Improvement of association between confidence and accuracy after integration of separate evidence over time

Zahra Azizi[1], Sajjad Zabbah[2], Azra Jahanitabesh[3], Reza Ebrahimpour[2,4*]

[1] Institute for Cognitive Science Studies, Department of Cognitive Modeling, Tehran, Iran.

[2] Institute for Research in Fundamental Sciences, School of Cognitive Sciences, Tehran, Iran.

[3] Department of Psychology, University of California, Davis, California, United States.

[4] Department of Artificial Intelligence, Faculty of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.

*** Correspondence**:

Reza Ebrahimpour

ebrahimpour@ipm.ir

Keywords: Confidence, metacognition, discrete pieces of evidence, perceptual decision-making, pupillometry, ERP.

Abstract

When making decisions in real-life, we may receive discrete evidence during a time period. Although participants can integrate information from separate cues to improve their accuracy, it is still debatable how confidence changes after receiving discrete information. Nevertheless, based on the strong positive relationship between accuracy and confidence, we predicted that similar to what is observed in accuracy, confidence would improve following the integration of separate pieces of information. We used a Random-dot-motion discrimination task in which one or two brief stimuli (i.e., pulse[s]) were presented, and participants had to indicate the predominant direction of dot motions by saccadic eye movement. Two pulses intervals (up to 1s) were randomly selected, where color-coded targets facilitated indicating confidence simultaneously. Using behavioral data, computational models, pupillometry, and EEG methodology, our data revealed that compared to single-pulse trials, in double-pulse trials, participants improve their confidence resolution rather than reporting higher confidence and that the EEG signal (centro-parietal-positivity) and pupil dynamics appear to follow confidence levels but only in double-pulse trials. Overall, our findings suggest that evidence integration helps to improve the associations between confidence and accuracy.

1 Introduction

To make everyday decisions, both humans and animals can integrate multiple separate pieces of information they receive from their environment. For example, imagine that a large bus is passing the street between you and a car moving at a distance. In this situation, you can easily decide whether the distant car is moving toward or away from you by integrating the separate information that can be seen through the bus windows. In this scenario, the more information you collect, the better your inference of the direction of the distant vehicle will be. Indeed, research has shown that the accuracy

of decisions can be significantly improved by integrating information from separate cues (Kiani, Churchland, & Shadlen, 2013; Kira, Yang, & Shadlen, 2015; tickle, Tsetsos, Speekenbrink, & Summerfield, 2020; Tohidi-Moghaddam, Zabbah, Olianezhad, & Ebrahimpour, 2019; Waskom & Kiani, 2018).

In addition to selecting a choice among some alternatives, our decisions are accompanied by confidence —a feeling that reflects the likelihood that the decision is correct (Kiani, Corthell, & Shadlen, 2014). For example, imagine that the scene in the previous scenario also includes foggy weather. In this case, low visibility may reduce the confidence of your judgments. This diminished confidence per se may lead to change your mind (Fleming, Putten, & Daw, 2018; Resulaj, Kiani, Wolpert, & Shadlen, 2009), impact your behavioral adjustments, and affect how quickly and accurately you make your consecutive decisions (Meyniel, Sigman, & Mainen, 2015; van den Berg, Zylberberg, Kiani, Shadlen, & Wolpert, 2016). Due to the potential effects of confidence on decision-making, in the last few years, considerable progress has been made to understand of the behavioral (Kiani et al., 2014; Zylberberg, Barttfeld, & Sigman, 2012; Zylberberg, Fetsch, & Shadlen, 2016) and the neuronal (Baranski et al., 2017; Gherman & Philiastides, 2015; Kiani & Shadlen, 2009) properties of confidence and its relation with perceptual decision-making. However, it is still unclear how confidence is established if time intervals separate the provided information.

It has been shown that when we need to decide based on the separate pieces of information, the accuracy of the decision is determined by integrating the information of pieces and exceeding expectations predicted by models (Kiani et al., 2013). Furthermore, performance did not depend on the interval between the cues (Kiani et al., 2013; Waskom & Kiani, 2018), and that the second cue had the larger leverage on decisions (Kiani et al., 2013; Tohidi-Moghaddam, Zabbah, Olianezhad, & Ebrahimpour, 2019). Accordingly, one may suggest that confidence would follow the same characteristics as accuracy: increased considerably after receiving separate pieces of information, not being affected by the time interval between the cues and be leveraged larger by the most recent pack of information.

Nevertheless, a large body of evidence (e.g., [Herce Castañón et al., 2019; Zylberberg et al., 2016]) determines that human observers do not consistently report their confidence with a similar pattern of their accuracy. From this standpoint, noise can be considered as the key parameter to clarify variations in confidence (Kiani et al., 2014; Zylberberg et al., 2012). For instance, an underestimation of sensory noise in decisions would lead to over and/or under-confidence (De Gardelle & Mamassian, 2015; Herce Castañón et al., 2019; Zylberberg, Roelfsema, & Sigman, 2014). Moreover, confidence ratings may not be originated solely from the available sensory evidence (Rahnev & Denison, 2018; Zylberberg et al., 2016). For example, the observers may integrate secondary evidence into their confidence rating, which may not have been used when making the decision—utilizing secondary evidence allows the observers to change their mind after the initiating of a response (Atiya et al., 2020; Resulaj et al., 2009).

To test the hypothetical relation between accuracy and confidence, in binary decisions, signal detection theory (SDT) can provide a method to characterize the observers' reliability in their confidence ratings by introducing metacognitive sensitivity and efficiency (Figure 1B; [Fleming, 2017; Maniscalco & Lau, 2012, 2014]). In fact, for many years, SDT has been used as a simple yet powerful methodology to distinguish between an observer's ability to categorize the stimulus and the behavioral response (Green & Swets, 1966) and to determine confidence resolution (Balsdon, Wyart, & Mamassian, 2020; Maniscalco & Lau, 2012, 2014).

This is a provisional file, not the final typeset article

Additionally, levels of confidence can be traced by behavioral, neural, and pupillometry signatures. For example, higher confidence is accompanied by faster and more accurate decisions (Kiani et al., 2014; van den Berg et al., 2016; Zylberberg et al., 2016). Moreover, motion stimulus fluctuations influence both the accuracy (Kiani et al., 2008; Resulaj et al., 2009) and confidence (Van Den Berg et al., 2016; Zylberberg et al., 2012) and inform us about the parts of the stimulus that bear more intensely on the choice and confidence (Kiani et al., 2013, 2008; Nienborg & Cumming, 2009). Furthermore, perceptual decision-making literature has introduced an EEG potential characterized by a centro-parietal positivity (CPP) as a neural correlate of sensory evidence accumulation (Kelly & O'Connell, 2013; O'connell, Dockree, & Kelly, 2012) and confidence (Boldt, Schiffer, Waszak, & Yeung, 2019; Herding, Ludwig, von Lautz, Spitzer, & Blankenburg, 2019; Tagliabue et al., 2019; Vafaei Shooshtari et al., 2019; Zizlsperger, Sauvigny, Händel, & Haarmeier, 2014). In particular, despite the differences between correct and incorrect decisions, CPP presents regardless of whether the decision is correct and incorrect (O'connell et al., 2012; Steinemann, O'Connell, & Kelly, 2018) and can reflect not only external evidence but also an internal decision quantity such as decision confidence. Along with CPP, levels of confidence can be traced by monitoring the pupil. The literature has suggested strong links between pupil response and both the decision (Murphy, Boonstra, & Nieuwenhuis, 2016) and confidence (Allen et al., 2016; Lempert, Chen, & Fleming, 2015; Urai, Braun, & Donner, 2017) via pupil-linked dynamics of the noradrenergic system (Laeng, Sirois, & Gredebäck, 2012). For example, pupillometry has provided some evidence that shows a partial dissociation between choice and confidence in decision-making (Balsdon et al., 2020). Therefore, considering the potential of response-time, CPP, and pupillometry signatures in capturing the distinction between accuracy and confidence, they can be considered informative paradigms to explore the confidence-accuracy association in decision-making.

To bridge the existing gap in confidence and decision-making when information integrated from temporally separated stimulus compared to when only one stimulus interval is provided, we implemented two experiments to explore, as the primary goal, whether, after receiving two separate pieces of evidence, confidence reflects decision accuracy. In addition, as the secondary goals, we examined whether confidence (1) depends on the time interval between the pieces, (2) is leveraged by the second piece, and (3) predicts implicit markers—response-time, CPP and pupillometry after receiving two separate pieces of evidence. To the best of our knowledge, how evidence accumulation processes improve the accuracy and confidence association using the combination of behavioral, neural, and pupillometry signatures was not examined before.

To clarify confidence, we required observers to make a two-alternative decision after viewing either one motion pulse (single-pulse) or two motion pulses (double-pulse) separated by four various temporal gaps (similar to [Kiani et al., 2013; Tohidi-Moghaddam et al., 2019]). Considering four temporal gaps allowed us to examine whether the participants' confidence report would vary if they respond immediately or up to 1s after the first pulse. We performed linear and logistic regression models to measure the impact of stimulus characteristics on choice and confidence. Also, we applied a set of computational models based on SDT to assess whether confidence can be predicted only based on decision information. We proposed that to predict confidence we need both decision and confidence information. Accordingly, first, we fitted the standard signal detection model to participants' choices and confidence to directly estimate each participant's confidence, stimulus sensitivity, and metacognitive sensitivity. Second, to explore our proposition based on the task's nature, a perfect integrator model and two optimized models were fitted based on choices and confidence of single-pulse trials to predict confidence of double-pulse trials. Then, in the second experiment, we used EEG methodology to examine the relation between participants' brain activity and their confidence. We expected that a neural indicator of perceptual decision making (CPP) would

show amplitude changes between the two levels of confidence. Participants' pupil response was also monitored across both experiments to examine the relation between participants' pupil response and their confidence. Considering the evidence that the accuracy of the decision is enhanced by integrating the information over time (Kiani et al., 2013), we hypothesized that confidence would follow the same characteristics as accuracy such that participants would integrate information from pulses invariant to the temporal gap to improve their confidence. In addition, based on the previous evidence that participants obtained more information from a second pulse (Kiani et al., 2013; Tohidi-Moghaddam et al., 2019), we predicted that confidence would be leveraged larger by more recent information. Furthermore, as the effectiveness of capturing the distinction between choice and confidence has been shown using implicit markers (e.g., for response-time, see Vafaei Shooshtari et al., 2019); for CPP, see Steinemann et al., 2018; for pupillometry, see Balsdon et al., 2020), we expected that the behavioral, neural, and pupillometry markers of confidence to be distinguishable on both single-pulse and double-pulse trials.

2 Materials and Methods

2.1 Participants

Consistent with methodological considerations in previous studies, a total of 19 observers participated in two separate studies. Six participants (two males; age range [27–38], M = 32.25, SD = 4.5) attended in our behavioral study —Experiment 1— and 13 participants (three males; age range [21–39], M = 31.41; SD = 5.56) took part in our EEG study —Experiment 2. All participants had a normal or corrected-to-normal vision, and they had no history of psychiatric and neurological disorders. Previous studies with similar paradigms in which a large number of trials were presented to a small number of participants (e.g., five participants in [Kiani et al., 2013]; six participants in [Kiani et al., 2014]; four participants in [van den Berg et al., 2016] and six participants in [Stine, Zylberberg, Ditterich, & Shadlen, 2020], assume that with extensive training, all participants would reach an acceptable level of performance. Accordingly, all participants received extensive training sessions on the Random-dot-motion discrimination task prior to data collection. To ensure that participants understood and followed the instruction correctly, they were asked to describe how they would complete the experiment and report their accuracy and confidence prior to experiments. In Experiment 1, one participant was excluded due to the difficulty in reporting decision and confidence simultaneously, and another participant withdrew from the experiment shortly after the start of the experiment. In addition, one participant was excluded from Experiment 2 because of the excessive noise in their EEG electrodes crucial to the analysis.

2.2 Stimuli

We explored confidence formation in a discrete environment with a random-dot-motion (RDM) discrimination task. Participants had to indicate the predominant motion direction of a cloud of moving dots (left or right) presented within a 5° circular aperture at the center of the screen. The dot density was 16.7 dots/degree^2/s and the displacement of the coherently moving dots produced an apparent speed of 6 deg/s. The RDM movies were generated by three interleaved sets of dots presented on consecutive video frames. Three video frames later, each dot was redrawn at a location consistent with the direction of motion or a random location within the stimulus space. More details can be found in previous studies (e.g. [Roitman & Shadlen, 2002]). The experiment code was programmed in MATLAB 2016a (The Mathworks Inc., USA) using PsychToolbox (Brainard & Vision, 1997; Kleiner, Brainard, & Pelli, 2007).

This is a provisional file, not the final typeset article

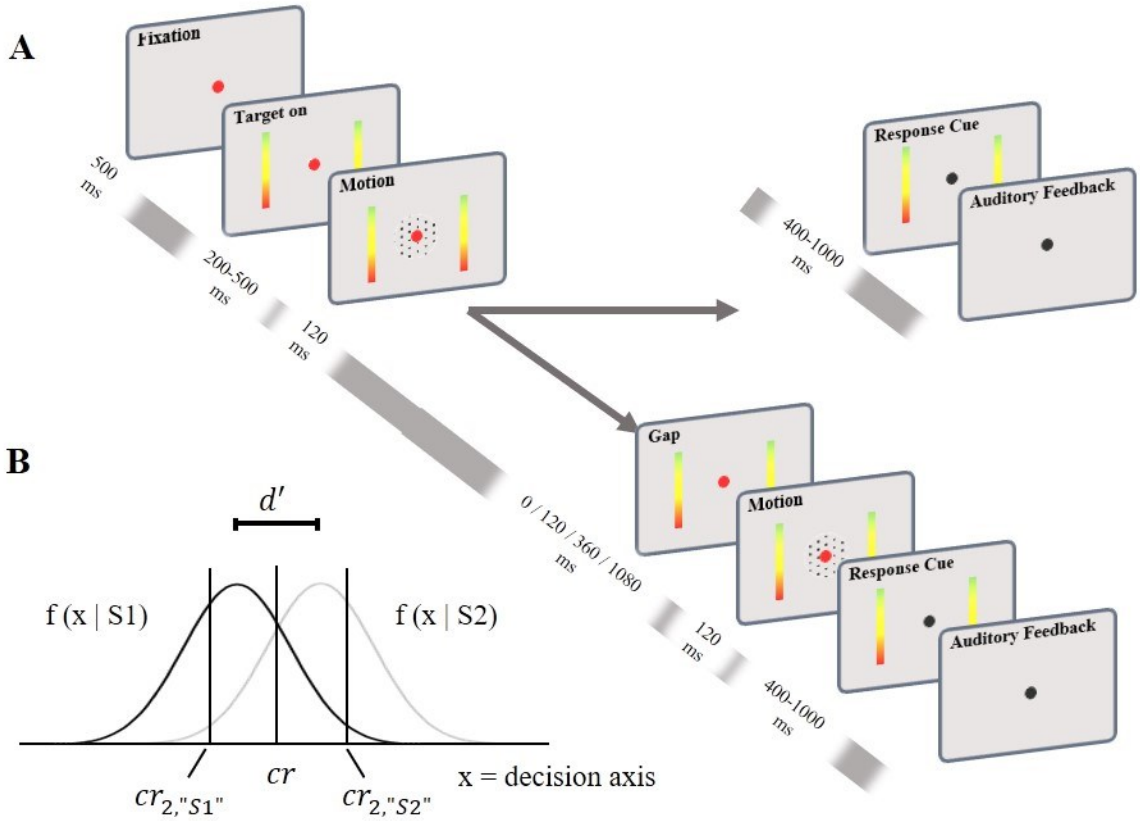


Figure 1. Task paradigm and Signal Detection Theory. **(A)** Participants had to indicate the predominant direction of motion of moving dots (left or right) by saccadic eye movement to one of the targets after receiving one or two pulse(s) of 120ms stimulus. The intervals between two pulses were selected randomly from 0 to 1080 ms and the direction of both pulses were the same. Color-coded targets enabled participants indicating their confidence simultaneously. **(B)** On each trial, a stimulus generates an internal response x within an observer, who must use x to decide whether the stimulus is S_1 or S_2, x is drawn from a normal distribution. The distance between these distributions is d', which measures the observer's ability to discriminate S_1 from S_2. The observer also rates decision confidence on a scale of high and low by comparing x to the additional response specific confidence criteria (cr_2 for each option). For details, see Supplementary Appendix 2 and refs (Fleming, 2017; Maniscalco & Lau, 2012, 2014).

