NEUROSCIENCE of Consciousness

Neuroscience of Consciousness, 2021, **7(1)**, 1–11

DOI: https://doi.org/10.1093/nc/niab039

Research Article

A robust confidence–accuracy dissociation via criterion attraction

Dobromir Rahnev*,†

School of Psychology, Georgia Institute of Technology, 654 Cherry Str. NW, Atlanta, GA 30332, USA

†Dobromir Rahnev, http://orcid.org/0000-0002-5265-2559

*Correspondence address. School of Psychology, Georgia Institute of Technology, 654 Cherry Str. NW, Atlanta, GA 30332, USA. E-mail: rahnev@psych.gatech.edu

Abstract

Many studies have shown that confidence and accuracy can be dissociated in a variety of tasks. However, most of these dissociations involve small effect sizes, occur only in a subset of participants, and include a reaction time (RT) confound. Here, I develop a new method for inducing confidence–accuracy dissociations that overcomes these limitations. The method uses an external noise manipulation and relies on the phenomenon of criterion attraction where criteria for different tasks become attracted to each other. Subjects judged the identity of stimuli generated with either low or high external noise. The results showed that the two conditions were matched on accuracy and RT but produced a large difference in confidence (effect appeared for 25 of 26 participants, effect size: Cohen's $d = 1.9$). Computational modeling confirmed that these results are consistent with a mechanism of criterion attraction. These findings establish a new method for creating conditions with large differences in confidence without differences in accuracy or RT. Unlike many previous studies, however, the current method does not lead to differences in subjective experience and instead produces robust confidence–accuracy dissociations by exploiting limitations in post-perceptual, cognitive processes.

Keywords: criterion attraction; confidence; metacognition; perceptual decision making

Highlights

- A new method is developed for inducing confidence–accuracy dissociations using external noise and criterion attraction.
- The method leads to large effects and consistent results across participants.
- The novel confidence–accuracy dissociation includes no reaction time confounds.
- Results point toward criterion attraction as a robust effect that is likely to have wide applicability.

Introduction

It is well known that across a variety of tasks, subjective ratings of confidence tend to closely track one's objective level of performance (Mamassian 2016). The close correspondence between confidence and accuracy has made studying subjective evaluation especially challenging because it has been difficult to separate it from objective performance on the main task (Morales *et al.* 2015a, 2019). One strategy for understanding the factors that specifically drive confidence has been to create conditions that are matched in objective performance (i.e. stimulus sensitivity) but differ in subjective performance (i.e. confidence).

Confidence–accuracy dissociations

The last decade has seen a proliferation of experiments on how subjective and objective performance can be dissociated. Most of this work has been done in the domain of perception. This research program has demonstrated that confidence–accuracy dissociations can be produced by a number of different factors including positive evidence bias (Zylberberg *et al.* 2012; Koizumi *et al.* 2015; Maniscalco *et al.* 2016; Samaha *et al.* 2016; Peters *et al.* 2017; Odegaard *et al.* 2018), stimulus variability (Zylberberg *et al.* 2014, 2016; de Gardelle and Mamassian 2015; Spence *et al.* 2016, 2018; Boldt *et al.* 2017, 2019; Desender *et al.* 2018), motor preparation and execution (Fleming *et al.* 2015; Gajdos *et al.* 2019), visual field location (Solovey *et al.* 2015; Li *et al.* 2018), transcranial magnetic stimulation (Rounis *et al.* 2010; Rahnev *et al.* 2012b, 2016; Shekhar and Rahnev 2018), varying pre-stimulus brain activity (Rahnev *et al.* 2012a; Samaha *et al.* 2017), confidence on the previous trial (Rahnev *et al.* 2015; Aguilar-Lleyda *et al.* 2021), attention (Wilimzig *et al.* 2008; Rahnev *et al.* 2011; Kurtz *et al.* 2017; Recht *et al.* 2019), arousal (Allen *et al.* 2016), unconsciously presented information (Vlassova *et al.* 2014), and stimulus visibility (Rausch *et al.* 2018). Many other factors that cause such dissociations are likely to be discovered in the coming years (Rahnev *et al.* 2021).

However, despite the existence of a very large number of confidence–accuracy dissociations, most previous manipulations are limited in three different ways. First, the dissociations are

Received: 8 June 2021; **Revised:** 28 September 2021; **Accepted:** 5 October 2021

typically small in magnitude. For example, if confidence is collected on a 4-point scale for two conditions matched on accuracy, most manipulations result in a confidence difference of about 0.1 and virtually never exceed 0.2. Second, many of the dissociations appear in some but not all participants. Third, while this is rarely explicitly reported, many dissociations include a reaction time (RT) confound such that the conditions that differ in confidence are matched in accuracy but not in RT. Indeed, when RT has been reported, it has often been shorter in the condition with high confidence (Samaha *et al.* 2016; Boldt *et al.* 2019). Such RT effects can even lead to questions as to whether these dissociations truly show a divergence of subjective confidence from objective performance because, in many theories, RTs are an integral part of the signal on which confidence ratings are based (Hanks *et al.* 2011; Fetsch *et al.* 2014a,b; Zylberberg *et al.* 2016).

While not every manipulation suffers from all three limitations, it is currently unclear whether any method induces confidence–accuracy dissociations that are large in magnitude, consistent across participants, and free of RT confounds. Given the importance of robustly dissociating confidence and accuracy for understanding the neural and computational mechanisms of confidence (Shekhar and Rahnev 2021a), it is important to exactly develop such manipulations.

Employing criterion attraction in an external noise paradigm

Here I explore whether a robust confidence–accuracy dissociation can be induced by exploiting the principle of criterion attraction and combining it with an external noise paradigm. Criterion attraction is the idea that when people perform two interleaved tasks that optimally require different criteria, the actual criteria used for the two tasks become 'attracted' toward each other. The phenomenon was first demonstrated by Gorea and Sagi in a series of experiments that suggested that in some conditions the criteria may even collapse onto the same unified criterion (Gorea and Sagi 2000, 2001, 2002). Whether the criterion is truly unified across conditions has been a source of controversy (Kontsevich *et al.* 2002; Denison *et al.* 2018; Lee *et al.* 2021) but criterion attraction can be a robust phenomenon even if the criteria from the different conditions never completely collapse onto each other.

It has been argued that to convincingly establish the location of one's internal criterion, it is necessary to employ an external noise paradigm (Denison *et al.* 2018; Lee *et al.* 2021). In external noise paradigms, the stimulus values of the to-be-discriminated categories are themselves sampled from a distribution (Dosher and Lu 1998; Gold *et al.* 2004; Lu and Dosher 2008; Qamar *et al.* 2013; Cabrera *et al.* 2015). The use of external noise in perceptual paradigms has a long history (Nagaraja 1964; Burgess *et al.* 1981; Legge *et al.* 1987) with its main advantage being that it makes it possible for the internal noise in the visual system to be benchmarked against an externally measurable quantity, which also often leads to more robust results that are consistent across observers (Lu and Dosher 2014).

In fact, Zak *et al.* (2012) used external noise to study criterion attraction. They demonstrated that for some participants the decision criteria across two conditions collapsed, although for other participants the criteria only became attracted to each other without becoming the same. The Zak *et al.*'s study thus demonstrates that criterion attraction can be studied using an external noise paradigm.

However, the Zak *et al.* (2012) study and previous research by Gorea and Sagi focused exclusively on criterion attraction for the decision criterion, which separates the main choice between the two stimulus categories. The principle of criterion attraction (mostly in its extreme form of complete criterion collapse) has been applied to confidence criteria in several studies to explain findings of confidence–accuracy dissociations (Rahnev *et al.* 2011, 2012a,b; Solovey *et al.* 2015; Morales *et al.* 2015b; Li *et al.* 2018), but these studies did not use external noise paradigms.

The current study

Here I specifically examine attraction for confidence criteria in an external noise paradigm to explore whether this approach can lead to a more robust confidence–accuracy dissociation. Full-contrast Gabor patches were presented with orientations sampled from two overlapping distributions. Participants judged which distribution was more likely to have generated each stimulus and provided a confidence rating. Critically, I included two different experimental conditions: in the low variability condition, the two distributions had similar means and low standard deviations (SDs), whereas in the high variability condition, the distributions had dissimilar means and high SDs (Fig. 1A). The means and SDs in the two conditions were proportionate so that the tasks were equally difficult (Fig. 1B). Furthermore, the optimal decision criterion for both conditions was identical [simply judge whether orientation is clockwise (CW) or counterclockwise (CCW) from vertical orientation]. However, in the presence of criterion attraction, the confidence criteria would move outward in the low variability condition (resulting in lower confidence) and inward in the high variability condition (resulting in higher confidence; Fig. 1C).

