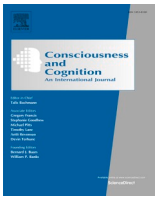
Consciousness and Cognition 136 (2025) 103942

Contents lists available at ScienceDirect

Consciousness and Cognition

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



Full Length Article

Similarities and differences in the effects of different stimulus manipulations on accuracy and confidence

Herrick Fung [a,*], Medha Shekhar [a,b], Kai Xue [a], Manuel Rausch [c,d,e], Dobromir Rahnev [a]

[a] School of Psychology, Georgia Institute of Technology, United States
[b] Center for Research in Cognition & Neurosciences, ULB Neuroscience Institute, Brussels, Belgium
[c] Rhine-Waal University of Applied Sciences, Cleves, Germany
[d] Catholic University of Eichstätt-Ingolstadt, Eichstätt, Germany
[e] University of Klagenfurt, Klagenfurt, Austria

ARTICLE INFO

Keywords:
Metacognition
Confidence
Perceptual decision making
Folded-X pattern

ABSTRACT

Visual stimuli can vary in multiple dimensions that affect accuracy and confidence in a perceptual decision-making task. However, previous studies have typically included just one or at most two manipulations, leaving it unclear whether each manipulation has a unique effect on accuracy vs. confidence. Subjects indicated whether a tilted Gabor patch was oriented clockwise or counter-clockwise from 45°. We included manipulations of the task-defining feature (tilt offset) and four auxiliary, non-task-defining features (size, duration, spatial frequency, and noise level). We found that the four auxiliary manipulations had fairly similar effects on accuracy and confidence. In contrast, the task-defining tilt offset manipulation stood out by affecting accuracy more strongly than confidence. In addition, tilt offset exhibited a supraadditive interaction with all other manipulations for both accuracy and confidence, whereas all auxiliary manipulations exhibited either no interactions or subadditive interactions with each other. Furthermore, tilt offset was the only manipulation for which confidence in incorrect trials decreased with increasing difficulty, while all auxiliary manipulations exhibited the opposite trend. Overall, our results reveal a noticeable similarity among the effects of all four auxiliary (non-task-defining) manipulations on accuracy and confidence, as well as a prominent difference between them and the task-defining manipulation (tilt offset). These results enable a priori predictions of how novel manipulations would affect accuracy and confidence.

1. Introduction

Confidence is the subjective sense of how likely our decisions are to be correct (Pouget et al., 2016; Sanders et al., 2016). Experiments designed to understand the mechanisms underlying confidence judgments typically include stimulus manipulations, such as lowering contrast (Shekhar & Rahnev, 2021b), infusing noise (Bang et al., 2019), and masking the stimulus (Lau & Passingham, 2006; Rausch et al., 2018). Such manipulations produce varying levels of perceptual performance, enabling us to investigate how confidence fluctuates with performance. Intuitively, confidence should align with performance. For instance, when recognizing a blurry face, both

* Corresponding author at: School of Psychology, Georgia Institute of Technology, 654 Cherry Str NW, Atlanta, GA 30332, United States.
E-mail address: herrickfung@gmail.com (H. Fung).

https://doi.org/10.1016/j.concog.2025.103942
Received 2 April 2025; Received in revised form 3 September 2025; Accepted 5 October 2025
Available online 11 October 2025

confidence and performance should decrease as the level of blurriness increases. Many papers have explored how confidence tracks performance (Busey et al., 2000; Yeung & Summerfield, 2012) and numerous factors that affect visual confidence have been identified (for review, see Rahnev & Denison, 2018; Shekhar & Rahnev, 2021a). Many computational models of confidence have been developed to explain different behavioral effects related to confidence (Boundy-Singer et al., 2023; Fleming & Daw, 2017; Hellmann et al., 2023, 2024; Maniscalco & Lau, 2016; Shekhar & Rahnev, 2021b, 2024; Xue et al., 2024a).

However, despite the variety of ways in which visual stimuli can be manipulated, most existing studies on visual confidence manipulated a single stimulus feature, which overlooks the potential differences in confidence and performance between various manipulations. While it is true that many manipulations can effectively vary perceptual performance, it is unclear whether all manipulations have unique effects on accuracy vs. confidence.

Among multiple different ways of manipulating a visual stimulus, one way of classifying them into meaningful subgroups is to distinguish between "task-defining" and "auxiliary" (non-task-defining) manipulations. Specifically, a task-defining manipulation affects the stimulus feature that directly determines the decision in a perceptual judgment task (e.g., adjusting the tilt values in an orientation discrimination task). In contrast, auxiliary manipulations affect the stimulus quality without affecting the visual feature on which the decision is based (e.g., blurring the stimulus, infusing noise, reducing stimulus duration in an orientation or face discrimination task).

This dichotomy has often been studied using an ensemble average task, in which manipulation of the sampling mean of the ensemble constitutes a task-defining manipulation, whereas the variance constitutes an auxiliary manipulation. Studies have shown that manipulating variance, as compared to manipulating the mean, has a stronger effect on confidence than accuracy (Boldt et al., 2017; Spence et al., 2016). However, this is not always the case (Zylberberg et al., 2014) and the effect may be subject to individual differences (de Gardelle and Mamassian, 2015). Building on this distinction, more recent work has extended the task from ensemble averaging to orientation and motion direction discrimination tasks. Increasing the contrast of a Gabor stimulus and increasing motion coherence (both auxiliary manipulations) increased confidence more than accuracy, whereas manipulating the tilt offset of a Gabor and motion direction (both task-defining manipulations) boosted accuracy but not confidence (Xue et al., 2025). Overall, these findings suggest that task-defining and auxiliary manipulations can differentially shape the relationship between confidence and accuracy. However, the scope of prior work has been limited by the small number of stimulus manipulations typically examined, leaving open the question of how confidence changes when multiple stimulus manipulations are presented within a single experiment and whether different auxiliary manipulations act in similar ways or not.

Having multiple stimulus manipulations within a single experiment would also enable investigation into potential interactions between task-defining and auxiliary manipulations—an underexplored area due to the limited number of stimulus manipulations in most existing studies in the field. To date, only one study has investigated such interactions for confidence, reporting no significant interaction effect between manipulating the mean and manipulating the variance in a color ensemble judgement task (Boldt et al., 2017). Further, no study to date has examined interactions among different auxiliary manipulations. Therefore, it remains largely unknown how different stimulus manipulations interact with each other and whether these interactions are the same or different for confidence and accuracy.

Further, emerging evidence suggests that different stimulus manipulations may produce different patterns of how confidence for correct and error trials changes with difficulty. Generally, it is thought confidence ratings produce a "folded-X" pattern where easier trials tend to increase confidence for correct trials but decrease it for error trials (Hangya et al., 2016; Sanders et al., 2016). However, some studies suggest that certain manipulations violate the canonical folded-X pattern and instead produce a pattern where easier trials increase confidence for both correct and error trials (Hellmann et al., 2023; Kiani et al., 2014; Rausch et al., 2018, 2021; Rausch & Zehetleitner, 2019; Shekhar & Rahnev, 2024; Xue et al., 2025). However, no study to date has investigated the effects of multiple task-defining and auxiliary manipulations on the folded-X pattern.

Here we conduct two experiments to explore the similarities and differences in the effects of various manipulations on confidence and accuracy. Both experiments featured four distinct manipulations within a single experiment in an orientation discrimination task. In both experiments, we manipulated one task-defining feature (tilt offset) and three other auxiliary (non-task-defining) manipulations—size, noise, and either duration (Experiment 1) or spatial frequency (Experiment 2). Because the duration manipulation had very little effect on both confidence and accuracy in Experiment 1, we replaced the duration manipulation with a spatial frequency manipulation in Experiment 2. Across both experiments, we found that all auxiliary manipulations produced very similar effects on confidence and accuracy. Equally notably, the task-defining tilt offset manipulation differed from all auxiliary manipulations in having a larger effect on accuracy than confidence. In addition, tilt offset was the only manipulation that exhibited a supraadditive interaction with any other manipulation and also the only manipulation that yielded a canonical folded-X pattern. Altogether, these findings reveal a very consistent profile of the confidence and accuracy effects of four visually distinctive auxiliary manipulations, as well as a fundamental difference between the auxiliary and task-defining manipulations, thus opening the door for predicting how novel manipulations would affect accuracy and confidence.