2.3 Experimental Tasks

Participants performed the RDM task in blocks of 200 trials. Each trial started with participants fixating a small red point (diameter 0.3°) at the screen center. After 500 ms, two choice-targets appeared to the left and right of the fixation point (10° eccentricity; Figure 1A). Each target was shaped as a gradient rectangle (9° length and 0.5° width). After a variable duration of 200 - 500 ms (truncated exponential distribution), the RDM was presented. Participants had to indicate their choice after receiving one or two pulses of 120ms of motion pulses. The interpulse interval of double-pulse trials was selected randomly from 0, 120, 360, and 1080ms. On single-pulse trials, motion coherence was randomly selected from these six values: 0%, 3.2%, 6.4%, 12.8%, 25.6%, and 51.2%, whereas,

on double-pulse trials, motion coherence of each pulse was randomly chosen from three values: 3.2%, 6.4%, and 12.8%. Both pulses had the same net direction of motion and participants were aware of it. In total, there were six single-pulse and 9 × 4 double-pulse trial types. After the offset of one or two motion pulse(s), a 400 to 1000 ms delay period (truncated exponential) was imposed before the Go signal appeared on the screen. In each trial, participants were required to indicate their response by directing the gaze to one of the targets, the upper extreme of targets representing full decision confidence and the lower extreme representing guessing (Figure 1A). We constructed a list of all possible conditions of motion coherences and gaps, then shuffled the listed conditions and assigned them randomly to the trials in each block to provide the approximate balance within the trials. Participants were instructed to achieve high performance. Distinctive auditory feedback (Beep Tones) was provided for correct and incorrect responses. The type of feedback of 0% coherence trials was selected randomly by a uniform distribution. In Experiment 1, each participant performed the task across multiple blocks on different days (12-20 blocks). Experiment 2 contained the same paradigm as Experiment 1. All variables of stimulus remained constant except, in Experiment 2, the EEG data were also recorded. In Experiment 2, each participant completed a session of 4-5 blocks.

2.4 EEG Recording and pre-processing

We used a 32-channel amplifier for the EEG signal recording (eWave, produced by ScienceBeam, http://www.sciencebeam.com/) which provided 1K sample/s of time resolution. EEG was recorded at 31 scalp sites (Fp1, Fp2, AF3, AF4, C3, C4, P3, P4, O1, O2, F7, F8, T7, T8, P7, P8, FPz, Fz, Cz, Pz, Oz, POz, FC1, FC2, CP1, CP2, FC5, FC6, CP5, CP6). The EEG signals were referenced to the right mastoid. The recorded data were taken to Matlab (The Mathworks Inc., USA) and pre-processed as follows. The signals were filtered using a band-pass filter from 0.1 Hz to 40 Hz (Zizlsperger et al., 2014) for removing high frequency and independent cognitive noises. Then, all trials were inspected, and those containing Electromyography (EMG) or other artifacts were identified and manually removed. The second artifact rejection step included independent components analysis (ICA) using the EEGLAB toolbox (Delorme & Makeig, 2004). To select the removable ICA component, the ADJUST plugin (Mognon, Jovicich, Bruzzone, & Buiatti, 2011) was used.

2.5 Pupillometry Recording and pre-processing

The eye data were collected using an EyeLink 1000 infrared eye-tracker system (SR Research Ltd. Ontario, Canada). This device allowed a 1000-Hz sampling rate and was controlled by a dedicated host PC. The system was calibrated and validated before each block by presenting nine targets at the display monitor's center, edges, and corners of the. The left eye's data was recorded and passed to the host PC via an Ethernet link during data collection.

Missing data and blinks, as detected by the EyeLink software, were padded and interpolated. Additional blinks were spotted using peak detection on the pupil signal's velocity and then linearly interpolated (Mathôt, 2013).

2.6 Experimental procedure

Participants were given a consent form in which the experiment was described in general terms. After providing written informed consent, participants completed the tasks in a semidark, sound-attenuating room to minimize distraction in both experiments. All instructions were presented and stimuli were displayed on a CRT monitor (17 inches; PF790; refresh rate, 75 Hz; screen resolution, 800 × 600). A head and chin rest confirmed that the distance between the participants' eyes and the

This is a provisional file, not the final typeset article

monitor's screen was 57 cm throughout the experiments. Participants were presented with demographic questions followed by training sessions and main sessions, respectively. The experimental protocol was approved by the ethics committee of the Iran University of Medical Sciences.

2.7 Data Analysis

Data analysis was performed using Matlab 2019a (The MathWorks Inc., United States).

2.7.1 Behavioral analyses

Except where otherwise specified, we reported behavioral data of the first experiment; however, all the analyses were repeated for the EEG experiment and if the results were inconsistent, it has been admitted (Experiment 2 results were reported in Supplementary Figures 1, 2, 3, 4, 5, and 9).

We used linear and logistic regression models to evaluate the influence of the stimulus characteristics on choice, and confidence and to evaluate the null hypothesis that a single coefficient of a linear regression model is zero we ran *t*-tests. For logistic regression models, we used maximum likelihood under a binomial error model (i.e., a GLM) to evaluate the null hypothesis that one or more of the regression coefficients were equal to zero. $P_{correct}$ was the probability of correct response and $Logit[P_{correct}]$ indicated $\log \frac{P_{correct}}{1 - P_{correct}}$. Also, P_{high} was the probability of high confidence, $Logit[P_{high}]$ indicated $\log \frac{P_{high}}{1 - P_{high}}$ and β_i denoted fitted coefficients

For single-pulse trials, to evaluate whether choice accuracy improved with coherence, we used the following logistic regression:

$$\boldsymbol{Logit[P_{correct}] = \beta_0 + \beta_1 C}, \quad (1)$$

where C was motion strengths of the stimulus. The null hypotheses were that coherence would have no effect on choice accuracy ($H_0: \beta_1 = 0$). For single-pulse trials, we also used the following function to determine whether confidence improved consistently with coherence and accurate choices:

$$\boldsymbol{Conf = \beta_0 + \beta_1 C + \beta_2 A}, \quad (2)$$

where C was motion strengths of the pulse and A was the accuracy of the response (0 or 1 for incorrect and correct). The null hypotheses were that coherence and accuracy would have no effect on confidence ($H_0: \beta_{1-2} = 0$).

We also used linear regression to evaluate the effect of interpulse interval on confidence in double-pulse trials:

$$\boldsymbol{Conf = \beta_0 + \beta_1 C_1 + \beta_2 C_2 + \beta_3 T + \beta_4 C_1 T + \beta_5 C_2 T}, \quad (3)$$

where C_1 and C_2 were motion strengths of each pulse, and T was the interpulse time interval. For double-pulse trials with equal pulse strength ($C_1 = C_2$), the redundant regression terms (β_2, β_4) were omitted. The null hypothesis was that the interpulse interval would not affect reported confidence ($H_0: \beta_{3-5} = 0$). The similar equation was used to assess the relation of accuracy and the interpulse

interval. To evaluate the impact of pulse sequence on confidence, the following regression model was fitted:

$$\boldsymbol{Conf} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}[\boldsymbol{C_1} + \boldsymbol{C_2}] + \boldsymbol{\beta_2}[\boldsymbol{C_2} - \boldsymbol{C_1}], \quad (4)$$

where C_1 and C_2 were corresponding motion strengths of each pulse. β_2 indicated how the confidence varied from trials in which $C_1 > C_2$ to trials with a reversed sequence of motion pulses $C_1 < C_2$. The null hypothesis was that the sequence of motion pulses did not influence the confidence (H_0: $\beta_2 = 0$).

In addition, to investigate the variation of confidence in double-pulse trials compared to single-pulse trials, we subtracted confidence of double-pulse trials from corresponding confidence in single-pulse trials for each participant. For example, the confidence of a sequence of 3.2%, 6.4% motion strength trial, subtract separately once from 3.2% and once from 6.4% corresponding confidence in single-pulse trials. The process repeated for the data of correct and incorrect trials too. Moreover, the same method was used to compare the accuracy of double-pulse and single-pulse trials. To assess the effect of choice accuracy on the variation of confidence in double-pulse and single-pulse trials, we fitted the following:

$$\boldsymbol{S_{Conf}} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 A}, \quad (5)$$

where the S_{Conf} was the subtraction of confidence in double-pulse trials from corresponding single-pulse trials and A was the accuracy of the response (0 or 1 for incorrect and correct). The null hypothesis was the choice accuracy did not affect the variation of S_{Conf} (H_0: $\beta_1 = 0$).

Moreover, we compared accuracy and confidence between single-pulse and double-pulse trials directly by fitted linear regression models using indicator variables for type of trials (1 for single-pulse and 0 for double-pulse). Also, as a sensitivity analysis we randomly chose the same numbers of trials from each coherence of single and double pulse trials of each participants' data where the reported confidence was between 40 to 60. Then, we compared the probability of correct choice — accuracy—of the selected data. This range was chosen to make sure there were enough numbers of trials for all conditions.

2.7.1.1 Response-time analysis

In the current study, response-time was referred to the time interval between the response cue onset and the participant's response. To represent the trace of high and low confidence from behavioral data, we have selected an equal number of single/double-pulse trials from each participant's data for 100 times. Then individual response-time were rank-ordered and binned into four quintiles. Next, we calculated he accuracy of high and low confidence trials in each bin. We expected to see a noteworthy difference between the accuracy of each bin grouped by levels of confidence. We only included motion strength of 3.2%, 6.4%, 12.8% of single-pulse trials (similar to coherence used in double-pulse trials) to control the impact of coherence on response-time. We used a linear model with separate intercepts for each participant on response-time to check whether the difference of response-time in high and low confidence can be traced in both single-pulse and double-pulse trials:

$$\boldsymbol{Conf} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 R} + \sum_{\boldsymbol{i=1}}^{\boldsymbol{N_{pt}}} \boldsymbol{\beta}_{\mathrm{i}} I_{pt}, \quad (6)$$

This is a provisional file, not the final typeset article

where R was the response-time of each trial and I_{pt} is an indicator variable that takes a value of 1 if the trial was completed by participant pt and 0 otherwise. The null hypothesis was that confidence did not depend on the response-time ($H_0: \beta_1 = 0$). Moreover, as a sensitivity analysis to evaluate the relation of delay-time imposed before the cue onset and response-time, we used a linear regression model to see if the delay times would predict the response-time. The null hypothesis was that the response-time did not rest on the delay-time.

2.7.2 Motion energy analysis

Random dot stimulus is stochastic, so the sensory evidence fluctuated within and across trials but around the nominal coherence level. To examine the fluctuations in motion during each trial, we filtered the sequence of random dots by using two pairs of quadrature spatiotemporal filters, as specified in previous studies (Adelson & Bergen, 1985; Kiani, Hanks, & Shadlen, 2008; Zylberberg et al., 2012). Since we aimed to understand the temporal course of confidence, we summed the energies across trials for each pulse in single/double-pulse trials.

We used a linear regression to test whether the confidence was more influenced by the second pulse's motion energy than that of the first pulse in double-pulse trials. We tested double-pulse trials with unequal motion strength using the following linear regression model:

$$\boldsymbol{Conf = \beta_0 + \beta_1 C_1 + \beta_2 C_2 + \beta_3 (M_1 + M_2) + \beta_4 M_2}, \quad (7)$$

where M_1 and M_2 were the motion energy of each pulse and C_1, C_2 were corresponding motion strengths of each pulse. The null hypothesis was ($H_0: \beta_4 = 0$). For double-pulse trials with equal pulse strength ($C_1 = C_2$), the redundant regression terms (β_2) were omitted. We used only double-pulse trials with nonzero interpulse interval in this analysis to avoid overlap in the windows for the first and second pulse. To certify that our findings were not relevant to participant performance, we analyzed motion energy in the two confidence levels only for correct responses. To evaluate the relation of confidence and motion energy in single-pulse trials, we fitted a linear regression model as follows:

$$\boldsymbol{Conf = \beta_0 + \beta_1 C + \beta_2 M}, \quad (8)$$

where M was the motion energy of the presented motion stimulus, C was the pulse coherence and the null hypothesis was that confidence did not depend on the motion energy ($H_0: \beta_2 = 0$).

2.7.3 General computational modeling approach

We implemented a set of computational models based on signal detection theory to provide a mechanistic explanation of the experimental data. According to SDT, observers set a decision criterion (cr) to discriminate between two stimuli (e.g., labeled as S_1 and S_2). They also set criteria $cr_{2,"S1"}$ and $cr_{2,"S2"}$ to determine confidence ratings around the decision criterion cr (Figure 1B; for more details, see Supplementary Appendix 2). In order to fit confidence data to SDT, we categorized the confidence ratings as either high or low. Since the participants were told to choose the upper part of the bar as high confidence and the lower part as low confidence (ranged from 0 to 100), we considered reported confidence higher than midline as high confidence and lower than midline as low confidence respectively. However, in addition to the midline, we tested various binary level set methods for categorizing participants' high and low confidence ratings. First, the highest 55% and 45% of each participant's confidence reports were considered high confidence (similar to [Zylberberg, Wolpert, & Shadlen, 2018]). Then, the mean of each participant's confidence was

calculated separately, and the confidence ratings above the mean were considered high ratings. Using these methods did not significantly alter reported confidence categorization (see Supplementary Figure 8).

We computed stimulus sensitivity (d'), confidence criteria, and the measures of metacognitive ability ($Meta\text{-}d'$, $Meta\text{-}d'/d'$) of single-pulse and double-pulse trials. We used code provided by Maniscalco and Lau (Maniscalco & Lau, 2012) in which metacognitive sensitivity ($Meta\text{-}d'$) is computed by setting the d' value that would produce the observed confidence. In addition, $Meta\text{-}d'/d'$ was calculated by normalizing $Meta\text{-}d'$ by d' through division. To support the fact that our findings were not relevant to variation of coherence of single and double-pulse trials, we only included single-pulse trials with motion strength of 3.2%, 6.4%, 12.8%. In addition, we fitted the SDT model with trials simulated by a perfect integrator model (the model is described later). However, one difference between groups was that they might not be matched for the number of trials: the single-pulse included on average fewer trials for each coherence per participant compared to double-pulse trials. Previous research has suggested that the number of trials could bias measures of metacognitive ability (Fleming, 2017). Therefore, in a control analysis, we created 100 sets of trials randomly from the single/double-pulse trials and from trials simulated by the perfect integrator model. Each set contained the same number of trials for each participant. We then averaged the metacognitive scores obtained from these 100 sets and repeated the comparison procedure.

2.7.4 Perfect integrator Model

To estimate the expected confidence ($P_{e(high)}$) in double-pulses trials, we assumed that each trial's confidence was achieved based on evidence integrating from both pulses by using a perfect integrator model. In the perfect integrator model, the expected accuracy ($P_{e(correct)}$) for double-pulse trials computed as the following (Kiani et al., 2013):

$$\boldsymbol{P_{e(correct)} = 1 - \phi(0, e_1 + e_2, \sqrt{2})}, \quad (9)$$

where e_1 and e_2 were the pieces of evidence that underlie by P_1 and P_2 (the probabilities of the correct choices in corresponding single-pulse trials) and were computed as:

$$\boldsymbol{e_i = \phi^{-1}(P_i, 0, 1), i = 1, 2} \quad (10)$$

where ϕ^{-1} was inverse ϕ, which represented the cumulative Gaussian distribution (Kiani et al., 2013).

To predict the confidence of double-pulse trials by this model, cr and d' were computed based on integrated evidence (Eq.9, 10, for more details see Supplementary Appendix 2). Then, confidence on correct and error trials were calculated based on confidence performance from corresponding single-pulse trials (similar to Eq.9, 10). Accordingly, confidence (for both correct response or incorrect response) would be predicted by the perfect integrator model. Besides, the model parameters, including confidence criteria along with d', $Meta\text{-}d'$ and, $Meta\text{-}d'/d'$ were computed.

2.7.5 Optimized models

We considered two optimized models. The purpose of these two models was to address how improving stimulus or metacognitive sensitivity could influence confidence prediction. In the type 1 optimized model, we optimized the perfect integrator model by letting the difference of d' (the measures of task performance in SDT) from the perfect integrator and double-pulse trials minimized

This is a provisional file, not the final typeset article

for each participant using the fmincon function in the optimization toolbox for Matlab. In the type 2 optimized model, using the similar optimization technique, we optimized the $Meta\text{-}d'$ in the perfect integrator model by providing each participant's $Meta\text{-}d'$ computed from behavioral data. In other words, in type 1 optimized model, stimulus sensitivity was optimized, whereas in the type 2 optimized model, metacognitive sensitivity was optimized. Thus, we examined how improving stimulus or metacognitive sensitivity could influence confidence prediction.

