Research Article

Psychological Science
2021, Vol. 32(7) 1157–1168
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0956797621994214
www.psychologicalscience.org/PS

Response Bias Reflects Individual Differences in Sensory Encoding

Dobromir Rahnev
School of Psychology, Georgia Institute of Technology

Abstract

Humans exhibit substantial biases in their decision making even in simple two-choice tasks, but the origin of these biases remains unclear. I hypothesized that one source of bias could be individual differences in sensory encoding. Specifically, if one stimulus category gives rise to an internal-evidence distribution with higher variability, then responses should optimally be biased against that stimulus category. Therefore, response bias may reflect a previously unappreciated subject-to-subject difference in the variance of the internal-evidence distributions. I tested this possibility by analyzing data from three different two-choice tasks (*n*s = 443, 443, and 498). For all three tasks, response bias moved in the direction of the optimal criterion determined by each subject's idiosyncratic internal-evidence variability. These results demonstrate that seemingly random variations in response bias can be driven by individual differences in sensory encoding and are thus partly explained by normative strategies.

Keywords

bias, criterion, perceptual decision making, sensory encoding, confidence, individual differences, response bias, visual perception, open data, open materials

Received 4/3/20; Revision accepted 12/15/20

Humans have the propensity to prefer one stimulus category over another even in the context of simple perceptual tasks. Recent research has revealed that such perceptual biases have a substantial sensory component. For example, Linares et al. (2019) showed that response bias in a two-choice task is driven by both sensory and decisional factors, the sensory factor being larger in magnitude. Other studies have further demonstrated that response bias depends on neuronal excitability (Iemi & Busch, 2018), can be manipulated by optogenetic manipulation of primary visual cortex (V1; Jin & Glickfeld, 2019) and middle temporal area (Fetsch et al., 2018), and can be predicted from the empirical form of the stimulus sensitivity (Wei & Stocker, 2017). Collectively, this line of research demonstrates that response bias in perceptual experiments has a substantial sensory component.

However, previous studies provide limited insight into perhaps the most confounding aspect of response bias: the fact that different people can exhibit opposite biases for the exact same task. For example, people have been shown to have idiosyncratic perceptual biases that change with the spatial location of a stimulus on a display (Afraz et al., 2010; Finlayson et al., 2017; Kosovicheva & Whitney, 2017; Moutsiana et al., 2016; Wexler et al., 2015). Critically, these biases are stable over time but vary greatly from person to person. Similar findings have been reported with simpler tasks such as the existence of stable but idiosyncratic biases for presentation order of stimuli in two-interval forced-choice tasks (García-Pérez & Alcalá-Quintana, 2011) and even in basic two-choice tasks that require subjects to identify whether a single stimulus is tilted clockwise or counterclockwise from vertical (Rahnev & Denison, 2018; Rahnev et al., 2016).

What causes such persistent and idiosyncratic biases in simple perceptual tasks? In a previous article, a colleague and I proposed that individual differences in stimulus encoding may partly explain the response

Corresponding Author:
Dobromir Rahnev, Georgia Institute of Technology, School of Psychology
E-mail: rahnev@psych.gatech.edu

biases in two-choice perceptual experiments (Rahnev & Denison, 2018). In such experiments, bias is typically assessed using signal detection theory, which can separate sensitivity (quantified by the measure of sensitivity, d') from the response bias (quantified by the measure of criterion, c; Green & Swets, 1966). In the absence of uneven priors or payoffs, accuracy is maximized by setting the criterion to 0, which corresponds to an unbiased response strategy. However, this approach assumes that the underlying distributions for the two stimulus categories have equal variance. If the equal-variance assumption is violated, then the optimal strategy is no longer to place a criterion at 0. On the basis of these considerations, we reasoned that individual variability in the variability of the internal distributions for each stimulus category may drive at least some of the idiosyncratic response bias observed in two-choice tasks (Rahnev & Denison, 2018).

What is the exact relationship between the optimal response strategy and the variance of the internal distributions of evidence? Let the distributions of internal evidence for Categories 1 and 2 have standard deviations of σ_1 and σ_2, respectively. Assume that σ_1 is smaller than σ_2. In this scenario, the two distributions no longer intersect halfway between their means; instead, they intersect at a point that is closer to the mean for Category 2 (Fig. 1). (Note that there is also a second point of intersection that does not lie between the two distributions' means. This point should be taken into account by the ideal observer, but ignoring it typically has only a minor effect on subjects' responses. I will return to this issue in the Discussion section.) If the difference in variability is ignored and equal variance is assumed during data analysis, then the optimal strategy, which is to place the criterion at the first point of intersection, would appear biased. Specifically, the optimal criterion, c_{opt}, would be greater than zero (Fig. 1). On the other hand, when σ_1 is larger than σ_2, the same logic dictates that c_{opt} will be smaller than 0. Failing to consider the possibility of internal distributions of unequal variance would thus result in categorizing the otherwise optimal criterion, c_{opt}, as "biased" because it will deviate from the value of 0 that is optimal under the assumption of equal variance.

How can one test whether the response bias of individual subjects indeed reflects their idiosyncratic sensory encoding as measured by the variance of the internal distributions of evidence? As Figure 1 demonstrates, one can expect a simple relationship between the ratio of the two standard deviations, $s = \sigma_1/\sigma_2$ (which can be computed in a straightforward manner for experiments that collect confidence ratings; see Macmillan & Creelman, 2005), and the optimal criterion c_{opt}: As s increases, c_{opt} should decrease. Therefore, whether the actual criterion c reflects the sensory encoding specified by the parameter s can be assessed by conducting an across-subjects correlation between $\log(s)$ and c. Note that because the parameter s is a ratio and thus not normally distributed, it is more appropriate to use its natural logarithm, $\log(s)$, for statistical tests.

On the basis of power calculations, which suggested the need for experiments with at least 250 subjects and 150 trials per subject (see the Method section), I selected three tasks for analysis from the recently published Confidence Database (Rahnev et al., 2020). All three tasks had a sample size greater than 400 and included simple two-choice perceptual judgments, thus making them ideal to test the current hypothesis. To anticipate the findings, I confirmed the presence of stable idiosyncratic biases and found that $\log(s)$ and c were negatively correlated in all three tasks, suggesting that response bias indeed reflects individual differences in sensory encoding.

Statement of Relevance

What is the origin of our decision biases? Centuries of research have shown that humans form preferences for one stimulus category over another in tasks that range from choosing a vacation destination to simple perceptual judgments. Such biases are typically seen as deficiencies to be avoided. However, here I examined whether response bias can reflect normative computations. Using perceptual decision making as a model system, I tested whether decision bias reflects individual differences in sensory encoding. I analyzed the data from three tasks (ns = 443, 443, and 498) reported in previously published articles and applied mathematical modeling to determine the idiosyncratic way in which two stimulus categories are encoded by each subject. The results showed that the response bias reflects the idiosyncratic sensory encoding in each individual, demonstrating that our decision biases are not necessarily deficiencies to be avoided but can normatively reflect how information is internally represented.

Method

Data-set selection

I searched for data sets in which subjects performed two-choice perceptual tasks with confidence ratings.