To anticipate, I found that both stimulus sensitivity (d') and RT were matched across the low and high variability conditions. However, there was a robust confidence dissociation with the low variability condition resulting in lower confidence for 25 of the 26 participants compared to the high variability condition. Estimation of the criterion locations for each condition and additional computational modeling suggested that this effect is consistent with a mechanism of criterion attraction.

Methods

Participants

Twenty-eight participants took part in the experiment. Two participants were excluded because they had a negative correlation between their confidence and accuracy, indicating that they may not have given confidence ratings as instructed. Therefore, all analyses were based on the remaining 26 participants (15 females, age range = 18–22). All participants had normal or corrected-to-normal vision and provided informed consent. The experiment was approved by the Georgia Tech Institutional Review Board.

Procedure

Participants indicated whether a grating of full contrast was drawn from one of two partially overlapping distributions. Distribution 1 tended to generate gratings with CCW orientations, whereas Distribution 2 tended to generate gratings with CW orientations. The experiment included two conditions. In the low variability condition, Distributions 1 and 2 were normal distributions, $N(\mu, \sigma^2)$, with means, μ, of –2.4° and 2.4°, respectively, and a standard deviation (SD), σ, of 6° (Fig. 1). Note that here 0° indicates vertical orientation, negative numbers indicate CCW orientations, and positive numbers indicate CW orientations. In the high variability condition, Distributions 1 and 2 were simply scaled by a factor of 3 such that they were normal distributions

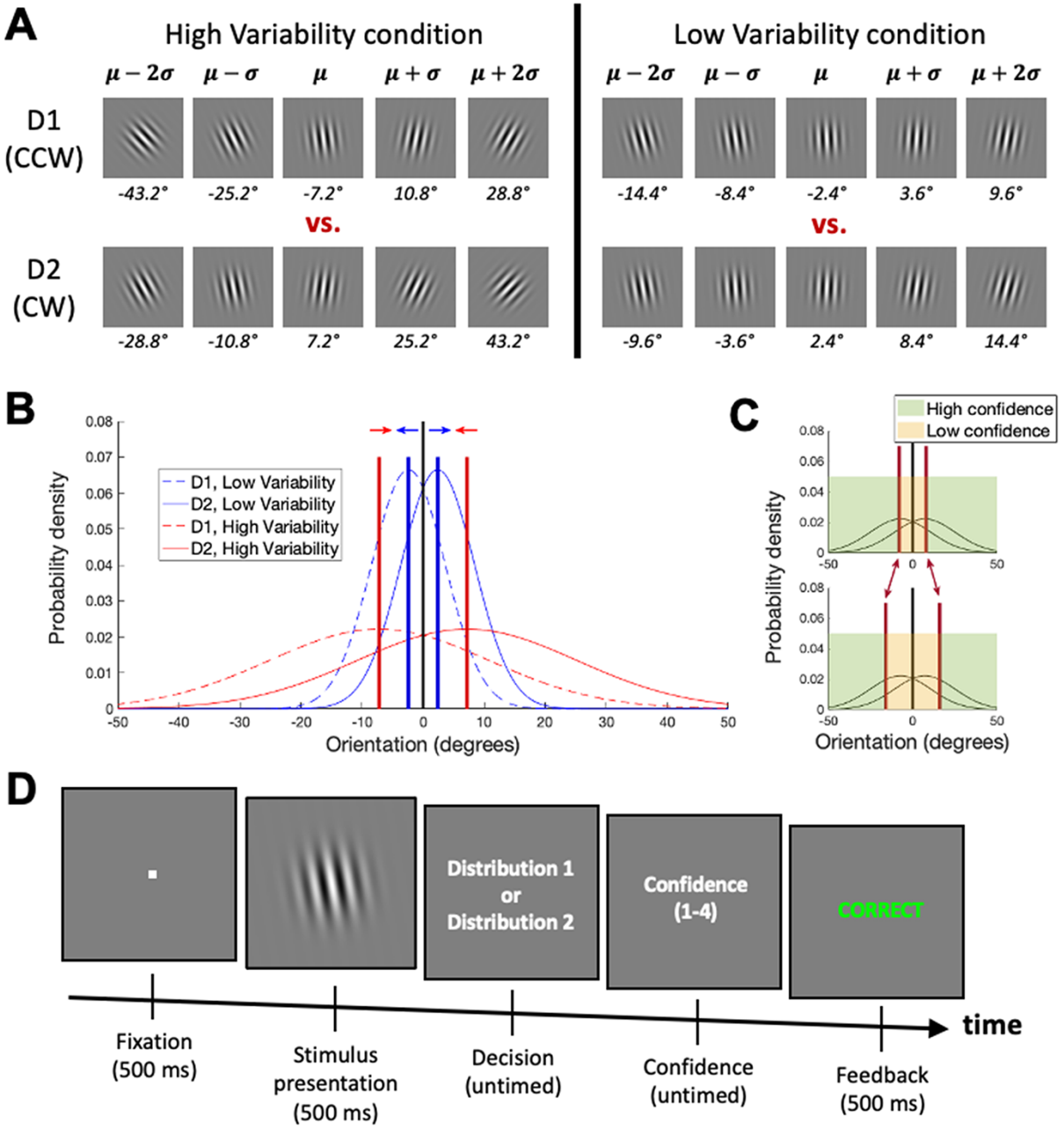

Figure 1. Experimental paradigm. (A) Participants judged the distribution (D1 vs. D2) that generated a stimulus with a given orientation. D1 was biased toward CCW orientations, while D2 was biased toward CW orientations. In the high variability condition, D1 and D2 had high means ($\mu = \pm 7.2°$) and SDs ($\sigma = 18°$), whereas in the low variability condition, D1 and D2 had low means ($\mu = \pm 2.4°$) and SDs ($\sigma = 6°$). The means and SDs in the high variability condition were thus exactly three times higher than in the low variability condition, resulting in identical maximum performance of $d' = 0.8$ in both cases. For each distribution, the figure displays the orientations corresponding to $\mu - 2\sigma$, $\mu - \sigma$, μ, $\mu + \sigma$, and $\mu + 2\sigma$. CCW (CW) orientations are indicated with negative (positive) numbers. (B) The two sets of distributions, together with the locations of confidence criteria that result in equivalent confidence ratings in the two conditions. If the criteria in the two conditions become 'attracted' to each other (see arrows), then the confidence criteria would become conservative for the low variability condition (resulting in low confidence) and liberal for the high variability condition (resulting in high confidence). (C) Depiction of a criterion shift. An outward move of the criteria (transition from top to bottom graph, corresponding to the expected shift for the low variability condition) results in a smaller area of high confidence. Conversely, an inward movement of the criteria (transition from bottom to top graph, corresponding to the expected shift for the high variability condition) results in a larger area of high confidence. (D) Task. On each trial, participants indicated the likely generating distribution and gave a confidence rating on a 4-point scale. Trial-by-trial feedback was provided throughout the whole experiment. The low and high variability distributions were presented in clearly marked, alternating blocks of 50 trials each

with means of –7.2° and 7.2°, respectively, and SD of 18°. Therefore, the overlap between the two distributions (measured as the difference between their means divided by their SD) was identical for the two conditions and resulted in a maximum sensitivity of $d' = \frac{2.4-(-2.4)}{6} = \frac{7.2-(-7.2)}{18} = 0.8$. Participants were explicitly given all of this information. Additionally, during the initial training,

they were presented with a series of 25 randomly generated grating orientations from each of these four distributions in order to aid them in building their understanding of the task.

Each trial started with a fixation period (500 ms), followed by stimulus presentation (500 ms), untimed decision period, untimed confidence period, and a feedback screen presented for 500 ms (Fig. 1D). The grating (100% contrast, 5° diameter) was presented at fixation, and on each trial its orientation was randomly generated from Distribution 1 or 2. Confidence was provided on a 4-point scale where 1 indicates low confidence and 4 indicates high confidence. Participants selected the generating distribution using the left and right arrows on a computer keyboard with their right hand and gave a confidence rating using the 1–4 keys using their left hand. Trial-by-trial feedback was provided throughout the whole experiment.