2.7.6 Model evaluation

We evaluated the models qualitatively (i.e., parameter recovery exercises) and quantitatively (i.e., maximum likelihood estimation).

In the qualitative method, based on the calculated parameters of the model, the probability of choosing high confidence for all combinations of motion strength for each participant was calculated ([Maniscalco & Lau, 2014]; see Supplementary Appendix 2). We compared the expected high confidence predicted by models to the observed high confidence in double-pulse trials using a linear regression model, as follows:

$$\boldsymbol{P_{high} = \beta_1 P_{e(high)} + \beta_0}, \tag{11}$$

where $P_{e(high)}$ was the expected probability of high confidence while $\boldsymbol{P_{high}}$ was observed high confidence in double-pulse trials. We compared slope (β_1) against the 1:1 line in each model.

In addition, to compare models quantitatively, an equal number of trials from each participant's trials was selected randomly and then each model fitted to the selected data. This procedure was repeated for 100 times, then the computed MLEs of each model were averaged.

2.7.7 Confidence suboptimality

The optimal decision-making is disrupted by several sources of suboptimality (Balsdon et al., 2020). In SDT, an added noise, ξ_n, represents a potential loss of information between sensory decision information and metacognitive information, such as confidence rating. This noise has a Gaussian distribution with zero mean and standard deviation σ (Maniscalco & Lau, 2014). The parameter σ determines how much the metacognitive variable is noisier than the decision variable (Maniscalco & Lau, 2014).

$$\boldsymbol{\xi_n = N\,(0, \sigma))} \tag{12}$$

This noise is related to metacognitive efficiency —$Meta\text{-}d'/d'$ (Maniscalco & Lau, 2014). We simulated trials using the same parameter values resulting from the perfect integrator model, except for increased noise, to consider this confidence suboptimality.

2.7.8 EEG analysis

In current study, the EEG analysis focused on a neural marker of perceptual decision-making linked with stimulus preparation and stimulus processing. First, we epoched the EEG responses for averages over the entire trial for all coherence. To characterize the time-course of confidence in the ERP data, time courses were baseline-corrected throughout 100 ms before stimulus onset and split by high and low confidence. Next, the component we focused on was the centro-parietal positivity (CPP) which possibly is identical to the classic P300 component (Herding et al., 2019; Twomey, Murphy, Kelly, & O'connell, 2015). The CPP is associated with the sampling of available evidence in perceptual

decisions and confidence rating at the time period of 200-500 ms after stimulus onset (Herding et al., 2019; Rausch, Zehetleitner, Steinhauser, & Maier, 2020; Vafaei Shooshtari et al., 2019; Zizlsperger et al., 2014) or at the time of the response (Boldt et al., 2019). In current study, CPP amplitude was measured as the mean amplitude in a time-window ranging from 200 ms to 500 ms after stimulus onset in an electrode cluster containing the electrodes CP1, CP2, Cz, and Pz (Boldt et al., 2019; Herding et al., 2019; Rausch et al., 2020; Tagliabue et al., 2019; Twomey, Kelly, & O'Connell, 2016; Vafaei Shooshtari et al., 2019; Zizlsperger et al., 2014). We epoched the EEG responses aligned to the stimulus onset, from 200 ms pre-stimulus to 500 ms post-stimulus for each pulse. Then, these epochs were baselined to a window -100 ms to stimulus-locked to prevent differences in the visual response to the stimulus affecting the baseline. To avoid overlap in the windows for the first and second pulse, we used only double-pulse trials with 360 and 1080 ms interpulse interval in this analysis.

The ERP signals were examined for levels of confidence separately in double-pulse trials and single-pulse trials. To support the fact that our findings were not relevant to motion pulse strength in double-pulse trials, we tested double-pulse trials with equal motion strength pulses. To test if CPP amplitude could distinguish between high and low confidence trials, we performed an independent samples *t*-test comparing the average of the high and low confidence trials at each time sample on the EEG signal of the four CPP electrodes at a significance threshold corresponding to a p-value of 0.05. Correction for multiple comparisons was performed using a cluster-based method (Maris and Oostenveld, 2007). Each cluster was characterized by the sum of the t-statistics comparing the average high and low confidence CPPs. These summed t-statistics were then compared to clusters of t-statistics obtained from a Monte Carlo simulation with 3000 iterations. To make sure that our findings were not relevant to participants' performance, as a secondary check, we analyzed the CPP signals in two confidence levels only for correct trials.

To present a statistical comparison of the single-pulse and double-pulse trials and to examine whether CPP amplitude could predict confidence for both single-pulse and double-pulse trials, we averaged the time courses of electrodes CP1, CP2, Cz, and Pz and then performed a linear regression between the amplitude of this signal of each participant and their mean confidence:

$$\boldsymbol{S = \beta_0 + \beta_1 Conf + \beta_2 GConf}\,, \tag{13}$$

where S was the mean amplitude of time period of 200-500 ms after stimulus onset for each participant and G was an indicator variable that took a value of 1 if the trial was double- and 0 otherwise. The null hypothesis was that the type of trials did not affect the impact between confidence and CPP signal ($H_0: \beta_2 = 0$).

2.7.9 Pupillometry analysis

First, to characterize the time-course of confidence in the pupillometry data, time courses were baseline-corrected throughout 200 ms before stimulus onset and split by high and low confidence. Previous studies showed that pupil dilation response after choice and before feedback reflected decision uncertainty (Colizoli, De Gee, Urai, & Donner, 2018; Urai et al., 2017). In current study, as confidence is an uncertainty complement (Hebart, Schriever, Donner, & Haynes, 2014; Kepecs & Mainen, 2012), the method was implemented to study the relation of confidence and pupil dilation response. Accordingly, the baseline-corrected pupil signal throughout 200 ms before feedback was calculated as our single-trial measure of pupil response. We epoched trials and baselined each trial by subtracting the mean pupil diameter 50 ms before the response. We included all trials of both

This is a provisional file, not the final typeset article

experiments in the analyses reported in this paper. To control the impact of coherence on pupil response, we only included motion strength of 3.2%, 6.4%, 12.8% of single-pulse trials.

According to the temporal low-pass characteristics of the slow peripheral pupil apparatus (Hoeks & Ellenbroek, 1993) even in the absence of amplitude variations in the underlying neural responses, trial-to-trial variations in response-time can impact pupil responses (Urai et al., 2017). Therefore, we removed the trial-to-trial fluctuations in pupil responses via a linear regression (De Gee, Urai, & Donner, 2018; Urai et al., 2017) and used it for analyses reported in this study. Statistical inferences were performed using a cluster-based procedure to examine whether pupil response could distinguish between high and low confidence trials (Maris and Oostenveld, 2007). Significant clusters were found using a *t*-test at each time-point (at a statistical threshold of $p < 0.05$) and comparing the sum of the t-statistics in these clusters to those obtained over 3000 permutations. As a secondary check, we analyzed the pupil response signals in two confidence levels only for correct trials to support the fact that our findings were not relevant to participants' performance.

To present a statistical comparison of the single-pulse and double-pulse trials and to examine whether pupil response could predict confidence for both single-pulse and double-pulse trials, we performed a linear regression between the average of pupil responses for each participant and their mean confidence:

$$\boldsymbol{P = \beta_0 + \beta_1 Conf + \beta_2 GConf}, \tag{14}$$

where P was mean the average of pupil response for each participant and G was an indicator variable that took a value of 1 if the trial is double-pulse trials and 0 otherwise. The null hypothesis was that the type of trials did not affect the relationship between confidence and pupil response signal ($H_0: \beta_2 = 0$).

2.7.10 General statistical analysis

To test our hypotheses, a series of regression analyses were run after confirming the assumptions of linear and logistic regression were met. Effect sizes were reported and as suggested, here, we considered small ($f^2 = .02$), medium ($f^2 = .15$), and large ($f^2 = .35$) effect sizes (see [Cohen, 1970]) at the alpha level of 5%.

Moreover, we used repeated-measures *t*-tests. As suggested, we considered small (d = .2), medium (d = .5), and large (d = .8) effect sizes for this assessment (see [Cohen, 1970]) and the statistical significance for *t*-tests was set to 0.05.

Throughout the analysis, Bonferroni corrections were applied to p-values less than 0.05 when more than one statistical test was carried out per hypothesis (with this indicated by bonf*[number of tests corrected for] in subscript), whereas non-significant p-values were reported uncorrected.

For tests of pupil response signals and ERPs between two levels of confidence, statistical inferences were performed using a cluster-based procedure (Maris, E. & Oostenveld, R, 2007). All cluster-based permutations were done using the FieldTrip toolbox (Oostenveld et al., 2011) in MATLAB (at a statistical threshold of $p < 0.05$).

3 Results

We tested our predictions in two studies that applied the same paradigm (Figure 1A). The first study used behavioral measures and pupillometry analyses, whereas for the second experiment, we recorded EEG signals as well. Participants decided about the direction of the RDM motion based on brief motion pulses. The task design contained different conditions which allowed us to compare participants' behavior in (i) one or two brief pulse(s) of random dot motion, (ii) different coherence of motion stimulus, and (iii) four distinct interpulse intervals.

3.1 Behavioral results

We used the single-pulse trials to benchmark the effect of coherence on choice accuracy and confidence. As shown in Figure 2A, for single-pulse trials, participants were more confident for high coherence stimuli (Figure 2A; Eq.2; $\beta_1 = .06$, $p_{bonf*2} < .001$, 95% CI = [.04, .08], $f^2 = .23$), ranged from 43 for 3.2% to 96 for 51.2% coherence. Also, accuracy improved with motion strength reached from 56% for 3.2% to 99% for 51.2% (Figure 2B, black line; Eq.1; $\beta_1 = .10$, $p < .001$, 95% CI = [.08, .12], $f^2 = .36$). In general, they also had better performance whenever they had reported higher confidence comparing to lower confidence (Figure 2B, red and green; Eq.2; $\beta_2 = .12$, $p_{bonf*2} < .001$, 95% CI = [.08, 0.15], $f^2 = .23$) except in weak coherence (3.2%) that the difference of high and low confidence was not significant (Figure 2B, red and green; Eq.2; $\beta_2 = .01$, $p = .85$, 95% CI = [-.07, 0.08], $f^2 = .01$).

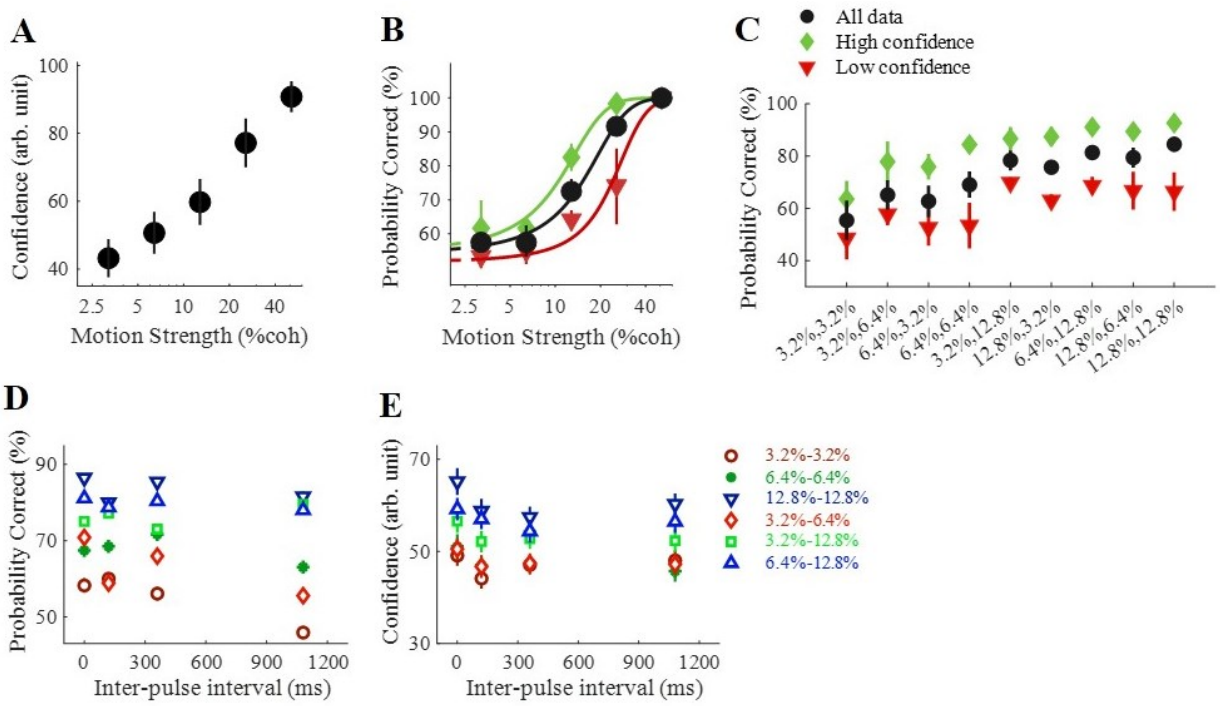


Figure 2. The interplay between confidence, accuracy, and coherence in single/double-pulse trials, and the interpulse interval in double-pulse trials. **(A)** Average confidence reported as a function of the motion coherence in single-pulse trials. **(B)** Accuracy in single-pulse trials and **(C)** double-pulse trials in all trials (black), split by high (green) and low (red) confidence decisions. **(D)** Choice accuracy for double-pulse trials grouping in all possible interpulse intervals. **(E)** Confidence of double-pulse trials was calculated by pooling data across all interpulse intervals. In (B) curves are model fits. In (D) and (E) each data point reports pooled data from indicated

This is a provisional file, not the final typeset article

sequence pulse and its reverse order (e.g., 12.8– 3.2% and 3.2 –12.8%). In all panels, data are represented as group mean ± SEM.

Moreover, in double-pulse trials, the accuracy improved with motion strength (Figure 2C, black dots) and participants were more accurate while reporting higher confidence (Figure 2C, green dots). Along with accuracy (Figure 2D; all *ps* > .1, [Kiani et al., 2013; Tohidi-Moghaddam et al., 2019], see Supplementary Figure 1A for Experiment 2 data), the confidence was largely unaffected by interpulse interval in double-pulse trials (Figure 2E; Eq.3; all *ps* > .1; Supplementary Table 2, results of individual participants; see Supplementary Figure 1B for Experiment 2 data). The two pulses separated by up to 1 s supported a level of confidence that was indistinguishable from a pair of pulses separated by no gap. To make sure that the trials with zero interpulse interval did not differ from trials with nonzero interpulse intervals, similar analyses were performed on data including trials with zero interpulse interval and just one of the nonzero interpulse intervals. Results revealed no significant difference (Eq.3, all *ps* > .1; Supplementary Table 3). Further, confidence was unaffected by the interpulse interval for double-pulse trials across each possible coherence sequences (Supplementary Table 4).

Direct comparison between participants' accuracy of single-pulse and double-pulse trials, along with previous studies (Kiani et al., 2013; Tohidi-Moghaddam et al., 2019), showed that participants' accuracy significantly differed ($\beta = .09$, $p < .001$, 95% CI = [.08, .11], $f^2 = .21$). However, in double-pulse trials participants were not more confident compared to single-pulse trials ($\beta = .02$, $p = .15$, 95% CI = [-.01, .04], $f^2 = .01$).

Furthermore, although the order of the pulses affected accuracy (Figure 3A, [Kiani et al., 2013; Tohidi-Moghaddam et al., 2019]), participants were not more confident in double-pulse trials with unequal pulse strength where the stronger motion appeared in a second order (Figure 3B; Eq. 4; $\beta_2 = .01$, $p = .51$, 95% CI = [.00, .02], $f^2 = .01$; see Supplementary Figure 2B for Experiment 2 data). Also, we explored whether if the first pulse would have less of an impact on confidence after a longer delay. According to the results, the order of the pulses failed to show significant effects on confidence on any of interpulse intervals (Supplementary Table 5).

