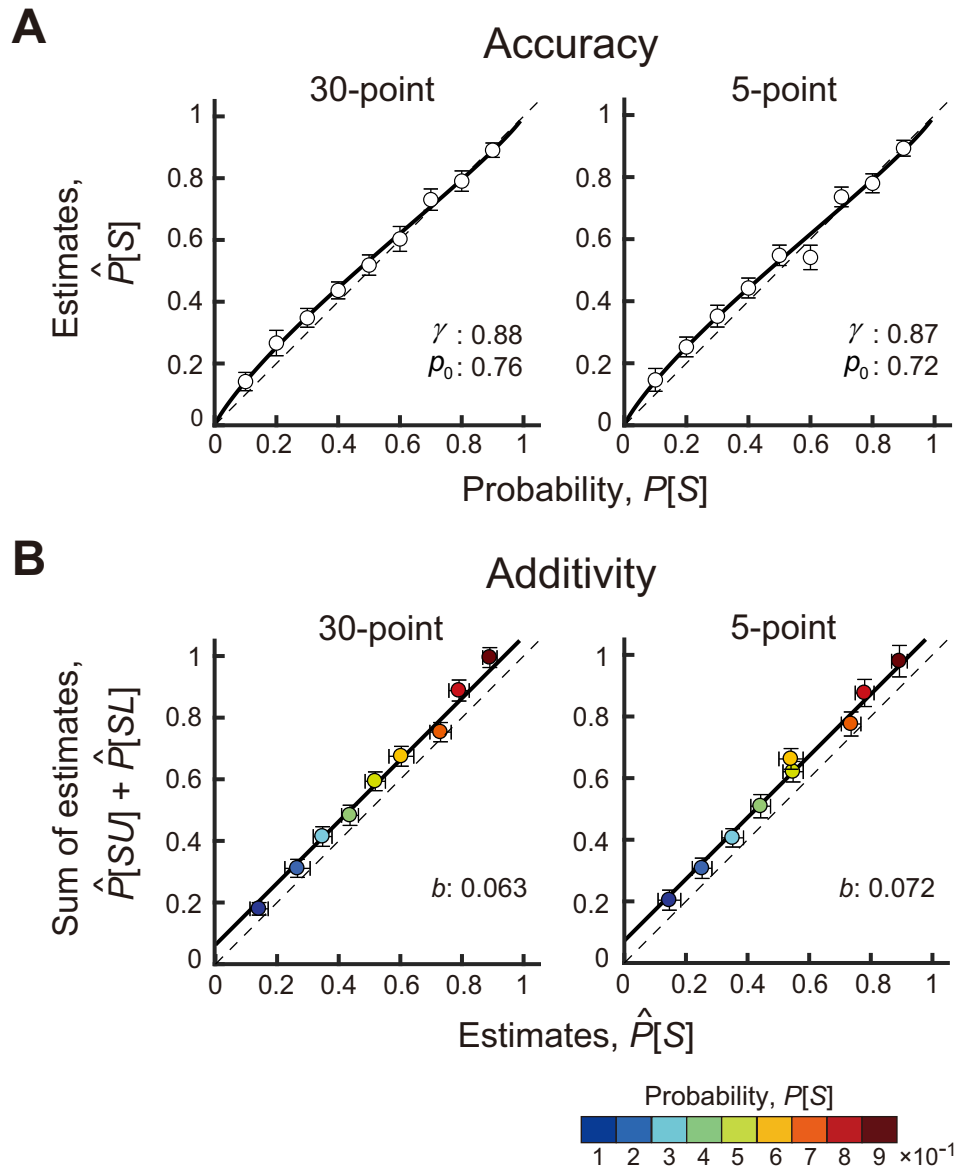


Fig 5. Results of the interval estimation tasks. **A.** Accuracy. The participant’s mean estimates of probability in the symmetric interval $\hat{P}[S]$ are plotted against the correct probabilities $P[S]$. The black thick curve is the maximum likelihood estimate of a linear-in-log-odds function fitted to the data. See text. **B.** Additivity. The sum of estimates in the upper and lower halves of the symmetric interval $\hat{P}[SU] + \hat{P}[SL]$ is plotted against the estimates of the symmetric interval $\hat{P}[S]$. The color scale of the circle indicates the correct probability between 0.1 and 0.9. The black thick line is the best-fit estimate by a super additive function. For A & B, data are averaged across the participants, and the error bars indicate ± 2 s.e.m.

<https://doi.org/10.1371/journal.pcbi.1011999.g005>

odds of the probability and p_0 is the "crossover point" where the probability distortion function crosses the identity line (see Zhang & Maloney [37] for details).

We also fit other probability distortion functions taken from [36] and [35] and found that the LLO function provided consistently better fits to the data (S1 Text). See [37,40] and Zhang et al [41] for additional discussion of these models and their near equivalence.

We fit the LLO Model to the mean estimates across participants (Fig 5A) using maximum-likelihood estimation. The thick curves in Fig 5A show the estimated LLO functions. The null hypothesis of no distortion is (in terms of the LLO parameters) $\gamma = 1$ with p_0 set to any value. S1 Table summarizes the results of AICc model comparisons. We found that the LLO function was 629 times (in the 30-point condition) and 2.2 times (in the 5-point condition) more likely than the null hypothesis of no distortion. In brief, there was considerable support for probability distortion in the 30-point condition but not in the 5-point condition.

The estimated parameter values ($\gamma = 0.88$ and $p_0 = 0.76$ in the 30-point condition, $\gamma = 0.87$ and $p_0 = 0.72$ in the 5-point condition) indicate that the participants overestimated small probabilities. The probability weighting function found in the decision under risk literature typically has a value $\gamma = 0.5\sim 0.8$ [42] and the cross-over point where the probability distortion curve crosses the identity line is found to be around $0.3\sim 0.4$. For instance, the γ parameter of Tversky & Kahneman [36] is 0.61 for gains and 0.69 for losses. This results in a slightly more concave function below the cross-over point (overestimating small probabilities) and more convex function above the cross-over point (underestimating moderate to large probabilities). Our results show less probability distortion in this respect than economic decision tasks in Tversky & Kahneman [36] but similar values are found in visuo-motor and visuo-cognitive tasks ($\gamma = 0.7\sim 0.9$ and $p_0 = 0.4\sim 0.7$ in Zhang & Maloney [37]). The estimated values of γ we find are only roughly consistent with the decision under risk literature; the estimated values of p_0 are markedly larger. The individual plots for the estimates for each participant are available in S3 Fig.

We repeated the analysis for *SU* and *SL*, the two halves the of symmetric interval. Probabilities were overestimated (S4 Fig). The LLO model fit the estimates of *SU* and *SL* best and the recovered values of γ and p_0 were similar to those in the estimates of the symmetric interval (S1 Table). To summarize, we found distortion (“overestimation”) in the probability estimates based on the sample when the induced probability was small to medium; the estimates were close to accurate for values of probability near 1. The pattern of distortion was similar regardless of the number of samples provided to the participants or whether the participant judged the full interval or one of the half intervals *SU* or *SL*.

Test of additivity. The participants overall misestimated the probability associated with target regions: we next test whether they can accurately sum these erroneous probabilities across disjoint target regions or whether they make an additional error, failing additivity. In Fig 5B, we plotted the sum of the mean estimates (across the participants and trials) in the two disjoint regions $\hat{P}[SU] + \hat{P}[SL]$ against the mean estimates in the single symmetric region $\hat{P}[S]$. We know that these estimates are distorted but we wish to test whether the participant’s estimates of the sum of the participant’s own estimates are systematically sub-additive or super-additive. The individual plot shows super-additivity for many participants (but not all; S5 Fig). We tested for failures of additivity consistent with the model:

$$\hat{P}[SU] + \hat{P}[SL] = \hat{P}[S] + b \quad [5]$$

The null hypothesis of additivity is (in terms of the parameters) $b = 0$. A failure of additivity can be super-additive ($b > 0$) or sub-additive ($b < 0$).

We fit the three models to the mean estimates across participants (Fig 5B), and found that the super-additive model outperformed the other models (S2 Table). The super-additive model was 1660 times and 9336 times more likely than the null hypothesis of additivity for the 30-point and 5-point conditions, respectively.

The thick curves in Fig 5B show the super-additivity functions obtained by the fit for the 30 point and 5 point conditions. On average, the sum of $\hat{P}[SU]$ and $\hat{P}[SL]$ was 6.3% larger than the estimates of $\hat{P}[S]$ in the 30-point condition and 7.2% in the 5-point condition.

Effects of probability distortions and super-additivity on movement planning tasks

The results showed that the participants have highly patterned failures in both tests of accuracy and additivity but these failures are small in magnitude. We might ask, how would the observed failures affect typical movement planning tasks? We examine how the pattern of probability distortion we find would have affected performance in analogues of two experimental tasks reported in the BDT literature.

In Fig 6A, we show a red penalty region with -100 points and a green reward region with +100 points from one of the experiments by Trommershäuser et al [10]. Participants attempt 100 movements (black circles) toward an aim point (red diamond). We set the aim point shown in Fig 6A to maximize the expected reward of a typical participant given his motor variance. The objective probabilities of hitting within the red region, red and green region, green region, and outside of both regions are 0.002, 0.035, 0.879, and 0.084, respectively. The resulting expected gain is 87.31 points. If a hypothetical participant misestimates probabilities in a fashion consistent with our results, those probabilities would be transformed into 0.005, 0.058, 0.868, and 0.123, respectively, by the observed linear log-odds function in Fig 5A. An overestimation of the probability of hitting the red region along with an underestimation of the probability of hitting the green region reduces the expected gain to 85.64 points (a loss of 1.91%). The effect on the optimal aim point is slight. A green diamond shows the optimal aim point taking the observed LLO function into account and it almost completely overlaps with the red diamond. The failures of accuracy we observe would have essentially no effect on human performance.

In Fig 6B, we consider a variant on a task used by Zhang et al [22] in their experiment 1. Participants were asked to complete a two-alternative, forced-choice task where one of the target options comprised three disjoint smaller rectangles (hitting anywhere in any of the three rectangles earned the reward) and the other was a larger single rectangle. We simplify their task to have two equally sized disjoint targets (Fig 6B; left). The objective chance of hitting the double target is 0.534. We scale the width of a larger single target so that the objective chances become equal. If a hypothetical participant is super-additive, the probability of hitting the double target would inflate to 0.597. The estimates for the single target is 0.564 based on the LLO probability distortion function. While the normative BDT decision maker picks the double and single targets equally often, human decision maker with failures of accuracy and additivity will pick the double target more often than normative. The single target would have to increase in width by 14.7% (from 4.453 to 5.107 mm; this increase can be seen in a light red border) to restore indifference between the targets.