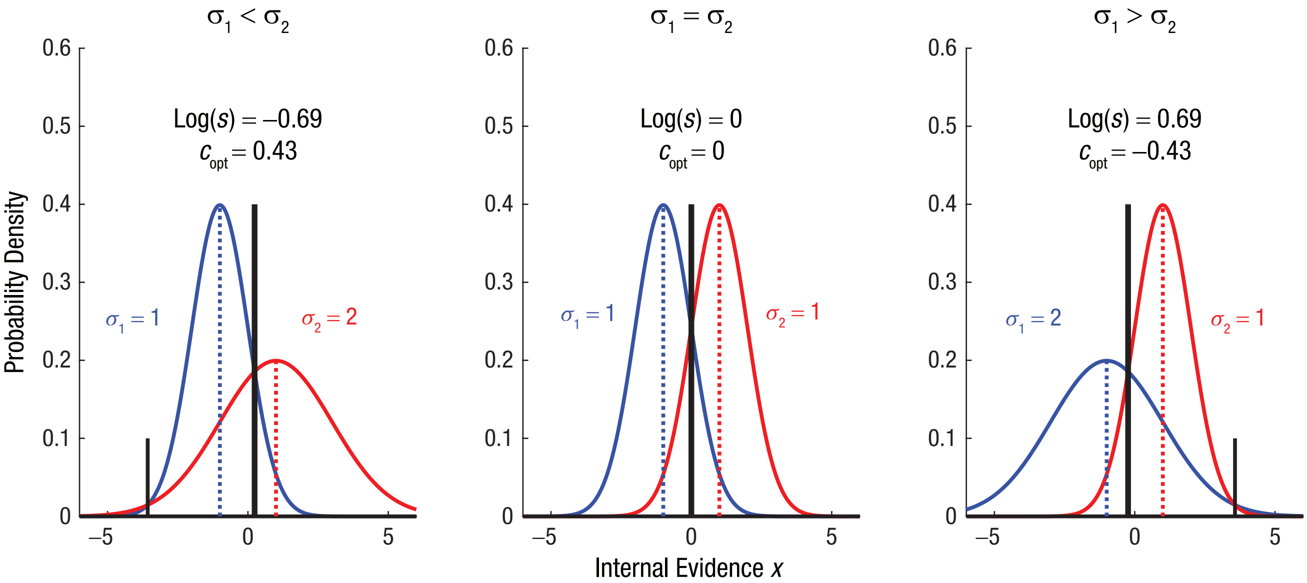


Fig. 1. Dependence of the optimal criterion on sensory encoding. The three graphs depict internal Gaussian distributions of evidence (x) for Categories 1 and 2 with standard deviations σ_1 and σ_2, respectively. The ratio of the two standard deviations, $s = \sigma_1/\sigma_2$, varies from 0.5 in the left graph to 1 in the middle graph to 2 in the right graph. The logarithm of s therefore changes from −0.69 to 0 to 0.69. The corresponding optimal criterion, c_{opt} (long solid vertical line; for details on its computation, see the Method section), changes from 0.43 to 0 to −0.43. The reason for the opposite relationship between log(s) and c_{opt} is that a distribution with a higher standard deviation (e.g., σ_2 in the left graph) brings the point of intersection between the two Gaussian distributions toward its own mean. Taken together, the three graphs demonstrate the negative relationship between log(s) and c_{opt}. Note the existence of a second point of intersection in the left and right panels (short vertical line), which is discussed in greater detail in the Method section.

Critically, on the basis of power analyses (see below), I required the data sets to have at least 250 subjects who each completed at least 150 trials per task. I first examined the 150 data sets included in the recently published Confidence Database (Rahnev et al., 2020) as of October 2020. Two data sets in the Confidence Database met the above criteria: “Haddara_2020_Expt1” and “Rouault_2018_Expt1.” I additionally conducted an extensive literature search for additional data sets but could not find any others that met the above criteria.

Therefore, I performed all analyses on the tasks from these two data sets. The first set, originally reported by Haddara and Rahnev (2020), consists of data from two separate tasks, considered here as Task 1 and Task 2. The second data set, originally reported as Experiment 1 by Rouault et al. (2018), consists of data from a single task, considered here as Task 3. The Haddara and Rahnev data set has 443 subjects, all of whom were included in the Confidence Database regardless of data quality. On the other hand, the Rouault et al. data set originally had 663 subjects, but the authors removed 165 subjects (24.9%) on the basis of exclusion criteria involving task comprehension and performance, leaving them with 498 subjects. The data from these 498 subjects were posted on the Confidence Database; therefore, here I analyzed only these data. All subjects provided informed consent, and the experiments were approved by the local institutional review boards.

Experimental designs

Complete details about Tasks 1 and 2 are available in the original article (Haddara & Rahnev, 2020). Briefly, subjects indicated whether the letter X or O (Task 1) or the color red or blue (Task 2) occurred more frequently in a 7 × 7 grid. In both tasks, each trial began with a fixation period (500 ms), followed by stimulus presentation (500 ms), an untimed perceptual judgment, and an untimed confidence rating provided on a 4-point scale (Figs. 2a and 2b). The two tasks were adapted from the study by Rahnev et al. (2015). In Task 1, approximately half of the subjects received trial-by-trial feedback, and the other half received no such feedback. The feedback screen consisted of the word “Correct” or “Wrong” and was presented for 500 ms. The group that did not receive trial-by-trial feedback saw a fixation cross for 500 ms instead of the feedback screen. All subjects were analyzed together regardless of whether they received feedback. No subject received feedback in Task 2. The more frequent stimulus within the 7 × 7 grid was presented in 30 locations in Task 1 and 27 locations in Task 2. Task 1 consisted of 330 trials