The experiment was organized in four runs each consisting of four 50-trial blocks for a total of 800 trials. Each block consisted of only low or high variability trials and the condition was clearly indicated before the beginning of the block. The low and high variability blocks alternated with the identity of the first block randomly chosen for each participant. Successive blocks were separated by 15-s breaks, while successive runs were separated by self-paced breaks. Before the beginning of the experiment, participants were given four blocks of training with a total of 100 trials.

It should be emphasized that participants were informed about all aspects of the experimental design, as well as about the optimal strategy of keeping two separate sets of criteria for the low and high variability conditions. An alternative design option would have been not to inform participants about the existence of low and high variability conditions at all. This would have likely resulted in a complete collapse of the confidence criteria for the two conditions (an extreme form of criterion attraction) but such an effect would not have reflected a limitation of human decision-making but rather a rational response strategy. The current design where participants are informed about all aspects of the experiment gives participants all information needed for optimal responses and can thus reveal the inherent limitations of human decision-making.

Participants completed the experiment on a 21.5-inch iMac monitor in a dark room. The distance between the monitor and the participants was 60 cm. The stimuli were created in MATLAB using Psychtoolbox 3 (Brainard 1997).

Analyses

For the main analyses, I computed the mean confidence and RT for each participant. Furthermore, to compare performance across participants, I computed the signal detection theory (SDT) measure d'. To do so, I calculated the hit rate (HR) and false alarm rate (FAR) by treating Distribution 2 as the target. The value of d' was computed separately for each condition using the formula:

$$d' = \Phi^{-1}(HR) - \Phi^{-1}(FAR)$$

where Φ^{-1} is the inverse of the cumulative standard normal distribution that transforms HR and FAR into z-scores.

To assess the presence of criterion attraction, I computed the locations of the decision thresholds, t_i, expressed in stimulus space in each condition using SDT by applying the formula:

$$t_i = -\frac{1}{2}\left(\Phi^{-1}(HR_i) + \Phi^{-1}(FAR_i)\right) \times \sigma$$

where t_i, HR_i, and FAR_i are the criterion location, hit rate, and false alarm rate for the i^{th} criterion ($i = -3, -2, \ldots, 3$), and σ is the standard deviation of the generating distributions. Note that t_0 is the criterion that separates decisions for Distribution 1 versus Distribution 2, while the confidence criteria $\pm t_k$ for $k = \{1, 2, 3\}$ separate the confidence ratings of k and $k+1$ (Macmillan and Creelman 2005). Negative criterion locations (i.e. $i < 0$) separate successive confidence ratings for 'Distribution 1' decisions, while positive criterion locations (i.e. $i > 0$) separate successive confidence ratings for 'Distribution 2' decisions. HR_i was estimated such that for positive values of i, HR_i is the proportion of trials where Distribution 2 was used to generate the stimulus and the participant chose Distribution 2 with a confidence rating higher than i. For negative values of i, HR_i is the proportion of trials where Distribution 2 was used to generate the stimulus and the participant chose either Distribution 2 regardless of confidence or Distribution 1 with confidence rating lower than or equal to $-i$. FAR_i was computed equivalently but for trials where Distribution 1 was used to generate the stimulus.

To assess the presence of criterion attraction, I computed the ratio $r_i = \frac{t_{i,HighVar}}{t_{i,LowVar}}$ for each pair of confidence criteria in the high and low variability conditions for each confidence criterion $i = -3, -2, -1, 1, 2, 3$. Since the stimulus distributions for the low and high variability conditions were identical except for scaling by a factor of 3, in the absence of internal noise, optimal confidence placement requires that the confidence criteria are also offset such that $r_i = 3$ for all i. On the other hand, criterion attraction would result in the confidence criteria being offset by a factor smaller than 3 (Fig. 1B). It should be noted that the optimal ratios r_i also depend on the internal noise associated with the perception of the stimuli, which I did not measure. However, previous research suggests that the internal noise for orientation detection of Gabor patches of full contrast presented for a long period (500 ms in the current experiment) is very small and typically lower than 1° (Sally and Gurnsey 2004; Beaudot and Mullen 2006; Bang *et al.* 2019). This is further corroborated by the fact that the observed d' levels were very close to the maximum possible value of 0.8 (see Results). Importantly, internal noise of 1° would have an almost negligible effect on the optimal ratios r_i. Indeed, such internal noise would increase the SD of the high and Low variability conditions to $\sqrt{18^2 + 1^2} = 18.03°$ and $\sqrt{6^2 + 1^2} = 6.08°$ and would therefore result in a negligible reduction of the optimal r_i from 3 to 2.964. Therefore, realistic levels of internal noise would have a very small effect on the optimal ratios r_i and are thus not considered further here.

Because the stimulus orientation values were drawn randomly from the generating distributions, there were slight variations in the maximum accuracy for the two conditions across participants. To remove such variability, for each participant, I computed the number of congruent trials (where the stimulus orientation had the same polarity as the mean of the generating distribution) for each of the two conditions. If the low (high) variability condition had m more congruent trials, then I excluded m congruent and trials from the low (high) variability condition and m incongruent trials from the high (low) variability condition. The process ensured that there were equal numbers of congruent and incongruent trials in the two conditions. The excluded trials were removed from the end of the experiment for each participant. On average, this procedure resulted in excluding 20 trials per participant for a 2.5% exclusion rate.

Statistical tests included standard frequentist tests such as t-tests and correlations. In addition, where appropriate, I also

performed Bayesian tests to quantify the evidence for either the null or alternative hypotheses. Conventionally, the null hypothesis is considered supported for Bayes factor values of $BF_{01} > 3$, whereas the alternative hypothesis is considered supported for $BF_{10} > 3$.

Model fitting

The main analyses above can show qualitatively the existence of criterion attraction but cannot be used to quantify its strength. Indeed, although the strength of the criterion attraction can be directly computed for each confidence criterion, there is no principled way in which these values can be combined to arrive at a single quantitative estimate per participant. For example, some participants have relatively few high-confidence responses, making the estimation of the extreme criteria (e.g. t_{-3} and t_3) relatively noisy. Therefore, a simple formula, such as averaging the estimated criterion attraction values for all confidence criteria, is likely to result in an imprecise estimate of the true criterion attraction.

Therefore, to precisely quantify the degree of criterion attraction, I specified a simple process model of criterion attraction. Fitting the model to the raw data allowed me to arrive at the best quantitative estimate of the strength of criterion attraction. According to the model, participants start with a set of criteria, t_i, used in the low variability condition with $i = \{-3, -2, \ldots, 3\}$. Optimally, if the criteria t_i are used in the low variability condition, then the same criteria should be scaled by a factor of 3 and used in the high variability condition. However, in the presence of criterion attraction, the criteria in the low variability condition increase by a factor of α and become equal to $\alpha \times t_i$, while the criteria in the high variability condition decrease by a factor of α and become equal to $\frac{3 \times t_i}{\alpha}$. Note that a value of α determines the strength of criterion attraction with $\alpha = 1$ corresponding to no criterion attraction and $\alpha = \sqrt{3} = 1.73$, corresponding to a situation where the two sets of criteria collapse onto a single set of identical values.

The model was fit to the data from each participant using the Bayesian Adaptive Direct Search toolbox, version 1.0.5 (Acerbi and Ma 2017). The model had seven free parameters: t_i for $i = \{-3, -2, -1, 1, 2, 3\}$ and α. For simplicity, t_0 was set to 0. The fitting was performed on the actual stimulus orientations encountered by each individual participant. To find the best fit, I computed the log-likelihood value associated with the full distribution of probabilities of each response type, as done previously (Yeon and Rahnev 2020; Shekhar and Rahnev 2021b):

$$Log\ likelihood = \sum_{i,j,k} \log(p_{ijk}) * n_{ijk}$$

where p_{ijk} and n_{ijk} are the response probability and the number of trials, respectively, associated with the Distribution $i \in \{1, 2\}$, confidence rating $j \in \{1, 2, 3, 4\}$, and condition k, where $k = 1$ corresponds to the low variability condition and $k = 2$ corresponds to the high variability condition. The best fit was determined as the set of parameters that maximized the log-likelihood value. Finally, I examined the resulting α values for each participant as the estimate of the strength of criterion attraction and performed t-tests to compare the resulting values to 1.