The failures we find in the test of accuracy and additivity are small but patterned and have only a modest effect on the expected value in the tasks considered. These results are also roughly consistent with those of Zhang et al [41]. They used Shannon information lost due to probability distortion and found that people pick parameters values for their model that nearly maximized information transmitted in going from external objective values of probabilities to internal estimates.

Measuring influence

We designed a separate task to measure influence, the decision task. As in the interval estimation tasks, a sample of white points ($N = 5$ or 30) is drawn from an anisotropic bivariate

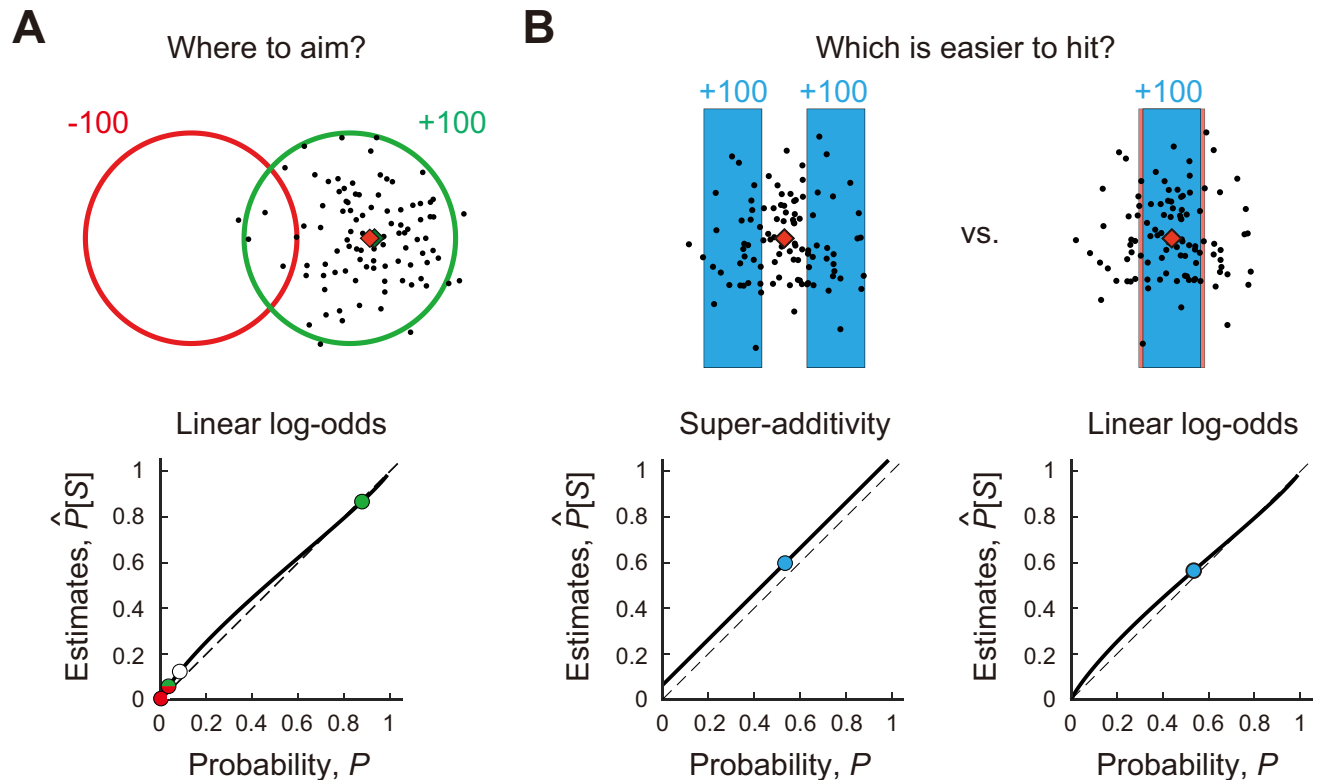


Fig 6. Hypothetical costs due to failures of accuracy and additivity in two previous experiments. **A.** A stimulus from Trommershäuser et al. [10]. We chose the median stimulus from their experiment. In their task, the participant made speeded reaching movements to the reward region (green circle). The red circle denotes the penalty region. The distance between the two circles is 1.5 times the radius of the circle. The radii of circles were 8.97 mm. A touch within the green region earns +100, within the red, -100, and within the green and red, 0. Hitting outside of both regions earns nothing. Black circles denote a possible isotropic bivariate Gaussian distribution of end points around the aim point (SD 3.89 mm, the average SD in Trommershäuser et al [10]). Given the standard deviation of the bivariate Gaussian distribution, the optimal aim point maximizing the expected reward was calculated and is shown as a red diamond. The objective probabilities of hitting each region with possible end points are plotted against the subjective probabilities as circles. We consider a hypothetical participant who overestimates small probabilities by the linear-in-log-odds function in Fig 6A (30 points) with $\gamma = 0.88, p_0 = 0.76$. The probability distortion slightly shifts the optimal aim point (with the new aim point shown as a green diamond almost completely covered by the red diamond). **B.** A stimulus similar to the stimuli in Experiment 1 used by Zhang et al [22]. They used a two-alternative forced-choice task. One of the options was a large, single rectangle target and the other comprised three disjoint smaller rectangles. To simplify our example, we replace the triple target with a double target. Hitting in either colored bar of the double target earned a full reward. Participants decided which target (single or double) to attempt to hit and made speeded reaching movements to the center of the chosen target. Hitting within the rectangle earned the same reward. The standard deviation of the reaching movement was chosen to be 3.05 mm (the average of the participants' measured SDs in experiment 1 in Zhang et al [22]). The widths of the two rectangles are 1.5 times the SD and the gap between two rectangles is 0.75 times the width of that rectangle. These widths and gap correspond to a median value of the targets used in Zhang et al [22]. The heights of the rectangles are set so that the virtual participant's end points do not fall outside the vertical boundaries. The width of the single rectangle is adjusted so that the objective probability of hitting the single target is the same as that of hitting the double target. The normative decision maker would pick each target 50% of the time. As a consequence of distortion of probability and super-additive, the decision maker instead picks the double target more often. If the single target is slightly increased in width by 14.7% (shown in a light red border), the decision maker would pick them equally often though his chances of hitting the single target are objectively greater.

<https://doi.org/10.1371/journal.pcbi.1011999.g006>

Gaussian distribution (Fig 7). The population covariance randomly changed with each trial. In this task, the participants could rigidly move (translate) the visible sample to any location on the screen.

The participants first set the sample to an initial position (blue square) to start the trial. Next, a green penalty boundary was shown (180 mm above the start). The participants then decided where to set the sample on the screen (Fig 7). We recorded the final vertical coordinate of the digitized pen as the participant's set point in the trial. The sample could be moved to the left or the right but such horizontal movement had no effect on the reward or penalty incurred. The design was factorial: two sample sizes crossed with two values of penalty. The trials from

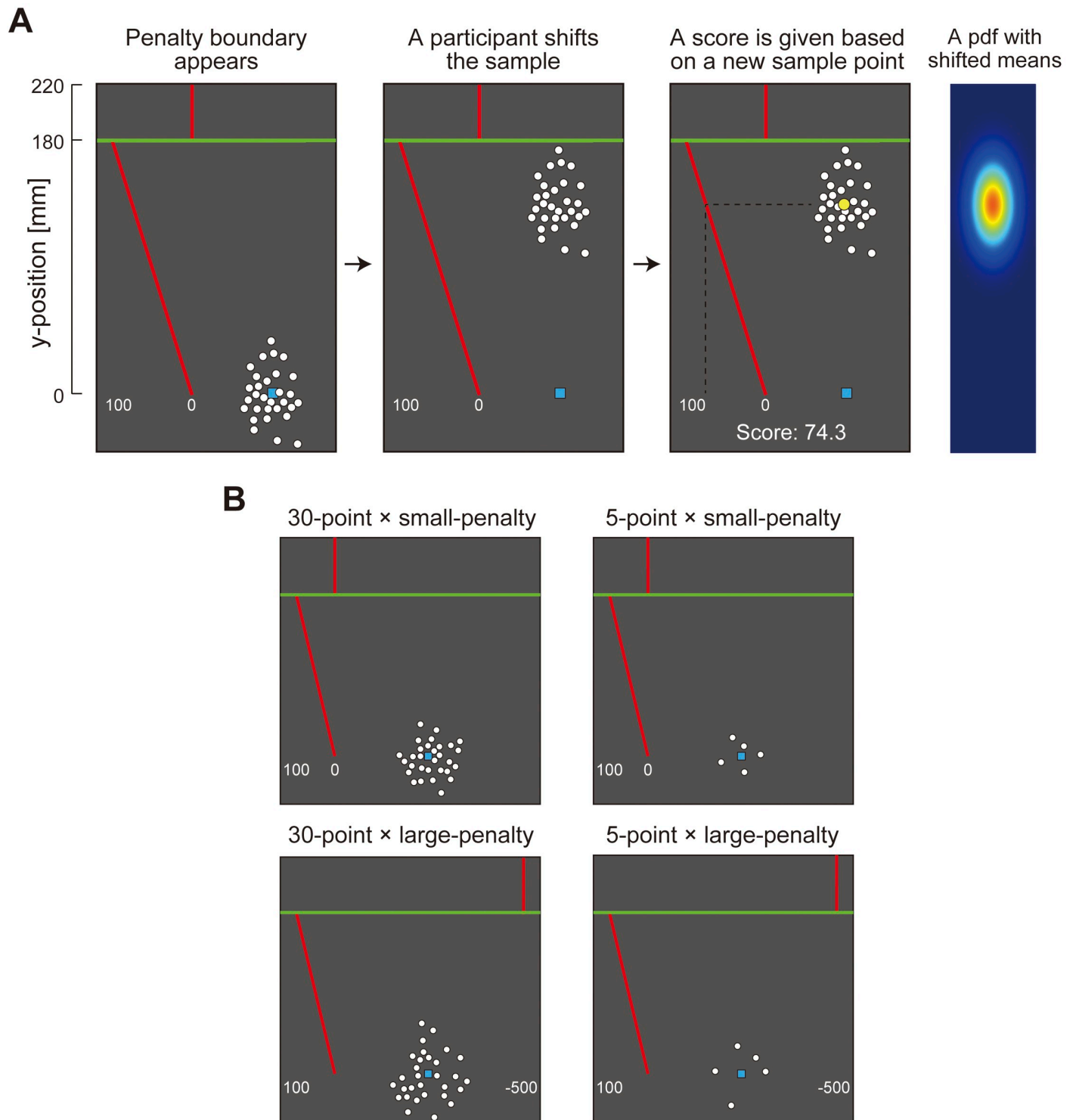


Fig 7. Decision task. (A) On each trial, a sample of 30 or 5 points was drawn from an invisible bivariate Gaussian distribution and shown on a visual display. A participant could rigidly shift the sample up and down from the starting point marked by a blue square. In moving the sample a participant also shifted the invisible pdf of the underlying distribution. After a participant set the location of the sample and its underlying pdf, one yellow point was drawn from the shifted distribution. A heat map illustrates a bivariate Gaussian distribution which underlies the sample after it is shifted from the starting point. The horizontal green line is the penalty boundary. If the new point appeared above the penalty boundary, a participant incurred a penalty, accompanied by an aversive sound. There were two penalty conditions, 0 and -500. If the yellow point fell on or below the green line and above the blue square, a participant received a reward proportional to the distance from the blue square of the starting point to the yellow point. If the yellow point fell at or below the blue square a participant received nothing. The rewards ranged from 0 (at or below the blue