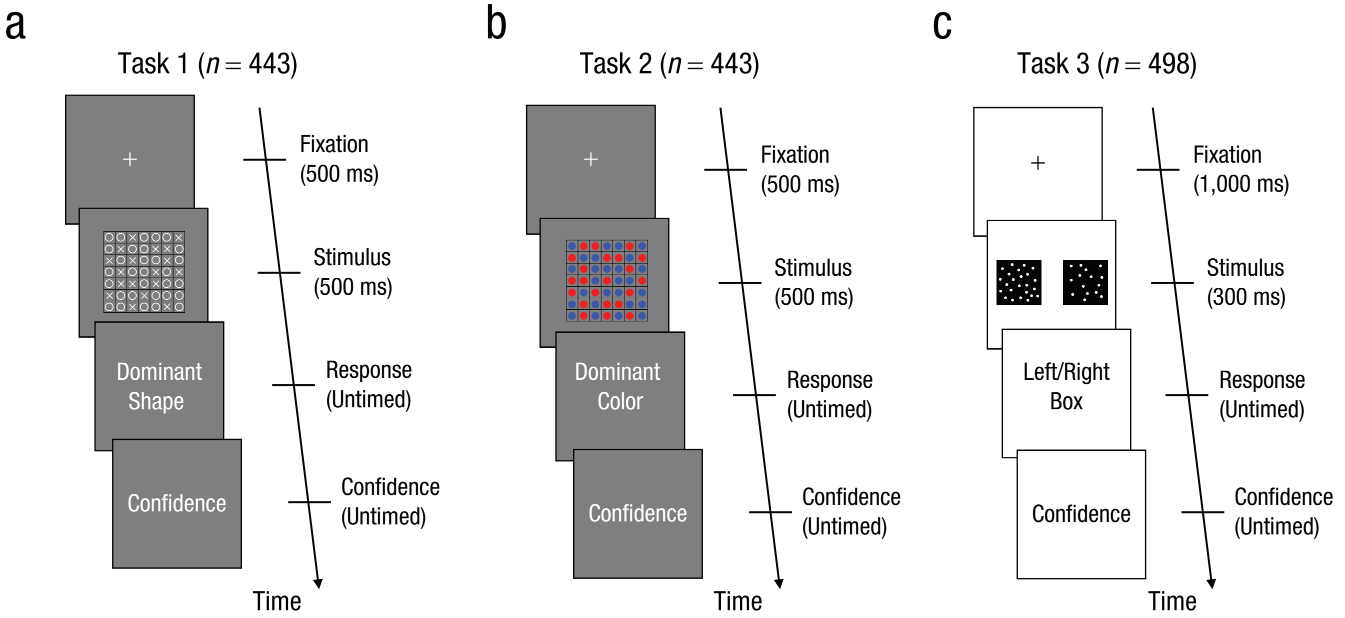


Fig. 2. Experimental tasks. Each task began with a fixation cross, after which a stimulus was presented for a brief period (either 500 or 300 ms). Finally, subjects indicated their perceptual judgment and confidence rating with separate, untimed button presses. Subjects judged (a) whether the letter X or O was more frequent (Task 1), (b) whether the color red or blue was more frequent (Task 2), or (c) whether the left or right box contained more white dots (Task 3). Confidence was given on a 4-point scale in Tasks 1 and 2 and on an 11-point scale in Task 3.

per subject, whereas Task 2 consisted of 150 trials per subject. Both tasks were organized in blocks of 30 trials, and subjects could take a break at the end of each block.

Complete details about Task 3 are available in the original article (Rouault et al., 2018). Briefly, subjects indicated which of two simultaneously presented black boxes had the highest number of white dots. Each trial began with a fixation period (1,000 ms), followed by stimulus presentation (300 ms), an untimed perceptual judgment, and an untimed confidence rating provided on an 11-point scale (Fig. 2c). No feedback was provided during the task. One box was always half-filled (313 dots out of 625 positions), whereas the other box contained an increment of +1 to +70 dots compared with the standard. Task 3 consisted of 210 trials per subject organized in five blocks.

All data were collected online using Amazon Mechanical Turk. The experiments were performed using *jsPsych* (Version 5.0.3 for Tasks 1 and 2, Version 4.3 for Task 3; de Leeuw, 2015).

Subject selection

The analyses here necessitated the estimation of the parameters of a signal detection theory model with unequal variance. However, the parameters of this model are difficult to estimate in the presence of extreme biases or if performance is too low or too high. Therefore, I excluded subjects who gave the same response or the same confidence rating on more than 95% of all trials or had accuracy lower than 55% or higher than 95% correct. In addition, I also excluded individual trials with response times that were faster than 200 ms or slower than 2 s. These criteria were identical to the exclusion criteria used by Haddara and Rahnev (2020).

The selection criteria resulted in the exclusion of 68 subjects in Task 1 (15.3%), 79 subjects in Task 2 (17.8%), and five subjects in Task 3 (1%). The exclusion rate was much lower in Task 3 because the original authors had already excluded a number of subjects using somewhat overlapping exclusion criteria. The exclusions in Tasks 1 and 2 were made independently for each task.

Analyses

I was interested in whether subjects can take into account the idiosyncratic shapes of their internal distributions when making perceptual decisions. As explained in the introduction, this would be reflected in a negative correlation between the criterion, c, and the logarithm of the standard-deviation ratio of the internal distributions, log(s). Note that Haddara and Rahnev (2020) also examined the response criterion using the same data but focused on how trial-by-trial feedback affects bias. They found that bias was reduced in the group that received feedback, but they did not examine the source of the idiosyncratic differences in this bias.

To address whether individual differences in sensory encoding affect response bias, I first computed the parameters *d*′ and *c*, which assume that the Gaussian distributions of evidence for the two stimulus categories have equal variance. To do so, I calculated the hit rate (HR) and false-alarm rate (FAR) by treating the letter X in Task 1, the blue color in Task 2, and the right box in Task 3 as the targets. Then, *d*′ and *c* were calculated using the following formulas:

$$d' = \Phi^{-1}\left(\mathrm{HR}\right) - \Phi^{-1}\left(\mathrm{FAR}\right) \tag{1}$$

and

$$c = -\frac{1}{2}\left(\Phi^{-1}\left(\mathrm{HR}\right) + \Phi^{-1}\left(\mathrm{FAR}\right)\right), \tag{2}$$

where Φ^{-1} is the inverse of the cumulative standard normal distribution that transforms HR and FAR into *z* scores. In cases where HR or FAR were equal to 0 or 1, a standard correction was applied such that if estimated from *k* trials, values of 0 were replaced with 1/(2*k*), whereas values of 1 were replaced with 1 – 1/(2*k*) (Macmillan & Creelman, 2005). Note that negative *c* values indicate a bias for the letter X (Task 1), the color blue (Task 2), and the right box (Task 3), whereas positive *c* values indicate a bias for the letter O (Task 1), the color red (Task 2), and the left box (Task 3).

The computations above assume that both stimulus categories produce internal-evidence distributions with equal variance. However, this assumption does not hold in general. For example, the target distribution has a higher variance in both detection and memory tasks (Macmillan & Creelman, 2005; Rahnev et al., 2011). More generally, even when neither stimulus category results in higher variability across the whole group, it is likely that individual subjects exhibit higher variability for one or the other stimulus category and that the category changes from subject to subject.

To compute the relative variance of the distributions for each stimulus category, I used standard techniques based on subjects' confidence ratings (Green & Swets, 1966). For each confidence and decision criterion, I computed the HR and FAR. I then computed *z*HR and *z*FAR as $\Phi^{-1}\left(\mathrm{HR}\right)$ and $\Phi^{-1}\left(\mathrm{FAR}\right)$, respectively, and found the line of best fit for the plot of *z*HR against *z*FAR values. As has been demonstrated previously (Macmillan & Creelman, 2005), the slope, *s*, of this line is equal to the ratio of the standard deviations of the internal distributions for Categories 1 and 2: $s = \sigma_1/\sigma_2$. Because only the ratio of the two standard deviations is fixed and their actual values do not matter, without loss of generality, one can set σ_1 equal to *s* and σ_2 equal to 1. Given these standard deviations, one can also obtain the distance between the means of the two Gaussian distributions as the intercept of the line of best fit above. Because only the distance between the two means matters, without loss of generality, one can set the mean of the Gaussian distribution for Category 1 to 0 (i.e., $\mu_1 = 0$), and therefore, the mean of the Gaussian distribution for Category 2 becomes the intercept of the line of best fit, which one can denote as μ (i.e., $\mu_2 = \mu$).