Data, materials, and code

All data and codes for the analyses have been made freely available at https://osf.io/g32tv/. The repository also includes the codes used to collect the data and can be reused by anyone who may want to employ this method of generating confidence–accuracy dissociations. In addition, the data have been uploaded to the Confidence Database (Rahnev *et al.* 2020).

Results

I investigated whether a robust confidence–accuracy dissociation could be achieved by exploiting the principle of criterion attraction in an external noise paradigm. Participants judged which of the two distributions generated a given stimulus and provided a confidence rating on a 4-point scale (Fig. 1D). Critically, two conditions—low variability and high variability (that were simply scaled versions of each other; Fig. 1A and B)—were presented in blocks of 50 trials. If the confidence criteria for the two conditions become attracted to each other, this would produce a confidence–accuracy dissociation.

Low and high variability conditions exhibit robust confidence–accuracy dissociation

I first compared the sensitivity (as computed by the SDT measure d') and RT between the two conditions. I found that participants performed equally well in the low variability ($d' = 0.78$) and high variability ($d' = 0.79$) conditions [$t(25) = 0.56$, $P = 0.58$, Cohen's $d = 0.11$, $BF_{01} = 4.17$; Fig. 2A]. Furthermore, the performance in each condition was indistinguishable from the maximum possible value of $d' = 0.8$ [low variability condition: $t(25) = -0.83$, $P = 0.4$, Cohen's $d = -0.16$, $BF_{01} = 3.53$; high variability condition: $t(25) = -0.38$, $P = 0.7$, Cohen's $d = -0.08$, $BF_{01} = 4.51$], suggesting that the internal noise associated with the perception of the Gabor patches must have been minimal. Similar to d', RT was also equated between the low variability (mean RT = 859 ms) and high variability (mean RT = 852 ms) conditions [$t(25) = -0.52$, $P = 0.61$, Cohen's $d = -0.10$, $BF_{01} = 4.26$; Fig. 2B]. These results show that the two conditions were very well matched in difficulty as measured both by stimulus sensitivity and RT.

Despite the close match in performance, there was a large difference in confidence between the two conditions (Fig. 2C). Specifically, the low variability condition resulted in much lower average confidence (2.43) than the high variability condition (2.81) with the difference exhibiting a very large effect size [$t(25) = 9.7$, $P = 5.8 \times 10^{-10}$, Cohen's $d = 1.9$; $BF_{10} = 1.9 \times 10^7$]. Furthermore, the higher confidence in the high variability condition was present in 25 of the 26 participants (96%), demonstrating that the difference in confidence was extremely consistent across participants.

To further establish that the difference in confidence is independent of any effects on performance, I examined whether participants who tended to have larger confidence effects also exhibited a corresponding difference in d' or RT. I found that this was not the case. Specifically, participants who exhibited large confidence effects (i.e. large difference between the confidence in the high and low variability conditions, $conf_{HighVar} - conf_{LowVar}$) were not more likely to have a large difference in d' (correlation between $conf_{HighVar} - conf_{LowVar}$ and $d'_{HighVar} - d'_{LowVar}$: $r = -0.11$, $P = 0.59$; $BF_{01} = 5.73$) or in RT (correlation between $conf_{HighVar} - conf_{LowVar}$ and $RT_{HighVar} - RT_{LowVar}$: $r = -0.18$, $P = 0.37$; $BF_{01} = 4.43$; Fig. 3). Therefore, the difference in confidence between the low and high variability conditions was not driven by the objective performance of the participants.

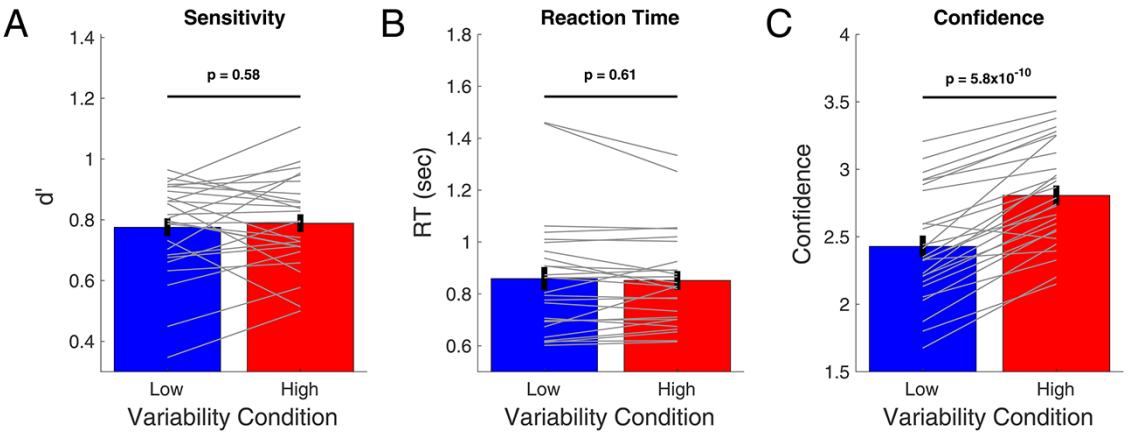

Figure 2. Confidence–accuracy dissociation between the low and high variability conditions. The low and high variability conditions were matched in terms of both d' (A) and RT (B), as also confirmed by a Bayes factors analysis. However, confidence was substantially higher in the high variability condition with the effect appearing in 25 of the 26 participants (C). Gray lines show individual participant data, and error bars show SEM

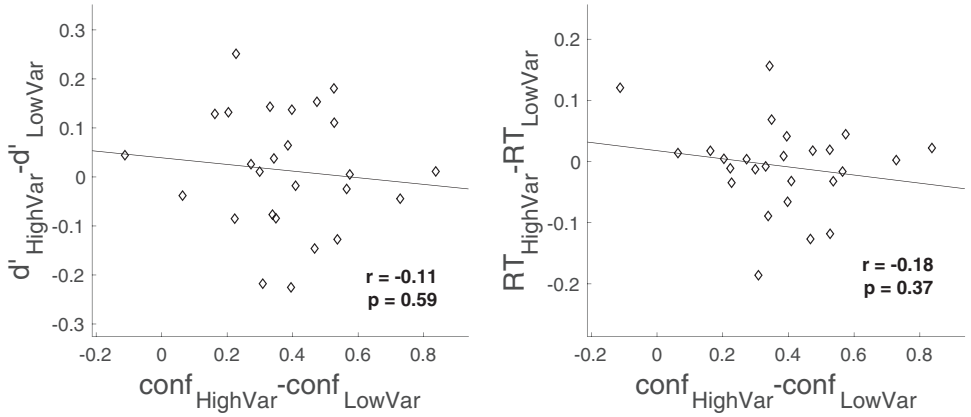

Figure 3. Confidence effects are not linked to d' or RT effects. Individual differences in the confidence effect (difference in confidence in the high and low variability conditions) were not related to individual differences in the d' or RT effects (difference in d' or RT in the high and low variability conditions). Diamonds represent individual participants; the black line depicts the line of best fit

Time course of the effects

It is important to emphasize that the confidence–accuracy dissociation results above were obtained despite the fact that the low and high variability conditions were presented in separate blocks of 50 trials. Given this blocked design, one may expect that the criterion attraction would be larger at the beginning of each block when the influence of the previous block is likely to be the strongest. Indeed, I found that confidence increased from the first to the second half of the blocks for the low variability condition [first half = 2.40, second half = 2.46; $t(25) = -2.80$, $P = 0.01$, Cohen's $d = -0.55$, $BF_{10} = 4.80$; Fig. 4, left] but decreased for the high variability condition [first half = 2.84, second half = 2.77; $t(25) = 3.02$, $P = 0.006$, Cohen's $d = 0.59$, $BF_{10} = 7.60$]. Thus, the difference in confidence between the two conditions was significantly smaller in the second half of the blocks [first half difference = 0.45, second half difference = 0.31; $t(25) = 4.07$, $P = 0.0004$, Cohen's $d = 0.80$, $BF_{10} = 75.25$]. These results demonstrate that the results were largest shortly after a switch from the previous block and suggest that these effects should be even larger in interleaved designs.