square) to 100 points (just below the green line). Once the additional yellow point exceeded the green boundary line, the rewards fell to 0 points or -500 points. The red slanted line is a plot of reward as a function of vertical location. The short red line at the top just marks the penalty region. A participant had to trade off the increased probability of a penalty if they moved the sample upwards and a reduction in reward if they moved it downwards. In the figure, a participant receives 74.3 points. (B) We combined the two values of the penalty (0 points and -500 points) with the two sample sizes (30 points and 5 points), resulting in four conditions in total. All of the tasks considered depended only on the vertical coordinates of the sample and we could in principle have used univariate Gaussian sampled distributed along a vertical line. We used bivariate samples simply to reduce the chance that sample points would overlap and occlude one another.

<https://doi.org/10.1371/journal.pcbi.1011999.g007>

the resulting four conditions were interleaved and the participants completed 50 trials in each condition for a total of 200 trials. The task is illustrated and explained in Fig 7 and its accompanying legend.

Calculating the optimal set point. The decision maker is given only the sample data from the population pdf and does not know the population pdf or its parameters. Thus the normative decision maker obtains the estimates $\hat{\mu}$, $\hat{\Sigma}$ based on the observation of sample mean (\bar{x}, \bar{y}) and sample covariance (s_x^2, s_y^2, s_{xy}) and they chooses a setpoint to maximize the expected reward. But beware: the decision maker cannot treat the estimated parameters as if they were the true population parameters. The normative decision maker must take into account the uncertainty in the estimates $\hat{\mu}$, $\hat{\Sigma}$ and allow for the difference in number of points in the 5-point and 30-point samples. See *Normative BDT model* in [Methods](#) for how to maximize the expected reward given $\hat{\mu}$, $\hat{\Sigma}$ and the number of points.

There were no trends in decision maker's set point from the beginning to the end of the task (S6 Fig). Therefore we computed the average set point across trials (S7A Fig). A two-way within-participant ANOVA showed a main effect of the penalty condition ($F [1, 16] = 42.68$, $p = 0.001$, $\eta^2 = 0.52$) and a main effect of the number of points ($F [1, 16] = 53.08$, $p = 0.001$, $\eta^2 = 0.20$). There was no significant interaction ($F [1, 16] = 0.99$, $p = 0.34$, $\eta^2 = 0.001$). The participants effectively made a riskier decision in the 5-point conditions compared with the 30-point conditions and made a safer decision in the large penalty conditions (-500 points) compared with the small penalty conditions (0 points). To further compare human performance to the normative BDT, we analyzed how each point in the sample influenced the participant's set point relative to the normative set point.

The normative influence of each point on the normative set point $I(P) = \partial S / \partial P$ was computed numerically. The influence of each point on the human decision maker's set point $\hat{I}(P) = \partial \hat{S} / \partial P$ was measured by regression analysis (See *Measuring influence* in [Methods](#)). Fig 8A shows the normative influence and measured influence for each penalty condition and sample size. The sample points are sorted and assigned an order index ranging from 1 (lowest sample point) to either 5 or 30 (highest sample point) depending on sample size. The normative influence is skew-symmetric: higher points nearer the penalty region are assigned negative influences large in magnitude while sample points furthest from the penalty region are also assigned influences large in magnitude but opposite in sign. Measured influence in contrast is largest in magnitude for sample points near the penalty region but points far from the penalty region have almost negligible influence. It might seem plausible that points far from the penalty region should be assigned little influence but the normative BDT decision maker does not do so. For the normative decision maker, the point nearest the penalty region and the point furthest from the penalty region have the largest influences (in magnitude) though opposite in sign.

Fig 8B shows the corresponding influence ratios $\phi(P) = \hat{I}(P) / I(P)$. The sample points are again sorted and assigned an order index. We could not reliably estimate the ratio near the median point because the normative influence measures (denominator) there were almost zero (black line in Fig 8A) and we omit them. The normative decision would have values of

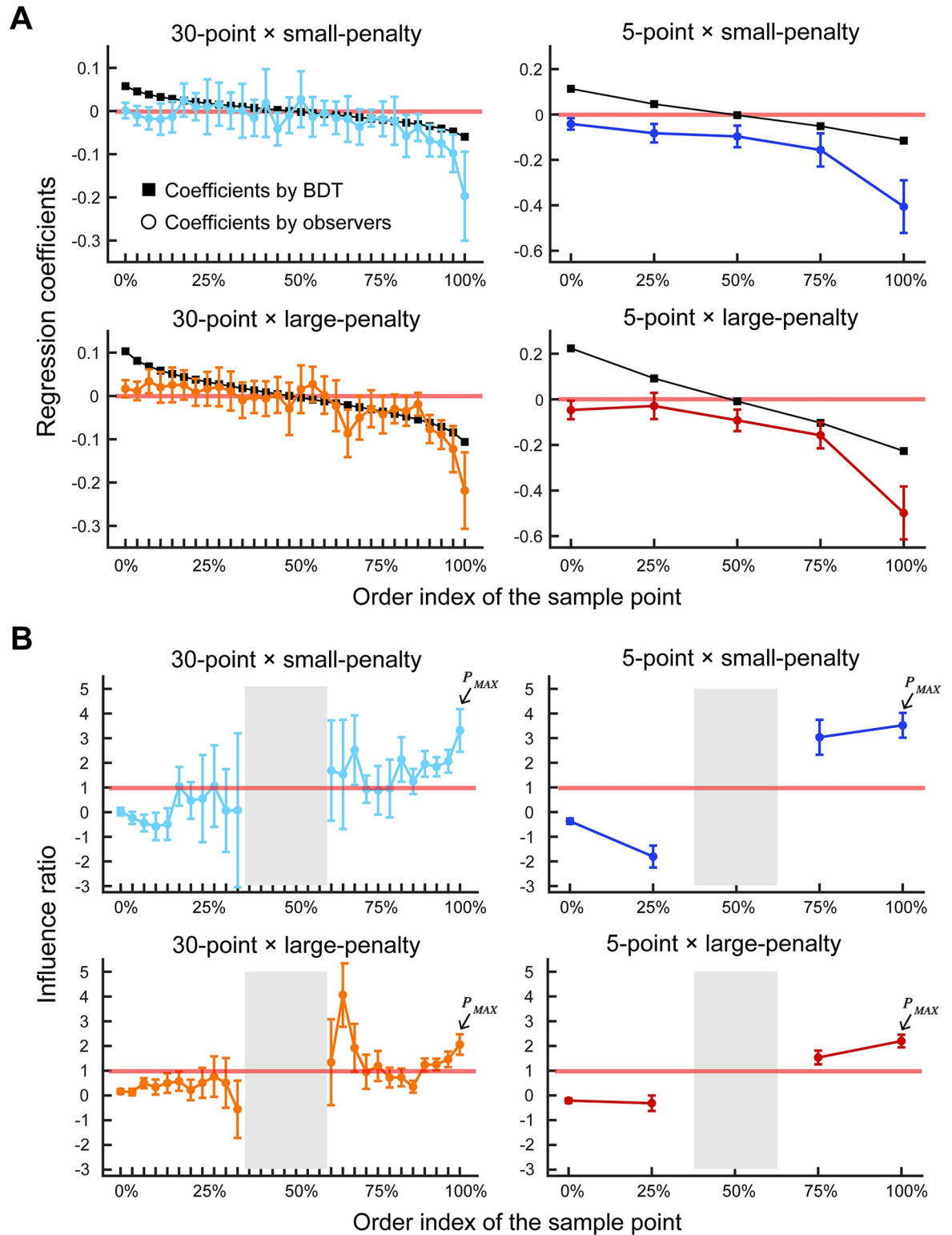


Fig 8. Measured influence and the influence ratio. A. Normative BDT influence and measured influence. We order the sample points for each sample point from 1 (the lowest) to either 5 or 30 (the highest) depending on sample size. We plot the mean across participants of the estimated influence of each sample point on the participant's actual set point versus its order index (colored circles). We plot the influence expected for the normative BDT decision-maker versus order index (black squares). The weights of each sample point were estimated using a ridge regression (see [Methods](#)) for each participant. The regression coefficients were then averaged across the

participants. The error bars indicate ± 2 s.e.m. Negative influence indicates that a set point is set further away from a penalty boundary when a sample point is generated close to a penalty boundary relative to a starting point. The influence measures for the normative BDT model are skew-symmetric and a thin red line marks the axis of symmetry. The highest points (near the penalty region) and the lowest points (farthest from the penalty region) have influence equal and opposite in sign. The middle points have less influence. In contrast, the human decision maker has influence measures that roughly decrease in magnitude as we go away from the penalty region. The lowest points in the sample have little or no influence. Sample points distant from the penalty region have little influence. **B. Influence ratios.** The average across participants of influence ratios (measured influence divided by normative BDT influence) for each sample point is plotted versus the order index of the sample point. Error bars denote ± 1 s.e.m. A value of 1 indicates that measured influence was identical to the normative BDT influence. The influence ratios deviate from 1 (marked by a thin red line). For the sample points nearest the penalty region the influence ratios are too large but they approach 0 for sample points far from the penalty region. The gray-shaded central range could not reliably be estimated due to the denominator (i.e., normative BDT influence) being near zero.