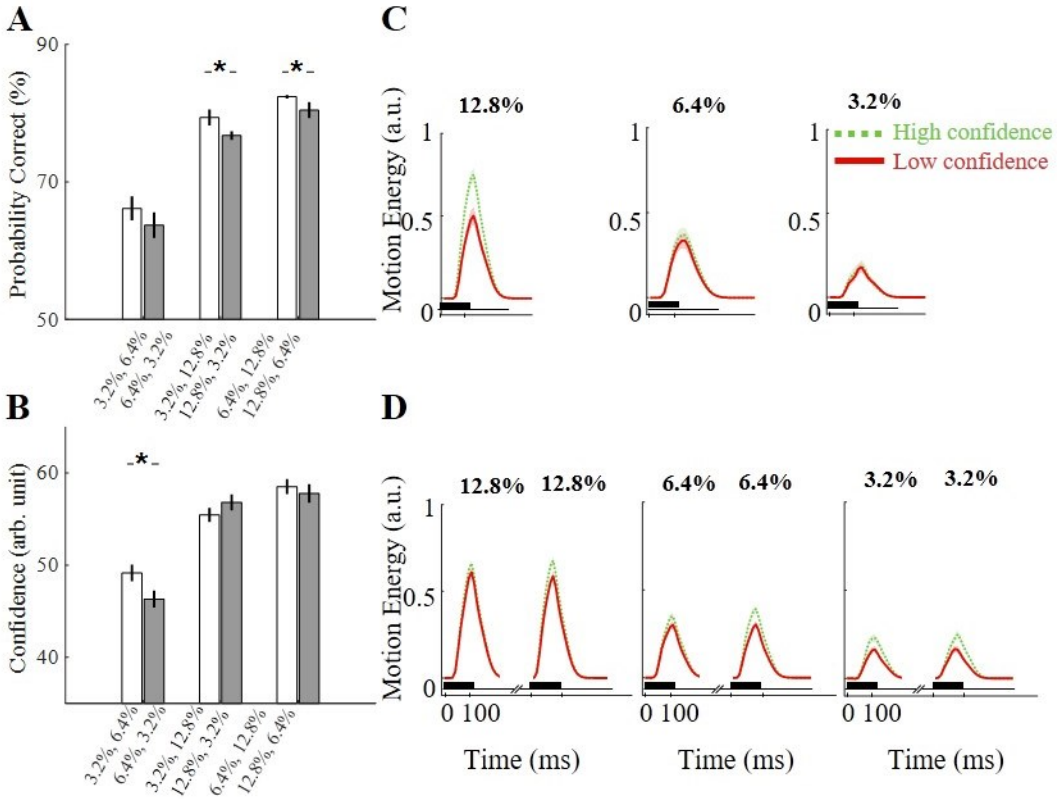


Figure 3. Choice confidence did not depend on the sequence of motion pulses. **(A)** The weak–strong pulse sequence contributed higher accuracy than the strong–weak sequence. Accuracy for double-pulse trials with unequal pulse strength was calculated by pooling data across all intertrial intervals. **(B)** The weak–strong pulse sequence did not contribute higher confidence than the strong–weak sequence. In all panels, data are represented as group mean ± SEM. (*p<0.05) **(C)** In single-pulse trials, low and high confidence cannot be determined by motion energy profiles in weaker pulses **(D)** The second pulse had slightly more impact on confidence. Data were pooled for all nonzero interpulse intervals while only correct trials with equal pulse strength are included. In (C) and (D), the shaded region around the mean indicates SEM. The black horizontal bars show the duration of the stimulus display. The units of motion energy are arbitrary and the same for all motion strengths.

3.2 Motion energy results

To yield a precise estimate of the decision-relevant sensory evidence accommodated in the stochastic stimuli, we employed motion energy filtering to the random dot motion stimuli. Figure 3D displays the average motion energy in double-pulse trials with equal pulse strength. A linear regression confirmed the influence of trial-to-trial fluctuations of motion energy on confidence (Eq.7; $\beta_2 = .10$, $p_{bonf*4} < .001$, 95% CI = [.04, .16], $f^2 = .19$). However, the difference of the motion energy profiles for high and low confidence responses was not larger for the second pulse than the first pulse with equal pulse strength (Eq.7; $\beta_3 = .11$, $p_{bonf*4} = .15$, 95% CI = [.07, .15], $f^2 = .13$), as well as, with unequal pulse strength (Eq.7; $\beta_4 = .10$, $p = .06$, 95% CI = [.06, .14], $f^2 = .08$). Consequently, motion

This is a provisional file, not the final typeset article

energy analysis could not provide independent confirmation of the asymmetric effect of pulses for confidence.

Additionally, in single-pulse trials, the motion energy profiles for high and low confidence with stronger pulse strength (12.8%, 6.4%) were significantly different (Figure 3C; Eq.8; $\beta_2 = .41$, $p_{bonf*2} < .001$, 95% CI = [.23, .59], $f^2 = .24$). However, the motion energy profiles in weak motion pulses were not significantly different (Figure 3D; Eq.8; β2 = .17, p = .44, 95% CI = [-.26, .60], f2 = .10). Accordingly, motion energy analysis suggests that when the pulses' motion strengths are weak, the participants' confidence would not depend on the motion energy.

3.3 The Interplay between confidence in single vs. double pulse(s) trials

To address accuracy and confidence variation in double-pulse from single-pulse trials, we subtracted $P_{correct}$ or confidence of each coherence (3.2%, 6.4% and 12.8%) in single-pulse trials from $P_{correct}$ or confidence of corresponding sequence in double-pulse trials. As we expected, $P_{correct}$ improved in all coherence combinations (Figure 4A). Additionally, considering all the trials, confidence increased when the other pulse was a strong pulse (12.8%) (Figure 4B and Figure 4C for correct trials), while decreased or did not change whenever the other pulse was a weak motion strength (3.2%, 6.4%). Interestingly, in error trials, confidence on double-pulse trials decreased compared to single-pulse trials for all conditions (Figure 4D). These results did not influence by the duration of interpulse intervals (Supplementary Figure 6A, B, C, D).

In other words, the participants reported lower confidence in double-pulse trials compared to single-pulse trials for incorrect choices but reported higher confidence for correct choices (Figure 4 Eq.5, $\beta_1 = .15$, $p < .001$, 95% CI = [.13, .17], $f^2 = .29$; Supplementary Figure 3 for Experiment 2; Supplementary Table 1, results of individual participants). This data is in line with the fact that good metacognitive sensitivity will provide higher confidence for correct responses, and lower for incorrect ones (Fleming, 2017; Maniscalco & Lau, 2012, 2014). Accordingly, we implemented computational models in order to study the relation of accuracy and confidence (confidence resolution) in single-pulse and double-pulse trials.

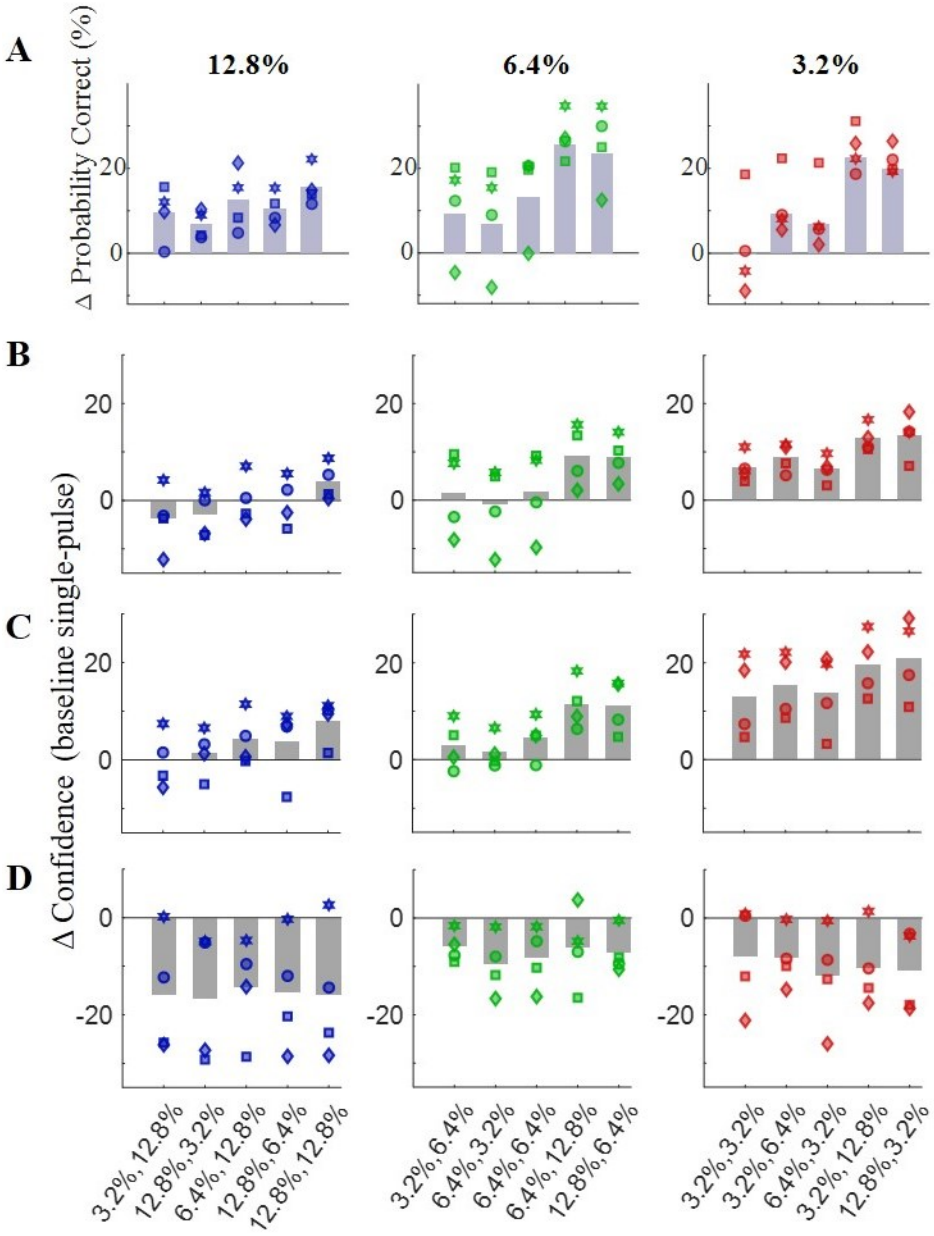


Figure 4. The variation of accuracy or confidence in double-pulse trials baselined by single-pulse trials. **(A)** Considering all the trials, the accuracy improved in most of the combinations in double-pulse trials. **(B)** Considering all the trials, the confidence improved in the combination with stronger pulses while the confidence in sequence with a weaker pulse either decreased or remained constant. **(C)** In correct-choice trials, the increasing effect of stronger pulses was enhanced and the confidence even slightly improved in combination with weaker pulses compared to corresponding baseline. **(D)** Interestingly, in incorrect trials, the confidence decreased in every condition. The gray bars represent mean data while each dot shows the data of each participant.

3.4 Computational models

This is a provisional file, not the final typeset article

In SDT models, d' is stimulus sensitivity and has relation to the choice performance. In general, as we expected, d' in double-pulse trials surpasses the d' in single-pulse trials (Figure 5C left, Supplementary Figure 4 left). Furthermore, compared to single-pulse trials, metacognitive sensitivity, $Meta\text{-}d'$, increased in double-pulse trials (Figure 5C middle, Supplementary Figure 4 middle). We also computed metacognitive efficiency ($Meta\text{-}d'/d'$), as another index of the metacognitive ability which attempt to equate for underlying accuracy. Here, $Meta\text{-}d'/d'$ computed from double-pulse trials was higher for most of the participants (Figure 5C right, Supplementary Figure 4 right). Pairwise comparisons of d', $Meta\text{-}d'$, and $Meta\text{-}d'/d'$ across models was shown in Supplementary Table 6. Furthermore, d', $Meta\text{-}d'$, and $Meta\text{-}d'/d'$ estimated for double-pulse trials were independent of the interpulse interval in both experiments (see Supplementary Figure 9 and Supplementary Table 7).

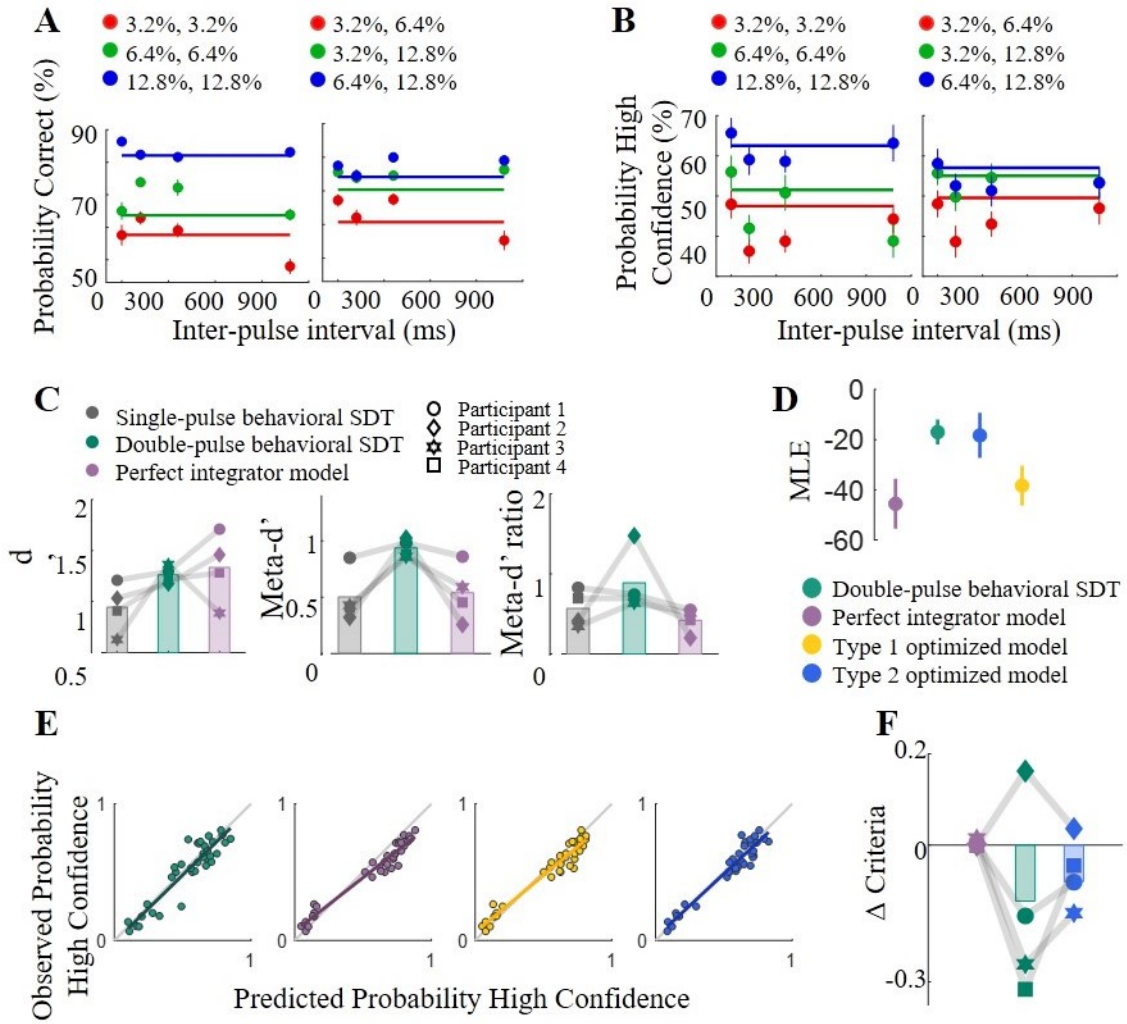


Figure 5. Comparison of the models and human behavior. **(A)** Accuracy in double-pulse trials. Horizontal lines show accuracy prediction by the perfect integrator model. **(B)** Confidence in double-pulse trials. Horizontal lines show confidence prediction by the perfect integrator model. In (A) and (B) each data point represents pooled data from the pulse sequence indicated by the legend and its reverse order. **(C)** Stimulus sensitivity (d'), metacognitive sensitivity ($Meta\text{-}d'$), metacognitive efficiency ($Meta\text{-}d'/d'$) estimated for single-pulse trials, double-pulse trials and the perfect integrator models for each participant. **(D)** Model comparison suggests strong evidence

in favor of the type 2 optimized model over the perfect integrator model and the type 1 optimized model. **(E)** Relation of the observed confidence and the predicted confidence by simulating the model of double-pulse trials (green), the perfect integrator model (purple), the type 1 optimized model (yellow), and the type 2 optimized model (blue). Colored lines indicate the best-fitting slope of a linear regression analysis. Each data point represents pooled data from each sequence of pulses for each participant. **(F)** Variation of confidence criteria comparing to single-pulse trials in the perfect integrator model, the type 2 optimized model and, the model fitted to double-pulse trials. For panels A, B, and D data are represented as group mean ± SEM. For panels C and F, each dot represents the data of each participant

To further quantify the improvement of metacognitive ability in double-pulse trials —a more accurate appraisal of decisions—, we randomly chose the same numbers of trials from each coherence of single-pulse and double-pulse trials from each participants' data where the reported confidence was between 40 to 60. Comparison among the probability of correct choice of these single-pulse and double-pulse trials showed higher performance in double-pulse trials with similar confidence ratings (see Supplementary Figure 7).