Having determined the means and standard deviations for the Gaussian distributions for each stimulus category, one can now compute the location of the optimal decision criterion, x_{opt}. This is the location where the two Gaussian distributions intersect, and therefore, their probability density functions are equal to each other:

$$\frac{1}{\sqrt{2\pi\sigma_1^2}}e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} = \frac{1}{\sqrt{2\pi\sigma_2^2}}e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2}.$$

To compute x_{opt}, one can solve for *x*:

$$\frac{\sqrt{2\pi\sigma_2^2}}{\sqrt{2\pi\sigma_1^2}} = e^{\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2}$$

$$\log\left(\frac{\sigma_2}{\sigma_1}\right) = \frac{\left(x-\mu_1\right)^2\sigma_2^2 - \left(x-\mu_2\right)^2\sigma_1^2}{2\sigma_1^2\sigma_2^2}$$

$$2\sigma_1^2\sigma_2^2\log\left(\frac{\sigma_2}{\sigma_1}\right) = x^2\left(\sigma_2^2-\sigma_1^2\right) + 2x\left(\sigma_1^2\mu_2 - \sigma_2^2\mu_1\right) + \left(\mu_1^2\sigma_2^2 - \mu_2^2\sigma_1^2\right)$$

$$x^2\left(\sigma_2^2-\sigma_1^2\right) + 2x\left(\sigma_1^2\mu_2 - \sigma_2^2\mu_1\right) + \left(\mu_1^2\sigma_2^2 - \mu_2^2\sigma_1^2 - 2\sigma_1^2\sigma_2^2\log\left(\frac{\sigma_2}{\sigma_1}\right)\right) = 0.$$

When σ_1 is equal to σ_2, one obtains the following: $x = (\mu_1 + \mu_2)/2$. This corresponds to the familiar case of equal variance, where the optimal criterion is located halfway between the peaks of the two distributions. However, when σ_1 is not equal to σ_2, the equation above has two solutions corresponding to the two locations where the two Gaussian distributions intercept:

$$x_{1,2} = \frac{\left(\sigma_2^2\mu_1 - \sigma_1^2\mu_2\right) \pm \sqrt{\left(\sigma_1^2\mu_2 - \sigma_2^2\mu_1\right)^2 - \left(\sigma_2^2 - \sigma_1^2\right)\left(\mu_1^2\sigma_2^2 - \mu_2^2\sigma_1^2 - 2\sigma_1^2\sigma_2^2 \log\left(\frac{\sigma_2}{\sigma_1}\right)\right)}}{\sigma_2^2 - \sigma_1^2}. \tag{4}$$

Note that if one considers the case where σ_1 is equal to s, σ_2 is equal to 1, μ_1 is equal to 0, and μ_2 is equal to μ, Equation 4 simplifies to the following:

$$x_{1,2} = \frac{-\mu s^2 \pm \sqrt{\mu^2 s^4 + \left(1 - s^2\right)\left(\mu^2 s^2 - 2s^2 \log\left(s\right)\right)}}{1 - s^2}. \tag{5}$$

Importantly, the solution x_1 (where the two expressions in the nominator are added) lies between the two means, whereas the solution x_2 (where the two expressions are subtracted) is typically an outlier and lies far from both means: It is a large negative value when s is less than 1 and a large positive value when s is greater than 1. The optimal decision strategy is to place criteria at both x_1 and x_2 and then choose one stimulus category for values between x_1 and x_2 and the other category for values outside of this range. However, across all subjects in the three tasks examined here, only 0.09% (i.e., the extreme tail) of the area of the internal distributions lay beyond x_2. Thus, an ideal observer (Knoblauch & Maloney, 2012) would respond differently, on average, once every 1,097 trials compared with an observer who simply ignores x_2. Given that it is theoretically questionable whether human observers can implement the decision strategy prescribed by the optimal observer (Macmillan & Creelman, 2005) and that the ideal observer would anyway produce a very similar pattern of responses, I implemented an observer model that places a criterion only at the location of the first solution, x_1. Therefore, I refer to x_1 as the location of the optimal criterion and call it x_{opt}.

One can then compute the value, c_{opt}, that the optimal criterion would have in standard analyses that assume equal variance for the two Gaussian distributions:

$$c_{\text{opt}} = -\frac{1}{2}\left(\frac{\mu_2 - x_{\text{opt}}}{\sigma_2} + \frac{\mu_1 - x_{\text{opt}}}{\sigma_1}\right). \tag{6}$$

Specifically, c_{opt} is the value of the optimal criterion if bias is to be computed in the standard way using Equation 2.

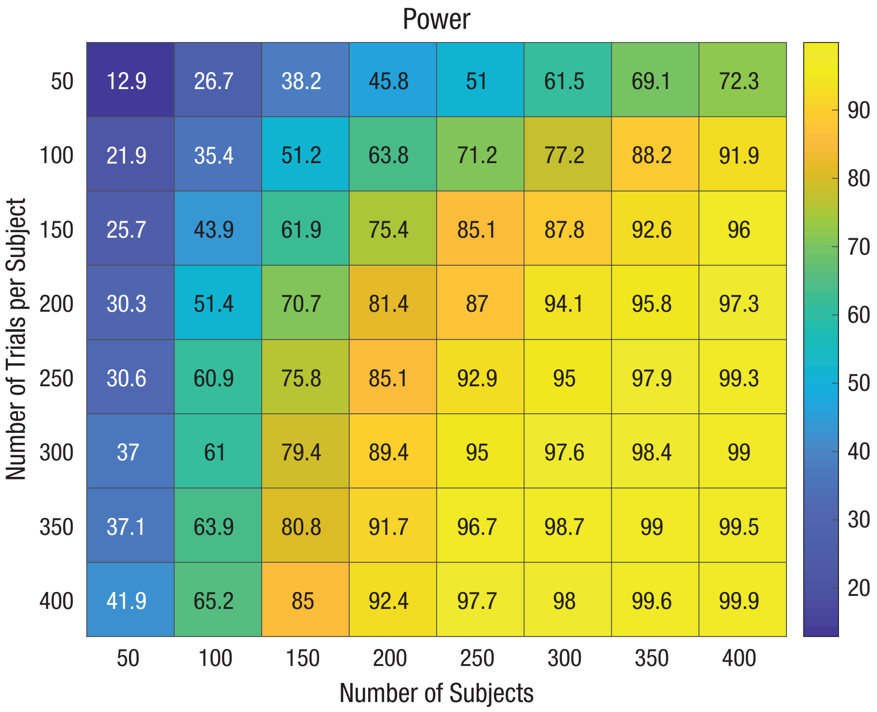
Having obtained the ratio of the standard deviations for the two categories, s, and the value of the optimal criterion, c_{opt}, I could then explore how these two quantities relate to the actual location of the decision criterion, c. Specifically, if subjects have access to the fact that the two stimulus categories result in internal distributions of different variances, then the criterion c that they use should correlate negatively with the ratio s but positively with the optimal criterion c_{opt}.

To test whether this is indeed the case, for each of the three tasks, I correlated c with both log(s) and c_{opt}. I used log(s) because s is on a log scale where $s = y$ and $s = 1/y$ correspond to equivalent scenarios (since the category labels are simply switched), which is correctly captured when taking its logarithm—because $\log(y) = -\log(1/y)$. In addition, the fact that s is a ratio means that its distribution has a heavy skew to the right, whereas the distribution of log(s) is symmetric and approximately Gaussian. Finally, I excluded values of s smaller than 1/3 and larger than 3 as outliers. These exclusions resulted in removing three subjects from Task 1 (0.8%), 16 subjects from Task 2 (4.4%), and five subjects from Task 3 (1%).