2. Methods

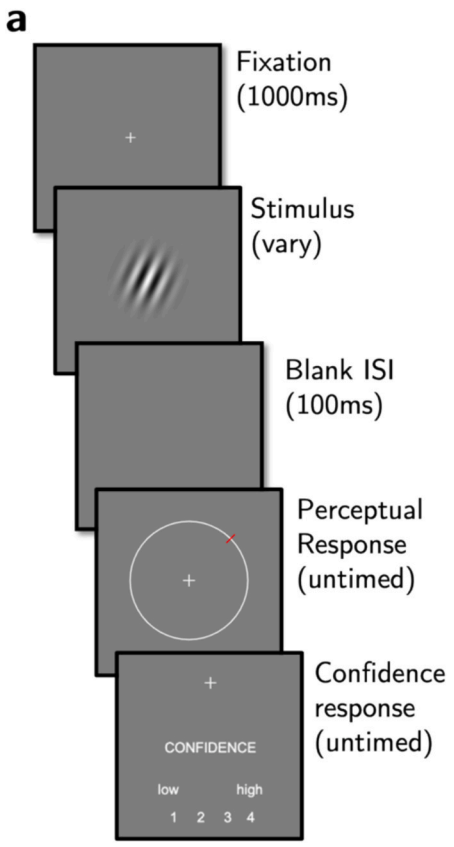
2.1. Subjects

Sixty-two subjects participated in the two experiments for monetary compensation ($10/hour). Subjects were recruited using Prolific (https://www.prolific.com/), an online crowdsourcing platform. For each experiment, we targeted for 30 subjects, a number that meets or exceeds the sample size used in previous studies on the confidence-accuracy relationship (Boldt et al., 2017; de Gardelle

and Mamassian, 2015; Moran et al., 2015; Rausch et al., 2018; Shekhar & Rahnev, 2021b; Xue et al., 2024b). Experiment 1 had 30 subjects (6 females) aged between 19–38 (M=26.6, SD=5.47). Experiment 2 had 32 subjects (9 females) aged between 18–37 (M=27.0, SD=5.39). None of the subjects participated in both experiments. Based on previous studies from our lab (Haddara & Rahnev, 2022; Rafiei et al., 2024), we excluded subjects with an overall accuracy less than 55 % or greater than 95 % or overall average confidence above 3.7, which led to the exclusion of two subjects from Experiment 1 and three subjects from Experiment 2. All subjects were naïve to the purpose of the study, reported normal or corrected-to-normal vision, and provided informed consent. Ethical approval for the study was obtained from the Georgia Tech Institutional Review Board.

2.2. Task

Each trial began with a fixation for 1 s and was followed by a Gabor patch stimulus (Fig. 1a). After a 100 ms interstimulus interval (ISI) with a blank screen, subjects first indicated whether the Gabor patch was tilted left (counterclockwise) or right (clockwise) relative to a 45° tilt. On the perceptual response screen, a red (#FF0000) line was drawn on a circle of the same size as the Gabor patch to indicate a 45° tilt reference. Subjects were instructed to press the left and right arrow on the keyboard with their right hand to indicate left and right tilt, respectively. Subjects subsequently rated their confidence on a 4-point scale, where 1 corresponds to low confidence and 4 corresponds to high confidence. They were encouraged to use the whole confidence scale. Subjects provided the confidence response using the keyboard with their left hand. Both the perceptual and confidence responses were untimed, and the next trial began after both responses were made. The experiment was programmed in JavaScript with the jsPsych library 7.3.3 (de Leeuw, 2015). The virtual chinrest method was used to control differences in viewing distance, display size, and stimulus size. Specifically, subjects were first asked to resize the stimuli on the screen to match the size of a credit card, adjusting for discrepancies in display size.



b

Experiment 1

Manipulations	Hard (H)	Medium (M)	Easy (E)
Size	2.5°	5°	7.5°
Duration	33ms	100ms	500ms
Noise	0.9	0.75	0.1
Tilt offset	0.5T	T	2T

Manipulations	Baseline condition	Hard conditions				Easy conditions			
	1	2	3	4	5	6	7	8	9
Size	M	H	M	M	M	E	M	M	M
Duration	M	M	H	M	M	M	E	M	M
Noise	M	M	M	H	M	M	M	E	M
Tilt offset	M	M	M	M	H	M	M	M	E

c

Experiment 2

Manipulations	Hard	Easy
Size	3°	7.5°
Spatial frequency	3.5 cycles	8 cycles
Noise	0.85	0.75
Tilt offset	T	1.8T

Fig. 1. Task and experimental design. (a) Each trial began with a fixation for 1 s and was followed by a Gabor patch stimulus. After a 100 ms interstimulus interval (ISI) with a blank screen, subjects first indicated whether the Gabor patch was tilted left (counterclockwise) or right (clockwise) relative to a 45° tilt and subsequently rated their confidence on a 4-point scale. Both responses were untimed, and the next trial began after both responses were made. (b) Stimulus parameters and experimental design of Experiment 1. Four stimulus features (size, duration, noise, and tilt offset) were manipulated with three levels of difficulty. "T" indicates the threshold estimated in a staircase procedure. The upper panel presents the exact stimulus parameters for each stimulus manipulation. The lower panel depicts the design of 9 experimental conditions. The level of difficulty of each stimulus manipulation is indicated by H, M, and E, referring to Hard, Medium, and Easy, respectively. (c) Stimulus parameters for Experiment 2. Experiment 2 had the same manipulations except that we used spatial frequency instead of duration. Unlike Experiment 1, here we used a full factorial design with a total of 16 conditions with each condition featuring a different combination of the four manipulations.

Then, subjects fixated on a central point and tracked the lateral movement of the target in their periphery, pressing the spacebar when the target disappeared. The procedure was repeated three times to estimate the viewing distance, assuming the blind spot is located at 13.5° of the visual angle (Li et al., 2020).

2.3. Stimuli and design

2.3.1. Experiment 1

In Experiment 1, we independently manipulated 4 stimulus features: the size, duration, noise, and tilt offset of the Gabor patches (Fig. 1b). For each feature, we sampled 3 levels of difficulty based on a pilot study. For the size manipulation, the three levels of difficulty were 2.5° (hard), 5° (medium), and 7.5° (easy) of visual angle. The three levels of duration were 33 ms (hard), 100 ms (medium), and 500 ms (easy). Noise was manipulated by blending (computing the weighted sum of two images) the Gabor patch with a background of random noise. A 90 % noise manipulation would indicate 10 % weighting of the Gabor patch and 90 % weighting of the noisy background. Three levels of difficulty for the noise manipulation were 90 % (hard), 75 % (medium), and 10 % (easy). Finally, tilt refers to the degree of offset from 45° and was manipulated based on an individualized threshold (T) obtained by a staircase procedure as described below. The three levels of tilt included 0.5 T (hard), T (medium), and 2 T (easy). Other stimulus parameters of the Gabor patches were kept constant (Spatial frequency=8 cycles; phase=0).

To manipulate these stimulus features independently, we first created a baseline condition that included the medium level for all four stimulus features. Other conditions were created by altering only the feature of interest while keeping other features at their baseline levels. For instance, the size-hard condition was created by changing the size-medium parameter from the baseline condition to the size-hard parameter while keeping other features unchanged. This approach creates eight additional conditions, derived from the four stimulus features each having easy and difficult levels. Including the baseline condition, we had a total of nine experimental conditions (Fig. 1b). Each condition was tested for 72 trials, except for the baseline condition, which was tested for 144 trials, resulting in a total of 720 trials. All experimental conditions were randomly interleaved throughout the experiment.

Subjects completed the 720 experimental trials in the form of 5 runs, each containing 6 blocks of 24 trials. Subjects were advised to take short breaks after each run and block. Before the start of the main experiment, subjects received 3 practice blocks and 2 staircase blocks. In all these blocks, the stimulus parameters for size, duration, and noise were set at medium difficulty. The first practice block contained 15 trials with a tilt offset of 10°. Then, subjects completed 40 trials with the tilt offset decreasing every 10 trials (tilt offset=5°, 3°, 2°, and 1°). Feedback on perceptual performance was given for the first two practice blocks. In the third practice block, subjects completed 20 trials without feedback with tilt offsets of 4°, 2°, and 1° randomly interleaved. Subjects then performed 2 staircase blocks to obtain the individualized threshold (T). The average individualized threshold was 3.39±0.90°. The staircase procedure was a three-down-one-up staircase with an asymmetric step size. The initial tilt offset was 2° and the step size was 0.2° implemented with a down/up factor of 0.7393 (García-Pérez, 1998). The staircase procedure terminates whenever the number of reversals exceeds 14 or the number of trials exceeds 60. The threshold was then obtained by averaging all reversal values except for the first reversal. The final threshold was obtained by averaging the threshold result from the 2 staircase blocks and served as the individualized threshold (T) throughout the experiment. On average, the individualized threshold was 3.38°. Subjects were not prompted to provide confidence responses during the staircase procedure. The whole experimental procedure took about 60 min to complete.

2.3.2. Experiment 2

Experiment 2 employed a full-factorial design on four stimulus features: size, spatial frequency, noise, and tilt offset of the Gabor patches with 2 levels of difficulty for each feature (Fig. 1c). Stimulus duration was kept constant at 200 ms. This 2x2x2x2 factorial design resulted in a total of 16 experimental conditions. Each condition was tested for 100 trials, for a total of 1600 trials. All conditions were randomly interleaved throughout the experiment.

We collected data from each subject over two different days. On each day, subjects completed 800 trials in the form of 5 runs, each containing 5 blocks of 32 trials. On the first day, subjects received the same 3 practice and 2 staircase blocks as in Experiment 1 before the main experiment to familiarize them with the task and obtain the individualized threshold. The average individualized threshold was 3.35°. On the second day, we only included two recall blocks. The first recall block contained 10 trials with a tilt offset of 10° and perceptual feedback. Then, subjects completed a block of 24 trials in their individualized threshold without receiving feedback. In all these blocks, only the tilt offset was manipulated, other stimulus parameters were set as follows: size=5°, spatial frequency=8 cycles, and noise=75 %. Each day of experimental session took about 60 min to complete.