We estimated the expected accuracy and confidence that would be achieved in double-pulse trials under the assumption of perfect integration. The accuracy in double-pulse trials surpasses the expectation measured by the perfect integrator model (Figure 5A; [Kiani et al., 2013]). Considering the strong positive relation of accuracy and confidence (Kiani et al., 2014; Vafaei Shooshtari et al., 2019), we expected the observed confidence would exceed the predicted confidence calculated by the perfect integrator model (Eq.9), but it did not (Figure 5B). Further, we predicted the d' in double-pulse trials by the perfect integrator model based on single-pulse trials performance. Accordingly, in general, d' in double-pulse trials surpasses the predictions but if participants' performance (d') in single-pulse trials failed, the predicted performance for double-pulse trials missed the observed data (Figure 5C left, Supplementary Figure 4 left). Furthermore, in line with confidence, the perfect integrator model failed to imitate the $Meta\text{-}d'$ in double-pulse trials (Figure 5C middle, Supplementary Figure 4 middle). Likewise, $Meta\text{-}d'/d'$ simulated from the perfect integrator model failed to track $Meta\text{-}d'/d'$ observed in double-pulse trials (Figure 5C right, Supplementary Figure 4). Altogether, the perfect integrator failed to employ observed confidence and metacognitive ability in double-pulse trials.

As a control investigation, we examined whether the differences in estimated metacognitive ability between models could result from the different number of trials. We averaged the metacognitive scores obtained from randomly selected equal numbers of samples, and found very similar results. Thus, the difference in the estimated metacognitive efficiency cannot be explained by the difference in the number of trials between the single-pulse, double-pulse, and perfect integrator models (Supplementary Figure 10).

To sum, the perfect integrator model failed to follow observed metacognitive ability in double-pulse trials, and the β_1 (slope in Eq.11) differed from 1:1 line which means predicted confidence failed to account for observed confidence (Figure 5E purple; Eq.11, $\beta_1 = .77$, $p < .001$, 95% CI = [.71, .83], $f^2 = 15.66$). Consequently, we introduced two models that optimized stimulus sensitivity (d') or metacognitive sensitivity ($Meta\text{-}d'$) of the perfect integrator model. In type 1 optimized model, where the stimulus sensitivity (related to decision performance) optimized, the predicted confidence slightly improved (Figure 5E yellow, Eq.11, $\beta_1 = .82$, $p < .001$, 95% CI = [.79, .85], $f^2 = 12.01$). In

This is a provisional file, not the final typeset article

the type 2 optimized model, the stimulus parameter (d') remained similar to the perfect integrator model whereas the placement of confidence criteria was optimized to follow the $Meta\text{-}d'$ of the observed data. In this way, the predicted confidence improved considerably (Figure 5E blue, Eq.11, $\beta_1 = .98$, $p < .001$, 95% CI = [.88, 1.08], $f^2 = 10.11$). As an extra measure, we take the confidence criteria of the single-pulse SDT model as the baseline and measure the variation of criteria of the perfect integrator model, the behavioral model and the type 2 optimized model of double-pulse trials. The type 2 optimized model's variation has changed compared to the perfect integrator model to follow the behavioral model (Figure 5F, and Supplementary Figure 5 for Experiment 2). Accordingly, these results suggest the inability to predict the proper change in confidence criteria as a factor that made the perfect integrator model unable to estimate the confidence in double-pulse trials from single-pulse trials.

In addition, to investigate whether the suboptimality in confidence reporting and inability to predict the proper change in confidence criteria is related to confidence noise, we simulated data using the perfect integrator model's parameters while setting higher confidence noise (Eq.12). The predicted confidence from this simulation improved (Eq.11, $\beta_1 = .97$, $p < .001$, 95% CI = [.83, 1.07], $f^2 = 9.00$). Consequently, one problem of the perfect integrator model is ignoring the effect of confidence noise.

3.4.1 Models' evaluation

We conducted parameter recovery simulations to evaluate models. In single-pulse trials, linear regression indicated that there was a significant effect between the predicted and observed confidence, (Eq.11, $\beta_1 = 1.04$, $p < .001$, 95% CI = [.90, 1.18], $f^2 = 8.09$). In double-pulse trials, regression coefficient was statistically significant and close to 1:1 line (Figure 5E green; Eq.11, $\beta_1 = 1.03$, $p < .001$, 95% CI = [.91, 1.15], $f^2 = 11.05$) meaning predicted confidence by classic SDT also explained a significant proportion of variance in the observed confidence. Moreover, as discussed before the type 2 optimized model intimate the behavioral data (Figure 5E blue), unlike the perfect integrator model (Figure 5E purple) and the type 1 optimized model (Figure 5E yellow).

A quantitative model comparison unsurprisingly favored the type 2 optimized model (mean MLE = -18.31) and the behavioral double-pulse model (mean MLE = -16.98) over the perfect integrator model (mean MLE = -45.53) (Figure 5D) and type 1 optimized model (mean MLE = -38.31).

3.5 Response-time analysis

As confidence was traced by both evidence and response-time in perceptual decision-making (Kiani et al., 2014; van den Berg et al., 2016; Zylberberg et al., 2016), we compared accuracy grouped by high and low confidence in four response-time bins. Accordingly, in double-pulse trials, the difference of accuracy grouped by high and low confidence was noteworthy in all response-time bins in double-pulse trials (Figure 6A) but not in single-pulse trials (Figure 6B). A linear regression confirmed the influence of response-time on confidence in double-pulse ($\beta_1 = -.04$, $p < .01$, 95% CI = [-.05, -.03], $f^2 = .10$) but not in single-pulse trials ($\beta_1 = -.12$, $p = .62$, 95% CI = [-1.78, -.07], $f^2 = .04$). Additionally, in our double-pulse trials, participants decided faster than single-pulse trials in all interval durations (Figure 6C). As a check, we regressed the delay-time before cue onset (0.4 to 1 s truncated exponential) and response-time in both single-pulse and double-pulse trials to examine the effect of imposed delay-time on response-time. The effect was small in both single-pulse ($\beta = -.01 \times 10^{-5}$, $p = .004$, 95% CI = [0.00, 0.00], $f^2 = .00$) and double-pulse trials ($\beta = -.0003 \times 10^{-6}$, $p < .001$, 95% CI = [0.00, 0.00], $f^2 = .01$).

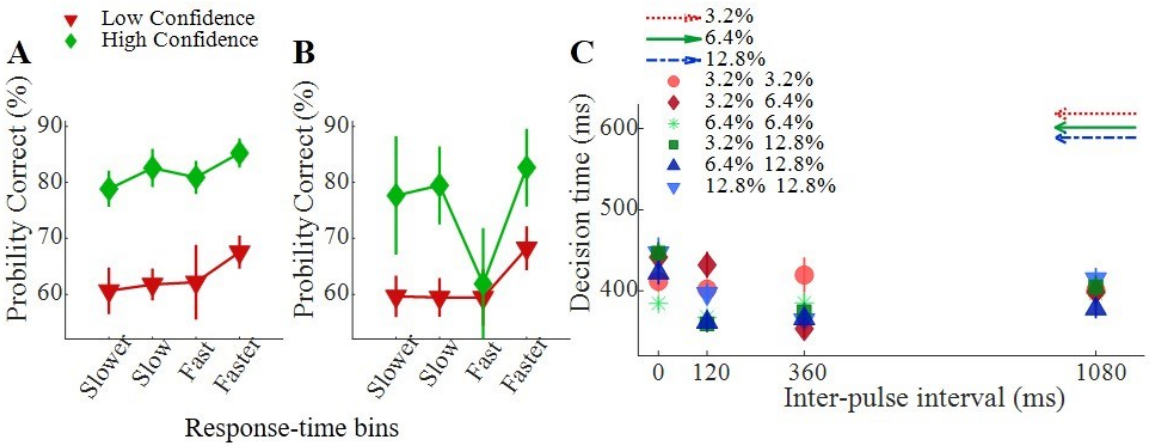


Figure 6. Response-time in single-pulse and double-pulse trials. **(A, B)** Accuracy as a function of response-time split by high (green) and low (red) confidence in double-pulse trials (A) and single-pulse trials (B). **(C)** Response-time of all coherence combinations clustered by interpulse intervals in double-pulse trials (dots) compared to single-pulse trials (arrows). In all panels, data are represented as group mean ± SEM.

3.6 EEG Analysis

We derived the ERPs of averaged signals for two confidence levels to verify whether there was a significant difference in the CPP across confidence levels. Figure 7 exhibits ERPs and scalp topographies for confidence levels time-locked to the stimulus onset in single-pulse trials. ERPs in the two confidence levels showed the ramping activities after 250 ms of the stimulus onset (Figure 7A). However, contrary to what we expected (Vafaei Shooshtari et al., 2019; Zizlsperger, Sauvigny, Händel, & Haarmeier, 2014), high confidence trials failed to confirm higher CPP amplitude in single-pulse trials. Accordingly, while, in single-pulse trials with low coherence (3.2% and 6.4%), there was no significant difference between CPP amplitude for low and high confidence trials (Figure 7C, 7D), there was a significant difference in 12.8% coherence (Figure 7B).

This is a provisional file, not the final typeset article

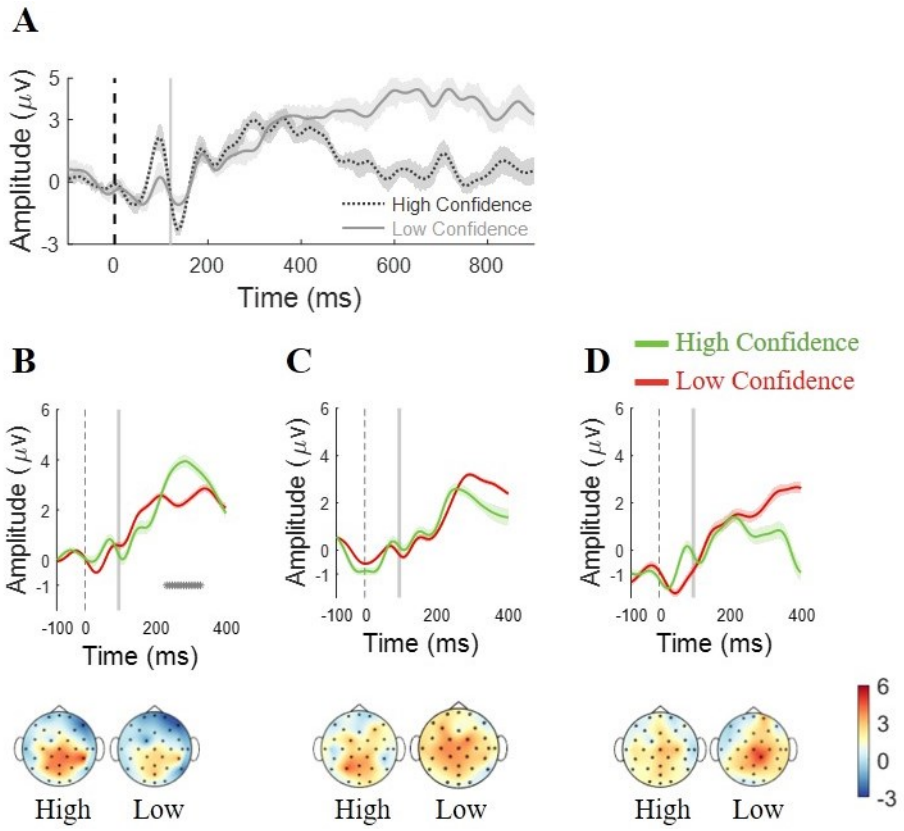


Figure 7. ERPs and scalp topographies in single-pulse trials. **(A)** ERP over the entire trial grouped by high and low confidence**.** ERPs and Scalp topographies in the two levels of confidence of single-pulse trials show significant difference in **(B)** 12.8% coherence, and insignificant difference in **(C)** 6.4% coherence, and **(D)** 3.2% coherence. Scalp topographies are represented in two levels of confidence (the mean amplitude in a time-window ranging from 200 ms to 500 ms after stimulus onset of second pulse). The shading region around the mean indicates SEM. * indicate p<.05 from a cluster-corrected *t*-test, of the difference between the two-time.

We also derived the ERPs of averaged signals for two confidence levels in double-pulse trials. ERPs in the two confidence levels showed the ramping activities after 250 ms of the stimulus onset and as we expected, high confidence trials confirmed higher CPP amplitude in double-pulse trials for all different interpulse intervals (see Figure 8A, 8B, 8C, and 8D over the entire trial). Moreover, Figure 8E, 8F, and 8G exhibit significant differences in CPP amplitude for two confidence levels time-locked to the stimulus onset in each pulse of double-pulse trials. As literature has confirmed that CPP amplitude is associated with confidence at the time of the response (Boldt et al., 2019), although ERPs for both pulses are presented, we only reported significancy for the second pulses. In adition, the previous findings distinguished the correlation between CPP signal and coherence (Kelly and O'Connell, 2013) and coherence is the decisive characteristic of confidence (Baranski and Petrusic, 1998; Van denBerg et al., 2016b). Accordingly, to verify that our findings are not relevant to different coherence, we only included double-pulse trials with equal motion strengths. Accordingly, ERPs and scalp topographies showed that even in trials with low coherence of double-pulse trials, there were significant difference between CPP amplitude for low and high confidence trials.

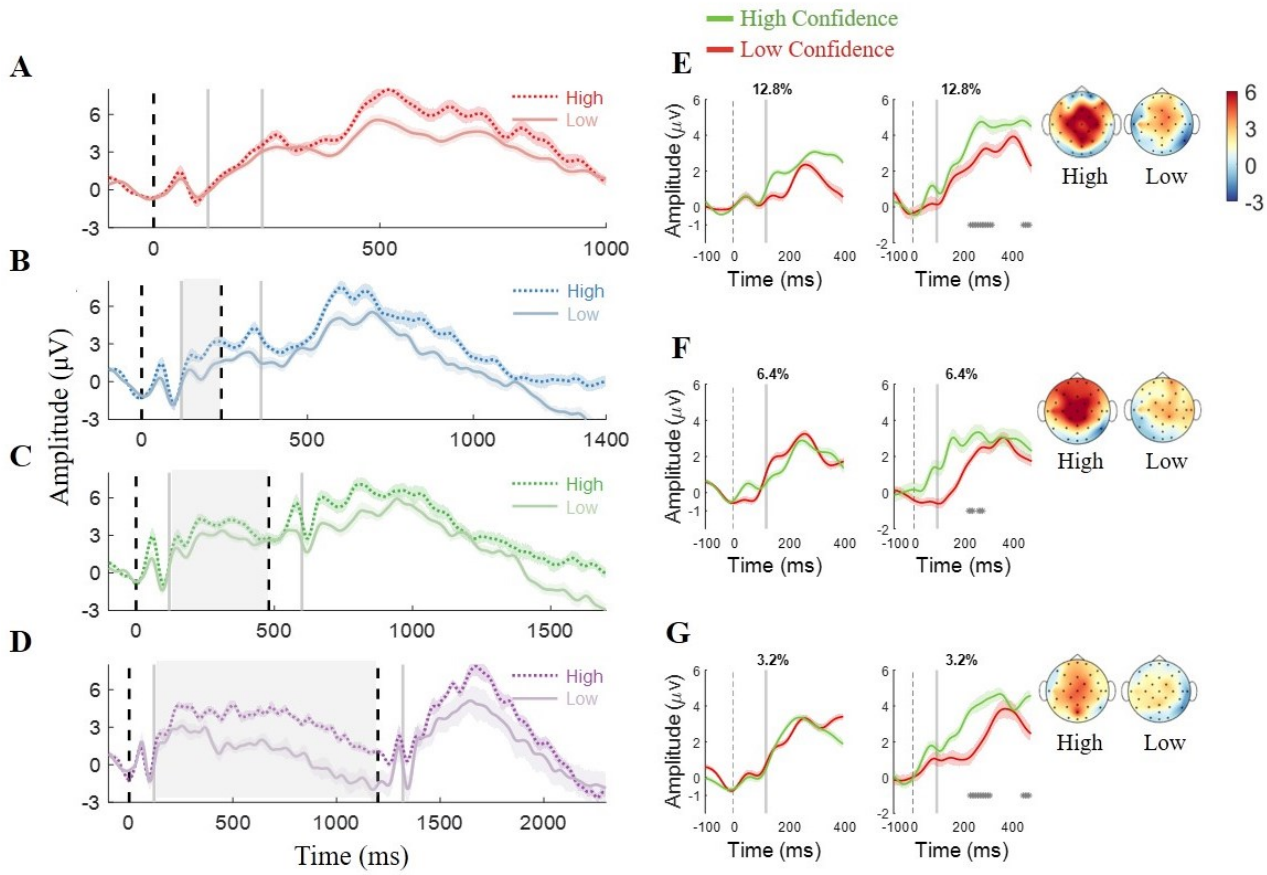


Figure 8. ERPs and scalp topographies in double-pulse trials. (**A-D**) ERPs over the entire trial grouped by high and low confidence in trials with zero, 120 ms, 360 ms, and 1080 ms interpulse intervals. **(E-G)** ERPs and Scalp topographies in the two levels of confidence are distinct after the stimulus onset. Motion strength are indicated at the top of each panel: 12.8%, 6.4%, and 3.2%. Scalp topographies are represented in two levels of confidence (the mean amplitude in a time-window ranging from 200 ms to 500 ms after stimulus onset of second pulse). The shading region around the mean indicates SEM. * indicate p<.05 from a t-test, of the difference between the two-time.