Analyzing each block in even finer intervals of 10 trials each revealed that this effect was driven mostly by a very large effect at the beginning of each block but without any evidence that the effect disappears by the end of the block (Fig. 4, right). Indeed, the difference in confidence between the two conditions was 0.54, 0.43, 0.24, 0.30, and 0.38, respectively, for the five sets of 10 trials constituting one 50-trial block. The smallest difference in confidence thus occurs toward the middle rather than at the end of the block. These results show that the criterion attraction effect is particularly strong at the very beginning of a block, but remains relatively stable afterward and does not disappear by the end of the block. The effects of criterion attraction thus appear relatively long-lasting and must extend for well over 50 trials.

Confidence criterion locations

The results so far demonstrate that the low and high variability conditions give rise to a robust confidence–accuracy dissociation. To further examine the nature of this dissociation, I computed the location, t_i, of each decision criterion for both the low and high variability conditions in orientation space (see Methods).

As could be expected, the Criterion t_0 used to distinguish between Distributions 1 and 2 was very close to 0° (vertical) in both the low variability [$t_0 = 0.089°$, $t(25) = 0.67$, $P = 0.51$, Cohen's $d = 0.13$, $BF_{01} = 3.93$] and the high variability conditions [$t_0 = 0.083°$, $t(25) = 0.40$, $P = 0.69$, Cohen's $d = 0.08$, $BF_{01} = 4.49$], with no difference between the two conditions [$t(25) = -0.04$, $P = 0.97$, Cohen's $d = -0.01$, $BF_{01} = 4.82$]. These results suggest that, as expected, participants made the primary decision by using a largely unbiased criterion centered on vertical orientation.

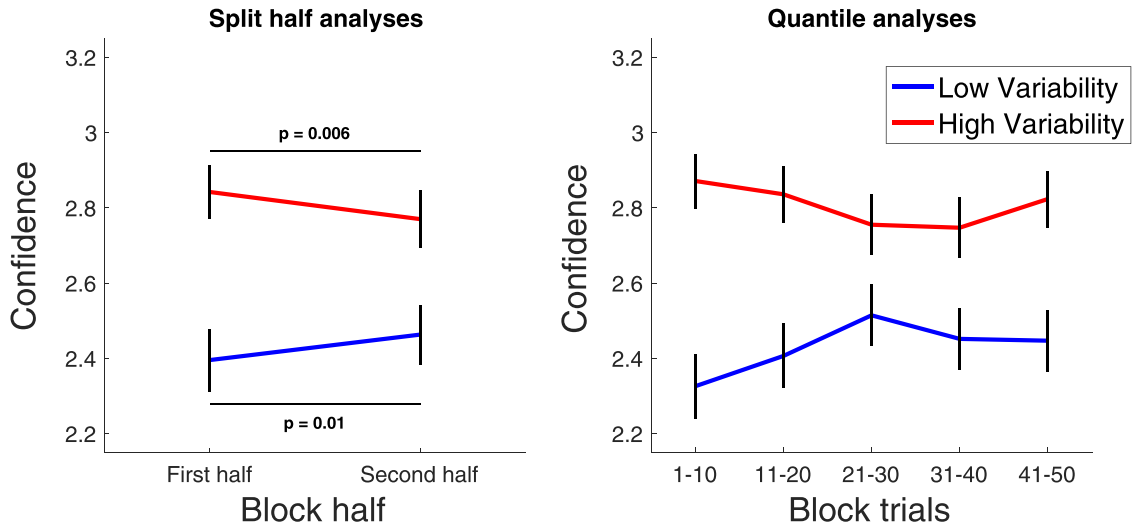

Figure 4. Time course of confidence effects. The difference in confidence between the low and high variability conditions was larger in the first than in the second half of the blocks. A finer split of each 50-trial block into five epochs shows that the difference in confidence is largest immediately at the start of a new block but remains relatively stable afterward, thus suggesting that the effects of criterion attraction can be relatively long-lasting

Critically, I examined the locations of the confidence criteria. In the absence of criterion attraction, the criterion locations in the high variability condition should be three times higher than in the low variability condition (Fig. 1B; note that this ratio could be expected to be slightly smaller due to internal noise; see Methods). However, I found the ratio, r, of their locations to be around 2 for all criteria ($r_{t_{-3}} = 2.03$, $r_{t_{-2}} = 1.95$, $r_{t_{-1}} = 2.24$, $r_{t_1} = 2.19$, $r_{t_2} = 2.05$, $r_{t_3} = 2.07$; Fig. 5). Since the ratios r are not normally distributed, I performed nonparametric signed rank tests that confirmed that all of these ratios were significantly lower than 3 (all P values < 0.0035). To avoid the issue of having to use non-parametric tests, I also compared the criteria in the high variability condition with the criteria in the low variability condition multiplied by a factor of 3. I found that the scaled criteria in the low variability condition were always more extreme (either more negative or more positive) than the criteria in the high variability condition [t_{-3}: $t(25) = 7.31$, $P = 1.2 \times 10^{-7}$, Cohen's $d = 1.43$, $BF_{10} = 1.2^{*}10^{5}$; t_{-2}: $t(25) = 6.08$, $P = 2.4 \times 10^{-6}$, Cohen's $d = 1.19$, $BF_{10} = 7.8 \times 10^{3}$; t_{-1}: $t(25) = 4.06$, $P = 0.0004$, Cohen's $d = 0.80$, $BF_{10} = 70.6$; t_1: $t(25) = 4.41$, $P = 0.0002$, Cohen's $d = 0.87$, $BF_{10} = 166.7$; t_2: $t(25) = 6.74$, $P = 4.6 \times 10^{-7}$, Cohen's $d = 1.32$, $BF_{10} = 3.6 \times 10^{4}$; t_3: $t(25) = 6.86$, $P = 3.5 \times 10^{-7}$, Cohen's $d = 1.34$, $BF_{10} = 4.6 \times 10^{4}$], thus confirming the existence of substantial criterion attraction.

Computational modeling

Finally, I developed a simple computational model to more precisely quantify the strength of criterion attraction. The model assumed that the criteria in the low and high variability conditions are multiplicatively attracted to each other by a constant factor α. Fitting the model to the data revealed that the criteria moved by an average factor of $\alpha = 1.244$ (range = 1.025–1.473, SD = 0.126), which was significantly larger than the factor of $\alpha = 1$ corresponding to a lack of criterion attraction [$t(25) = 9.86$, $P = 4.2 \times 10^{-10}$, Cohen's $d = 1.93$, $BF_{10} = 2.5 \times 10^{7}$; Fig. 6A]. Note that if two criteria are originally separated by a factor of 3 (e.g. 10° and 30°), then after each of them is attracted to the other by a factor of 1.244 and they become separated by a factor of $\frac{3}{1.244^2} = 1.94$ (e.g. 12.4° and 24.1°). Importantly, the parameter α was estimated to be higher than 1 for all participants, suggesting that all 26 participants may have exhibited some level of criterion attraction.

As may be expected, the estimated strength of criterion attraction was strongly correlated with the confidence effect (across-subject correlation between α and $conf_{HighVar} - conf_{LowVar}$: $r = 0.75$, $P = 8.6 \times 10^{-6}$; Fig. 6B). Nevertheless, even though the criterion attraction was found for all participants and had a relatively large magnitude, the attraction fell far short of causing the same set of criteria to be used for both the low and high variability conditions. Indeed, a unique set of criteria across the two conditions would correspond to an α level of $\sqrt{3} = 1.732$ or a 73.2% change in the locations of the criteria. The observed average $\alpha = 1.244$ corresponds to a 24.4% change or exactly one-third of the strength of criterion attraction that would result in a unique set of criteria for the two conditions.