<https://doi.org/10.1371/journal.pcbi.1011999.g008>

normalized influence consistently near 1 with no patterned deviations. Instead, there is a trend: the points close to the maximum point have equal or greater influence compared to normative whereas the points close to the minimum point have almost no influence or negative influence (the influence is in the opposite direction from normative influence). In particular, the influence of the maximum point is two to three times greater than normative (marked by arrows). The pattern of measured influence deviates markedly from the expected pattern derived from the normative BDT decision maker.

An alternative heuristic strategy. The results of the influence analysis suggest that participants gave considerable (and inappropriate) influence to the points closest to the penalty region above the green line. We consider an alternative heuristic strategy (the *max-point strategy*) where the participants set the P_{MAX} point to a reference boundary internalized in their visual system (Fig 9A). The mental reference boundary could be anywhere on the screen and could be above the penalty boundary. This is a free parameter for the model fit. In the models, the set point is determined by moving the P_{MAX} point until it reaches the mental reference boundary B as below $S = B - P_{MAX}$ where the location of P_{MAX} refers to the location when the decision makers set the sample at the starting position. The mental reference boundary is set to be identical for 30-point samples and 5-point samples. As a consequence, the Max-Point model predicts higher mean settings for the 5-point sample than 30-point sample (Fig 9A). The average mental reference boundary recovered by the model fit across participants was 179.0 ± 7.3 mm for the small penalty condition and 168.7 ± 9.6 mm for the large penalty condition.

Therefore, the Max-Point model predicts a higher set point in the 5-points than the 30-points condition (Fig 9A). In Fig 9C and 9D, we show correlation plots between the actual set point and the model prediction of normative BDT and the heuristic model. The normative BDT clearly failed to capture the behavioral pattern whereas the Max-Point model well matched. We calculated the AICc for each model, each participant, and each condition (Fig 9B). On average across conditions and participants, the Max-Point model was 1.2×10^{12} times more likely than the normative BDT (average AICc = 586.2 for BDT, average AICc = 549.3 for Max-point). In the max point strategy, we estimated a mental reference boundary that is identical for the two sample sizes (5 and 30) but different for the two penalty magnitudes (0 and -500 points). This two-parameter model fits the observers' data quite well (Fig 9D). This suggests that observers pick the same criterion for the reference boundary regardless of the number of points in the sample.

We repeated the model comparison using an alternative criterion for model comparison, the Bayesian Information Criterion (BIC). The differences between AICc and BIC were minor (average BIC = 586.2 for BDT, average BIC = 551.1 for Max-point). These results suggest that, in their decision, the decision makers primarily relied on the extreme point P_{MAX} rather than the parametric estimates of the population pdf.

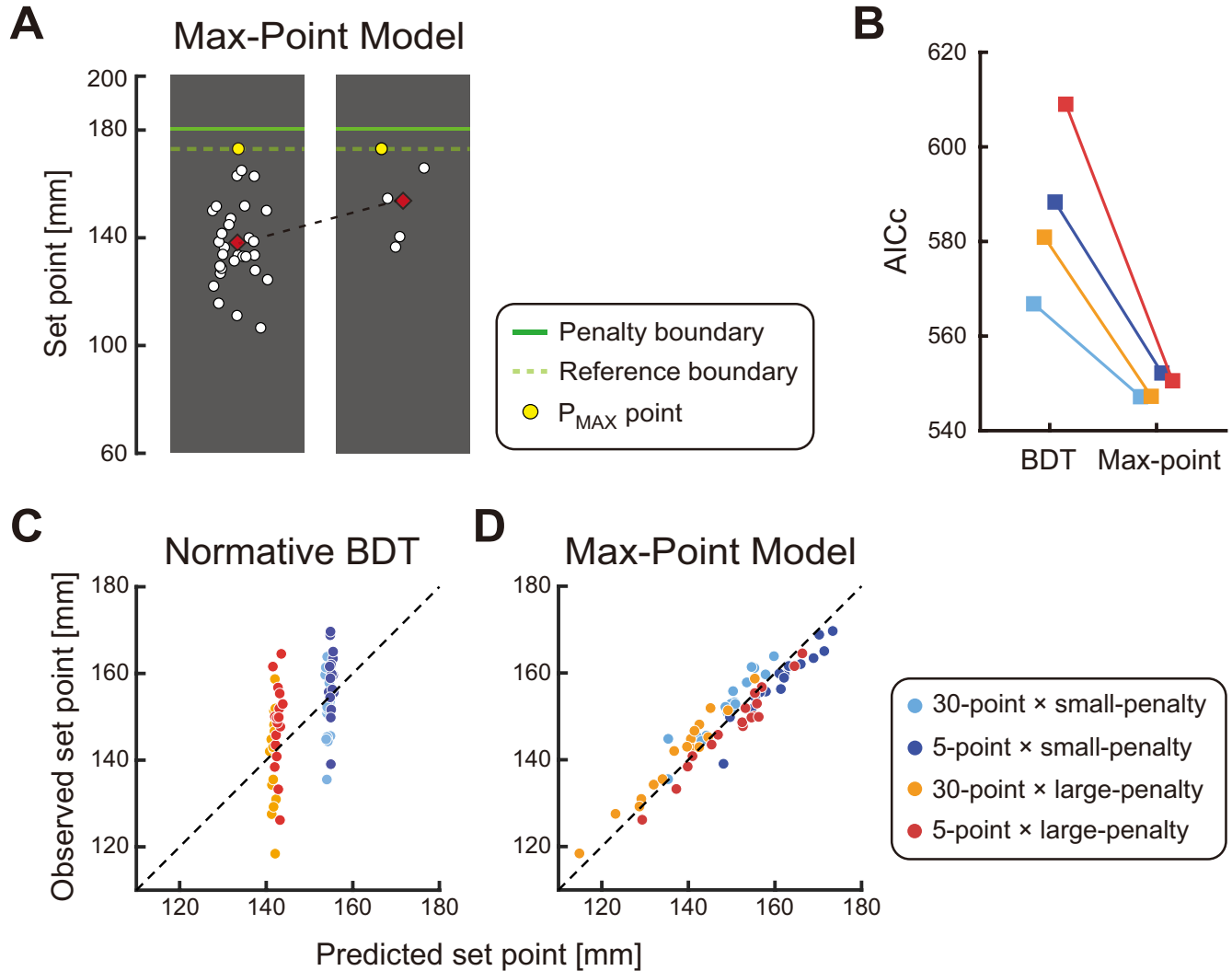


Fig 9. The Max-Point Model. **A.** Illustration of the Max-point model. Two samples drawn from the same pdf are shown, one of size 5 and the other of size 30. The decision maker sets each sample so that the maximum point P_{MAX} falls on a criterion boundary (dashed green line) chosen by the decision maker. The mental reference boundary is identical for 30-point samples and 5-point samples and as a consequence, the mean settings (red diamonds) for the 5-point sample are markedly higher than those for the 30-point samples. **B.** An AICc model comparison of the normative BDT model and the Max-Point Model. The lower the AICc, the better the model. **C.** A plot of mean settings for each participant versus the predictions of the normative BDT model. **D.** A plot of mean settings for each participant versus the predictions of the Max-point model. In sum, the Max-Point Model outperformed the normative BDT and reproduced the participant's set point fairly well.

<https://doi.org/10.1371/journal.pcbi.1011999.g009>

To summarize, the influence estimates we obtain from human decision makers deviate from normative. The sample points nearer to the green boundary are assigned much more influence than normative Gaussian BDT model would assign. Points distant from the penalty region are assigned influences near 0. The decision makers are making their judgments on a subset of the sample points near the penalty region, ignoring those further away. Their error was costly, leading to a larger number of penalty trials in the 5-point than 30-point condition (S7C Fig). The actual total score was on average 82.7% (57.0% ~ 95.7%) relative to the BDT theoretical maximum score (S7B Fig). Human decision makers have an efficiency of 82.7%: their choices of strategy cost them 17.3% of their potential winnings.

Discussion

Bayesian Decision Theory (BDT) is used to model ideal performance in a wide variety of experiment tasks in human perception, movement planning, and cognition [1,2,23]. It captures important aspects of human performance in everyday tasks: the outcomes of these tasks entail rewards or penalties for the decision maker and the decision maker is only partially in control of the outcome. The decision maker can choose an action (make a decision) but the choice of action does not determine the outcome. Marr [43] famously argued that human perception and cognition could be analyzed on any of three levels: computational, algorithmic or implementational. BDT as a normative model exists on the computational level and we test only whether its predictions match human performance. There is promising work on BDT and probabilistic models on levels other than the computational [44]. For example, some cognitive models propose hypothetical sampling processes as part of an algorithm in executing Bayesian inference [45,46].

The interval estimation tasks and the decision task we considered were chosen to test three essential properties of Bayesian decision theory. As its heart, BDT is a collection of rules for combining information about the probabilities of outcome with information about the values of outcomes to arrive at a decision maximizing the expected Bayes gain [1,47]. We are particularly concerned with tasks—visuo-motor or visual cognitive—where participant’s knowledge about uncertainty is based on samples taken from probability density functions (pdf’s). The interval estimation tasks test human ability to work with uncertainty in the form of pdf’s estimated from samples. The decision task introduces a value function and tests human ability to combine their knowledge about uncertainty with the values of outcomes about actions.

BDT maximizes expected reward. Experimental tests of BDT often estimate the *efficiency* [10,11,48] of the human decision maker—the ratio of human winnings to the maximum expected possible winnings predicted by BDT, arguably the key measure to consider in evaluating human performance. If efficiency is substantially less than 1 then we have evidence that the BDT model is not appropriate as a model of human performance. Past experimental tests of the BDT model in a wide variety of cognitive, perceptual and motor tasks have decidedly mixed outcomes, some finding that overall human performance approaches the efficiency limit dictated by BDT while other find a marked gap between human performance and optimal (see Maloney & Zhang [1] and Rahnev & Denison [23] for reviews).

Maloney & Mamassian [20] and Zhang et al [21] argue that a comparison of overall winnings to the maximum possible is not a powerful test of the hypothesis that BDT describes how human decision makers make decisions. Heuristic rules different from BDT can achieve efficiencies as close as we like to 1 [20]. Good overall performance may be due to an unwitting choice of task by the experimenter [21]. The brain may appear to be Bayesian [49,50] but in reality, it may be doing something else. We suspect that reports that human performance is not significantly different from optimal are just Type II errors: more participants or a more sensitive measure (e.g., influence reported in this study) would disclose significant deviations from optimal.