Furthermore, a linear regression model confirmed the influence of confidence on CPP amplitude in double-pulse trials (β_1 = .88, p < .05, 95% CI = [.39, 2.20]; f^2 = .10), whereas there was not sufficient evidence to show the influence of confidence on CPP amplitude in single-pulse trials (β_1 = -6.58, p = .10, 95% CI = [-13.40, .24]; f^2 = .28). Additionally, a statistical comparison of the single-pulse and double-pulse trials represented that the difference of the CPP amplitude for high and low confidence responses depended on the type of trials (Eq.13, β_3 = 7.46, p_{bonf*3} < .05, 95% CI = [3.09, 11.83], f^2 = .25).

To certify that our findings are not relevant to participant performance, we only analyzed the ERP signals in the two confidence levels only for correct responses**.** Supplementary Figure 11 expresses that the ERPs of the two confidence levels were independent of accuracy coding in the centro-parietal area, and they were associated with a degree of certainty.

This is a provisional file, not the final typeset article

3.7 Pupil responses

As shown in Figure 10A, 10B, 10C, 10D, and 10E, during decision formation in both single-pulse and double-pulse trials, pupil dilation in response to the confidence levels was not clearly distinct after the stimulus onset, this is in line with previous works (Vafaei Shooshtari et al., 2019). However, pupil responses were peaking immediately after making the choice, similar to the previous findings (Urai et al., 2017).

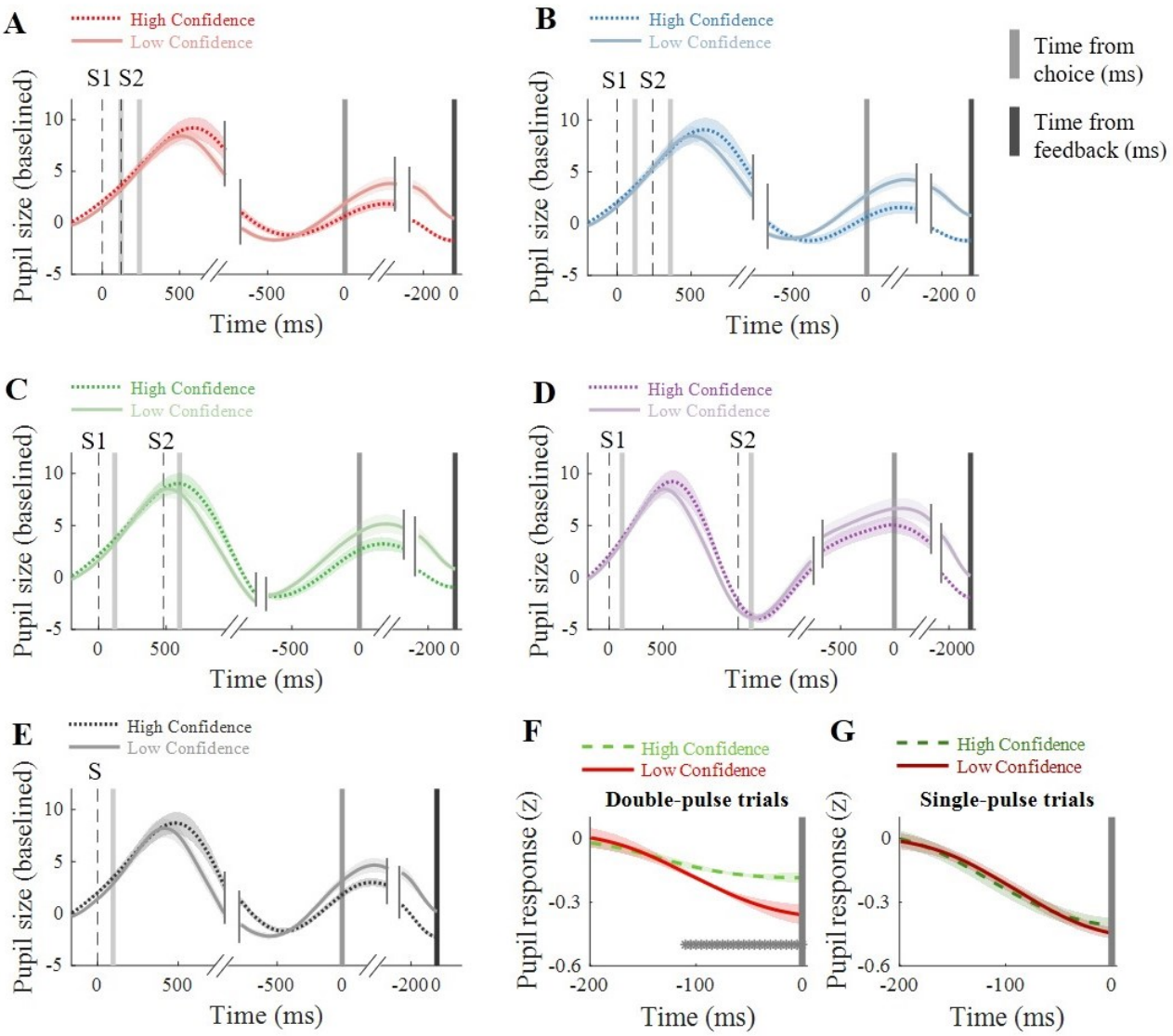


Figure 9. Pupil responses in single-pulse and double-pulse trials. **(A-D)** Time course of pupil responses throughout the entire trial grouped by high and low confidence in trials with zero, 120 ms, 360 ms, and 1080 ms interpulse intervals. **(E)** Time course of pupil responses throughout the entire trial grouped by high and low confidence in single-pulse trials. **(F, G)** Standardized pupil response across time-window aligned to the feedback, high confidence trials (green) vs. low confidence trials (red) in double-pulse trials and single-pulse trials. The shading region around the mean indicates SEM. * indicate $p < .05$ from a *t*-test in F and G.

Additionally, to quantify the temporal evolution of confidence in the pupil, we took the mean baseline-corrected pupil signal during 200 ms before feedback delivery as our measure of pupil response. In line with previous work (Colizoli, De Gee, Urai, & Donner, 2018; Urai et al., 2017) pupil responses reflected decision confidence in our double-pulse trials (Figure 10F; β_1 = -.15, p < .001, 95% CI = [-.25, -.05], f^2 = .28), whereas in single-pulse trials, pupil responses did not reflect decision confidence (Figure 10G; β_1 = - .12, p = .21, 95% CI = [-.21, -.03], f^2 = .20). Furthermore, a statistical comparison of the single-pulse and double-pulse trials represented that the difference of the pupil response for high and low confidence responses depended on the type of trials (Eq.14, β_3 =.33, p_{bonf*3} < .05, 95% CI = [-.09, .76], f^2 = .23). To certify that our findings are not relevant to participant performance, we only analyzed the pupil response signals in the two confidence levels for correct responses. Supplementary Figure 12 expresses that the pupil responses during 200 ms before feedback delivery of the two confidence levels were independent of accuracy, and they were associated with confidence in double-pulse trials.

4 Discussion

The current study was designed to clarify the confidence of decisions in a more similar context to the real world, where the evidence arrives separately. Using an experimental design adopting either single or double pulse(s) of RDM stimuli, in two separate studies, we integrated behavioral, EEG responses, and pupillometry to examine how people combined separate pieces of information to form their decision. We were particularly interested in how accuracy and confidence are related to each other in the context of decision-making. We predicted that similar to accuracy, confidence would improve when people integrate separate pieces of information, confidence would be leveraged larger by more recent information, and that using behavioral, neural, and pupillometry markers of confidence would be distinguishable in both single-pulse trials and double-pulse trials. In summary, we found that although participants' confidence was invariant primarily to the interpulse interval, confidence scoring was not noticeably enhanced in double-pulse trials compared to single-pulse trials. Instead, participants only reported their confidence with higher resolution suggesting that adding an extra pack of information only increases participants' metacognitive sensitivity. Moreover, participants used both pulses to decide about their confidence irrespective of their recency or primacy orders. Results also revealed that using RT, EEG, and pupillometry analysis allows a conceivable track of the confidence in double-pulse trials but not in single-pulse trials. Findings are discussed separately for behavioral, computational, and implicit confidence markers in the following sub-sections.

4.1 Behavioral and motion energy findings

Contrary to our hypothesis, and unlike accuracy, confidence ratings in double-pulse trials did not increase compared to single-pulse trials. In addition, our data failed to support our prediction that participants mainly trust the evidence of one of the pulses and ignore the other one. The trusted pulse can either be the first or second pulse; it also can simply be the stronger pulse. Nevertheless, the effect of sequence and interaction of pulses on confidence was examined and no effect was observed. Inconsistent with the previous claims that participants obtained more information from a second pulse (Kiani et al., 2013; Tohidi-Moghaddam et al., 2019), our motion energy analysis did not confirm the asymmetric influences of the two pulses for confidence. Yet our findings are in line with a large body of evidence (De Gardelle & Mamassian, 2015; Herce Castañón et al., 2019; Rahnev & Denison, 2018; Zylberberg et al., 2016, 2014) that shows that the observers do not make their

This is a provisional file, not the final typeset article

decisions precisely following the confidence rating. Indeed, comparing confidence in double-pulse trials grouped by accuracy, we found that the participants had lower confidence in double-pulse trials than single-pulse trials for incorrect choices but higher confidence for correct choices. In other words, compared with single-pulse trials, in double-pulse trials, participants adjusted their confidence by enhancing their confidence resolution or metacognitive sensitivity. Typically, confidence facilitates evidence accumulation and drives a confirmation bias in perceptual decision-making (Rollwage et al., 2020). Likewise, we suggest that an extra brief and weak evidence can validate confidence and improve metacognitive sensitivity.

4.2 Computational modeling findings

To understand the nature of the differences of metacognitive sensitivity in double-pulse vs. single-pulse trials, we compared corresponding estimated metacognitive parameters for each participant. Likewise, we included the expected parameters that would be achieved in double-pulse trials under the assumption of perfect integration. Accordingly, we computed $Meta\text{-}d'/d'$ as a measure of 'metacognitive efficiency'. In the case of $Meta\text{-}d' = d'$, the observer is metacognitively 'ideal'. Indeed, all the information available for the decision would be used to report the confidence. Yet, in many cases, we might find that $Meta\text{-}d' < d'$, along with some degree of noise or suboptimality (Fleming & Lau, 2014; Maniscalco & Lau, 2012). Conversely, we may find that $Meta\text{-}d' > d'$, if participants are able to draw on additional information such as hunches (Rausch & Zehetleitner, 2016), further processing of stimulus information (Charles, Van Opstal, Marti, & Dehaene, 2013) or extra prior knowledge on the task. In the model fitted to double-pulse trials, $Meta\text{-}d'/d'$ was around .8 and near to ideal for almost all participants. However, as it varied considerably between participants in single-pulse trials, the value could not be adjusted in the perfect integrator model too.

Previously, the better-than-expected performance in double-pulse trials was explained by underperformance in single-pulse trials (Kiani et al., 2013). In current study, metacognitive sensitivity in double-pulse trials surpasses the value predicted by the perfect integrator model (Figure 5C and Supplementary Figure 4). This effect can be followed in all of our participants (except one of the participants from Experiment 2) and can be explained by low confidence resolution in single-pulse trials.

Metacognitive noise is the noise that affects confidence estimates but not perceptual decisions (De Martino, Fleming, Garrett, & Dolan, 2013; Jang, Wallsten, & Huber, 2012; Maniscalco & Lau, 2016; Mueller & Weidemann, 2008; Rahnev, Nee, Riddle, Larson, & D'Esposito, 2016; Shekhar & Rahnev, 2018; Van den Berg, Yoo, & Ma, 2017). A recent work categorized sources of metacognitive inefficiency (Shekhar & Rahnev, 2020). Accordingly, metacognitive noise is a superordinate term for all noise sources that impact the confidence formation process (Shekhar & Rahnev, 2020, 2021), ranging from systematic to nonsystematic input and computation. Nevertheless, the exact source of metacognitive noise remains unclear (Shekhar & Rahnev, 2020). This noise can be traced in our perfect integrator model, which was capable of accumulating decision evidence perfectly but could not predict confidence formation in our task. We suggest that the perfect integrator model was unable to adjust to confidence criteria when predicting confidence in double-pulse trials. However, an improved SDT capable of addressing metacognitive noise might be able to empower the employed perfect integrator model. Furthermore, SDT is not the only available model to implement a perfect integrator model. Previous studies suggested attractor models as a candidate model to implement the perfect integrator model (Kiani et al., 2013; Waskom & Kiani, 2018). Attractor models are a group of networks that form a bridge between cognitive theory and biological data which exploits inhibition to achieve to compete among alternatives (Wang, 2002; Wong & Huk,

2008). Although these models can integrate momentary evidence to establish a decision, they have specific failure behaviors that would be apparent when gaps separate the sources of evidence in time (Kiani et al., 2013). Besides, when the stimulus is very short, mostly, none of the attractors can be reached, and the network would revert back to the resting state after the stimulus offset (Wang, 2002). Therefore, the choice would be assigned randomly. However, our experiments' data represent a noteworthy performance in single-pulse trials, not supporting this expectation. Consequently, to implement a perfect integrator model by implementing an attractor model, a mechanism for simulating a very short stimulus might be considered. Moreover, our behavioral assays highlighted different relationships between confidence and accuracy in the different conditions of the task. So, a dedicated neural module with a plausible circuit of confidence might be a better option to implement a perfect integrator model. Recently, multi-layer recurrent network models have been developed to account for decision confidence mechanisms (Atiya, Rañó, Prasad, & Wong-Lin, 2019; Paz, Insabato, Zylberberg, Deco, & Sigman, 2016). These models consist of multiple layers of neural integrators and align with neural evidence of decision confidence (Kepecs, Uchida, Zariwala, & Mainen, 2008; Murphy, Robertson, Harty, & O'Connell, 2015), they are suggested to justify the observed behavior.

Furthermore, perceptual decisions are often modeled using ideal observers (e.g., SDT). However, a source of suboptimal behavior in decision-making is 'lapse' (Gold & Ding, 2013; Pisupati, Chartarifsky-Lynn, Khanal, & Churchland, 2021). The lapses are an additional constant rate of errors independent of the evidence strength (Gold & Ding, 2013; Pisupati et al., 2021). Lapse rate has been shown to increase with higher perceptual uncertainty (Pisupati et al., 2021) and would be accounted for by fitting extra parameters to psychometrics models. Accordingly, as the perfect integrator model was based on SDT, ignoring lapse in the single-pulse trials might lead to mis-estimation of decision parameters in double-pulse trials. Consequently, further models including the lapse parameters (Pisupati et al., 2021), may improve the perfect integrator model's predictivity.

4.3 Implicit confidence markers

Although research suggests faster decisions accompanied by higher confidence (Kiani et al., 2014; Vafaei Shooshtari et al., 2019; van den Berg et al., 2016; Zylberberg et al., 2016), our results do not show such an association in the presence of a brief piece of evidence. Moreover, our participants decided much faster in double-pulse trials compared single-pulse trials. Therefore, we proposed that the decrease of response-time in double-pulse trials would be reflected with higher internal confidence. However, another hypothesis of this variation pointed to the extra time duration in double-pulse trials, which can be used to increase readiness to decide. Therefore, to explore the hypothesis, we regressed the delay-time before response cue onset and response-time in both single-pulse and double-pulse trials. If the variation of response-time was primarily dependent on extra delay time, the delay time should have had a considerable effect on response-time, especially in our 120ms single-pulse trials when the stimulus duration was concise and the delay time varied. Nevertheless, the effects in both double-pulse and single-pulse trials are weak. Accordingly, the hypothesis that faster decisions reflect higher confidence in double-pulse trials is supported. In addition, response-time in double-pulse trials traces the confidence unlike in single-pulse trials.