2.4. Analyses

We first computed subjects' performance (d') and mean confidence for each stimulus manipulation. d' was calculated using the following formula:

$$d' = \Phi^{-1}(\text{Hit rate}) - \Phi^{-1}(\text{False alarm rate})$$

where Φ^{-1} is the inverse of the cumulative standard normal distribution that transforms hit rate and false alarm rate into z scores. Note that since Experiment 2 employed a full-factorial design, we computed the d' and mean confidence for each stimulus manipulation by combining trials that shared the same level of difficulty on the manipulation of interest (8 different experimental conditions for each level of difficulty). This allowed us to investigate the main effect of each stimulus manipulation.

We performed two different analyses to quantify the dissociable relationship between confidence and d' for all four manipulations in both experiments. For each manipulation, we first plotted confidence against d' and fit a linear regression. We then computed the slope (β) of the linear regression, which shows the change in confidence for each unit increase in d'. A larger slope denotes a bigger change in confidence for each unit increase in d' (relatively larger impact on confidence than accuracy), whereas a lower slope denotes a smaller change in confidence for each unit increase in d' (relatively smaller effect on confidence than accuracy). However, we could not perform this analysis for the duration manipulation because it had a very small effect on both confidence and accuracy and also eight subjects in Experiment 1 showed slightly higher accuracy in the short than long duration. Therefore, to avoid excluding an excessive number of subjects due to the issues related to the duration manipulation, we instead removed the duration manipulation

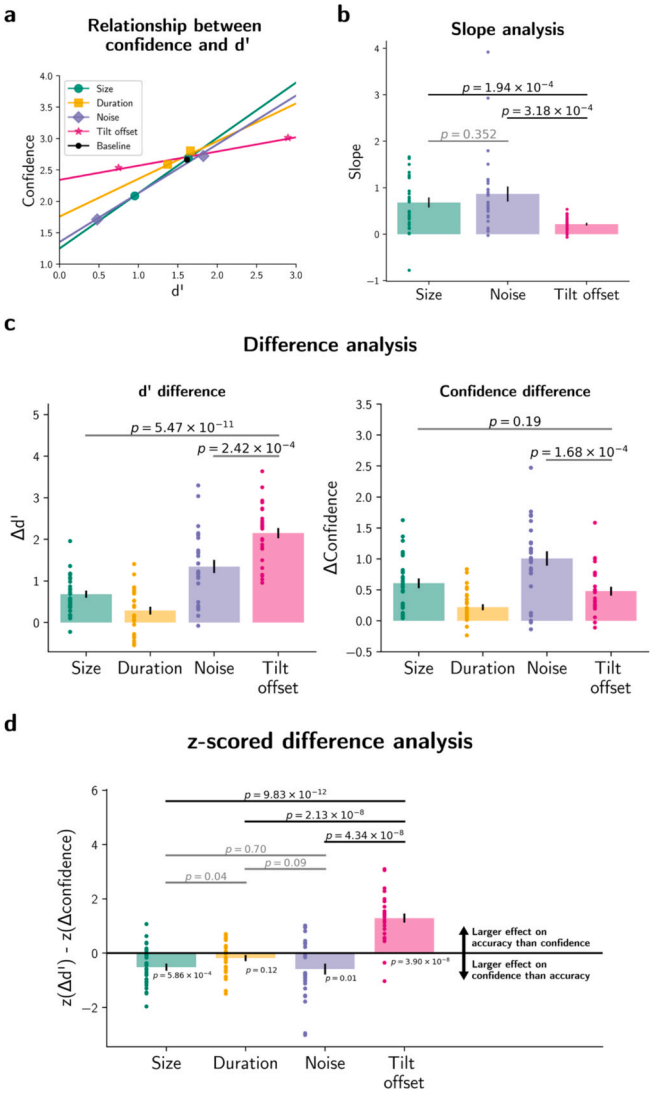


Fig. 2. Experiment 1 behavioral results. (a) Relationship between confidence ratings and sensitivity (d'). The relationship between confidence and d' appeared to be very similar for the size, duration, and noise manipulations but was very different for the tilt offset manipulation. The figure plot average d' vs. average confidence across subjects, as well as a linear regression on these average values for visualization only. In the actual analyses, we fitted separate linear regressions for each subject for each stimulus manipulation. (b) The slope of the linear regression line plotting confidence against d' for the size, noise, and tilt offset manipulations. Larger slopes denote a larger increase in confidence for each unit increase in d'. We found a similarly large slope for both the size and noise manipulations. Both of which were significantly larger than the slope of the tilt offset manipulation. (c) Differences in d' and confidence between each stimulus manipulation's easy and hard conditions. The tilt offset manipulation produced a significantly larger d' difference, as compared to the size and noise manipulation, but yielded a smaller confidence difference. Note that the figure only highlights comparisons that change direction for d' vs. confidence. (d) *z*-score differences between d' and confidence for each stimulus manipulation. A positive value denotes relatively larger effect on accuracy than confidence. The tilt offset manipulation is the only manipulation that yielded a larger effect on accuracy than confidence, which was significantly different from the size, duration, and noise manipulations. *p*-value below the zero line correspond to the results of one-sample t-tests against 0 for each stimulus manipulation. Dots show individual subjects. Error bars represent SEM.

from this slope analysis. Slopes from different manipulations were then compared using paired sample t-tests. We further performed linear mixed-effect modelling to examine how confidence and stimulus difficulty predicted trial-level accuracy and also accounting for individual differences across subjects (see Supplementary results). We also performed Bayes Factor analyses to assess the strength of evidence for our hypotheses. All statistical analyses were performed in Python3.9 with the Pingouin package (https://pingouin-stats.org/build/html/index.html). We used the default Cauchy scale factor ($r = \sqrt{2}/2$) for all Bayes factor analyses.

In addition, we employed another method for quantifying the relative effects of each manipulation on d' and confidence. The goal of this alternative method was to more directly illustrate the difference between d' and confidence. Specifically, we computed the differences in d' and confidence between the easy and hard conditions for each manipulation and compared these differences between pairs of manipulations using paired sample t-tests. Since different stimulus manipulations produce varying differences between the easy and hard condition for both d' and confidence, to enable comparison between these differences across multiple stimulus manipulations, we first z-scored the differences in d' and confidence to bring them onto the same scale and compute the difference between d' and confidence: $z(\Delta d') - z(\Delta \text{confidence})$. The resulting metric indicates the relative effect on accuracy versus confidence, where a positive value corresponds to a larger effect on accuracy than confidence, while a negative value corresponds to a larger effect on confidence than accuracy. We then compared this z-scored difference to zero using one-sided t-tests to test for the significance between d' and confidence difference for each manipulation and compared these z-scored differences between pairs of manipulations using paired sample t-tests.

Apart from examining the relationship between confidence and d', we also examined how confidence in correct and error trials changes with stimulus discriminability. Prior studies coined the term "folded-X pattern" to describe a pattern of results where easy conditions lead to a confidence increase for correct trials but a confidence decrease for error trials (Hangya et al., 2016; Sanders et al., 2016). To quantify the folded-X pattern, we fit linear regressions to predict confidence by stimulus difficulty separately for correct and error trials in each stimulus manipulation. We then computed the slope (β) of the linear regression, in which a positive β indicates a confidence increase with stimulus discriminability, whereas a negative β indicates a confidence decrease with stimulus discriminability. The folded-X pattern corresponds to a positive β for correct trials and a negative β for error trials. The obtained β values were compared to zero using one-sided t-tests.

In Experiment 2, we also investigated the two-way interaction effects among the four stimulus manipulations. We investigated the interaction effects separately for d' and confidence by conducting two-way repeated-measures analyses of variance (ANOVAs).

2.5. Data and code availability

All data and codes are available at https://github.com/herrickfung/4m_data_code/.

3. Results

We explored the similarities and differences in the effects of five different manipulations on confidence and accuracy. Subjects indicated whether a Gabor patch was tilted left or right relative to a 45° tilt and provided confidence on a 4-point scale. In Experiment 1, we independently manipulated the size, duration, noise, and the tilt offset of the Gabor patch, while in Experiment 2, we jointly manipulated the size, spatial frequency, noise, and the tilt offset. We then performed extensive analyses to investigate the behavioral effects of these manipulations on confidence and performance.

3.1. Effects of each manipulation on average confidence and accuracy

3.1.1. Experiment 1

We first examined how each of the four manipulations in Experiment 1 affected average confidence and average performance. To do so, we plotted confidence against performance (d') (Fig. 2a). Qualitatively, confidence increased with d' for all four manipulations. Critically, the relationship between confidence and d' appeared to be almost identical for all auxiliary manipulations, including size, duration, and noise, but very different for the task-defining tilt offset manipulation. We examine this qualitative impression in two different ways below.

First, to quantitatively examine the relative effects of each manipulation on d' and confidence, we computed the slope (β) of the linear regression, which shows the change in confidence for each unit increase in d' (Fig. 2b and Supplementary Table 1). Since the duration manipulation had a very small effect on both confidence and accuracy, we excluded the duration manipulation from this analysis (see Methods). The results for the size, noise, and tilt offset manipulations confirmed the qualitative impressions from Fig. 2a. Specifically, the size and noise manipulations exhibited very similar, large slopes (Size: $\beta = 0.68$, Noise: $\beta = 0.87$; $t(27)=0.328$, $p=0.745$, $BF_{01}=4.59$, $d=0.064$). In contrast, the tilt offset manipulation exhibited a substantially smaller slope ($\beta = 0.22$) compared to both size ($t(27)=4.31$, $p=1.94\text{x}10^{-4}$, $BF_{10}=144$, $d=0.814$) and noise ($t(27)=4.12$, $p=3.18\text{x}10^{-4}$, $BF_{10}=93.2$, $d=0.779$). These findings demonstrate that the size and noise manipulations have similar, large effects on confidence compared to their effects on d', whereas the tilt offset manipulation has a substantially smaller effect on confidence compared to its effect on d'.