We break down the Bayesian computation into elementary operations and test human ability to carry out three of these operations. We considered visual cognitive tasks where the human decision maker is given a sample from a bivariate Gaussian probability density function (pdf) and must use it normatively (Fig 2A–2D). The transition from sample to pdf is a key step in the BDT computation because the use of the estimated pdf allows parametric decision-making based on a handful of estimated parameters and ignores any accidental structure in the sample. We first tested *accuracy*, the ability to correctly estimate the probability that an additional point from the specified pdf will fall into any specified region (Fig 4). This ability is

essential to the BDT computation when different regions carry different penalties (Fig 6A). We then tested *additivity*, the ability to estimate the probability of landing in a region composed of multiple disjoint subregions (Fig 4). We found small but patterned failures of both accuracy and additivity (Fig 5) but argued that the costs of these observed failures in accuracy and additivity were minor or almost negligible in motor tasks used in the literature of BDT (Fig 6).

We last measured the *influence* of each point in the sample on the participants' performance (Fig 7). The normative transformation from the sample to the estimated pdf leads to skew-symmetric influence: the highest point and the lowest point are assigned large influences in magnitude and opposite in sign (Fig 8A). In contrast, the individual's use of sample information in making decisions deviated markedly from skew-symmetric (Fig 8A and 8B). Although the brain may appear to be Bayesian [49,50], in some tasks it is doing something else.

How can we understand human performance when it deviates from normative? One approach assumes that decision-makers transform from the external objective values of outcomes to the internal subjective values, utility. However, it is not plausible that a non-linear utility function exhibits any significant curvature over the small amounts of money involved in our experiment [51]. If there is significant curvature for small amounts of money, we would expect to find much greater loss aversion for the large penalty condition than we do. The loss aversion we find does not likely support this argument—the number of sample points that fell into the penalty region was about 0–2 points and was roughly the same even though the size of the penalty increased from 0 points to -500 points (S7D Fig).

Another approach assumes that decision-makers use a different family of probability density functions that is not the correct family [23]. Some studies, for example, assume that the decision maker's internal model is based on pdfs from a “high-tailed” distribution family or a family of skewed distributions [21,52,53]. Maloney & Thomas [54], for example, investigate signal detection theory models where true SDT pdfs are not Gaussian but are drawn from pdf families with higher or lower tails than the Gaussian. All of the authors just cited assume that decision-makers are still engaged in estimating parameters from a sample (a key step in the BDT computation) but they are using the wrong pdf family (Fig 2A–2D).

Examination of the influence of individual points suggests that human decision-makers in the decision task do not estimate the parameters of a pdf. The Max-Point model we developed is not based on such parametric decision-making. Rather the model assumes a heuristic decision rule based on just one sample point closest to the penalty region. Nevertheless, the model led to a much better fit to human data than BDT (Fig 9).

The failures of accuracy and additivity do not invalidate the claim that normative BDT is a useful approximation to human performance and a reliable model of how human decision makers will behave in experiments. However, the discrepancy between measured influence and normative indicates that the human decision maker is using information differently than the normative decision maker and even ignoring sample points far from the penalty region that the normative decision maker assigns great weight to. Our finding is broadly consistent with the fact that the visual information is weighted differently by the order of stimulus presentation [55], reference stimulus [56], or visual appearance [57]. In this study, however, we developed a method to measure how much *each piece* of information in the sample influenced the decision maker's action. We showed the evident asymmetry in influence: decision-makers assign weight on outlying sample points close to the penalty region two or three times greater than the normative while they ignore sample points far from the penalty region. A heuristic model based on the maximum sample point is consistent with the asymmetry in influence we found. We reject the normative Gaussian BDT model we began with, even as an approximation to human behavior.

Methods

Ethics statement

This study was approved by the University Committee on Activities Involving Human Subjects of New York University (IRB-FY2018-2006) and was carried out in accordance with their guidance. The participants provided written informed consent before the experiment.

Participants

Seventeen participants (mean age 21.9, range 19–30, 5 males) completed the experiments that consisted of the estimation tasks (1st part of the experiment) and the decision task (2nd part of the experiment). Informed consent was given by each participant before the experiment. All participants were not aware of the hypothesis under test. The participants received US\$12 per hour plus a performance-related bonus (average: \$9.7, range: \$4.3–\$11.7).

Apparatus

Stimuli were displayed on a vertical monitor (VPIXX, VIEWPIXX, 514 mm × 288 mm). The monitor resolution was 1920 × 1080 pixels with a 60-Hz refresh rate. The participants were seated at a viewing distance of 60 cm. A pen-tablet was set in front of the monitor (Wacom Intuos Pro Large, workspace: 311 × 216 mm). The participants manipulated a digitized pen to carry out the tasks. The horizontal-vertical coordinates of the digitized pen were recorded at 60 Hz. All stimuli were controlled using the Matlab Psychophysics Toolbox [58,59].

The estimation tasks

The tests of accuracy and additivity were carried out in the same session. In both tasks, the stimulus was a sample of white points drawn from an anisotropic bivariate Gaussian distribution with mean $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance $\Sigma = \begin{pmatrix} \sigma^2/2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$. The origin was set at the vertical and horizontal center of the screen. On each trial, the population variance σ^2 was drawn from a uniform distribution on the interval 100 mm to 400 mm. The points were round with a radius of 1.5 mm. There were two sample sizes as the number of white points produced was $N = 5$ or 30. Two white triangles were presented at left and right sides of the area where the points were displayed at the vertical mean μ_y of the Gaussian pdf (Fig 4A). After the presentation of the sample, two green horizontal lines (width: 1 mm) across the distribution of the white points appeared (Fig 4B). The participants were asked to judge the probability that a new point drawn from the same distribution would fall into the region delimited by the two lines. The participants recorded their probability by moving the digitized pen horizontally on a pen tablet.

There were three configurations with respect to the center of the screen. The lines covered the range from $+w\sigma$ to $-w\sigma$ in the symmetric interval S , from the center of the screen to $+w\sigma$ in the upper half of symmetric interval SU , and from the center of the screen to $-w\sigma$ in the lower half of the symmetric interval SL , where w means the interval width and σ means the population standard deviation. We varied the interval width to make nine probability conditions, $[0.126\sigma, 0.253\sigma, 0.385\sigma, 0.524\sigma, 0.674\sigma, 0.842\sigma, 1.036\sigma, 1.282\sigma, 1.645\sigma]$. The value of $P[S]$ spanned the range 0.1 and 0.9 and $P[SU]$ and $P[SL]$ spanned the range 0.05 and 0.45. See Fig 4C.

There were $54 = 2 \times 3 \times 9$ conditions created by combining two sample sizes (5 or 30 points), three configurations (S , SU , SL), and nine probabilities. The participants repeated the

task in each condition of the 54 conditions 5 times (270 trials in total). We divided the experimental session into six blocks of 45 trials each. In the first three blocks, we presented the 30-point conditions, in the remaining three blocks we presented the 5-point conditions. Within each block, the configuration of horizontal lines was fixed but its order was balanced across participants.

The decision task

The measurements of influence were carried out after the interval estimation tasks. The same participants completed this task after completing the previous estimation tasks. A sample of white points ($N = 5$ or 30) was drawn from an anisotropic bivariate Gaussian distribution with mean $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance $\Sigma = \begin{pmatrix} \sigma^2/2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$ on each trial. The population variance σ^2 was drawn from a uniform distribution on the interval 100 mm to 400 mm.

In this task the participants could move the sample rigidly to any place on the screen (Fig 7). To begin a trial, the participants moved the sample to a blue initial position. After holding for 1 sec, a white penalty boundary (PB) appeared 180 mm above the initial position. The PB turned green after a random interval (0.4–0.8 sec), which signaled the start of the trial. The participants moved the sample up or down deciding where to set the sample on the screen. They signaled their decision by pressing the button. After a button press, a new yellow point drawn from the same population pdf appeared and the vertical position of the yellow point with respect to the green line determined the participant's reward or penalty. If the yellow point was above the green line then the participant incurred the penalty (0 or -500 points). If the position was on or beneath the green line by $0 \leq \Delta \leq 180$ mm then the participant received $(180 - \Delta)/1.8$ points. That is, if the yellow point was exactly on the green line ($\Delta = 0$) then the participant received 100 points. If it was 180 millimeters below the green line that participant received 0 points. The red lines in Fig 7 sketch the reward function and it was also shown on the screen as well as the score in the trial and the total score. The participants were instructed to maximize the total score and told that the obtained total score would convert into a bonus payment at the end of the experiment (75 cents per 1000 points).

There were four experimental conditions, two sample sizes (5 or 30 points) and two penalties (0 or -500 points). Each condition was repeated for 50 trials, resulting in 200 trials in total. We divided the experimental session into 5 blocks of 40 trials. In each trial, either one of the four conditions was randomly chosen. The participants received an average bonus of \$9.70 (range \$4.30 –\$11.70).

Normative BDT model

We modeled the normative setting point based on Bayesian Decision Theory [1,4]. A data sample (P_1, P_2, \dots, P_N) was randomly generated from the population pdf $\phi(x, y | \mu, \Sigma)$. A decision rule in the normative BDT is based on the estimated pdf $\phi(x, y | \hat{\mu}, \hat{\Sigma}, N)$. Although the population pdf is a bivariate Gaussian distribution, the horizontal coordinate is not relevant information as the reward is based on the vertical coordinate of a new yellow point. Therefore, we treated our model as a univariate pdf in the vertical dimension.

On each trial, the normative decision maker observes N data points in the sample (P_1, P_2, \dots, P_N) . They convert the sample to the sample mean \bar{y} and sample variance s^2 , and utilize these observations \bar{y} , s^2 , and N to estimate the true population mean μ and population variance σ^2 . We denote the estimates of population mean and that of population variance as $\hat{\mu}$ and $\hat{\sigma}^2$, respectively.

To obtain those estimates, we assumed that the normative decision maker uses the knowledge of the generative model which produced the sample. In our experiment, we set the mean of the population pdf to a constant value. We can thus leave the estimates of population mean and focus on the estimates of population variance.

The prior probability of the estimated population variance can be written as a uniform probability density function:

$$P(\hat{\sigma}^2) = \begin{cases} \frac{1}{300}, & \text{if } \hat{\sigma}^2 \in [100, 400] \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

See [S8A Fig](#) for this prior distribution.