Our findings furthermore suggest that reported confidence might not follow the confidence marker in EEG response. We focused on the CPP —a neural correlate of perceptual processing believed to reflect evidence accumulation and correlated to confidence (Boldt et al., 2019; Herding et al., 2019; Rausch et al., 2020; Vafaei Shooshtari et al., 2019; Zizlsperger et al., 2014). However, our findings suggest that in the presence of a brief and weak stimulus, entirely unlike in double-pulse trials, CPP

This is a provisional file, not the final typeset article

amplitudes show no significant variation in a high and low level of confidence. As confidence in single and double-pulse trials did not vary significantly, we suggest that variation of CPP amplitude shares more commonalities with implicit confidence measures rather than explicit confidence measures like ratings. Moreover, we propose that pupil response relation to confidence rating varies as the task condition changes; when participants access brief and weak stimuli, no association is detected, unlike in the presence of a pair of separated stimuli. Our current observations are not easily reconciled with existing theoretical accounts of the impact of the confidence level on pupil response (Allen et al., 2016; Lempert et al., 2015; Urai et al., 2017). Therefore, our study has thus theoretical implications for the confidence and decision formation and provides a new insight into the relationship between confidence and accuracy in perceptual decision making. In line with the theoretical account, the findings of this study can clear up some of the ambiguities of real-world decision-making when the evidence is separated in time. Considering the fact that it would be more probable that people change their mind when decisions are accompanied by lower confidence (Fleming, Putten, & Daw, 2018; Resulaj, Kiani, Wolpert, & Shadlen, 2009), knowing how to make decisions in a discrete environment and confidence, can help encourage more sustainable choices in marketing and behavioral economics. Also, people can integrate information from themselves and others based on their confidence (Bahrami, B. et al., 2010; Bang, D. et al., 2017). Accordingly, confidence can affect the extent of people pursue others or be persuaded by them (Petty, Briñol, & Tormala, 2002). To sum up, when participants access brief and mainly weak stimuli, the confidence ratings are not reliable, and confidence could not be traced from response-time, pupil, and EEG response. In other words, implicit confidence markers, in some cases, might be incapable of following the conscious confidence rating. This is in line with innovative findings abstracting implicit confidence measures from explicit confidence measures (Logan & Crump, 2010).

4.4 Limitations and future directions

Our results were grounded in assumptions of integration strategy in decision-making. However, this insight has recently been reconsidered (Carland, Marcos, Thura, & Cisek, 2016; Stine et al., 2020). An urgency-gate model might explain participants' decisions better (Evans, Hawkins, Boehm, Wagenmakers, & Brown, 2017; Thura, Beauregard-Racine, Fradet, & Cisek, 2012) rather than an integration strategy such as perfect integration. A participant's strategy could be something between no integration and perfect integration or in a completely different space of models (Stine et al., 2020) and might be changed depending on the task paradigm or even the participant's internal state (Evans & Hawkins, 2019; Najafi & Churchland, 2018; Tsetsos, Gao, McClelland, & Usher, 2012). Consequently, further models to discuss the decision strategy in separated pulses could guide future works. In addition, future experiments could develop computational approaches and attempt to implement other scenarios in a discrete environment to study choice and confidence formation and examine the involved processes. In addition, although the vast number of trials for each participant allowed us to do a robust subject-wise analysis and our EEG study replicated the same behavioral and modeling data, our small sample size prevents us from making general claims. Nevertheless, future research might capitalize on our paradigm to provide a situation in which confidence remains persistent but metacognitive sensitivity improves. In this way, future research continues studying the neural basis of metacognitive ability and consciousness in addition to previous works (Feuerriegel, Blom, & Hogendoorn, 2021; Fleming & Dolan, 2012).

5 Conclusion

To sum, despite its limitations, the present study sheds new light on confidence formation, especially in perceptual decision-making when diverse temporal gaps separate a pair of visual cues. Our data suggest that accumulated evidence from both pulses shape confidence invariant to the interpulse intervals up to 1 s and the association of accuracy and confidence improved. However, the confidence would not follow the same characteristics as accuracy. For example, confidence would not be leveraged larger by more recent information. Moreover, we showed that the classic perfect integrator model merely highlighted evidence accumulation and ignored the effect of the metacognitive noise that affects confidence, so it failed to predict confidence. Finally, integrating evidence from two separated pieces of information supports the confidence to be traced from implicit markers like RT, EEG, and pupil responses, unlike the situation in which participants have to decide based on a brief and weak pulse of information.

6 Conflict of Interest

The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

7 Author Contributions

ZA: conceptualization, data acquisition, analysis, visualization, writing - original draft, writing - review and editing; SZ: conceptualization, supervision, writing - review and editing; AJ: supervision, writing - review and editing; RE: conceptualization, supervision, writing - review and editing.

8 Acknowledgments

Data were recorded in the Cognitive Science Laboratory of Shahid Rajaee University. The authors would like to thank Samuel Klein for their constructive comments and proofreading of the manuscript. We thank all our participant and finally, special thanks to the open science movement and generous researchers who we had their helpful comments during the implementation and analysis.

9 Ethics

The ethics committee of the Iran University of Medical Sciences (protocol #IR.IUMS.REC1399648) approved the experimental protocol, and participants gave written informed consent.

10 Supplementary Material

The Supplementary Material for this article can be found online at:

11 Data Availability Statement

The datasets generated for this study are available on request to the corresponding author.

12 References

This is a provisional file, not the final typeset article

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Josa A*, *2*(2), 284–299.

Allen, M., Frank, D., Schwarzkopf, D. S., Fardo, F., Winston, J. S., Hauser, T. U., & Rees, G. (2016). Unexpected arousal modulates the influence of sensory noise on confidence. *Elife*, *5*, e18103, https://doi.org/10.7554/eLife.18103.

Atiya, N. A. A., Rañó, I., Prasad, G., & Wong-Lin, K. F. (2019). A neural circuit model of decision uncertainty and change-of-mind. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-10316-8, https://doi.org/10.1038/s41467-019-10316-8.

Atiya, N. A. A., Zgonnikov, A., O'Hora, D., Schoemann, M., Scherbaum, S., & Wong-Lin, K. (2020). Changes-of-mind in the absence of new post-decision evidence. *PLoS Computational Biology*, *16*(2), e1007149, https://doi.org/10.1371/journal.pcbi.1007149.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. Science, 329(5995), 1081-1085. https://doi.org/10.1126/science.1185718

Balsdon, T., Wyart, V., & Mamassian, P. (2020). Confidence controls perceptual evidence accumulation. *Nature Communications*, *11*(1), 1–11, https://doi.org/10.1038/s41467-020-15561-w.

Bang, D., Aitchison, L., Moran, R. et al. Confidence matching in group decision-making. Nat Hum Behav 1, 0117 (2017). https://doi.org/10.1038/s41562-017-0117

Baranski, J. V, Petrusic, W. M., Peters, M. A. K., Thesen, T., Ko, Y. D., Maniscalco, B., … Dolan, R. J. (2017). The relationship between perceptual decision variables and confidence in the human brain. *Journal of Neuroscience*, *32*(18), 412–428, https://doi.org/10.1093/cercor/bhu181.

Boldt, A., Schiffer, A.-M., Waszak, F., & Yeung, N. (2019). Confidence predictions affect performance confidence and neural preparation in perceptual decision making. *Scientific Reports*, *9*(1), 1–17, https://doi.org/10.1038/s41598-019-40681-9.

Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Carland, M. A., Marcos, E., Thura, D., & Cisek, P. (2016). Evidence against perfect integration of sensory information during perceptual decision making. *Physiology*, *115*(1), 1–13, https://doi.org/10.1152/jn.00264.2015.

Charles, L., Van Opstal, F., Marti, S., & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *Neuroimage*, *73*, 80–94, https://doi.org/10.1016/j.neuroimage.2013.01.054.

Cohen, J. (1970). Approximate power and sample size determination for common one-sample and two-sample hypothesis tests. *Educational and Psychological Measurement*, *30*(4), 811–831, https://doi.org/10.1177/001316447003000404.

Colizoli, O., De Gee, J. W., Urai, A. E., & Donner, T. H. (2018). Task-evoked pupil responses reflect internal belief states. *Scientific Reports*, *8*(1), 1–13, https://doi.org/10.1038/s41598-018-31985-3.

De Gardelle, V., & Mamassian, P. (2015). Weighting mean and variability during confidence judgments. *PLoS ONE*, *10*(3). https://doi.org/10.1371/journal.pone.0120870.

De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105, https://doi.org/10.1038/nn.3279.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21, https://doi.org/10.1016/j.jneumeth.2003.10.009.

Evans, N. J., & Hawkins, G. E. (2019). When humans behave like monkeys: Feedback delays and extensive practice increase the efficiency of speeded decisions. *Cognition*, *184*, 11–18, https://doi.org/10.1016/j.cognition.2018.11.014.

Evans, N. J., Hawkins, G. E., Boehm, U., Wagenmakers, E.-J., & Brown, S. D. (2017). The computations that support simple decision-making: A comparison between the diffusion and urgency-gating models. *Scientific Reports*, *7*(1), 1–13, https://doi.org/10.1038/s41598-017-16694-7.

Feuerriegel, D., Blom, T., & Hogendoorn, H. (2021). Predictive activation of sensory representations as a source of evidence in perceptual decision-making. *Cortex*, *136*, 140–146, https://doi.org/10.1016/j.cortex.2020.12.008.

Fleming, S. M. (2017). HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, *2017*(1), https://doi.org/10.1093/nc/nix007.

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1338–1349, https://doi.org/10.1098/rstb.2011.0417.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443, https://doi.org/10.3389/fnhum.2014.00443.

Fleming, S. M., Putten, E. J., & Daw, N. D. (2018). Neural mediators of changes of mind about perceptual decisions. *Nature Neuroscience*, 1, https://doi.org/10.1038/s41593-018-0104-6.

Gherman, S., & Philiastides, M. G. (2015). Neural representations of confidence emerge from the process of decision formation during perceptual choices. *Neuroimage*, *106*, 134–143, https://doi.org/10.1016/j.neuroimage.2014.11.036.

Gold, J. I., & Ding, L. (2013). How mechanisms of perceptual decision-making affect the psychometric function. *Progress in Neurobiology*, *103*, 98–114, https://doi.org/10.1016/j.pneurobio.2012.05.008.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

Hebart, M. N., Schriever, Y., Donner, T. H., & Haynes, J.-D. (2014). The relationship between perceptual decision variables and confidence in the human brain. *Cerebral Cortex*, *26*(1), 118–130, https://doi.org/10.1093/cercor/bhu181.

This is a provisional file, not the final typeset article

Herce Castañón, S., Moran, R., Ding, J., Egner, T., Bang, D., & Summerfield, C. (2019). Human noise blindness drives suboptimal cognitive inference. *Nature Communications*, *10*(1), 1–11. https://doi.org/10.1038/s41467-019-09330-7

Herding, J., Ludwig, S., von Lautz, A., Spitzer, B., & Blankenburg, F. (2019). Centro-parietal EEG potentials index subjective evidence and confidence during perceptual decision making. *NeuroImage*, *201*, 116011, https://doi.org/10.1016/j.neuroimage.2019.116011.

Hoeks, B., & Ellenbroek, B. A. (1993). A neural basis for a quantitative pupillary model. *Journal of Psychophysiology*, *7*, 315.

Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review*, *119*(1), 186, https://doi.org/10.1037/a0025960.

Kelly, S. P., & O'Connell, R. G. (2013). Internal and external influences on the rate of sensory evidence accumulation in the human brain. *Journal of Neuroscience*, *33*(50), 19434–19441, https://doi.org/10.1523/JNEUROSCI.3355-13.2013.

Kepecs, A., & Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society of London 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, https://doi.org/10.1038/nature07200.

Kiani, R., Churchland, A. K., & Shadlen, M. N. (2013). Integration of direction cues is invariant to the temporal gap between them. *Journal of Neuroscience*, *33*(42), 16483–16489, https://doi.org/10.1523/JNEUROSCI.2094-13.2013.

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.

Kiani, R., Hanks, T. D., & Shadlen, M. N. (2008). Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. *Journal of Neuroscience*, *28*(12), 3017–3029, https://doi.org/10.1523/JNEUROSCI.4761-07.2008.

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, *324*(5928), 759–764. https://doi.org/10.1126/science.1169405

Kira, S., Yang, T., & Shadlen, M. N. (2015). A neural implementation of Wald's sequential probability ratio test. *Neuron*, *85*(4), 861–873, https://doi.org/10.1016/j.neuron.2015.01.007.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3?

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science*, *7*(1), 18–27, https://doi.org/10.1177/1745691611427305.

Lempert, K. M., Chen, Y. L., & Fleming, S. M. (2015). Relating pupil dilation and metacognitive confidence during auditory decision-making. *PLoS One*, *10*(5), e0126588, https://doi.org/10.1371/journal.pone.0126588.

Logan, G. D., & Crump, M. J. C. (2010). Cognitive illusions of authorship reveal hierarchical error detection in skilled typists. *Science*, *330*(6004), 683–686, https://doi.org/10.1126/science.1190483.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430, https://doi.org/10.1016/j.concog.2011.09.021.

Maniscalco, B., & Lau, H. (2014). Signal detection theory analysis of type 1 and type 2 data: meta-d′, response-specific meta-d′, and the unequal variance SDT model. In *The cognitive neuroscience of metacognition* (pp. 25–66). Springer, https://doi.org/10.1007/978-3-642-45190-4_3.

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

Mathôt, S. (2013). A simple way to reconstruct pupil size during eye blinks. *Retrieved From*, *10*, m9.

Maris, E. & Oostenveld, R (2007). Nonparametric statistical testing of EEG-and MEG-data. *J. Neurosci. Methods 1641*, 177–190.

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as Bayesian probability: From neural origins to behavior. *Neuron*, *88*(1), 78–92.

Mognon, A., Jovicich, J., Bruzzone, L., & Buiatti, M. (2011). ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology*, *48*(2), 229–240, https://doi.org/10.1111/j.1469-8986.2010.01061.x.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review*, *15*(3), 465–494, https://doi.org/10.3758/PBR.15.3.465.

Murphy, P. R., Boonstra, E., & Nieuwenhuis, S. (2016). Global gain modulation generates time-dependent urgency during perceptual choice in humans. *Nature Communications*, *7*, 13526, https://doi.org/10.1038/ncomms13526.

Murphy, P. R., Robertson, I. H., Harty, S., & O'Connell, R. G. (2015). Neural evidence accumulation persists after choice to inform metacognitive judgments. *Elife*, *4*, e11946, https://doi.org/10.7554/eLife.11946..

Najafi, F., & Churchland, A. K. (2018). Perceptual Decision-Making: A Field in the Midst of a Transformation. *Neuron*, *100*(2), 453–462, https://doi.org/10.1016/j.neuron.2018.10.017.

Nienborg, H., & Cumming, B. G. (2009). Decision-related activity in sensory neurons reflects more than a neuron's causal effect. *Nature*, *459*(7243), 89–92, https://doi.org/10.1038/nature07821.

This is a provisional file, not the final typeset article

O'connell, R. G., Dockree, P. M., & Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature Neuroscience*, *15*(12), 1729, https://doi.org/10.1038/nn.3248.

Oostenveld R, Fries P, Maris E, Schoffelen JM (2011) FieldTrip: open-source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci* 2011:156869, https://doi.org/10.1155/2011/156869.

Paz, L., Insabato, A., Zylberberg, A., Deco, G., & Sigman, M. (2016). Confidence through consensus: a neural mechanism for uncertainty monitoring. *Scientific Reports*, *6*, 21830, https://doi.org/10.1038/srep21830.

Petty, R. E., Briñol, P., & Tormala, Z. L. (2002). Thought confidence as a determinant of persuasion: the self-validation hypothesis. Journal of personality and social psychology, 82(5), 722.

Pisupati, S., Chartarifsky-Lynn, L., Khanal, A., & Churchland, A. K. (2021). Lapses in perceptual decisions reflect exploration. *Elife*, *10*, e55490, https://doi.org/10.7554/eLife.55490.

Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences*, *41*.

Rahnev, D., Nee, D. E., Riddle, J., Larson, A. S., & D'Esposito, M. (2016). Causal evidence for frontal cortex organization for perceptual decision making. *Proceedings of the National Academy of Sciences*, *113*(21), 6059–6064, https://doi.org/10.1038/s41467-020-17661-z.