Although my analyses were conducted within the framework of signal detection theory, I suspect that similar results would be obtained if the analyses were conducted within different frameworks, such as sequential sampling (Forstmann et al., 2016) or exemplar-storage models (McKinley & Nosofsky, 1995). Therefore, the critical idea here is not the use of the specific signal detection metrics but rather the relating of idiosyncratic biases to individual differences in sensory encoding. Further, in line with more than 60 years of research (Green & Swets, 1966), my analyses assume that each stimulus category gives rise to a Gaussian internal distribution. It is possible that these results would change if distributions of very different shapes were assumed, but such distributions would be at odds with a wide variety of findings that point toward Gaussian variability (Macmillan & Creelman, 2005) and are therefore not entertained here.

Power analyses

Demonstrating a significant negative correlation between log(s) and c faces at least two challenges. First, both log(s) and c require a large number of trials per subject for accurate estimation. Second, most subjects exhibit values of log(s) and c close to zero, which makes it hard to establish how these two quantities correlate with each other (the limited ranges make it harder to detect a significant correlation). To gain insight into the severity of these challenges, I performed simulations to quantify the power of different experimental setups



Number of Trials per Subject	50	100	150	200	250	300	350	400
50	12.9	26.7	38.2	45.8	51	61.5	69.1	72.3
100	21.9	35.4	51.2	63.8	71.2	77.2	88.2	91.9
150	25.7	43.9	61.9	75.4	85.1	87.8	92.6	96
200	30.3	51.4	70.7	81.4	87	94.1	95.8	97.3
250	30.6	60.9	75.8	85.1	92.9	95	97.9	99.3
300	37	61	79.4	89.4	95	97.6	98.4	99
350	37.1	63.9	80.8	91.7	96.7	98.7	99	99.5
400	41.9	65.2	85	92.4	97.7	98	99.6	99.9

Fig. 3. Power for observing a significantly negative correlation between log(*s*) and *c* when the criterion *c* reflects the optimal criterion c_{opt}. Each cell represents the percentage of expected significant correlations for a given combination of number of subjects and number of trials per subject. The values were obtained via simulations, which is why there are several inversions where a cell with a higher number of trials shows lower power than a cell with a lower number of trials.

to reveal a significant correlation between log(*s*) and *c* for cases in which the criterion *c* closely reflected the optimal criterion c_{op}. I ran 64 sets of simulations by varying both the number of subjects and the number of trials per subject from 50 to 400 in steps of 50. For each set of simulations, I generated 1,000 individual experiments with the corresponding number of subjects and number of trials per subject.

For each subject of each simulated experiment, without loss of generality, I set μ_1 equal to 0, and μ_2 equal to μ, σ_1 equal to *s*, σ_2 equal to 1. The values of μ were then sampled from a normal distribution with a mean of 1.5 and standard deviation of 0.5, whereas the values of *s* were sampled so that log(*s*) came from a normal distribution with a mean of 0 and standard deviation of 0.15. I chose these values on the basis of the values empirically observed in the three tasks reported here. In addition, the location of the decision criterion was sampled from a normal distribution centered on the optimal criterion location for the subject and a standard deviation of 0.3. Finally, I determined the locations of three confidence criteria (so that confidence would be given on a 4-point scale) for each subject so that the location of each criterion was greater than the preceding criterion by a value sampled from a uniform distribution U(0,1).

For each simulated experiment, I computed log(*s*) and *c* on the basis of the simulated data alone (without any reference to the true generating parameters). I then correlated these two values across subjects and reported the percentage of simulated experiments (among the 1,000 for each simulation set) for which *r* is less than 0 and *p* is less than .05 (Fig. 3). On the basis of the estimated power, it appears that experiments need at least 250 subjects and at least 150 trials per subject to have sufficient power to uncover a significant negative correlation between log(*s*) and *c*.

The values in Figure 3 should monotonically increase within each row and column. However, because these values were produced using simulations, several inversions can be observed. Finally, I note that the obtained power estimates depend on the values of the parameters chosen for the simulations—mainly, the variability of log(*s*) and the variability of the location of the actual criterion around the optimal criterion. Nevertheless, given that all of the values chosen here are based on

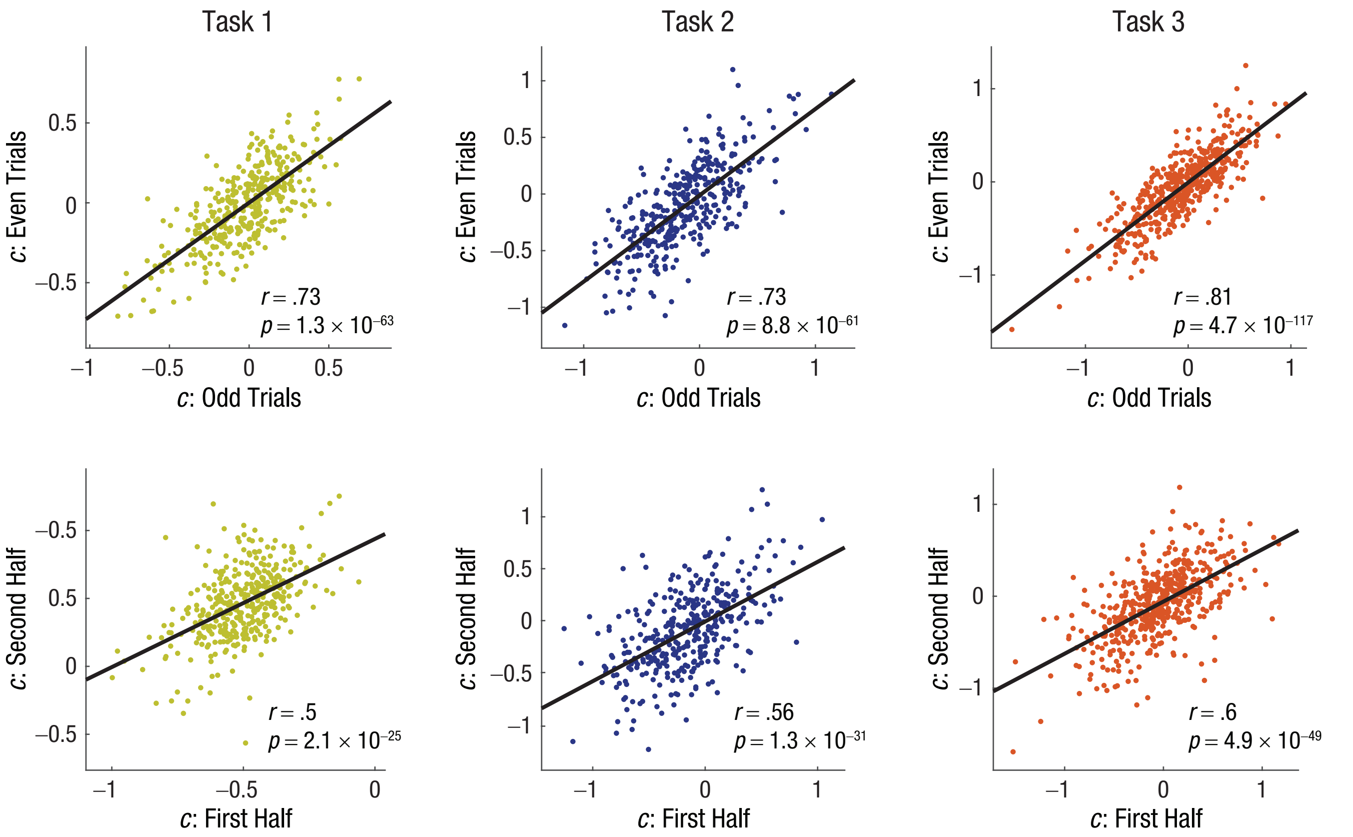


Fig. 4. Scatterplots showing the association between criterion, *c*, on odd and even trials (top row) and in the first half and second half of the experiment (bottom row), separately for each of the three tasks. Diagonal lines indicate best-fitting regressions.

the empirically observed data, the obtained power for each set of simulations is likely to be a reasonable approximation of the true power.