The likelihood function of the estimated population variance depends on the number of points in the sample N and the sample variance s^2 . The estimates of population variance are distributed as chi-squared random variables with $N-1$ degrees of freedom and is described as a chi-square probability density function:

$$P(s^2|\hat{\sigma}^2, N) = \chi_{N-1}^2(\hat{\sigma}^2; s^2) = \frac{N-1}{2^{N-\frac{1}{2}}\Gamma(\frac{N-1}{2})s^2} \frac{\hat{\sigma}^2(N-1)^{\frac{N-1}{2}-1}}{s^2} e^{-\frac{\hat{\sigma}^2(N-1)}{2s^2}} \tag{7}$$

where $\Gamma()$ is the standard gamma function. [S8B Fig](#) shows the example likelihood function when $N = 5$ and $s^2 = 200$. [S9 Fig](#) also illustrates the examples with varying N and s^2 .

The normative decision maker computes the posterior probability of the estimated population variance using Bayes rule.

$$P(\hat{\sigma}^2|s^2, N) \propto P(s^2|\hat{\sigma}^2, N)P(\hat{\sigma}^2) \tag{8}$$

The posterior is proportional to the product of the likelihood and the prior. [S8C Fig](#) shows the posterior probability distribution.

Once a particular population variance is estimated, we can determine the probability density function for producing the sample. Since the sample of 5 or 30 points was randomly drawn from a bivariate Gaussian distribution, we set the estimated pdf in the form of a Gaussian distribution

$$f(y|S, \hat{\sigma}^2) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left[-\frac{(y-S)^2}{2\hat{\sigma}^2}\right] \tag{9}$$

where y is a point in the vertical coordinate and S denotes a selected setpoint that shifts the center of the estimated pdf. [S8D Fig](#) illustrates three examples of the estimated pdf with varying $\hat{\sigma}^2$ and fixed S . [S8E Fig](#) also illustrates the products of example pdfs with the reward function.

Given a setpoint S , we can compute the expected reward by integrating the reward function $G(y)$ and the estimated pdf $f(y|S, \hat{\sigma}^2)$ for each possible estimated population variance ([S8F Fig](#)). We need to scale this expected reward by the posterior probability of the estimated population variance since the choice of $\hat{\sigma}^2$ for the estimated pdf $f(y|S, \hat{\sigma}^2)$ depends on $P(\hat{\sigma}^2|s^2, N)$. [S8G Fig](#) shows the scaled expected reward as a function of $\hat{\sigma}^2$.

The final expected reward can be obtained by integrating the scaled expected reward over the estimated population variance ([S8H Fig](#)) and is defined below.

$$EG(S) = \iint G(y)f(y|S, \hat{\sigma}^2)P(\hat{\sigma}^2|s^2, N)dyd\hat{\sigma}^2 \tag{10}$$

We changed the set point such that the expected reward can be maximized:

$$S^* = \underset{s}{\operatorname{argmax}} EG(S) \tag{11}$$

When P_1, \dots, P_N are i.i.d. sample drawn from a normal distribution $N(0, \sigma^2)$, the random variable $\frac{s^2(N-1)}{\sigma^2}$ is distributed according to the chi-square distribution with $N-1$ degrees of freedom. Therefore, the probability of observing the sample variance s^2 given the population variance σ^2 can be written as:

$$s^2 \sim \chi_{N-1}^2(s^2; \sigma^2) = \frac{N-1}{2^{N-\frac{1}{2}} \Gamma(\frac{N-1}{2}) \sigma^2} \frac{s^2(N-1)^{\frac{N-1}{2}-1}}{\sigma^2} e^{-\frac{s^2(N-1)}{2\sigma^2}} \tag{12}$$

We can thus obtain the likelihood function of the estimated population variance in Eq (7) by flipping s^2 and $\hat{\sigma}^2$ around. S10 Fig illustrates the optimal set point as a function of the sample variance given the number of sample points and the size of the penalty.

Measuring influence

Influence of each point on the human decision maker. We performed a regression analysis to estimate the influence of each point in the sample on the participant’s set point in the decision task. We ordered points in each sample ($N = 30$ or 5) in ascending order from the lowest sample point to 5 or 30 (highest sample point) depending on sample size (Fig 8). We used these explanatory variables to predict the final set point. We regressed the vector of participant’s set point \mathbf{S} by the matrix of explanatory variables \mathbf{P} (i.e., the order index of the sample point) containing a set of vectors from \mathbf{P}_{MIN} to \mathbf{P}_{MAX} to estimate the weight vector \mathbf{W} . Our regression model is thus simply $\mathbf{S} = \mathbf{P} \cdot \mathbf{W}^T$.

There are two issues if we used an ordinary least squares regression in this analysis. First, in the 30-point condition, the number of explanatory variables (i.e., 30 variables) is relatively large to the number of observations (i.e., 50 trials). Under such a condition, the ordinary least squares regression tends to grossly overfit the training data and does not generalize the model prediction to the unobserved new data [60]. Second, because the explanatory variables are order statistics, some variables are correlated with each other (e.g., the second maximum point possibly takes a large value when the maximum point is large whereas the second maximum point should be small when the maximum point is small). The regression coefficients can become poorly determined when there are many correlated variables in a linear regression model [61].

We therefore used a ridge regression (L2-norm regularization) to estimate the weight vector \mathbf{W} . Ridge regression is a machine learning technique which alleviates the problem of multicollinearity by adding the regularization term to the cost function [61] as follows

$$\text{Total Error} = \sum_j (S_j - \sum_i W_i P_{ij} - W_0)^2 + \lambda \sum_i W_i^2 \tag{13}$$

where S_j is a set point at j -th trial, P_{ij} is a i -th explanatory variable (the order index of the sample point) at j -th trial, W_i is a regression coefficient for i -th variable, W_0 is the coefficient for a constant term, and λ is a regularization parameter. In the ordinary least squares regression, the cost function to be minimized is solely the first term which is the sum of the squared difference between the data and model prediction. Ridge regression involves the second penalty term which is the sum of squared weight multiplied by regularization parameter λ . Ridge regression basically shrinks the magnitude of the regression weights by increasing the amount of λ . A large λ forces the regression coefficients to be closer to zero whereas a small λ has the opposite

effect and in the limit as $\lambda \rightarrow 0$ the estimated weights converge to ordinary least squares regression coefficients.

To avoid overfitting, we performed 10-fold cross-validation [60,61]. Specifically, for each participant and condition, we split the 50 data set into 10 equal-sized parts. Nine of ten parts (45 trials) were used for the training data set and the remaining part (5 trials) was used for the validation set. Given a particular λ , we trained our model on the training data set and then used the trained model to predict the total error on the left-out validation set. We repeated this process 10 times using different training and validation sets and we calculated the average error and standard error across 10 folds for each λ . We used the "one standard error rule" [61] for the choice of λ rather than the best λ with the minimized error, which results in selecting a more regularized model and preventing overfitting. The resulting estimates of regression weights are shown in Fig 8A.

Influence of each point on the normative decision-maker

We computed the influence on the normative set point as follows. We assume the normative set point S^* as being a linear combination of weights and sample points $\sum_{i=1}^N W_i^* P_i$. We induced a small change to a point P_i while keeping the other points same. If there is a non-zero weight on P_i , a setting should change corresponding to the changes in P_i ; otherwise the weight indicates zero. We denote the shifts in the normative set point with respect to the changes in a point as $\frac{\partial S^*}{\partial P_i}$ (i.e., a partial derivative). Since any other points are unchanged, the small changes in a particular point, ∂P_i , is the only contributor to shift the set point and its amount depends on the amount of weight W_i^* . Therefore $\partial P_i W_i^*$ is a measure of the amount of shifts with respect to that point (i.e., $\frac{\partial S^*}{\partial P_i}$) and we can derive the weight W_i^* by

$$\frac{\partial S^*}{\partial P_i} = \partial P_i W_i^* \iff W_i^* = \frac{\partial S^*}{\partial P_i} \cdot \frac{1}{\partial P_i}.$$

We performed a Monte Carlo simulation to calculate the normative weights. We first generated a Gaussian sample of 30 or 5 points and computed the normative set point given a statistics of the sample. In each sample, we induced changes in a point P_i by 1 mm (due to the limitation of computational precision) and recorded the shifts in the normative set point given the changed statistics. We repeated the same process by generating a different sample 1,000 times and we defined the normative influence as $W_i^* = E \left[\frac{\partial S^*}{\partial P_i} \right] \cdot \frac{1}{\partial P_i}$. The simulated normative weights are shown in Fig 8A.

Model comparison

For model comparison, we applied AICc—Akaike information criterion with a correction for finite sample size—to each participant and model as the information criterion for goodness-of-fit [62]. The formula for AICc adds an extra penalty term $\frac{2K^2+2K}{N-K-1}$ to the formula for AIC by taking into account the number of model parameters K and the number of data points N . This penalty avoids potential overfitting and helps select the models that have fewer parameters as Bayesian information criterion (BIC) does.

Supporting information

S1 Fig. Estimates of probability did not change systematically across trials. The participant's estimates of probability in the symmetric interval are plotted versus trial. Data is averaged across the participants. The color scale of the circle indicates the correct probability between 0.1 and 0.9. The estimates were retained consistently from the beginning to the end of

the task. Three-way within-participant ANOVA, using the correct probability (9), sample condition (2), and the number of trials (5) as independent variables, showed no significant main effect of the trial ($F [2.6, 40.9] = 1.54, p = 0.22, \eta^2 = 0.00$).

(TIF)

S2 Fig. Estimates of probability scatter around the correct probability. A small vertical line marks the observer's estimate of probability in each trial (i.e., each sample). The data are taken from the symmetric interval condition. The vertical grey lines in lower panels indicate the possible probabilities taken by the counting point strategy. With a 5-point sample, these could be from 0.0 to 1.0 in steps of 0.2. For instance, the estimate would be 0.2 if one point falls within the interval. The counting strategy fails to predict the observers' estimates of probability.

(TIF)

S3 Fig. Plot of accuracy in the symmetric interval for each participant. The participant's estimates of probability in the symmetric interval are plotted against the correct probability for each participant. Data is averaged across trials.

(TIF)

S4 Fig. Plot of accuracy in upper and lower halves of the symmetric interval. The participants' estimates of probability in the upper half (A) and lower half (B) of the symmetric interval are plotted against the correct probability. Each white circle denotes the estimates for a single participant and a filled circle is the average estimates across participants. The black thick curve is the best-fit estimate by a linear in log-odds (LLO) function. The LLO parameters for each fit are γ and p_0 .