Second, we also examined the relative effects of each manipulation on d' and confidence by computing the differences in d' and confidence between the easy and hard conditions for each manipulation (Fig. 2c). We found that the duration, size, and noise manipulations had increasingly larger effects on both d' and confidence (all 6p's<0.05 for all pairwise comparisons between pairs of manipulations for both d' and confidence), suggesting that all three auxiliary manipulations had similar relative effects on

performance and confidence. In contrast, the task-defining tilt offset manipulation had a relatively large effect on d' but a relatively small effect on confidence. Indeed, the tilt offset manipulation had a significantly larger effect on d' compared to both the noise ($t(27)$=4.23, p=2.42 x 10^{-4}, BF_{10}=119, d=0.799) and size manipulations ($t(27)$=10.5, p=5.47x10^{-11}, BF_{10}=1.76x10^{8}, d = 1.97). However, tilt offset had a significantly smaller effect on confidence compared to the noise manipulation ($t(27)$=-4.36, p=1.68x10^{-4}, BF_{10}=165, d=-0.825) and a numerically smaller but not significantly different effect on confidence from the size manipulation ($t(27)$=-1.33, p=0.19, BF_{01}=2.26, d=0.251).

To directly compare the effects of each manipulation on d' vs. confidence, we computed the z-score differences between d' and

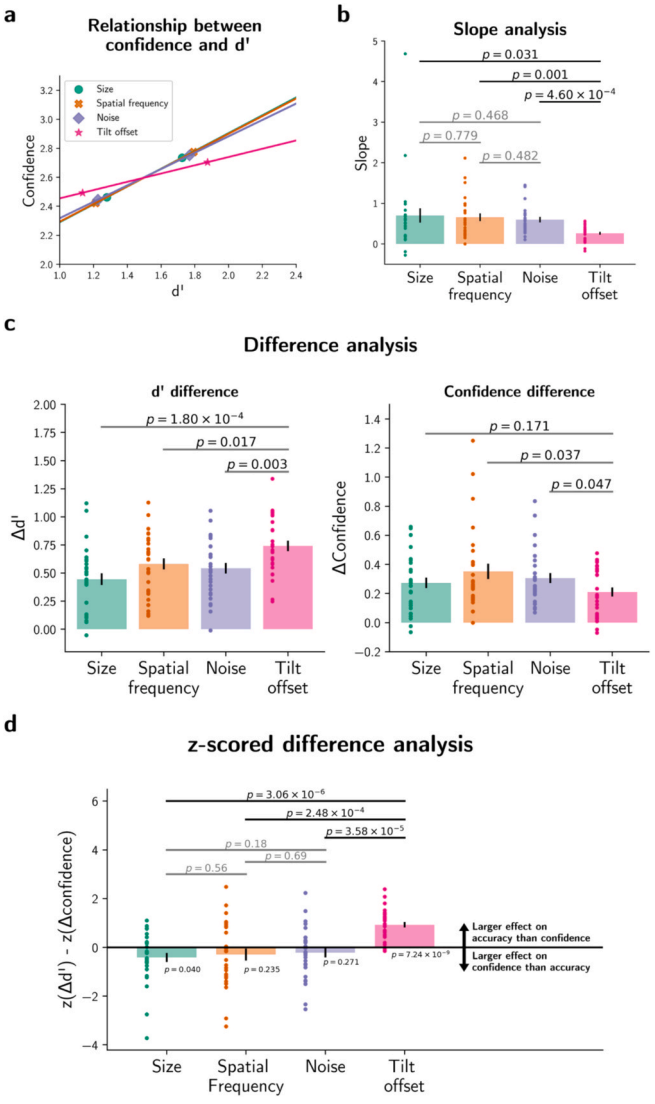


Fig. 3. Experiment 2 behavioral results. (a) Relationship between confidence ratings and sensitivity (d'). The relationship between confidence and d' was nearly identical for the size, spatial frequency, and noise manipulations, but was very different for the tilt offset manipulation. The figure plot average d' vs. average confidence across subjects, as well as a linear regression on these average values for visualization only. In the actual analyses, we fitted separate linear regressions for each subject for each stimulus manipulation. (b) The slope of the linear regression line plotting confidence against d' for each stimulus feature. A larger slope would denote a larger increase in confidence for each unit increase in d'. A similarly larger slope was found for size, spatial frequency, and noise manipulations. All of them were larger than the slope of the tilt offset manipulation. (c) d' and Confidence difference between each stimulus manipulation's easy and hard conditions. The tilt offset manipulation produced a significantly larger d' difference, as compared to the size, spatial frequency, and noise manipulation, but yielded no significant differences in confidence difference. Note that the figure only highlights comparisons that change direction for d' vs. confidence. (d) *z*-scored differences between d' and confidence for each stimulus manipulation. A positive value denotes relatively larger effect on accuracy than confidence. *p*-value below the zero line correspond to the results of one-sample t-tests against 0 for each stimulus manipulation. The tilt offset manipulation is the only manipulation that yielded larger effect on accuracy than confidence, which was significantly different from the size, spatial frequency, and noise manipulations. Dots show individual subjects. Error bars represent SEM.

confidence (Fig. 2d). Specifically, we *z*-scored the differences between the easy and hard conditions across manipulations for d' and confidence separately. We then examined the difference between these two z-scores: $z(\Delta d') - z(\Delta \text{confidence})$. We found that size and noise had significantly negative z-score differences (Size: $t(27)$=-3.89, p=5.86x10^{-4}, BF_{10}=54.1, d=-0.736; Noise: $t(27)$=-2.92, p=0.007, BF_{10}=6.28, d=-0.552), indicating a larger effect on confidence than accuracy. Duration also exhibited a negative z-score difference, but this effect was not significant ($t(27)$=-1.60, p=0.122, BF_{01}=1.61, d=-0.302). In contrast, tilt offset was the only manipulation that demonstrated a significantly positive z-score difference ($t(27)$=7.56, p=3.90x10^{-8}, BF_{10}=3.63x10^{5}, d=1.43), indicating a larger effect on accuracy than confidence. Critically, the z-score difference for tilt offset was significantly higher than for the size, noise, and duration manipulations (all three p's<0.001 and BF_{10}>10^{5} for all pairwise comparisons). Size, noise, and duration exhibited relatively similar z-score differences, though the z-score difference for duration was closer to 0 due to the fact that duration had negligible effects on both d' and confidence (Size vs duration: $t(27)$=2.19, p=0.04, BF_{10}=1.55, d=0.413; Size vs noise: $t(27)$=0.39, p=0.70, BF_{01}=4.65, d=0.073; Duration vs noise: $t(27)$=1.74, p=0.09, BF_{01}=1.32, d=0.328). Overall, these results suggest that all auxiliary manipulations have similar relative effects on d' and confidence, while the task-defining tilt offset manipulation impacts d' more than confidence compared to the other manipulations.

3.1.2. Experiment 2

In Experiment 1, we manipulated size, duration, noise, and tilt offset independently. Specifically, when manipulating one variable, all other variables were given a default, baseline value. This design allowed us to compare the different manipulations but leaves it unclear whether the manipulations interact with each other. To address this question, in Experiment 2, we jointly manipulated four stimulus features using a 2x2x2x2 experimental design. Since duration demonstrated minimal impact on both confidence and accuracy in Experiment 1, we replaced the duration manipulation with a spatial frequency manipulation in Experiment 2. We first examined the effects of each manipulation on accuracy and confidence (as in Experiment 1) and then analyzed the interactions between the different manipulations.

To qualitatively examine the effects of each manipulation on accuracy and confidence, we first plotted confidence against performance (d') for each manipulation while combining across the remaining three manipulations (Fig. 3a). As expected, confidence increased with d' for all four manipulations. Critically, the relationship between confidence and d' was close to identical for all auxiliary manipulations in Experiment 2, including size, spatial frequency, and noise, but was very different for the task-defining tilt offset manipulation.

As in Experiment 1, we quantitatively examined the relative effect of each manipulation on d' and confidence by computing the slope (β) of the linear regression, which shows the change in confidence for each unit increase in d' (Fig. 3b and Supplementary Table 2). We found that the size, spatial frequency, and noise manipulations exhibited very similar, large slopes (Size: $\beta = 0.70$, Spatial frequency: $\beta = 0.66$, Noise: $\beta = 0.60$; p>0.1 and BF_{01}>3 for all three pairwise comparisons; Fig. 3b). In contrast, the tilt offset manipulation exhibited a significantly smaller slope ($\beta = 0.26$), as compared to spatial frequency ($t(26)$=3.61, p=0.001, BF_{10}=27.5, d=1.04), noise ($t(26)$=4.0, p=4.60 x 10^{-4}, BF_{10}=67.6, d=1.14), and size ($t(26)$=2.28, p=0.031, BF_{10}=1.84, d=0.663). These findings replicated and extended results from Experiment 1, showing that all auxiliary manipulations have similarly large effects on confidence compared to their impact on d', whereas the task-defining tilt offset manipulation has a substantially smaller effect on confidence compared to its effect on d'.