Results

I investigated whether response bias reflects individual differences in how different stimulus categories are encoded by each subject. Denison and I previously predicted that response bias may be partly driven by the fact that the ratio, *s*, of the standard deviations of the two internal distributions in two-choice tasks likely varies between subjects (Rahnev & Denison, 2018). Specifically, a larger standard deviation for a given category would optimally result in a response criterion that is shifted toward the mean of that distribution (Fig. 1), thus resulting in a negative correlation between log(*s*) and the criterion *c*. I tested for this relationship using two different data sets (Haddara & Rahnev, 2020; Rouault et al., 2018), made available as part of the Confidence Database (Rahnev et al., 2020), that consisted of three separate tasks (*ns* = 443, 443, and 498).

I first examined whether there was evidence for the existence of stable idiosyncratic differences in response bias. To this end, I computed the criterion *c* for odd and even trials within each subject's data and correlated these values across subjects for each of the three tasks. I found that the two criterion values were highly correlated (Task 1: $r = .73$, $p = 1.3 \times 10^{-63}$, 95% confidence interval [CI] = [.68, .77]; Task 2: $r = .73$, $p = 8.8 \times 10^{-61}$, 95% CI = [.67, .77]; Task 3: $r = .81$, $p = 4.7 \times 10^{-117}$, 95% CI = [.78, .84]; Fig. 4, top row). These results clearly demonstrate that response bias can be reliably calculated and that similar biases emerge when independent sets of trials are analyzed. I further examined whether these criterion values remained stable over the time course of each experiment by computing the correlation between the criterion *c* derived separately for the first half and second half of all trials for a given subject. I again found positive correlations for all three tasks (Task 1: $r = .5$, $p = 2.1\times 10^{-25}$, 95% CI = [.42, .57]; Task 2: $r = .56$, $p = 1.3 \times 10^{-31}$, 95% CI = [.49, .63]; Task 3: $r = .6$, $p = 4.9 \times 10^{-49}$, 95% CI = [.54, .65]; Fig. 4, bottom row), although the strength of these correlations was predictably

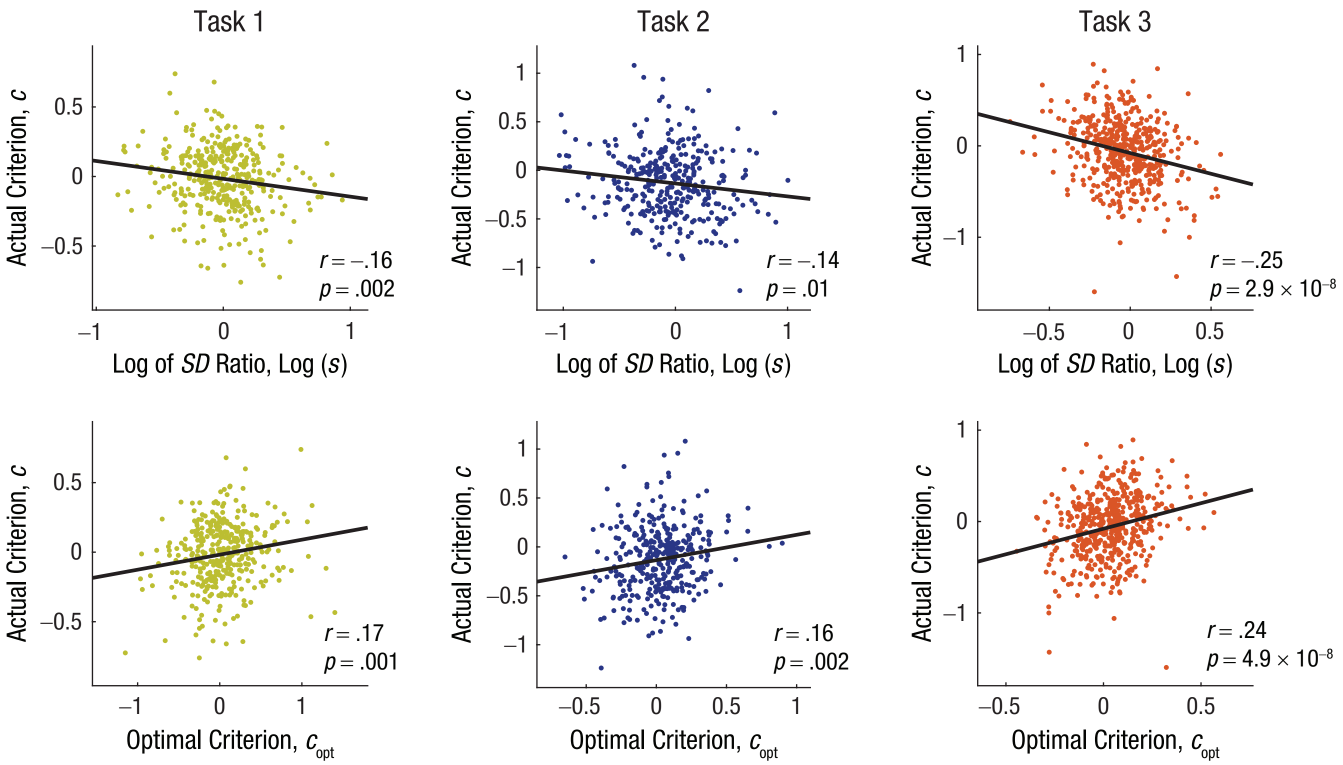


Fig. 5. Scatterplots showing the association between actual criterion, c, and the log of the standard-deviation ratio, log(s) (top row), and between actual criterion, c, and optimal criterion, c_{opt} (bottom row), separately for each of the three tasks. Diagonal lines indicate best-fitting regressions.

slightly lower compared with the correlations between odd and even trials. Overall, these results point to the existence of stable, idiosyncratic biases that vary substantially between different subjects.

Critically, I tested whether these stable, idiosyncratic response biases reflected individual differences in sensory encoding. As explained in Figure 1, subjects with higher values of log(s) would optimally have lower values of the criterion c, and subjects with lower values of log(s) would optimally have higher values of the criterion c. Therefore, if subjects were sensitive to their individual sensory encoding, then I expected to observe a negative across-subjects correlation between log(s) and c. Consistent with this prediction, correlations between log(s) and c were significantly negative for each one of the three tasks (Task 1: $r = -.16$, $p = .002$, 95% CI = [−.25, −.06]; Task 2: $r = -.14$, $p = .009$, 95% CI = [−.24, −.04]; Task 3: $r = -.25$, $p = 2.9 \times 10^{-8}$, 95% CI = [−.33, −.16]; Fig. 5, top row), thus providing direct support for the notion that response bias reflects individual differences in sensory encoding.