Discussion

Many studies in the last decade have demonstrated the existence of confidence–accuracy dissociations (Wilimzig *et al.* 2008; Zylberberg *et al.* 2012, 2014, 2016; Vlassova *et al.* 2014; de Gardelle and Mamassian 2015; Koizumi *et al.* 2015; Rahnev *et al.* 2015; Song *et al.* 2015; Samaha *et al.* 2016; Spence *et al.* 2016, 2018; Boldt *et al.* 2017; Desender *et al.* 2018). Such findings have been foundational in discovering the mechanisms of confidence computations and have informed debates regarding the optimality of confidence ratings (Aitchison *et al.* 2015; Navajas *et al.* 2017; Rahnev and Denison 2018). However, most previous confidence–accuracy dissociations were of relatively small magnitude, were relatively inconsistent across participants, and included an RT confound. Here I developed a new experimental design based on external noise and the principle of criterion attraction. The design produced a confidence–accuracy dissociation of large magnitude and effect size, which was consistent across participants and free of RT confounds. These results establish a new method of inducing robust confidence–accuracy dissociations and have important implications about the ongoing debate regarding the malleability of subjective criteria.

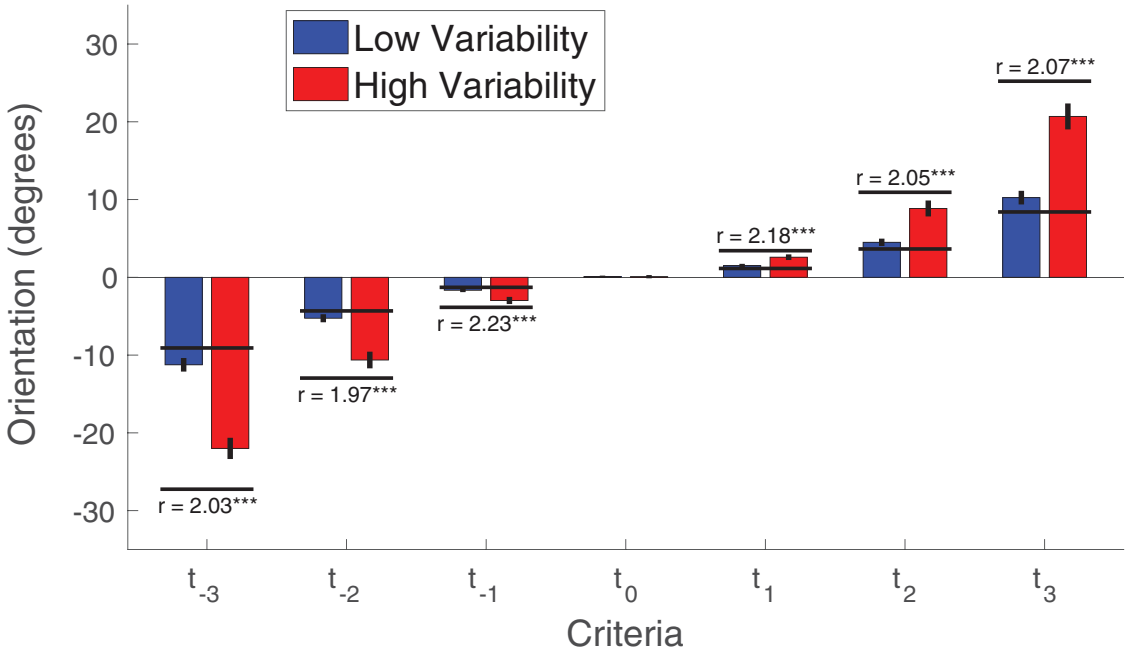

Figure 5. Criterion locations in the low and high variability conditions. The Criterion t_0 used to distinguish between Distributions 1 and 2 was not different between the two conditions. However, the ratio, *r*, between confidence criteria in the high and low variability conditions was significantly lower than 3 for all criteria. The black horizontal lines depict the expected criterion locations in the absence of criterion attraction. Error bars show SEM. *** $P < 0.001$

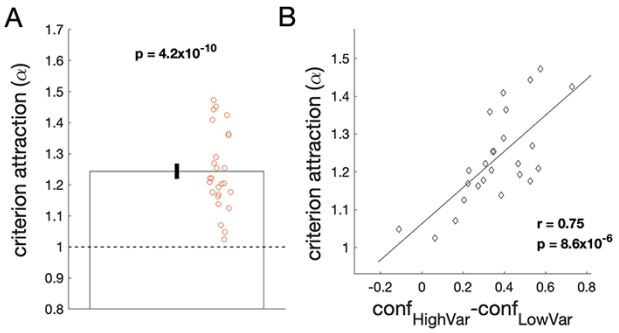

Figure 6. Quantifying the strength of criterion attraction. (A) Estimated criterion attraction α. The dashed horizontal line depicts $\alpha = 1$ that corresponds to a lack of criterion attraction. Circles depict individual participants, and the error bar depicts SEM. On average, the criteria shifted by 24.4%, which is one-third of the shift required for the criteria in the two conditions to become identical. (B) Scatterplot of the estimated criterion attraction (α) and the confidence effect ($conf_{HighVar} - conf_{LowVar}$) for each participant. The black line shows the line of best fit. The strong relationship between the two effects confirms that participants with high criterion attraction also had larger changes in confidence between the two conditions

Mechanisms of criterion attraction

Criterion attraction is usually interpreted as a limitation of criterion setting where humans are unable to maintain two completely separate sets of criteria (Gorea and Sagi 2005). The idea is that the two sets of criteria pull on each other, thus leading to criterion attraction. In theory, criterion attraction could occur either at a perceptual or a post-perceptual (i.e. cognitive) level. In the current study, criterion attraction occurs at a post-perceptual level. Indeed, the perceptual stimuli here were very clear (presented at full contrast for a relatively long time) and therefore perception itself was unlikely to be substantially affected by the two conditions of the experiment (low vs. high variability conditions). Therefore, the observed confidence–accuracy dissociation in the present study is due not to differences in subjective experience but to cognitive limitations in converting the subjective experience into an appropriate confidence rating. As such, the present effect cannot be used in studies that seek to produce differences in subjective experience in the absence of performance confounds (Morales *et al.* 2015a, 2019). What the current paradigm produces instead is differences in reported confidence in the absence of performance confounds, with the confidence difference stemming from post-perceptual factors. This effect may still be useful for understanding subjective experience by providing a control case where a robust difference in reported confidence is not based on a change in subjective experience.

It is also important to clarify that criterion attraction is a separate phenomenon from other established effects such as regression to the mean. Regression to the mean occurs when in a random sequence of events an extreme observation is followed, with high probability, by a less extreme one. As such, regression to the mean is a purely statistical effect with no human participant necessary. On the other hand, criterion attraction is not a statistical necessity and instead reflects limitations of human criterion setting (e.g. it would be easy to design an artificial agent with no criterion attraction).

Relationship to confidence–accuracy dissociation in previous research

As already reviewed in the Introduction, confidence–accuracy dissociations have been observed in many different studies using a variety of manipulations such as stimulus variability, attention, visual field location, and positive evidence bias. It is thus important to consider whether the current effects relate to any of these previous manipulations. For example, studies on

stimulus variability (Zylberberg *et al.* 2014, 2016; de Gardelle and Mamassian 2015; Spence *et al.* 2016; Boldt *et al.* 2017) or positive evidence bias (Zylberberg *et al.* 2012; Koizumi *et al.* 2015) used stimulus manipulations in ways that may seem similar to the external noise manipulation here. However, it is important to clarify that none of these previous studies had a true external noise manipulation—in each case, there was a one-to-one mapping between the generating stimulus category and each stimulus. However, no such one-to-one mapping exists for external noise paradigms like in the present study; instead, a given stimulus feature (e.g. orientation) could be generated from either stimulus category. As such, the external noise manipulation used here is qualitatively different from the manipulations in most previous research on confidence–accuracy dissociations.

The presence of robust confidence differences in the absence of corresponding differences in either accuracy or RT may appear at odds with proposals that confidence is determined by the accuracy and RT of decision (Kiani *et al.* 2014; Zylberberg *et al.* 2016). However, these results only falsify an extreme version of this proposal where confidence is determined exclusively by accuracy and RT. Indeed, the current results show that confidence is influenced by factors separate from accuracy and RT, but cannot be used to argue that accuracy or RT are not causally important for confidence ratings in other designs. Similarly, criterion attraction can be modeled using any framework (e.g. signal detection and accumulation to bound) and thus cannot be used to distinguish between different frameworks of perceptual decision-making.