(TIF)

S5 Fig. Plot of additivity in each participant. The sum of estimates in the upper and lower halves of the symmetric interval are plotted against the estimates of the symmetric interval for each participant. Data is averaged across trials.

(TIF)

S6 Fig. Trial-by-trial set point in the decision task. A trial-wise set point averaged over the participants is plotted for each condition. The horizontal dashed lines denote the mean set point across all trials. There is no evident pattern in the residuals.

(TIF)

S7 Fig. Performance index in the decision task. **A.** The decision makers' actual set points. Each circle denotes the individual data averaged across trials, and a rectangle denotes the average across all participants. The error bars indicate ± 2 s.e.m. **B.** The efficiency: the ratio of the actual total score to the maximum total score possible predicted by the normative decision model. The actual total scores were significantly smaller than the maximum possible scores in all conditions except the 30-point, -500 points condition (two-tailed paired-sample t -test: $t_s [16] > 3.25, p_s < 0.005, d_s > 1.15$, Bonferroni corrected for four conditions). ** indicates $p < 0.01$ from the maximum possible score. **C.** The number of penalty trials that the participants incurred in each condition. The error bars indicate ± 2 s.e.m. A two-way within-participant ANOVA showed a main effect of the penalty size ($F [1, 16] = 91.78, p = 0.001, \eta^2 = 0.20$) and a main effect of the sample size ($F [1, 16] = 22.73, p = 0.001, \eta^2 = 0.30$). The participants thus incurred a larger number of penalties with fewer samples and with a smaller penalty. **D.** The number of exception points that fall in the penalty region. An individual data point is the average across 50 trials for each condition and each participant.

(TIF)

S8 Fig. Illustration for the estimation procedure of the sufficient model. **A.** Prior distribution of the estimates of population variance $\hat{\sigma}^2$. On each trial, the population variance was uniformly chosen from the range between $\sigma^2 = 100\text{mm}$ and $\sigma^2 = 400\text{mm}$. The prior distribution was set to be the same distribution for generating the population variance. **B.** Likelihood function of the estimates of population variance. Given the number of points in the sample N and the sample variance s^2 , the likelihood function of the estimates of population variance as a chi-square distribution with $N-1$ degree of freedom can be estimated. A solid vertical line shows the sample variance in this trial. We set the number of points to $N = 5$. **C.** Posterior probability distribution of the estimates of population variance. The posterior is a product of the prior distribution and likelihood function. **D.** Estimated probability density function modeled as a Gaussian distribution. The width of the pdf depends on the estimated population variance. We show three examples of the estimated pdfs. In the decision task, the location of the estimated pdf can be shifted by a set point S (here we chose $S = 150\text{mm}$). A small gray inset shows a reward function $G(y)$. **E.** The function illustrates the product of the estimated pdf with the reward function. **F.** The ideal decision maker integrates the estimated pdf with the reward function given the set point and the estimated population variance, which produces the expected reward function as a function of the estimated population variance. **G.** The ideal decision maker then scales the expected reward function **F** by the posterior probability of the estimated population variance **C**. **H.** The final output of the expected reward can be obtained by integrating the function shown in **G** over $\hat{\sigma}^2$. The ideal decision maker chooses the ideal set point S^* maximizing expected reward.
(TIF)

S9 Fig. Likelihood function of the estimated population variance as a chi-square distribution. Given the population variance $\hat{\sigma}^2$, the random variable of the sample variance s^2 is distributed according to a chi-square distribution with $N-1$ degrees of freedom. Therefore, the likelihood function of the estimated population variance can be described as a chi-square probability density function. We show the likelihood functions when $N = 30$ (left) and $N = 5$ (right). The sample variance varied between 50 mm and 450 mm.
(TIF)

S10 Fig. Optimal set point as a function of the sample variance. The optimal set point was modelled for the variance in the sample. The optimal set point changes from trial to trial as the sample points are resampled in each trial and the sample variance changes. In Fig 9C, the optimal set point is averaged across trials and is plotted against the observer's average set point.
(TIF)

S1 Text. Model comparison in the test of accuracy. Details of the model formula in the test of accuracy.
(PDF)

S1 Table. Model comparison in the test of accuracy. Note—"No. Par." is the number of free parameters for each model. $\Delta\text{AICc} = \text{AICc}_{\text{H1}} - \text{AICc}_{\text{H0}}$. A positive AICc difference indicates a better-fitting model than H0 (no distortion). The evidence ratio is defined by $\exp\left(\frac{\Delta\text{AICc}}{2}\right)$ and is the relative likelihood of model pairs and represents the evidence about models as to which is better in a K-L information sense. A value of evidence ratio means how many times the data is more likely under the alternative than the null hypothesis. Evidence ratios greater than 10 represent strong evidence for the alternative hypothesis whereas evidence ratios less than 0.1 represent strong evidence for the null hypothesis. Evidence ratios greater than 3 (or less than 0.33) represent substantial evidence for the alternative hypothesis (or for the null hypothesis). A

value in the intermediate range (0.33 and 3) supports neither hypothesis. H1 (LLO function) is strongly supported. The right four columns summarize the free parameters that best describe the data: γ is the slope of the curve and p_0 is the crossover point.

(PDF)

S2 Table. Model comparison in the test of additivity. See the note in [S1 Table](#).

$\Delta AICc = AIC_{C_{H0}} - AIC_{C_{H1}}$. A positive AICc difference indicates a better-fitting model than a null hypothesis (H0). Three hypotheses concerning estimates for the test of additivity are below. A super-additive model (H1) outperformed the other models H0 : $\hat{P}[SU] + \hat{P}[SL] = \hat{P}[S]$ H1 : $\hat{P}[SU] + \hat{P}[SL] = \hat{P}[S] + b (b > 0)$ H2 : $\hat{P}[SU] + \hat{P}[SL] = \hat{P}[S] + b (b < 0)$.

(PDF)

Acknowledgments

We thank Jakob Phillips for his support in the experiment.

Author Contributions

Conceptualization: Keiji Ota, Laurence T. Maloney.

Data curation: Keiji Ota.

Formal analysis: Keiji Ota, Laurence T. Maloney.

Funding acquisition: Keiji Ota, Laurence T. Maloney.

Investigation: Keiji Ota.

Methodology: Keiji Ota, Laurence T. Maloney.

Project administration: Laurence T. Maloney.

Resources: Keiji Ota.

Software: Keiji Ota.

Supervision: Laurence T. Maloney.

Validation: Keiji Ota.

Visualization: Keiji Ota.

Writing – original draft: Keiji Ota, Laurence T. Maloney.

Writing – review & editing: Keiji Ota, Laurence T. Maloney.

References

1. Maloney LT, Zhang H. Decision-theoretic models of visual perception and action. *Vision Res.* 2010; 50(23):2362–74. Epub 2010/10/12. <https://doi.org/10.1016/j.visres.2010.09.031> PMID: 20932856.
2. Ma WJ. Bayesian Decision Models: A Primer. *Neuron.* 2019; 104(1):164–75. Epub 2019/10/11. <https://doi.org/10.1016/j.neuron.2019.09.037> PMID: 31600512.
3. Wald A. *Statistical decision functions.* New York: Wiley; 1950.
4. Berger JO. *Statistical Decision Theory and Bayesian Analysis.* 2nd ed. New York: Springer; 1985.
5. Green DM, Swets JA. *Signal detection theory and psychophysics.* New York: Wiley; 1966.
6. Knill DC, Richards W. *Perception as Bayesian inference.* New York: Cambridge University Press; 1996.
7. Battaglia PW, Schrater PR. Humans trade off viewing time and movement duration to improve visuomotor accuracy in a fast reaching task. *J Neurosci.* 2007; 27(26):6984–94. Epub 2007/06/29. <https://doi.org/10.1523/JNEUROSCI.1309-07.2007> PMID: 17596447; PubMed Central PMCID: PMC6672223.