It should be noted that the correlation between log(s) and c does not take the sensitivity of each observer into account. Therefore, to account for the varying sensitivity between subjects, I directly estimated the optimal criterion c_{opt} for each subject on the basis of both the ratio of the two standard deviations of the internal distributions, s, and the distance between their means, μ. I then correlated the actual criterion, c, with the optimal criterion, c_{opt}. I found significant positive correlations between these two quantities for each of the three tasks (Task 1: $r = .17$, $p = .001$, 95% CI = [.07, .27]; Task 2: $r = .16$, $p = .002$, 95% CI = [.06, .27]; Task 3: $r = .24$, $p = 4.9 \times 10^{-8}$, 95% CI = [.16, .33]; Fig. 5, bottom row). Together, these results demonstrate that subjects can take the nature of their idiosyncratic sensory encoding into account when making perceptual decisions.

Finally, I performed two control analyses. First, I confirmed that the pattern of results in Figure 5 could be replicated in my simulations. Using the same methods and parameters from the power computations in Figure 3, I simulated an experiment with 400 subjects and 200 trials per subject, thus roughly matching Tasks 1 to 3. I then analyzed the simulated data in the same way as the actual experimental data and confirmed that the resulting scatterplots appeared to be very similar to the results in Figure 5 (see Fig. S1 in the Supplemental Material available online). These simulations confirm that the results I observed could plausibly be obtained if subjects' criterion indeed followed the optimal criterion placement. Second, given the relatively small effect sizes obtained in Figure 5—average r values were −.18 for the correlations between log(s) and c and .19 for the correlations between c and c_{opt}—I sought to examine what a reasonable upper bound would be for the correlations in the tasks I was examining. I repeated the same simulations from the previous analysis 1,000 times and assessed the relationship between the estimated criterion c with the true underlying parameters s_{true} (used to generate the simulated data) and $c_{\text{true opt}}$ (computed directly on the basis of the value of s_{true}). I found average correlation coefficients (rs) of −.33 (SD = .05) for the correlation between log(s_{true}) and c and .36 (SD = .05) for the correlation between c and $c_{\text{true opt}}$ (see Fig. S2 in the Supplemental Material), suggesting that the maximum correlations that could be expected are only about twice as large as the ones actually observed. A major factor for this relatively low upper bound is the presence of estimation noise when computing signal detection theory parameters from limited data.

Discussion

It is well known that humans can appropriately place decision criteria on the basis of the specific task demands. This ability has been demonstrated even for cases in which the shapes and variabilities of the category distributions are experimentally manipulated to necessitate the use of complex, multidimensional criteria (Ashby & Maddox, 2005; McKinley & Nosofsky, 1995). Nevertheless, even when performing the same task, individual subjects tend to have stable, idiosyncratic biases. Here, I investigated whether individual differences in response bias reflect subject-by-subject idiosyncrasies in sensory encoding. I reanalyzed data from three different tasks (ns = 443, 443, and 498) from two previous articles (Haddara & Rahnev, 2020; Rouault et al., 2018). I confirmed the existence of stable individual differences in response bias and, critically, found that the response bias reflected the idiosyncratic sensory encoding of individual subjects. These results demonstrate that response bias is not simply a bug (Summerfield & Li, 2018) but that it follows normative principles that inform how one should respond given how sensory information is represented internally.

The sources of idiosyncratic bias in perceptual decision making

Several previous studies have investigated idiosyncratic biases and have proposed different sources for them. One line of research has suggested that individual biases may arise from inhomogeneities in how visual information is processed in different neuronal populations in the visual cortex (Afraz et al., 2010). This proposal can explain various idiosyncratic perceptual biases that depend on the spatial location of a stimulus, such as biases (a) in gender and age identification (Afraz et al., 2010), (b) in judgments of the direction of optic flow (Wexler et al., 2015), and (c) in object localization (Kosovicheva & Whitney, 2017). In this view, response bias is an inevitable phenomenon arising directly from the limitations in sensory processing, and therefore, this proposal is conceptually different from the view that response biases may partly reflect normative principles.

An alternative account of the existence of stable response biases is that they are due to stable dispositions that can be characterized as individual traits (Kantner & Lindsay, 2012, 2014). This view is based on studies that show that individuals have similar biases in old/new memory tasks across time and even across different tasks.

Here, I argue that beyond being an inevitable consequence of inhomogeneous neural processing or an individual predisposition, bias also reflects normative strategies based on the idiosyncrasies of information encoding for each subject. However, these three explanations of response bias are not mutually exclusive. For example, it is possible that bias for choosing red over blue (to pick one example) could be simultaneously due to (a) inevitable individual differences in the strength of neural activations induced by each color, (b) personal preference for choosing one color over the other that is fully independent of the sensory information, and (c) normative strategies that consider the form of the internal encoding for blue and red colors. Therefore, my findings should not be taken to imply that response bias is optimal or that it is exclusively driven by normative principles. However, my results demonstrate that response bias is not simply a failing of the sensory or decisional systems and can, in fact, be the result of an adaptive process that is well adjusted to the individual differences in information processing.

Differences from the ideal-observer model

It should be noted that my analyses implemented the standard assumption that subjects place a single decision criterion directly on the evidence axis. However, this analysis ignores the fact that in cases of unequal variance, the two Gaussian distributions intersect at two different points, and therefore, the ideal observer places two separate decision criteria on the evidence axis

(Knoblauch & Maloney, 2012). Nevertheless, for the subjects in these experiments, placing two criteria would result in a different response only once every 1,097 trials and is thus unlikely to meaningfully alter my results. More importantly, it is questionable whether human subjects can implement the complex decision strategy required by the optimal observer, especially in experiments with confidence ratings (Macmillan & Creelman, 2005; Rahnev & Denison, 2018). Therefore, I do not claim that human subjects implement the optimal decision strategy, and this question is outside the scope of this article. Instead, what my results show is that humans are sensitive to their idiosyncratic sensory encoding even if their response strategy falls short of optimality.

Limitations

Although my results were replicated in three independent tasks, the actual effect sizes observed were modest. Indeed, the correlation between the actual criterion and optimal criterion was, on average, only .19. This modest correlation suggests that although humans do take their idiosyncratic sensory encoding into account when setting their decision criterion, this may have only a small influence on the final decision criterion. Nevertheless, additional simulations suggest that because of small numbers of trials per subject, and the accompanying estimation noise, the maximum correlation one could expect is only about twice as large (see Fig. S2). Therefore, the modest effect size in my data is not a reliable indicator of the importance of the internal sensory distributions in the setting of the response criterion (ideally, what is needed for a reliable estimate is a study with both a very large number of subjects and a very large number of trials per subject). Ultimately, although these results are a strong indicator that subjects are sensitive to the idiosyncratic differences in sensory encoding, it is possible that the influence of these idiosyncrasies is relatively small and that the other factors already discussed have a larger influence on the criterion.