Further, as in Experiment 1, we also examined the relative effects of each manipulation on d' and confidence by computing the difference in d' and confidence between the easy and hard conditions for each manipulation while combining across all remaining manipulations (Fig. 3c). We found that, compared to each other, the size, noise, and spatial frequency manipulations had increasingly larger effects on both d' and confidence (p<0.05 for all 6 pairwise comparisons), suggesting that all three of these manipulations had similar relative effects on performance and confidence. In contrast, the tilt offset manipulation had a larger effect on d' but a smaller effect on confidence. Specifically, the tilt offset manipulation had a significantly larger effect on d' compared to size ($t(28)$=4.31, p=1.80x10^{-4}, BF_{10}=154, d=1.11), spatial frequency ($t(28)$=2.53, p=0.017, BF_{10}=2.90, d=0.615), and noise ($t(28)$=3.27, p=0.003, BF_{10}=13.3, d=0.778). However, tilt offset had a significantly smaller effect on confidence compared to spatial frequency ($t(28)$=2.18, p=0.037, BF_{10}=1.54, d=0.605) and noise ($t(28)$=2.08, p=0.047, BF_{10}=1.27, d=0.538) and a numerically smaller effect (though non-significant) on confidence compared to size ($t(28)$=1.41, p=0.171, BF_{01}=2.09, d=0.337).

Similarly, we computed the *z*-score differences between d' and confidence (Fig. 3d). We found that size demonstrated significantly negative z-score differences ($t(28)$=-2.16, p=0.04, BF_{10}=1.45, d=-0.430), indicating a larger effect on confidence than accuracy. Spatial frequency and noise also exhibited a negative z-score difference, but the effects were not significant (Spatial frequency: $t(28)$=-1.21, p=0.23, BF_{01}=2.60, d=-0.258; Noise: $t(28)$=-1.12, p=0.27, BF_{01}=2.86, d=-0.249). In contrast, tilt offset is the only manipulation that demonstrated significantly positive z-score differences ($t(28)$=8.14, p=7.24x10^{-9}, BF_{10}=1.76x10^{-6}, d=1.10), indicating a larger effect on accuracy than confidence. Critically, the z-score difference for tilt offset was significantly higher than for the size, noise, and spatial frequency manipulations (all three p's<0.001 and BF_{10}>80 for all pairwise comparisons). Size, noise and spatial frequency showed relatively similar z-score differences (all three p's>0.1 and BF_{01}>2 for all pairwise comparisons). These results demonstrate that the size, spatial frequency, and noise manipulations have similar effects on d' and confidence, while the tilt offset manipulation stood out in affecting d' more than confidence compared to all other manipulations. Overall, these results from Experiment 2 successfully replicated and expanded the results from Experiment 1, showing that all auxiliary manipulations exhibit similar relationship between d' and confidence, whereas the task-defining tilt offset manipulation was different from all other manipulations by affecting d' more than confidence.

Critically, the design of Experiment 2 also allowed us to examine how the different manipulations interact with each other. To this end, we analyzed all six possible two-way interactions among the size, spatial frequency, noise, and tilt offset manipulations for both d'

and confidence (Fig. 4). We found that all two-way interactions involving tilt offset yielded significant supraadditive effects on both d' (Size vs. Tilt offset: $F(1,28)=21.0$, $p=8.69\text{x}10^{-5}$, $\eta_p^2=0.429$; Spatial frequency vs. Tilt offset: $F(1,28)=26.4$, $p=1.89\text{x}10^{-5}$, $\eta_p^2=0.485$; Noise vs. Tilt offset: $F(1,28)=26.5$, $p=1.87\text{x}10^{-5}$, $\eta_p^2=0.486$) and confidence (Size vs. Tilt offset: $F(1,28)=10.4$, $p=0.003$, $\eta_p^2=0.271$; Spatial frequency vs. Tilt offset: $F(1, 28)=31.7$, $p=5.03\text{x}10^{-6}$, $\eta_p^2=0.531$; Noise vs. Tilt offset: $F(1,28)=14.4$, $p=7.38\text{x}10^{-4}$, $\eta_p^2=0.339$). Specifically, the tilt offset manipulation increased d' and confidence more when the other manipulations (size, noise, or spatial frequency) were easy compared to when they were difficult.

Contrary to the supraadditive interactions we observed for the tilt offset manipulation, the remaining manipulations showed either subadditive interactions or no interactions between each other. Specifically, we found a subadditive interaction between the size and noise manipulations for both d' ($F(1,28)=18.2$, $p=2.03\text{x}10^{-4}$, $\eta_p^2=0.394$) and confidence ($F(1,28)=73.7$, $p=2.50\text{x}10^{-9}$, $\eta_p^2=0.725$). Specifically, the noise manipulation increased d' and confidence less when the size manipulation was easy compared to when they were difficult. We also found a subadditive interaction between the noise and spatial frequency but only for d' ($F(1,28)=7.16$, $p=0.012$, $\eta_p^2=0.204$). There was no interaction between noise and spatial frequency for confidence ($F(1,28)=1.19$, $p=0.28$, $\eta_p^2=0.041$). Finally, there was also no interaction between size and spatial frequency for either d' ($F(1,28)=0.09$, $p=0.76$, $\eta_p^2=0.003$) or confidence ($F(1,28)=0.62$, $p=0.44$, $\eta_p^2=0.022$). These results demonstrate that the tilt offset manipulation stands out in that it is the only manipulation that produces supraadditive interaction effects for either d' or confidence.

3.2. *Effects of each manipulation on confidence for correct vs. error trials (folded-X pattern)*

Confidence is typically thought to increase with easier trials for correct responses but to decrease with easier trials for error

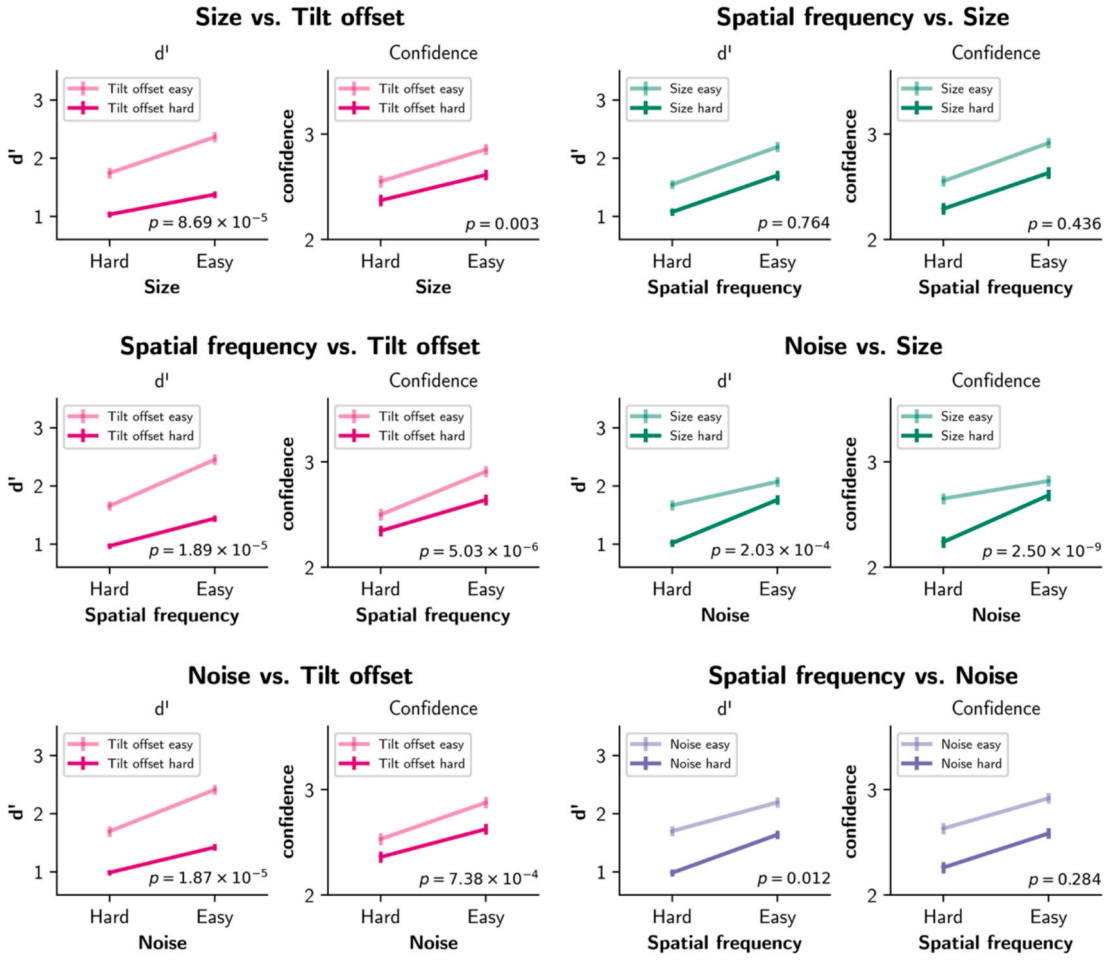


Fig. 4. All two-way interaction effects among stimulus manipulations for d' and confidence in Experiment 2. All interactions with the tilt offset manipulation (left) produced a consistent and similar supraadditive interaction for both d' and confidence. Interactions among all other stimulus manipulation (right) exhibited either no or subadditive interactions. Error bars represent SEM; p-values show interactions effects for two-way repeated measures ANOVAs.

responses (Hangya et al., 2016; Sanders et al., 2016). This pattern, known as the folded-X pattern, can be used to understand how different stimulus features interact with the confidence generation process (Rausch et al., 2018; Shekhar & Rahnev, 2024). Thus, we also analyzed how confidence changes for correct and error responses in each stimulus manipulation.