Criterion attraction in previous research

Many studies by Gorea and Sagi suggested the presence of criterion attraction for various interleaved conditions (Gorea and Sagi 2000, 2001, 2002, 2005; Gorea *et al.* 2005; Zak *et al.* 2012). In some of these studies, the criteria even appeared to collapse onto the same unified criterion. Consequently, several more recent studies have simply assumed the presence of a unique criterion for different interleaved conditions without directly testing this assumption (Rahnev *et al.* 2011, 2012a,b; Solovey *et al.* 2015; Morales *et al.* 2015b; Li *et al.* 2018). These studies invariably used internal noise designs where criterion locations cannot be expressed in stimulus parameters and thus the presence of a unique criterion is perhaps impossible to test directly. However, this practice has recently been strongly criticized (Lee *et al.* 2021), especially in the light of a recent study using external noise where the criteria across different conditions were found not to be identical (Denison *et al.* 2018).

The notion of the confidence criteria attracting each other but without becoming identical may provide a unifying account for all studies to date. Findings that have been explained by a fully unified criterion across conditions could generally be explained just as well by criterion attraction (Morales *et al.* 2015b). On the other hand, the study by Denison *et al.* (2018) could also have featured criterion attraction that was simply too difficult to detect due to the study's unique experimental design. Indeed, in that study, the optimal decision strategy in some conditions was not to give high confidence ratings at all, thus eliminating certain confidence criteria and making it difficult to test for criterion attraction. The possibility of an undetected criterion attraction in Denison *et al.*'s study is also supported by the fact that their data were best explained not by an optimal Bayesian model but by a heuristic one, which is consistent with the existence of decision biases.

These considerations suggest that the debate surrounding criterion interactions across conditions should perhaps move away from arguments of whether the criteria in different conditions are fully independent or fully unified. There already appears to be a substantial amount of evidence that the truth is somewhere in between. Therefore, future efforts would be better directed toward identifying the strength of criterion interactions and the factors that modulate this strength. By precisely quantifying the degree of criterion attraction in an external noise paradigm, the current paper makes a small step in this direction.

Criterion attraction in the real world

All studies discussed here featured participants performing tasks on artificial stimuli in laboratory conditions. Therefore, an important question concerns whether similar criterion attraction exists in real-world settings. I think that it does. Indeed, virtually all perception occurs in conditions that vary from context to context and is therefore likely subject to criterion attraction. For example, when attempting to detect a mosquito, one should optimally adopt conservative criteria for plain backgrounds and liberal criteria for patterned backgrounds. While the relevant study has not been performed, I suspect that such situations are accompanied by strong criterion attraction where people have a much stronger bias to detect a mosquito against a plain background compared to a patterned one.

In addition, criterion attraction is likely to be relevant beyond perceptual tasks. Consider a teacher grading essays by fifth and eighth graders mixed in a single pile (but still clearly identified). Will the teacher be able to use age-appropriate grading criteria or will she exhibit criterion attraction and give relatively lower scores to the fifth graders and relatively higher scores to the eighth graders? Although I can only speculate, it seems likely that criterion attraction would occur in any situation where a person evaluates two or more groups of different abilities in mixed order, potentially with important societal consequences.

Limitations

One limitation of the current study is that its design only points to the existence of criterion attraction but does not make it possible to measure the movement of the criteria in each condition independently. It is thus possible that, at least for some participants, the criteria from one condition remained constant and only the criteria in the other condition were attracted to the criteria in the first. I note that such an effect would still be a form of 'criterion attraction' and therefore would not change any of the conclusions of the current study. This question could be resolved by future studies that present each condition in isolation (ideally on separate days) before presenting them together in the same experimental session. Nevertheless, the split-half analyses already provide some evidence that both sets of criteria experienced attraction. Indeed, given that confidence increased from the first to the second half of blocks for the low variability condition but decreased from the first to the second half of blocks for the high variability condition, it appears that both sets of conditions experienced stronger criterion attraction in the first (compared to the second) half of blocks. Another limitation is that, as noted in the Methods, the current study did not measure the level of internal noise. Finally, the current study also leaves open the question of whether criterion attraction would have been even stronger in the absence of trial-by-trial feedback, or if the two conditions were interleaved in all blocks.

Conclusion

I report on the strongest, to my knowledge, confidence–accuracy dissociation to date: even though both d' and RT were well matched across the two conditions (both $BF_{01} > 3$), confidence was robustly different. The effect was consistent across subjects and very strong in both magnitude (0.38-point difference on a scale where the extremes, 1 and 4, are 3 points apart) and effect size (Cohen's $d = 1.9$). Therefore, the current study provides a robust method for inducing large confidence–accuracy dissociations.

Acknowledgements

I thank Mia Huff for her help with data collection.

Funding

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

Conflict of interest statement

None declared.

Author contributions

D.R. conceptualized the study idea, analyzed the data, and wrote the manuscript.

References

Acerbi L, Ma WJ. Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. *Adv Neural Inf Process Syst* 2017:1836–46. https://proceedings.neurips.cc/paper/2017/file/df0aab058ce179e4f7ab135ed4e641a9-Paper.pdf (14 October 2021, date last accessed).

Aguilar-Lleyda D, Konishi M, Sackur J *et al.* Confidence can be automatically integrated across two visual decisions. *J Exp Psychol* 2021;**47**:161–71.

Aitchison L, Bang D, Bahrami B *et al.* Doubly bayesian analysis of confidence in perceptual decision-making. *PLoS Comput Biol* 2015;**11**:e1004519.

Allen M, Frank D, Schwarzkopf DS *et al.* Unexpected arousal modulates the influence of sensory noise on confidence. *ELife* 2016;**5**:e18103.

Bang JW, Shekhar M, Rahnev D. Sensory noise increases metacognitive efficiency. *J Exp Psychol* 2019;**148**:437–52.

Beaudot WHA, Mullen KT. Orientation discrimination in human vision: psychophysics and modeling. *Vision Res* 2006;**46**:26–46.

Boldt A, de Gardelle V, Yeung N. The impact of evidence reliability on sensitivity and bias in decision confidence. *J Exp Psychol* 2017;**43**:1520–31.

Boldt A, Schiffer A-M, Waszak F *et al.* Confidence predictions affect performance confidence and neural preparation in perceptual decision making. *Sci Rep* 2019;**9**:4031.

Brainard DH. The psychophysics toolbox. *Spat Vis* 1997;**10**:433–6.

Burgess AE, Wagner RF, Jennings RJ *et al.* Efficiency of human visual signal discrimination. *Science* 1981;**214**:93–4.

Cabrera CA, Lu Z-L, Dosher BA. Separating decision and encoding noise in signal detection tasks. *Psychol Rev* 2015;**122**:429–60.

de Gardelle V, Mamassian P. Weighting mean and variability during confidence judgments. *PLoS One* 2015;**10**:e0120870.

Denison RN, Adler WT, Carrasco M *et al.* Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proc Natl Acad Sci* 2018;**115**:11090–5.

Desender K, Boldt A, Yeung N. Subjective confidence predicts information seeking in decision making. *Psychol Sci* 2018;**29**:761–78.

Dosher BA, Lu Z-L. Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proc Natl Acad Sci USA* 1998;**95**:13988–93.

Fetsch CR, Kiani R, Newsome WT *et al.* Effects of cortical microstimulation on confidence in a perceptual decision. *Neuron* 2014a;**83**:797–804.

Fetsch CR, Kiani R, Shadlen MN. Predicting the accuracy of a decision: a neural mechanism of confidence. *Cold Spring Harb Symp Quant Biol* 2014b;**79**:185–97.

Fleming SM, Maniscalco B, Ko Y *et al.* Action-specific disruption of perceptual confidence. *Psychol Sci* 2015;**26**:89–98.

Gajdos T, Fleming S, Garcia MS *et al.* Revealing subthreshold motor contributions to perceptual confidence. *Neurosci Conscious* 2019;**5**:niz001.

Gold JM, Sekuler AB, Bennett PJ. Characterizing perceptual learning with external noise. *Cogn Sci* 2004;**28**:167–207.

Gorea A, Caetta F, Sagi D. Criteria interactions across visual attributes. *Vision Res* 2005;**45**:2523–32.

Gorea A, Sagi D. Failure to handle more than one internal representation in visual detection tasks. *Proc Natl Acad Sci USA* 2000;**97**:12380–4.

Gorea A, Sagi D. Disentangling signal from noise in visual contrast discrimination. *Nat Neurosci* 2001;**4**:1146–50.