Another limitation of this study is that it cannot reveal exactly how subjects learned the relative variability of the internal distributions. Given that the different tasks had between 150 and 330 trials per subject, the learning process likely relied on heuristics that can be applied even after a few trials early in the experiment. Nevertheless, the exact mechanisms of how the criterion is learned over the course of an experiment remain to be described. Finally, although I have used standard techniques to estimate the relative variability of the two internal distributions (Macmillan & Creelman, 2005), future studies may benefit from an independent estimation of the relative variability using tasks such as magnitude estimation (Petzschner et al., 2015).

Conclusion

Using three large data sets (all *n*s > 400), I found that human perceptual decision making reflects one's idiosyncratic sensory encoding. These results demonstrate that normative considerations can explain, at least in part, why different subjects have different biases on the exact same task.

Transparency

Action Editor: Krishnankutty Sathian
Editor: Patricia J. Bauer
Author Contributions

D. Rahnev is the sole author of this article and is responsible for its content.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was supported by the National Institutes of Health (Award R01MH119189) and the Office of Naval Research (Award N00014-20-1-2622).

Open Practices

All data, materials, and analysis code have been made publicly available via OSF and can be accessed at https://osf.io/e4mbv. The design and analysis plan for the study were not preregistered. This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797621994214

References

Afraz, A., Pashkam, M. V., & Cavanagh, P. (2010). Spatial heterogeneity in the perception of face and form attributes. *Current Biology*, *20*(23), 2112–2116. https://doi.org/10.1016/j.cub.2010.11.017

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, *56*(1), 149–178. https://doi.org/10.1146/annurev.psych.56.091103.070217

de Leeuw, J. R. (2015). jsPsych: A Javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Fetsch, C. R., Odean, N. N., Jeurissen, D., El-Shamayleh, Y., Horwitz, G. D., & Shadlen, M. N. (2018). Focal optogenetic suppression in macaque area MT biases direction discrimination and decision confidence, but only transiently. *eLife*, *7*, Article e36523. https://doi.org/10.7554/eLife.36523

Finlayson, N. J., Papageorgiou, A., & Schwarzkopf, D. S. (2017). A new method for mapping perceptual biases across visual space. *Journal of Vision*, *17*(9), Article 5. https://doi.org/10.1167/17.9.5

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology*, *67*, 641–666. https://doi.org/10.1146/annurev-psych-122414-033645

García-Pérez, M. A., & Alcalá-Quintana, R. (2011). Interval bias in 2AFC detection tasks: Sorting out the artifacts. *Attention, Perception, & Psychophysics*, *73*(7), 2332–2352. https://doi.org/10.3758/s13414-011-0167-x

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. John Wiley.

Haddara, N., & Rahnev, D. (2020). *The impact of feedback on perceptual decision making and metacognition: Reduction in bias but no change in sensitivity*. PsyArXiv. https://doi.org/10.31234/OSF.IO/P8ZYW

Iemi, L., & Busch, N. A. (2018). Moment-to-moment fluctuations in neuronal excitability bias subjective perception rather than strategic decision-making. *eNeuro*, *5*(3), Article ENEURO.0430-17.2018. https://doi.org/10.1523/ENEURO.0430-17.2018

Jin, M., & Glickfeld, L. L. (2019). Contribution of sensory encoding to measured bias. *The Journal of Neuroscience*, *39*(26), 5115–5127. https://doi.org/10.1523/JNEUROSCI.0076-19.2019

Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition*, *40*(8), 1163–1177. https://doi.org/10.3758/s13421-012-0226-0

Kantner, J., & Lindsay, D. S. (2014). Cross-situational consistency in recognition memory response bias. *Psychonomic Bulletin & Review*, *21*(5), 1272–1280. https://doi.org/10.3758/s13423-014-0608-3

Knoblauch, K., & Maloney, L. T. (2012). *Modeling psychophysical data in R*. Springer. https://doi.org/10.1007/978-1-4614-4475-6

Kosovicheva, A., & Whitney, D. (2017). Stable individual signatures in object localization. *Current Biology*, *27*(14), R700–R701. https://doi.org/10.1016/j.cub.2017.06.001

Linares, D., Aguilar-Lleyda, D., & López-Moliner, J. (2019). Decoupling sensory from decisional choice biases in perceptual decision making. *eLife*, *8*, Article e43994. https://doi.org/10.7554/eLife.43994

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Erlbaum.

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(1), 128–148. https://doi.org/10.1037/0096-1523.21.1.128

Moutsiana, C., de Haas, B., Papageorgiou, A., van Dijk, J. A., Balraj, A., Greenwood, J. A., & Schwarzkopf, D. S. (2016). Cortical idiosyncrasies predict the perception of object size. *Nature Communications*, *7*(1), Article 12110. https://doi.org/10.1038/ncomms12110

Petzschner, F. H., Glasauer, S., & Stephan, K. E. (2015). A Bayesian perspective on magnitude estimation. *Trends in Cognitive Sciences*, *19*(5), 285–293. https://doi.org/10.1016/j.tics.2015.03.002

Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences*, *41*, Article e223. https://doi.org/10.1017/S0140525X18000936

Rahnev, D., Desender, K., Lee, A. L. F., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., Arbuzova, P., Atlas, L. Y., Balcı, F., Bang, J. W., Bègue, I., Birney, D. P., Brady, T. F., Calder-Travis, J., Chetverikov, A., Clark, T. K., Davranche, K., Denison, R. N., Dildine, T. C., . . . Zylberberg, A. (2020). The Confidence Database. *Nature Human Behaviour*, *4*(3), 317–325. https://doi.org/10.1038/s41562-019-0813-1

Rahnev, D., Koizumi, A., McCurdy, L. Y., D'Esposito, M., & Lau, H. (2015). Confidence leak in perceptual decision making. *Psychological Science*, *26*(11), 1664–1680. https://doi.org/10.1177/0956797615595037

Rahnev, D., Maniscalco, B., Graves, T., Huang, E., De Lange, F. P., & Lau, H. (2011). Attention induces conservative subjective biases in visual perception. *Nature Neuroscience*, *14*(12), 1513–1515. https://doi.org/10.1038/nn.2948

Rahnev, D., Nee, D. E., Riddle, J., Larson, A. S., & D'Esposito, M. (2016). Causal evidence for frontal cortex organization for perceptual decision making. *Proceedings of the National Academy of Sciences, USA*, *113*(20), 6059–6064. https://doi.org/10.1073/pnas.1522551113

Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric symptom dimensions are associated with dissociable shifts in metacognition but not task performance. *Biological Psychiatry*, *84*(6), 443–451. https://doi.org/10.1016/j.biopsych.2017.12.017

Summerfield, C., & Li, V. (2018). Perceptual suboptimality: Bug or feature? *Behavioral and Brain Sciences*, *41*, Article e245. https://doi.org/10.1017/S0140525X18001437

Wei, X.-X., & Stocker, A. A. (2017). Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences, USA*, *114*(38), 10244–10249. https://doi.org/10.1073/pnas.1619153114

Wexler, M., Duyck, M., & Mamassian, P. (2015). Persistent states in vision break universality and time invariance. *Proceedings of the National Academy of Sciences, USA*, *112*(48), 14990–14995. https://doi.org/10.1073/pnas.1508847112