Across both experiments, we observed violations of the canonical folded-X pattern for all stimulus manipulations except tilt offset (Fig. 5). In Experiment 1, for the size, duration, and noise manipulations, easier conditions increased confidence in both correct trials (Size: β=0.311, $t(26)$=7.03, p=1.80x10^{-7}; Duration: β=0.113, $t(26)$=4.82, p=5.40x10^{-5}; Noise: β=0.514, $t(26)$=8.75, p=3.18x10^{-9}) and error trials (Size: β=0.260, $t(26)$=7.47, p=6.23x10^{-8}; Duration: β=0.07, $t(26)$=1.88, p=0.07; Noise: β=0.374, $t(26)$=5.55, p=7.86x10^{-6}). This joint increase in confidence regardless of the response accuracy shows that size, duration, and noise conditions violated the folded-X pattern. However, for the tilt offset manipulation, easier conditions increased confidence in the correct trials (β=0.247, $t(26)$=7.32, p=8.97x10^{-8}), but not in the error trials (β=-0.05, $t(26)$=-1.21, p=0.24), which resembles the canonical folded-X pattern. The same pattern emerged in Experiment 2. Specifically, for size, spatial frequency, and noise manipulations, easier conditions increased confidence for both correct trials (Size: β=0.249, $t(28)$=6.61, p=3.62x10^{-7}; Spatial frequency: β=0.342, $t(28)$=6.14, p=1.24x10^{-6}; Noise: β=0.282, $t(28)$=8.87, p=1.25 x 10^{-9}) and error trials (Size: β=0.175, $t(28)$=6.07, p=1.50x10^{-6}; Spatial frequency: β=0.162, $t(28)$=3.12, p=0.004; Noise: β=0.176, $t(28)$=4.56, p=9.14x10^{-5}). In contrast, for the tilt offset manipulation, easier conditions increased confidence in the correct trials (β=0.215, $t(28)$=8.10, p=8.13x10^{-9}), but decreased confidence in the error trials (β=-0.105, $t(28)$=-5.57, p=5.79x10^{-6}). These results demonstrate that all auxiliary manipulations, including size, spatial frequency, duration, and noise manipulations, violated the folded-X pattern and instead showed a joint increase in confidence for correct and error trials. In contrast, the task-defining tilt offset manipulation stood out in being the only manipulation that demonstrated the canonical folded-X pattern.

3.3. Effects of each manipulation on reaction time (RT)

Since RT and confidence are related, we also investigated how different manipulations affect RT and whether the differences in RT would mimic differences in confidence. We found that even though the tilt offset manipulation had the smallest effect on confidence (Fig. 2c), its effect on RT was as large as or larger than the effect of the remaining manipulations (Supplementary Fig. 1). Further, we observed no obvious difference between the tilt offset manipulation and other stimulus manipulations on RT patterns for correct vs. error trials (Supplementary Fig. 2). Specifically, the results for tilt offset – similar to the remaining manipulations – were in line with the double-increase pattern of confidence. Overall, these results demonstrate the confidence results observed in this paper are not explainable by differences in RT, and that the RT effects in fact tend to follow the accuracy results more closely than the confidence results.

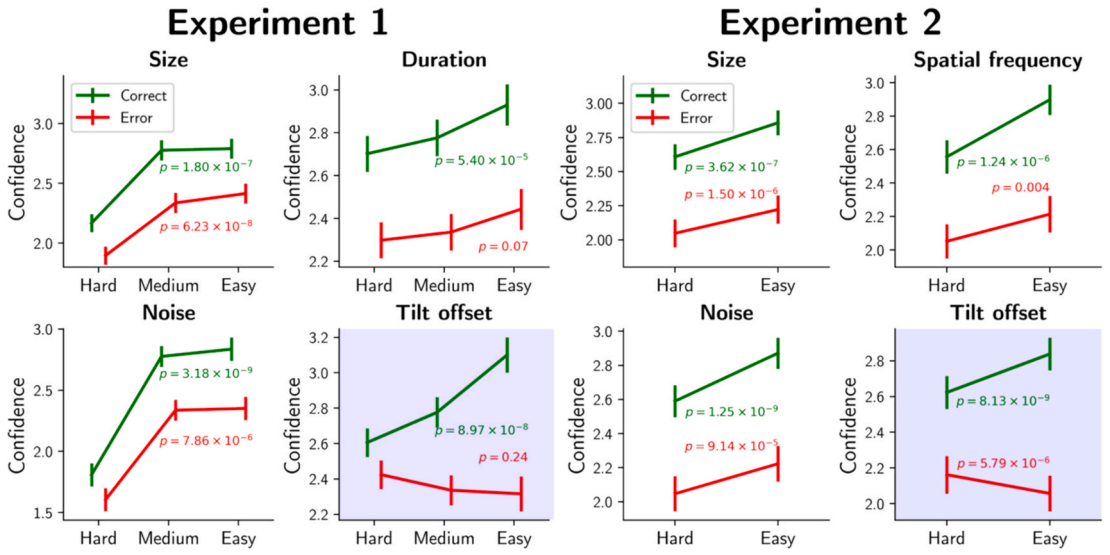


Fig. 5. Effects of each manipulation on confidence for correct vs. error trials (folded-X pattern). In both experiments, the canonical folded-X pattern (where easier conditions lead to a confidence increase for correct trials but a confidence decrease for error trials) only appeared for the tilt offset manipulation. All other stimulus manipulations (size, duration, noise, and spatial frequency) led to violations of the folded-X pattern, such that easier conditions led to increased confidence for both correct and error trials. The green and red lines denote the confidence for correct and incorrect trials, respectively. Error bars represent SEM. Statistical results show the significance of the one-sided *t*-test to determine if the average slope was significantly different from zero. The purple background highlights the tilt offset results, which are the only ones to exhibit the canonical folded-X pattern. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Discussion

We investigated the effects of multiple stimulus manipulations on confidence and accuracy. Across two experiments, we tested the effects of one task-defining manipulation (tilt offset) and four auxiliary manipulations (size, duration, spatial frequency, and noise). We found that the tilt offset manipulation stands out in three different ways. Specifically, it is the only manipulation that (1) affects accuracy more strongly than confidence, (2) demonstrates the canonical folded-X pattern, and (3) exhibits supraadditive interactions with the other manipulations for both confidence and accuracy. In contrast, the other four auxiliary manipulations showed fairly similar effects on confidence and accuracy despite the vast differences in how they impact the appearance of the stimuli. Overall, these results reveal very similar effects among four visually distinctive auxiliary manipulations and a prominent difference between them and the task-defining manipulation. These results can be used to predict how untested stimulus manipulation would affect confidence and accuracy.

Several studies have found that different manipulations may sometimes produce different effects on accuracy vs. confidence. For instance, manipulating the variance of the elements in an ensemble averaging task affects confidence more than accuracy compared to manipulating the mean of the ensemble (Boldt et al., 2017; Spence et al., 2016). Likewise, in an orientation discrimination task, a contrast manipulation affects confidence more than accuracy, whereas a tilt offset manipulation affects accuracy more than confidence (Xue et al., 2025). However, such studies do not reveal if each manipulation has its own distinctive influence on confidence vs. accuracy, or if most manipulations conform to only a few patterns. Our results demonstrate that multiple visually distinctive stimulus manipulations produced just two distinct patterns of confidence-accuracy relationship. Specifically, the task defining tilt offset manipulation affects accuracy more than confidence, whereas all auxiliary manipulations (i.e., size, duration, noise, and spatial frequency) have a relatively larger effect on confidence than accuracy. Critically, all auxiliary manipulations demonstrate very similar effects on confidence vs. accuracy as indicated by the consistent slopes across different manipulations. Our results thus extend previous research by suggesting that, despite the existence of many possible stimulus manipulations, there may be a limited number of possible effects on confidence vs. accuracy that such manipulations could produce.

The folded-X pattern has been proposed as the statistical signature of confidence (Hangya et al., 2016; Sanders et al., 2016). It has even been argued that whenever the folded-X pattern is detected in a psychological, neural, or physiological variable, that variable should be considered a correlate of confidence (Kepecs et al., 2008; Kepecs & Mainen, 2012; Sanders et al., 2016). Consequently, some studies have used the folded-X pattern to detect correlates of confidence empirically (Braun et al., 2018; Rolls et al., 2010; Urai et al., 2017; Voodla et al., 2025). Despite the frequent empirical observation of the folded-X pattern (Shekhar & Rahnev, 2024; Xue et al., 2025), numerous studies have reported violations of this pattern (Hellmann et al., 2023; Kiani et al., 2014; Rausch et al., 2018, 2021; Shekhar & Rahnev, 2024; Xue et al., 2025). Our results confirm that violations of the folded-X pattern are rather common. In fact, all four auxiliary manipulations (size, duration, noise, and spatial frequency) showed violations of the folded-X pattern. Instead, these manipulations exhibited a double-increase pattern where easier conditions led to a confidence increase for both correct and error trials. Consistent with previous results, our study suggests that the emergence of the folded-X pattern might be related to the nature of the stimulus manipulation. These results cast doubt on the assumption that the folded-X pattern can be regarded as a hallmark feature of confidence and raise concerns about using this pattern as a definitive criterion for identifying empirical correlates of confidence.