Gorea A, Sagi D. Natural extinction: a criterion shift phenomenon. *Vis Cogn* 2002;**9**:913–36.

Gorea A, Sagi D. Decision and attention. In: Itti L, Rees G, Tsotsos J. K. (eds.) *Neurobiology of Attention*. London: Academic Press, 2005, 152–9.

Hanks TD, Mazurek ME, Kiani R *et al.* Elapsed decision time affects the weighting of prior probability in a perceptual decision task. *J Neurosci* 2011;**31**:6339–52.

Kiani R, Cueva CJ, Reppas JB *et al.* Dynamics of neural population responses in prefrontal cortex indicate changes of mind on single trials. *Curr Biol* 2014;**24**:1542–7.

Koizumi A, Maniscalco B, Lau H. Does perceptual confidence facilitate cognitive control? *Atten Percept Psychophys* 2015;**77**:1295–306.

Kontsevich LL, Chen -C-C, Verghese P *et al.* The unique criterion constraint: a false alarm? *Nat Neurosci* 2002;**5**:707.

Kurtz P, Shapcott KA, Kaiser J *et al.* The influence of endogenous and exogenous spatial attention on decision confidence. *Sci Rep* 2017;**7**:6431.

Lee JL, Denison R, Ma WJ. *Assimilating Subjective Inflation to Post-perceptual Decision-making*. 2021. 10.31234/osf.io/ys8mb.

Legge GE, Kersten D, Burgess AE. Contrast discrimination in noise. *J Opt Soc Am A* 1987;**4**:391.

Li MK, Lau H, Odegaard B. An investigation of detection biases in the unattended periphery during simulated driving. *Atten Percept Psychophys* 2018;**80**:1325–32.

Lu Z-L, Dosher BA. Characterizing observers using external noise and observer models: assessing internal representations with external noise. *Psychol Rev* 2008;**115**:44–82.

Lu Z-L, Dosher BA. *Visual Psychophysics*. Cambridge, MA: MIT Press, 2014.

Macmillan NA, Creelman CD. *Detection Theory: A User's Guide*. 2nd edn. Mahwah, NJ: Erlbaum, 2005.

Mamassian P. Visual confidence. *Annu Rev Vision Sci* 2016;**2**: annurev-vision-111815-114630.

Maniscalco B, Peters MAK, Lau H. Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Atten Percept Psychophys* 2016;**78**:923–37.

Morales J, Chiang J, Lau H. Controlling for performance capacity confounds in neuroimaging studies of conscious awareness. *Neurosci Conscious* 2015a;**2015**:niv008.

Morales J, Odegaard B, Maniscalco B. The neural substrates of conscious perception without performance confounds. *PsyArXiv* 2019:1–29.

Morales J, Solovey G, Maniscalco B *et al.* Low attention impairs optimal incorporation of prior knowledge in perceptual decisions. *Atten Percept Psychophys* 2015b;**77**:2021–36.

Nagaraja NS. Effect of luminance noise on contrast thresholds. *J Opt Soc Am* 1964;**54**:950.

Navajas J, Hindocha C, Foda H *et al.* The idiosyncratic nature of confidence. *Nat Hum Behav* 2017;**1**:810–8.

Odegaard B, Grimaldi P, Cho SH *et al.* Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. *Proc Natl Acad Sci USA* 2018;**115**: E1588–E1597.

Peters MAK, Thesen T, Ko YD *et al.* Perceptual confidence neglects decision-incongruent evidence in the brain. *Nat Hum Behav* 2017;**1**:0139.

Qamar AT, Cotton RJ, George RG *et al.* Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. *Proc Natl Acad Sci USA* 2013;**110**:20332–7.

Rahnev D, Bahdo L, de Lange FP *et al.* Prestimulus hemodynamic activity in dorsal attention network is negatively associated with decision confidence in visual perception. *J Neurophysiol* 2012a;**108**:1529–36.

Rahnev D, Balsdon T, Charles L *et al.* Consensus goals for the field of visual metacognition. *PsyArXiv* 2021.

Rahnev D, Denison RN. Suboptimality in perceptual decision making. *Behav Brain Sci* 2018;**41**:1–66.

Rahnev D, Desender K, Lee ALF *et al.* The confidence database. *Nat Hum Behav* 2020;**4**:317–25.

Rahnev D, Koizumi A, McCurdy LY *et al.* Confidence leak in perceptual decision making. *Psychol Sci* 2015;**26**:1664–80.

Rahnev D, Maniscalco B, Graves T *et al.* Attention induces conservative subjective biases in visual perception. *Nat Neurosci* 2011;**14**:1513–5.

Rahnev D, Maniscalco B, Luber B *et al.* Direct injection of noise to the visual cortex decreases accuracy but increases decision confidence. *J Neurophysiol* 2012b;**107**:1556–63.

Rahnev D, Nee DE, Riddle J *et al.* Causal evidence for frontal cortex organization for perceptual decision making. *Proc Natl Acad Sci* 2016;**113**:6059–64.

Rausch M, Hellmann S, Zehetleitner M. Confidence in masked orientation judgments is informed by both evidence and visibility. *Atten Percept Psychophys* 2018;**80**:134–54.

Recht S, Mamassian P, de Gardelle V. Temporal attention causes systematic biases in visual confidence. *Sci Rep* 2019;**9**:11622.

Rounis E, Maniscalco B, Rothwell JC *et al.* Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cogn Neurosci* 2010;**1**:165–75.

Sally SL, Gurnsey R. Orientation discrimination across the visual field: matching perceived contrast near threshold. *Vision Res* 2004;**44**:2719–27.

Samaha J, Barrett JJ, Sheldon AD *et al.* Dissociating perceptual confidence from discrimination accuracy reveals no influence of metacognitive awareness on working memory. *Front Psychol* 2016;**7**:851.

Samaha J, Iemi L, Postle BR. Prestimulus alpha-band power biases visual discrimination confidence, but not accuracy. *Conscious Cogn* 2017;**54**:47–55.

Shekhar M, Rahnev D. Distinguishing the roles of dorsolateral and anterior PFC in visual metacognition. *J Neurosci* 2018;**38**:5078–87.

Shekhar M, Rahnev D. Sources of metacognitive inefficiency. *Trends Cogn Sci* 2021a;**25**:12–23.

Shekhar M, Rahnev D. The nature of metacognitive inefficiency in perceptual decision making. *Psychol Rev* 2021b;**128**:45–70.

Solovey G, Graney GG, Lau H. A decisional account of subjective inflation of visual perception at the periphery. *Atten Percept Psychophys* 2015;**77**:258–71.

Song A, Koizumi A, Lau H. A behavioral method to manipulate metacognitive awareness independent of stimulus awareness. In: Overgaard M (ed.), *Behavioral Methods in Consciousness Research*. Oxford: Oxford University Press, 2015,77–85.

Spence ML, Dux PE, Arnold DH. Computations underlying confidence in visual perception. *J Exp Psychol* 2016;**42**:671–82.

Spence ML, Mattingley JB, Dux PE. Uncertainty information that is irrelevant for report impacts confidence judgments. *J Exp Psychol* 2018;**44**:1981–94.

Vlassova A, Donkin C, Pearson J. Unconscious information changes decision accuracy but not confidence. *Proc Natl Acad Sci* 2014;**111**:16214–8.

Wilimzig C, Tsuchiya N, Fahle M *et al.* Spatial attention increases performance but not subjective confidence in a discrimination task. *J Vis* 2008;**8**:1–10.

Yeon J, Rahnev D. The suboptimality of perceptual decision making with multiple alternatives. *Nat Commun* 2020;**11**:3857.

Zak I, Katkov M, Gorea A *et al.* Decision criteria in dual discrimination tasks estimated using external-noise methods. *Atten Percept Psychophys* 2012;**74**:1042–55.

Zylberberg A, Barttfeld P, Sigman M. The construction of confidence in a perceptual decision. *Front Integr Neurosci* 2012;**6**:79.

Zylberberg A, Fetsch CR, Shadlen MN. The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. *ELife* 2016;**5**:e17688.

Zylberberg A, Roelfsema PR, Sigman M. Variance misperception explains illusions of confidence in simple perceptual decisions. *Conscious Cogn* 2014;**27**:246–53.