8. Dal Martello MF, Ota K, Pietralla DE, Maloney LT. Detecting visual texture patterns in binary sequences through pattern features. *J Vis.* 2023; 23(13):1. Epub 2023/11/01. <https://doi.org/10.1167/jov.23.13.1> PMID: 37910088; PubMed Central PMCID: PMC10627294.
9. Warren PA, Graf EW, Champion RA, Maloney LT. Visual extrapolation under risk: human observers estimate and compensate for exogenous uncertainty. *Proc Biol Sci.* 2012; 279(1736):2171–9. Epub 2012/02/03. <https://doi.org/10.1098/rspb.2011.2527> PMID: 22298845; PubMed Central PMCID: PMC3321716.
10. Trommershauser J, Maloney LT, Landy MS. Statistical decision theory and the selection of rapid, goal-directed movements. *J Opt Soc Am A Opt Image Sci Vis.* 2003; 20(7):1419–33. Epub 2003/07/19. <https://doi.org/10.1364/josaa.20.001419> PMID: 12868646.
11. Trommershauser J, Maloney LT, Landy MS. Statistical decision theory and trade-offs in the control of motor response. *Spat Vis.* 2003; 16(3–4):255–75. Epub 2003/07/16. <https://doi.org/10.1163/156856803322467527> PMID: 12858951.
12. Ota K, Shinya M, Kudo K. Motor planning under temporal uncertainty is suboptimal when the gain function is asymmetric. *Front Comput Neurosci.* 2015; 9:88. Epub 2015/08/04. <https://doi.org/10.3389/fncom.2015.00088> PMID: 26236227; PubMed Central PMCID: PMC4502360.
13. Ota K, Shinya M, Kudo K. Sub-optimality in motor planning is retained throughout 9 days practice of 2250 trials. *Sci Rep.* 2016; 6:37181. Epub 2016/11/22. <https://doi.org/10.1038/srep37181> PMID: 27869198; PubMed Central PMCID: PMC5116677.
14. Ota K, Shinya M, Maloney LT, Kudo K. Sub-optimality in motor planning is not improved by explicit observation of motor uncertainty. *Sci Rep.* 2019; 9(1):14850. Epub 2019/10/18. <https://doi.org/10.1038/s41598-019-50901-x> PMID: 31619756; PubMed Central PMCID: PMC6795881.
15. Kording KP, Wolpert DM. Bayesian integration in sensorimotor learning. *Nature.* 2004; 427(6971):244–7. Epub 2004/01/16. <https://doi.org/10.1038/nature02169> PMID: 14724638.
16. Hudson TE, Wolfe U, Maloney LT. Speeded reaching movements around invisible obstacles. *PLoS Comput Biol.* 2012; 8(9):e1002676. Epub 2012/10/03. <https://doi.org/10.1371/journal.pcbi.1002676> PMID: 23028276; PubMed Central PMCID: PMC3447970.
17. Zhang H, Morvan C, Etezzad-Heydari LA, Maloney LT. Very slow search and reach: failure to maximize expected gain in an eye-hand coordination task. *PLoS Comput Biol.* 2012; 8(10):e1002718. Epub 2012/10/17. <https://doi.org/10.1371/journal.pcbi.1002718> PMID: 23071430; PubMed Central PMCID: PMC3469464.
18. Juni MZ, Gureckis TM, Maloney LT. Information sampling behavior with explicit sampling costs. *Decision.* 2016; 3(3):147–68. Epub 2016/07/19. <https://doi.org/10.1037/dec0000045> PMID: 27429991; PubMed Central PMCID: PMC4942190.
19. Miyazaki M, Yamamoto S, Uchida S, Kitazawa S. Bayesian calibration of simultaneity in tactile temporal order judgment. *Nat Neurosci.* 2006; 9(7):875–7. Epub 2006/05/30. <https://doi.org/10.1038/nn1712> PMID: 16732276.
20. Maloney LT, Mamassian P. Bayesian decision theory as a model of human visual perception: testing Bayesian transfer. *Vis Neurosci.* 2009; 26(1):147–55. Epub 2009/02/06. <https://doi.org/10.1017/S0952523808080905> PMID: 19193251.
21. Zhang H, Daw ND, Maloney LT. Testing whether humans have an accurate model of their own motor uncertainty in a speeded reaching task. *PLoS Comput Biol.* 2013; 9(5):e1003080. Epub 2013/05/30. <https://doi.org/10.1371/journal.pcbi.1003080> PMID: 23717198; PubMed Central PMCID: PMC3662689.
22. Zhang H, Daw ND, Maloney LT. Human representation of visuo-motor uncertainty as mixtures of orthogonal basis distributions. *Nat Neurosci.* 2015; 18(8):1152–8. Epub 2015/06/30. <https://doi.org/10.1038/nn.4055> PMID: 26120962; PubMed Central PMCID: PMC4487408.
23. Rahnev D, Denison RN. Suboptimality in Perceptual Decision Making. *Behav Brain Sci.* 2018:1–107. Epub 2018/02/28. <https://doi.org/10.1017/S0140525X18000936> PMID: 29485020; PubMed Central PMCID: PMC6110994.
24. Gigerenzer G, Gaissmaier W. Heuristic decision making. *Annu Rev Psychol.* 2011; 62:451–82. Epub 2010/12/04. <https://doi.org/10.1146/annurev-psych-120709-145346> PMID: 21126183.
25. Trommershauser J, Gepshtein S, Maloney LT, Landy MS, Banks MS. Optimal compensation for changes in task-relevant movement variability. *J Neurosci.* 2005; 25(31):7169–78. Epub 2005/08/05. <https://doi.org/10.1523/JNEUROSCI.1906-05.2005> PMID: 16079399; PubMed Central PMCID: PMC6725228.
26. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993.
27. Wasserman L. All of statistics: a concise course in statistical inference. New York: Springer; 2004.
28. Lehmann EL. Theory of point estimation. 2nd ed. New York: Springer; 1998.

29. Hogg RV, Mckean JW, Craig AT. Introduction to mathematical statistics. 7th ed. London: Pearson Education; 2012.
30. Hampel FR. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*. 1974; 69(346). <https://doi.org/10.1080/01621459.1974.10482962>
31. Huber PJ, Ronchetti EM. Robust statistics. 2nd ed. New York: John Wiley & Sons; 2009.
32. Landy MS, Maloney LT, Johnston EB, Young M. Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Res*. 1995; 35(3):389–412. Epub 1995/02/01. [https://doi.org/10.1016/0042-6989\(94\)00176-m](https://doi.org/10.1016/0042-6989(94)00176-m) PMID: 7892735.
33. Maloney LT, Landy MS. A statistical framework for robust fusion of depth information. *Proceedings of the SPIE: Visual Communications and Image Processing IV*. 1989;(1199):1154–63.
34. Gonzalez R, Wu G. On the shape of the probability weighting function. *Cogn Psychol*. 1999; 38(1):129–66. Epub 1999/03/26. <https://doi.org/10.1006/cogp.1998.0710> PMID: 10090801.
35. Prelec D. The Probability Weighting Function. *Econometrica*. 1998; 66(3). <https://doi.org/10.2307/2998573>
36. Tversky A, Kahneman D. Advances in prospect theory: Cumulative representation of uncertainty *Journal of Risk and uncertainty*. 1992 5:297–323.
37. Zhang H, Maloney LT. Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Front Neurosci*. 2012; 6:1. Epub 2012/02/02. <https://doi.org/10.3389/fnins.2012.00001> PMID: 22294978; PubMed Central PMCID: PMC3261445.
38. Barnard GA. Statistical inference. *Journal of the Royal Statistical Society Series B Methodological*. 1949 11:115–49.
39. Berkson J. Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association*. 1944; 39(227):357–65. <https://doi.org/10.1080/01621459.1944.10500699>
40. Luce RD. Utility of gains and losses: Measurement-theoretical and experimental approaches. New Jersey: Lawrence Erlbaum Associate; 2000.
41. Zhang H, Ren X, Maloney LT. The bounded rationality of probability distortion. *Proc Natl Acad Sci U S A*. 2020; 117(36):22024–34. Epub 2020/08/28. <https://doi.org/10.1073/pnas.1922401117> PMID: 32843344; PubMed Central PMCID: PMC7486738.
42. Wu G, Gonzalez R. Curvature of the Probability Weighting Function. *Management Science*. 1996; 42(12):1676–90. <https://doi.org/10.1287/mnsc.42.12.1676>
43. Marr D. Vision: A computational investigation into the human representation and processing of visual information. Massachusetts: MIT Press; 1982.
44. Griffiths TL, Vul E, Sanborn AN. Bridging Levels of Analysis for Probabilistic Models of Cognition. *Current Directions in Psychological Science*. 2012; 21(4):263–8. <https://doi.org/10.1177/0963721412447619>
45. Gershman SJ, Vul E, Tenenbaum JB. Multistability and perceptual inference. *Neural Comput*. 2012; 24(1):1–24. Epub 2011/10/26. https://doi.org/10.1162/NECO_a_00226 PMID: 22023198.
46. Shi L, Griffiths TL, Feldman NH, Sanborn AN. Exemplar models as a mechanism for performing Bayesian inference. *Psychon Bull Rev*. 2010; 17(4):443–64. Epub 2010/08/13. <https://doi.org/10.3758/PBR.17.4.443> PMID: 20702863.
47. Gilboa I. Rational choice. Massachusetts: MIT press; 2012.
48. Geisler WS. Sequential ideal-observer analysis of visual discriminations. *Psychol Rev*. 1989; 96(2):267–314. Epub 1989/04/01. <https://doi.org/10.1037/0033-295x.96.2.267> PMID: 2652171.
49. Knill DC, Pouget A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci*. 2004; 27(12):712–9. Epub 2004/11/16. <https://doi.org/10.1016/j.tins.2004.10.007> PMID: 15541511.
50. Pouget A, Beck JM, Ma WJ, Latham PE. Probabilistic brains: knowns and unknowns. *Nat Neurosci*. 2013; 16(9):1170–8. Epub 2013/08/21. <https://doi.org/10.1038/nn.3495> PMID: 23955561; PubMed Central PMCID: PMC4487650.
51. Rabin M. Risk Aversion and Expected-utility Theory: A Calibration Theorem. *Econometrica*. 2000; 68(5):1281–92. <https://doi.org/10.1111/1468-0262.00158>
52. Acerbi L, Wolpert DM, Vijayakumar S. Internal representations of temporal statistics and feedback calibrate motor-sensory interval timing. *PLoS Comput Biol*. 2012; 8(11):e1002771. Epub 2012/12/05. <https://doi.org/10.1371/journal.pcbi.1002771> PMID: 23209386; PubMed Central PMCID: PMC3510049.
53. Stocker AA, Simoncelli EP. Noise characteristics and prior expectations in human visual speed perception. *Nat Neurosci*. 2006; 9(4):578–85. Epub 2006/03/21. <https://doi.org/10.1038/nn1669> PMID: 16547513.

54. Maloney LT, Thomas EAC. Distributional assumptions and observed conservatism in the theory of signal detectability. *Journal of Mathematical Psychology*. 1991; 35(4):443–70. [https://doi.org/10.1016/0022-2496\(91\)90043-s](https://doi.org/10.1016/0022-2496(91)90043-s)
55. Juni MZ, Gureckis TM, Maloney LT. Effective integration of serially presented stochastic cues. *J Vis*. 2012; 12(8). Epub 2012/08/23. <https://doi.org/10.1167/12.8.12> PMID: 22911906; PubMed Central PMCID: PMC3556466.
56. Li V, Hecce Castanon S, Solomon JA, Vandormael H, Summerfield C. Robust averaging protects decisions from noise in neural computations. *PLoS Comput Biol*. 2017; 13(8):e1005723. Epub 2017/08/26. <https://doi.org/10.1371/journal.pcbi.1005723> PMID: 28841644; PubMed Central PMCID: PMC5589265.
57. Spitzer B, Waschke L, Summerfield C. Selective overweighting of larger magnitudes during noisy numerical comparison. *Nat Hum Behav*. 2017; 1(8):145. Epub 2017/07/17. <https://doi.org/10.1038/s41562-017-0145> PMID: 32340412.
58. Brainard DH. The Psychophysics Toolbox. *Spatial Vision*. 1997; 10:433–6. <https://doi.org/10.1163/156856897X00357>. PMID: 9176952
59. Pelli DG. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*. 1997; 10(4):437–42. <https://doi.org/10.1163/156856897x00366> PMID: 9176953
60. Yarkoni T, Westfall J. Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspect Psychol Sci*. 2017; 12(6):1100–22. Epub 2017/08/26. <https://doi.org/10.1177/1745691617693393> PMID: 28841086; PubMed Central PMCID: PMC6603289.
61. Hastie T, Tibshirani R, Friedman JH. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. New York: Springer; 2009.
62. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference*. 2nd ed. New York: Springer; 2002.