Our results also suggest that there could be a relationship between whether a manipulation produces the folded-X or the double-increase pattern and whether that manipulation affects accuracy or confidence more strongly. Specifically, the tilt offset manipulation exhibited the folded-X pattern and showed the smallest effect on confidence among all stimulus manipulations. Analytically, this may be because, for easy conditions, the increased confidence in correct trials is offset by the decreased confidence in error trials, resulting in a smaller net increase in confidence. However, additional analyses revealed that the relatively smaller effect on confidence for the tilt offset manipulation (Fig. 2c) was also separately observed for both correct and error trials (though it was more pronounced in error trials; Supplementary Fig. 3). The fact that the tilt offset manipulation showed the same qualitative pattern when only correct trials are considered suggests that this effect is not exclusively driven by error trials and is therefore not statistically identical to the fact that only the tilt offset manipulation resulted in a folded-X pattern.

To our knowledge, only one previous study has examined the interactions between the effects of different manipulations on confidence. That study found no interaction between manipulating the mean color and manipulating the color variance in a color ensemble task (Boldt et al., 2017). In contrast, we observed numerous interactions on both accuracy and confidence. Specifically, the task-defining tilt offset manipulation demonstrated a supraadditive interaction with all auxiliary manipulations (i.e. size, noise, and spatial frequency) on both confidence and accuracy. In contrast, the auxiliary manipulations exhibited either subadditive interactions or no interactions on accuracy and confidence. Critically, these interaction effects are unlikely to be artifacts of floor or ceiling effects because both d' and confidence values remained far from the boundaries (confidence values were between 2.23 and 2.92; range=1–4; d' values were between 1.0 and 2.52; range=0–4.87). Our results demonstrate that interactions between different manipulations are likely to be rather common but do not yet clarify the mechanisms underlying these different interactions.

A central question raised by our work is what makes the tilt offset manipulations different from all others. Indeed, that manipulation had a unique confidence-accuracy relationship, it was the only one to exhibit the folded-X pattern, and it alone showed supraadditive interactions with other manipulations. We can think of three, non-exclusive possibilities for what sets the tilt offset manipulation apart.

First, it could be that the central reason for the difference between tilt offset and all other manipulations is that tilt offset is the task-defining manipulation, whereas all others are auxiliary manipulations. Note that while plausible, establishing that this is the driving force behind our effects would require further evidence where other features (besides tilt offset) are task-defining. The task-defining nature of stimulus tilt would also naturally lead to more attention being directed to the tilt offset compared to the features affected by

other manipulations. Thus, a prediction that stems from our results is that fundamentally different effects will emerge for manipulations that affect the task-defining features vs. manipulations that affect any non-task-defining stimulus feature.

Second, a separate and non-mutually exclusive possibility is that what differentiates tilt offset from all other manipulations is how noticeable the manipulation is. Indeed, the tilt offset manipulation was perceptually subtle (e.g., in Experiment 1, the average tilt values for clockwise stimuli were 46.75°, 48.5°, and 52°, which are not easy to tell apart; Fig. 6). In contrast, all auxiliary stimulus manipulations were readily apparent (e.g., durations of 33/100/500 ms and sizes of 2.5°/5°/7.5° are easy to tell apart). This difference was further confirmed by an additional experiment, which demonstrated that subjects were close to chance (32.7 %) when discriminating the three levels of the tilt offset manipulation used in Experiment 1 but performed much better for all other manipulations (all accuracy levels>70 %; see Supplementary Results and Supplementary Fig. 4). Thus, it is possible that this difference in how noticeable a stimulus manipulation is what drives the behavioral differences between the tilt offset and all other manipulations. In particular, subjects may use cues about the difficulty of the trial when making confidence judgements, as has been proposed by the weighted evidence and visibility (WEV) model (Hellmann et al., 2023; Rausch et al., 2018, 2021), though recent work shows that explicit cues about trial difficulty have only a minimal effect on confidence ratings (Xue et al., 2024b). Indeed, the WEV model has been shown to allow for the emergence of both the folded-X and the double-increase patterns (Rausch & Zehetleitner, 2019). Thus, it is possible that a computation that involves an implicit estimation of difficulty based on the perceived features of a stimulus could explain the observed difference between tilt offset and all other manipulations.

Third, both of our experiments treated tilt offset manipulation differently from the other manipulations prior to the beginning of the main experiments. Specifically, the tilt offset manipulation was practiced using a relatively wide range of values, whereas all other stimulus features were set to a constant level. In addition, the tilt offset manipulation is also the only manipulation on which we performed a staircase procedure to determine the individualized threshold. Since the staircase procedure was conducted before the experimental trials, from the subjects' perspectives it had a similar effect as the practice trials in that it exposed them to a wider range of values for that manipulation. To empirically test whether exposure to a wide range of values from a given manipulation can lead to the effects we observed, we conducted the slope and difference analyses separately on the first vs. second half of each experiment. The idea is that subjects have already been exposed to a wider range of values for each manipulation in the second half of the experiment. Thus, if the prior exposure to a range of values is critical to our findings, we should observe a difference between the first and second halves of the experiment. Our results were replicated in both splits of the data (Supplementary Fig. 5), suggesting that the effect of prior exposure to a range of values is unlikely to explain the differences we observed between the tilt manipulation and all remaining manipulations. Nevertheless, future experiments with tighter control over practice trials and staircasing procedures are needed to fully determine whether these factors can explain the results observed here.

Our results allow us to make predictions on how future manipulations would affect accuracy and confidence. Specifically, we speculate that manipulations of task-defining features that are also not readily noticeable would produce the effects associated with the tilt offset manipulation here. That is, such manipulations would have relatively more impact on accuracy than confidence, exhibit the folded-X pattern, and interact with all other non-task-defining stimulus manipulations supraadditively. In contrast, manipulations of non-task-defining features that are also readily noticeable would produce the effects associated with noise, duration, spatial frequency, and size. That is, such manipulations would have relatively less impact on accuracy than confidence, exhibit the double-increase pattern (violating the folded-X pattern), and show either no interactions or subadditive interactions among themselves. This proposed dichotomy is sufficiently robust to explain many existing findings with clearly distinguishable stimulus manipulations (Boldt et al., 2017; Kiani et al., 2014; Moran et al., 2015; Rausch et al., 2018, 2021; Sanders et al., 2016; Spence et al., 2016; Xue et al., 2025). For example, Kiani et al. (2014) found that manipulating the motion coherence (non-task-defining feature) in a motion direction discrimination task produced the double-increase pattern. Conversely, in a task where subjects were required to judge the dominant color in an ensemble containing an unequal amount of white and black squares, manipulating the proportion of the two colored squares (task-defining feature) created the canonical folded-X pattern (Moran et al., 2015). Overall, while still speculative, our

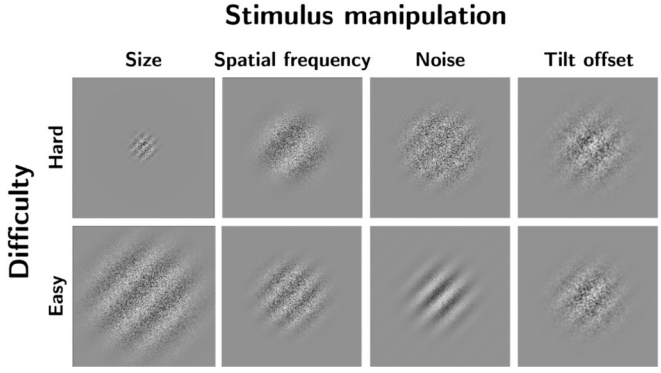


Fig. 6. Visualization of the levels of difficulty of the four stimulus manipulations used in Experiments 1 and 2. The images for size, noise, and tilt offset manipulation are based on the easy and hard levels in Experiment 1, whereas the images for the spatial frequency manipulation are based on the two levels in Experiment 2. As seen in this figure, the two level of difficulty of the tilt offset manipulation is the hardest to tell apart. Note that we could not visualize the extreme levels of the duration manipulation (33 vs. 500 ms), but those levels are also very easy to distinguish.

interpretation is general enough to encompass many different stimulus manipulations and can potentially enable predicting the effects of new, untested manipulations applied to a range of perceptual tasks.

Currently, the generalizability of our interpretations is currently limited by the fact that we only used a single task – the orientation discrimination task – with tilt offset as the sole task-defining feature. Therefore, it remains unclear whether our findings would be observed in other perceptual tasks or with other task-defining features. For instance, we would predict that for a size discrimination task where size is the task-defining feature, manipulating size – but not any other variable – would produce the same effects as the tilt offset manipulation in the current study. Future research should address this question by testing the effects of task-defining vs. auxiliary manipulations across a range of tasks.

Another limitation of our study is that to comprehensively explore the data, we conducted many hypothesis tests, which increases the risk that some of the significant results reported in the paper are false positives. However, we emphasize that most key effects were replicated across two independent experiments, which increases confidence that they are not spurious. Additionally, most tests were not meant to be interpreted in isolation, but rather to evaluate the overall pattern across multiple conditions. More importantly, the critical finding of a confidence-accuracy dissociation is confirmed by multiple types of analyses, including the slope analysis, difference analysis, z-scored difference analysis, and linear mixed effect modelling, all of which converge to the same conclusion.

In conclusion, we explored the effects of multiple stimulus manipulations on confidence and accuracy in an orientation discrimination task and found that the tilt offset manipulation stands out in three key ways. We propose that this distinction arises because tilt offset is both task-defining and less perceptually noticeable. This interpretation enables predictions about how untested stimulus manipulations might affect confidence and accuracy. While exploratory, this account offers a coherent explanation for both our results and prior findings, and it generates testable predictions for future research.

CRediT authorship contribution statement

Herrick Fung: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Medha Shekhar:** Investigation, Formal analysis. **Kai Xue:** Investigation. **Manuel Rausch:** Writing – review & editing, Investigation, Funding acquisition. **Dobromir Rahnev:** Writing – review & editing, Supervision, Investigation, Funding acquisition.

Declaration of competing interest

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

This work was funded by the National Institute of Health (R01MH119189) and the Office of Naval Research (N00014-20-1-2622) to D.R. M.R. was supported by the Deutsche Forschungsgemeinschaft (grants RA 2988/3-1 and RA 2988/4-1).

Author Contributions: All authors designed research and interpret the results. H.F. and M.S. performed the research and analyzed the data. H.F. collected the data and wrote the first draft of the paper. H.F. and D.R. edited the paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.concog.2025.103942.

Data availability

All data and codes are available at https://github.com/herrickfung/4m_data_code/.

References

Bang, J. W., Shekhar, M., & Rahnev, D. (2019). Sensory noise increases metacognitive efficiency. *Journal of Experimental Psychology: General, 148*(3), 437–452. https://doi.org/10.1037/xge0000511.supp

Boldt, A., de Gardelle, V., & Yeung, N. (2017). The impact of evidence reliability on sensitivity and bias in decision confidence. *Journal of Experimental Psychology: Human Perception and Performance, 43*(8), 1520–1531. https://doi.org/10.1037/xhp0000404

Boundy-Singer, Z. M., Ziemba, C. M., & Goris, R. L. T. (2023). Confidence reflects a noisy decision reliability estimate. *Nature Human Behaviour, 7*(1). https://doi.org/10.1038/s41562-022-01464-x

Braun, A., Urai, A. E., & Donner, T. H. (2018). Adaptive history biases result from confidence-weighted accumulation of past choices. *Journal of Neuroscience, 38*(10), 2418–2429. https://www.jneurosci.org/content/38/10/2418.abstract.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7*(1), 26–48. https://doi.org/10.3758/BF03210724

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

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review, 124*(1), 91–114. https://doi.org/10.1037/rev0000045

García-Pérez, M. A.. (1998). Forced-choice staircases with fixed step sizes: Asymptotic and small-sample properties. *Vision Research, 38*(12), 1861–1881. https://doi.org/10.1016/S0042-6989(97)00340-4

de Gardelle, V., & Mamassian, P. (2015). Weighting mean and Variability during Confidence Judgments. *PLoS One1, 10*(3), Article e0120870. https://doi.org/10.1371/journal.pone.0120870

Haddara, N., & Rahnev, D. (2022). The Impact of Feedback on Perceptual Decision-making and Metacognition: Reduction in Bias but No Change in Sensitivity. *Psychological Science, 33*(2), 259–275. https://doi.org/10.1177/09567976211032887.

Hangya, B., Sanders, J. I., & Kepecs, A. (2016). A Mathematical Framework for Statistical Decision Confidence. *Neural Computation, 28*(9), 1840–1858. https://doi.org/10.1162/NECO_a_00864

Hellmann, S., Zehetleitner, M., & Rausch, M. (2023). Simultaneous modeling of choice, confidence, and response time in visual perception. *Psychological Review, 130*(6), 1521–1543. https://doi.org/10.1037/rev0000411

Hellmann, S., Zehetleitner, M., & Rausch, M. (2024). Confidence is Influenced by evidence Accumulation Time in Dynamical Decision Models. *Computational Brain & Behavior, 7*(3), 287–313. https://doi.org/10.1007/s42113-024-00205-9

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1594), 1322–1337. https://doi.org/10.1098/rstb.2012.0037

Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature, 455*(7210), 227–231. https://doi.org/10.1038/nature07200

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice Certainty is Informed by both evidence and Decision Time. *Neuron, 84*(6), 1329–1342. https://doi.org/10.1016/j.neuron.2014.12.015

Lau, H. C., & Passingham, R. E. (2006). Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences, 103*(49), 18763–18768. https://doi.org/10.1073/pnas.0607716103

Li, Q., Joo, S. J., Yeatman, J. D., & Reinecke, K. (2020). Controlling for Participants' Viewing Distance in Large-Scale, Psychophysical Online Experiments using a Virtual Chinrest. *Scientific Reports, 10*(1). https://doi.org/10.1038/s41598-019-57204-1

Maniscalco, B., & Lau, H. (2016). The signal processing architecture underlying subjective reports of sensory awareness. *Neuroscience of Consciousness, 2016*(1), Article niw002. https://doi.org/10.1093/nc/niw002

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology, 78*, 99–147. https://doi.org/10.1016/j.cogpsych.2015.01.002

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience, 19*(3), 366–374. https://doi.org/10.1038/nn.4240

Rafiei, F., Shekhar, M., & Rahnev, D. (2024). The neural network RTNet exhibits the signatures of human perceptual decision-making. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-024-01914-8

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

Rausch, M., Hellmann, S., & Zehetleitner, M. (2018). Confidence in masked orientation judgments is informed by both evidence and visibility. *Attention, Perception, & Psychophysics, 80*(1), 134–154. https://doi.org/10.3758/s13414-017-1431-5

Rausch, M., Hellmann, S., & Zehetleitner, M. (2021). Modelling visibility judgments using models of decision confidence. *Attention, Perception, & Psychophysics, 83*(8), 3311–3336. https://doi.org/10.3758/s13414-021-02284-3

Rausch, M., & Zehetleitner, M. (2019). The folded X-pattern is not necessarily a statistical signature of decision confidence. *PLOS Computational Biology, 15*(10), Article e1007456. https://doi.org/10.1371/journal.pcbi.1007456

Rolls, E. T., Grabenhorst, F., & Deco, G. (2010). Decision-making, Errors, and confidence in the Brain. *Journal of Neurophysiology, 104*(5), 2359–2374. https://doi.org/10.1152/jn.00571.2010

Sanders, J. I., Hangya, B., & Kepecs, A. (2016). Signatures of a Statistical Computation in the Human sense of confidence. *Neuron, 90*(3), 499–506. https://doi.org/10.1016/j.neuron.2016.03.025

Shekhar, M., & Rahnev, D. (2021a). Sources of Metacognitive Inefficiency. *Trends in Cognitive Sciences, 25*(1), 12–23. https://doi.org/10.1016/j.tics.2020.10.007

Shekhar, M., & Rahnev, D. (2021b). The Nature of Metacognitive Inefficiency in Perceptual Decision making. *Psychological Review, 128*(1), 45–70. https://doi.org/10.1037/rev0000249

Shekhar, M., & Rahnev, D. (2024). How do humans give confidence? a comprehensive comparison of process models of perceptual metacognition. *Journal of Experimental Psychology: General, 153*(3), 656–688. https://doi.org/10.1037/xge0001524

Spence, M. L., Dux, P. E., & Arnold, D. H. (2016). Computations underlying confidence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance, 42*(5), 671. https://psycnet.apa.org/record/2015-53185-001.

Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications, 8*(1), 14637. https://doi.org/10.1038/ncomms14637

Voodla, A., Uusberg, A., & Desender, K. (2025). Metacognitive confidence and affect – two sides of the same coin? *Cognition and Emotion*, 1–18. https://doi.org/10.1080/02699931.2025.2451795

Xue, K., Shekhar, M., & Rahnev, D. (2024a). *A novel behavioral paradigm reveals the nature of confidence computation in multi-alternative perceptual decision making*. https://www.researchsquare.com/article/rs-5510856/latest.

Xue, K., Fung, H., & Rahnev, D. (2025). *Stimulus reliability but not boundary distance manipulations violate the folded-X pattern of confidence*. *PsyArXiv*. https://osf.io/preprints/psyarxiv/865fs_v2.

Xue, K., Shekhar, M., & Rahnev, D. (2024b). Challenging the Bayesian confidence hypothesis in perceptual decision-making. *Proceedings of the National Academy of Sciences, 121*(48), Article e2410487121. https://doi.org/10.1073/pnas.2410487121

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1594), 1310–1321. https://doi.org/10.1098/rstb.2011.0416

Zylberberg, A., Roelfsema, P. R., & Sigman, M. (2014). Variance misperception explains illusions of confidence in simple perceptual decisions. *Consciousness and Cognition, 27*, 246–253. https://doi.org/10.1016/j.concog.2014.05.012