Rausch, M., & Zehetleitner, M. (2016). Visibility is not equivalent to confidence in a low contrast orientation discrimination task. *Frontiers in Psychology*, *7*, 591, https://doi.org/10.3389/fpsyg.2016.00591.

Rausch, M., Zehetleitner, M., Steinhauser, M., & Maier, M. E. (2020). Cognitive modelling reveals distinct electrophysiological markers of decision confidence and error monitoring. *NeuroImage*, *218*, 116963, https://doi.org/10.1016/j.neuroimage.2020.116963.

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature*, *461*(7261), 263, https://doi.org/10.1038/nature08275.

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

Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., & Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature Communications*, *11*(1), 1–11, https://doi.org/10.1038/s41467-020-16278-6.

Shekhar, M., & Rahnev, D. (2018). Distinguishing the roles of dorsolateral and anterior PFC in visual metacognition. *Journal of Neuroscience*, *38*(22), 5078–5087, https://doi.org/10.1167/18.10.665.

Shekhar, M., & Rahnev, D. (2020). Sources of Metacognitive Inefficiency. *Trends in Cognitive Sciences,* https://doi.org/10.1016/j.tics.2020.10.007.

Shekhar, M., & Rahnev, D. (2021). The nature of metacognitive inefficiency in perceptual decision making. *Psychological Review*, *128*(1), 45, https://doi.org/10.1037/rev0000249.

Steinemann, N. A., O'Connell, R. G., & Kelly, S. P. (2018). Decisions are expedited through multiple neural adjustments spanning the sensorimotor hierarchy. *Nature Communications*, *9*(1), 1–13, https://doi.org/10.1038/s41467-018-06117-0.

Stine, G. M., Zylberberg, A., Ditterich, J., & Shadlen, M. N. (2020). Differentiating between integration and non-integration strategies in perceptual decision making. *Elife*, *9*, e55365, https://doi.org/10.7554/eLife.55365.

Tagliabue, C. F., Veniero, D., Benwell, C. S. Y., Cecere, R., Savazzi, S., & Thut, G. (2019). The EEG signature of sensory evidence accumulation during decision formation closely tracks subjective perceptual experience. *Scientific Reports*, *9*(1), 1–12, https://doi.org/10.1038/s41598-019-41024-4.

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

tickle, hannah, Tsetsos, K., Speekenbrink, M., & Summerfield, C. (2020). Optional Stopping in a Heteroscedastic World. *PsyArXiv*. https://doi.org/10.31234/OSF.IO/T7DN2

Tohidi-Moghaddam, M., Zabbah, S., Olianezhad, F., & Ebrahimpour, R. (2019). Sequence-dependent sensitivity explains the accuracy of decisions when cues are separated with a gap. *Attention, Perception, and Psychophysics*, *81*(8), 2745–2754. https://doi.org/10.3758/s13414-019-01810-8

Tsetsos, K., Gao, J., McClelland, J. L., & Usher, M. (2012). Using time-varying evidence to test models of decision dynamics: bounded diffusion vs. the leaky competing accumulator model. *Frontiers in Neuroscience*, *6*, 79, https://doi.org/10.1038/s41598-017-16694-7.

Twomey, D. M., Kelly, S. P., & O'Connell, R. G. (2016). Abstract and effector-selective decision signals exhibit qualitatively distinct dynamics before delayed perceptual reports. *Journal of Neuroscience*, *36*(28), 7346–7352, https://doi.org/10.1523/JNEUROSCI.4162-15.2016.

Twomey, D. M., Murphy, P. R., Kelly, S. P., & O'connell, R. G. (2015). The classic P300 encodes a build-to-threshold decision variable. *European Journal of Neuroscience*, *42*(1), 1636–1643, https://doi.org/10.1111/ejn.12936.

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), 1–11, https://doi.org/10.1038/ncomms14637.

Vafaei Shooshtari, S., Esmaily Sadrabadi, J., Azizi, Z., & Ebrahimpour, R. (2019). Confidence Representation of Perceptual Decision by EEG and Eye Data in a Random Dot Motion Task. *Neuroscience*, *406*. https://doi.org/10.1016/j.neuroscience.2019.03.031

Van Den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *ELife*, *5*(FEBRUARY2016), 1–21. https://doi.org/10.7554/eLife.12192

This is a provisional file, not the final typeset article

Van den Berg, R., Yoo, A. H., & Ma, W. J. (2017). Fechner's law in metacognition: A quantitative model of visual working memory confidence. *Psychological Review*, *124*(2), 197, https://doi.org/10.1037/rev0000060.

van den Berg, R., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). Confidence Is the Bridge between Multi-stage Decisions. *Current Biology*, *26*(23), 3157–3168. https://doi.org/10.1016/j.cub.2016.10.021

Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, *36*(5), 955–968, https://doi.org/10.1016/S0896-6273(02)01092-9.

Waskom, M. L., & Kiani, R. (2018). Decision Making through Integration of Sensory Evidence at Prolonged Timescales. *Current Biology*, *28*(23), 3850–3856.e9. https://doi.org/10.1016/j.cub.2018.10.021

Wong, K.-F., & Huk, A. C. (2008). Temporal dynamics underlying perceptual decision making: Insights from the interplay between an attractor model and parietal neurophysiology. *Frontiers in Neuroscience*, *2*, 28, https://doi.org/10.3389/neuro.01.028.2008.

Zizlsperger, L., Sauvigny, T., Händel, B., & Haarmeier, T. (2014). Cortical representations of confidence in a visual perceptual decision. *Nature Communications*, *5*, 3940, https://doi.org/10.1038/ncomms4940.

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, *6*, 79, https://doi.org/10.3389/fnint.2012.00079.

Zylberberg, A., Fetsch, C. R., & Shadlen, M. N. (2016). The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. *ELife*, *5*(OCTOBER2016), e17688. https://doi.org/10.7554/eLife.17688

Zylberberg, A., Roelfsema, P. R., & Sigman, M. (2014). Variance misperception explains illusions of confidence in simple perceptual decisions. *Consciousness and Cognition*, *27*(1), 246–253. https://doi.org/10.1016/j.concog.2014.05.012

Zylberberg, A., Wolpert, D. M., & Shadlen, M. N. (2018). Counterfactual reasoning underlies the learning of priors in decision making. *Neuron*, *99*(5), 1083–1097, https://doi.org/10.1016/j.neuron.2018.07.035.

Supplementary Material

13 Supplementary Appendix 1: Figures and Tables

13.1 Supplementary Tables

Supplementary Table 1. Subtraction of confidence in double-pulse trials from single-pulse trials was significantly affected by choice accuracy.

Participant	$\boldsymbol{\beta_1}$
$\mathbf{P_1}$	0.17 ($p < 0.01$) CI = [.15, .19]
$\mathbf{P_2}$	0.21 ($p < 0.01$) CI = [.19, .23]
$\mathbf{P_3}$	0.10 ($p < 0.01$) CI = [.08, .12]
$\mathbf{P_4}$	0.12 ($p < 0.01$) CI = [.10, .14]

Each row shows the coefficients of Equation 5, the corresponding p values, and 95% confidence interval.

This is a provisional file, not the final typeset article

Supplementary Table 2. Performance was largely unaffected by interpulse interval for double-pulse trials with equal pulse strength and unequal pulse strength for each participant.

	Equal strength		**Unequal strength**		
Participant	β_3	β_4	β_3	β_4	β_5
P_1	0.02 (p = 0.73)	-0.01 (p = 0.91)	-0.05 (p = 0.58)	0.01 (p = 0.59)	-0.01 (p = 0.86)
	CI = [-0.09, 0.13]	CI = [-0.02, 0.01]	CI = [-0.18, 0.10]	CI = [-0.01, 0.02]	CI = [-0.03, 0.01]
P_2	-0.08 (p = 0.28)	-0.01 (p = 0.95)	-0.08 (p = 0.57)	-0.01 (p = 0.44)	-0.01 (p = 0.24)
	CI = [-0.21, 0.07]	CI = [-0.02, 0.02]	CI = [-0.26, 0.09]	CI = [-0.02, 0.02]	CI = [-0.02, 0.01]
P_3	-0.01 (p = 0.92)	-0.01 (p = 0.87)	0.01 (p = 0.35)	0.01 (p = 0.94)	-0.01 (p = 0.40)
	CI = [-0.07, 0.06]	CI = [-0.02, 0.01]	CI = [-0.08, 0.11]	CI = [-0.01, 0.02]	CI = [-0.02, 0.01]
P_4	0.07 (p = 0.16)	-0.01 (p = 0.04)	0.10 (p = 0.14)	-0.01 (p = 0.11)	-0.01 (p = 0.23)
	CI = [-0.02, 0.17]	CI = [-0.02, 0.00]	CI = [-0.03, 0.22]	CI = [-0.02, 0.01]	CI = [-0.02, 0.00]

Each row shows the coefficients of Equation 3, the corresponding p values, and 95% confidence interval of β_i.

Supplementary Table 3. Performance was unaffected by interpulse interval for double-pulse trials with equal pulse strength and unequal pulse strength included trials with zero interpulse interval and one of the nonzero interpulse interval.

	Equal strength		**Unequal strength**		
	β_3	β_4	β_3	β_4	β_5
0 vs. 120	-0.14 (p = 0.61)	-0.03 (p = 0.33)	-0.45 (p = 0.22)	0.03 (p = 0.18)	-0.01 (p = 0.71)
	CI = [-0.69, 0.41]	CI = [-0.09, 0.32]	CI = [-1.19, 0.27]	CI = [-0.02, 0.09]	CI = [-0.06, 0.04]
0 vs. 360	0.20 (p = 0.08)	0.03 (p = 0.04)	-0.03 (p = 0.79)	0.01 (p = 0.83)	-0.01 (p = 0.13)
	CI = [0.01, 0.39]	CI = [0.01, 0.06]	CI = [-0.27, 0.21]	CI = [-0.02, 0.02]	CI = [-0.03, 0.01]
0 vs. 1080	0.01 (p = 0.91)	-0.01 (p = 0.21)	-0.01 (p = 0.73)	0.01 (p = 0.98)	-0.01 (p = 0.58)
	CI = [-0.06, 0.07]	CI = [-0.02, 0.01]	CI = [-0.09, 0.07]	CI = [-0.01, 0.02]	CI = [-0.02, 0.01]

Each row shows the coefficients of Equation 3, the corresponding p values, and 95% confidence interval of β_i.

This is a provisional file, not the final typeset article

1182 **Supplementary Table 4.** Confidence was unaffected by interpulse interval for double-pulse trials
1183 across all possible coherence sequence.

	β_1
3.2%, 3.2%	0.01 (p = 0.60)
	CI = [-0.03, 0.05]
3.2%, 6.4%	-0.01 (p = 0.87)
	CI = [-0.05, 0.04]
6.4%, 3.2%	-0.01 (p = 0.65)
	CI = [-0.06, 0.03]
6.4%, 6.4%	-0.03 (p = 0.16)
	CI = [-0.07, 0.01]
3.2%, 12.8%	-0.01 (p = 0.65)
	CI = [-0.06, 0.03]
12.8%, 3.2%	-0.02 (p = 0.39)
	CI = [-0.07, 0.02]
6.4%, 12.8%	-0.02 (p = 0.29)
	CI = [-0.06, 0.02]
12.8%, 6.4%	-0.01 (p = 0.48)
	CI = [-0.06, 0.03]
12.8%, 12.8%	-0.02 (p = 0.27)
	CI = [-0.07, 0.02]

1184 Each row shows the coefficients of linear regression on confidence and interpulse interval, their related
1185 p values and a 95% confidence interval of β.

1186 **Supplementary Table 5.** The order of the pulses failed to show significant effects on confidence on
1187 any of interpulse intervals.

	β_2
0	0.01 (p = 0.08)
	CI = [-0.01, 0.02]
120	-0.01 (p = 0.48)
	CI = [-0.02, 0.01]
360	-0.01 (p = 0.51)
	CI = [-0.02, 0.01]
1080	0.01 (p = 0.37)
	CI = [-0.01, 0.02]

1188 Each row shows the coefficients of Equation 4, the corresponding p values. and 95% confidence
1189 interval of β_i.

This is a provisional file, not the final typeset article

Supplementary Table 6. Pairwise comparisons across models (1: single-pulse trials, 2: double-pulse trials, 3: perfect integrator) for SDT parameters.

	d'			$Meta\text{-}d'$			$Meta\text{-}d'/d'$		
	1 vs. 2	**1 vs. 3**	**2 vs. 3**	**1 vs. 2**	**1 vs. 3**	**2 vs. 3**	**1 vs. 2**	**1 vs. 3**	**2 vs. 3**
tstat	**3.79**	**4.05**	**1.25**	**1.98**	**0.21**	**2.48**	**1.74**	**0.12**	**3.58**
df	**22**	**22**	**22**	**22**	**22**	**22**	**22**	**22**	**22**
pValue	**0.001**	**0.52×10^{-5}**	**0.22**	**0.05**	**0.83**	**0.02**	**0.09**	**0.90**	**0.002**
95% CI	**[0.13, 0.44]**	**[0.20, 0.62]**	**[-.08, 0.32]**	**[-.02, 1.12]**	**[-.66, 0.82]**	**[.10, 1.15]**	**[-.17, 1.98]**	**[-1.09, 1.23]**	**[.35, 1.32]**
Cohen's *d*	**1.55**	**1.65**	**- 0.51**	**0.81**	**0.09**	**1.01**	**0.71**	**0.05**	**1.46**

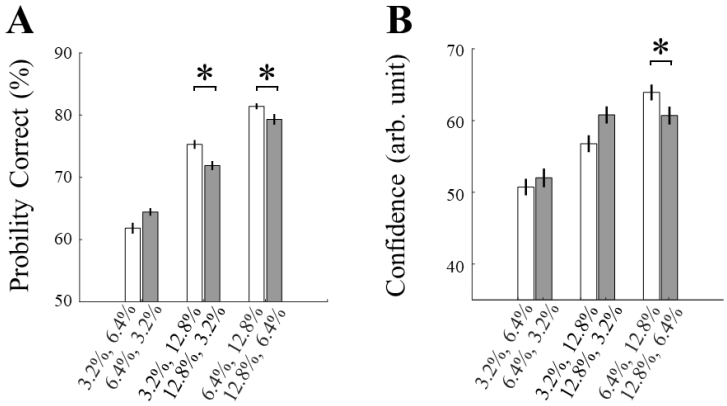
1193 **Supplementary Table 7.** d', $Meta\text{-}d'$, *and* $Meta\text{-}d'/d'$ was unaffected by interpulse interval for
1194 double-pulse trials in both experiments.

		d'	$Meta\text{-}d'$	$Meta\text{-}d'/d'$
Experiment 1	β_i	-0.08 (p = 0.33)	-0.23 (p = 0.09)	-0.18 (p = 0.16)
		CI = [-0.22, 0.07]	CI = [-0.48, 0.03]	CI = [-0.43, 0.07]
Experiment 2	β_i	-0.07 (p = 0.38)	0.09 (p = 0.49)	0.15 (p = 0.30)
		CI = [-0.23, 0.08]	CI = [-0.17, 0.35]	CI = [-0.13, 0.43]

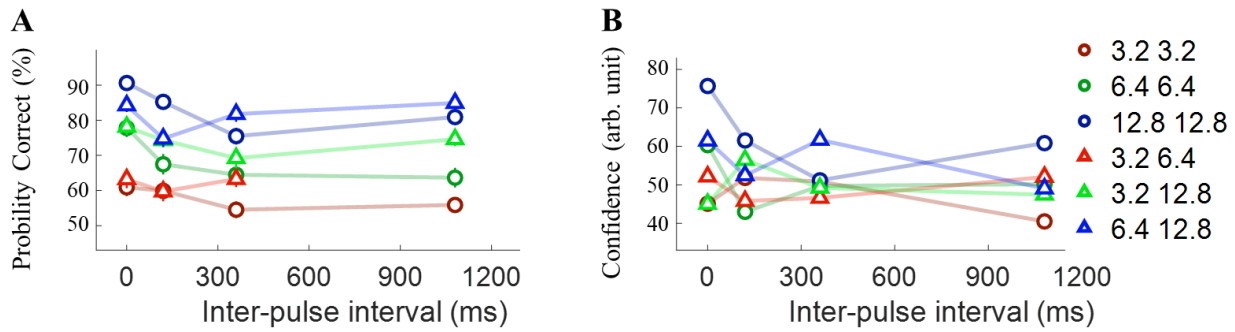
1195 Each cell shows the coefficients of linear regression on SDT's parameters and interpulse interval, the
1196 corresponding p values and 95% confidence interval of β_i.

This is a provisional file, not the final typeset article

13.2 Supplementary Figures



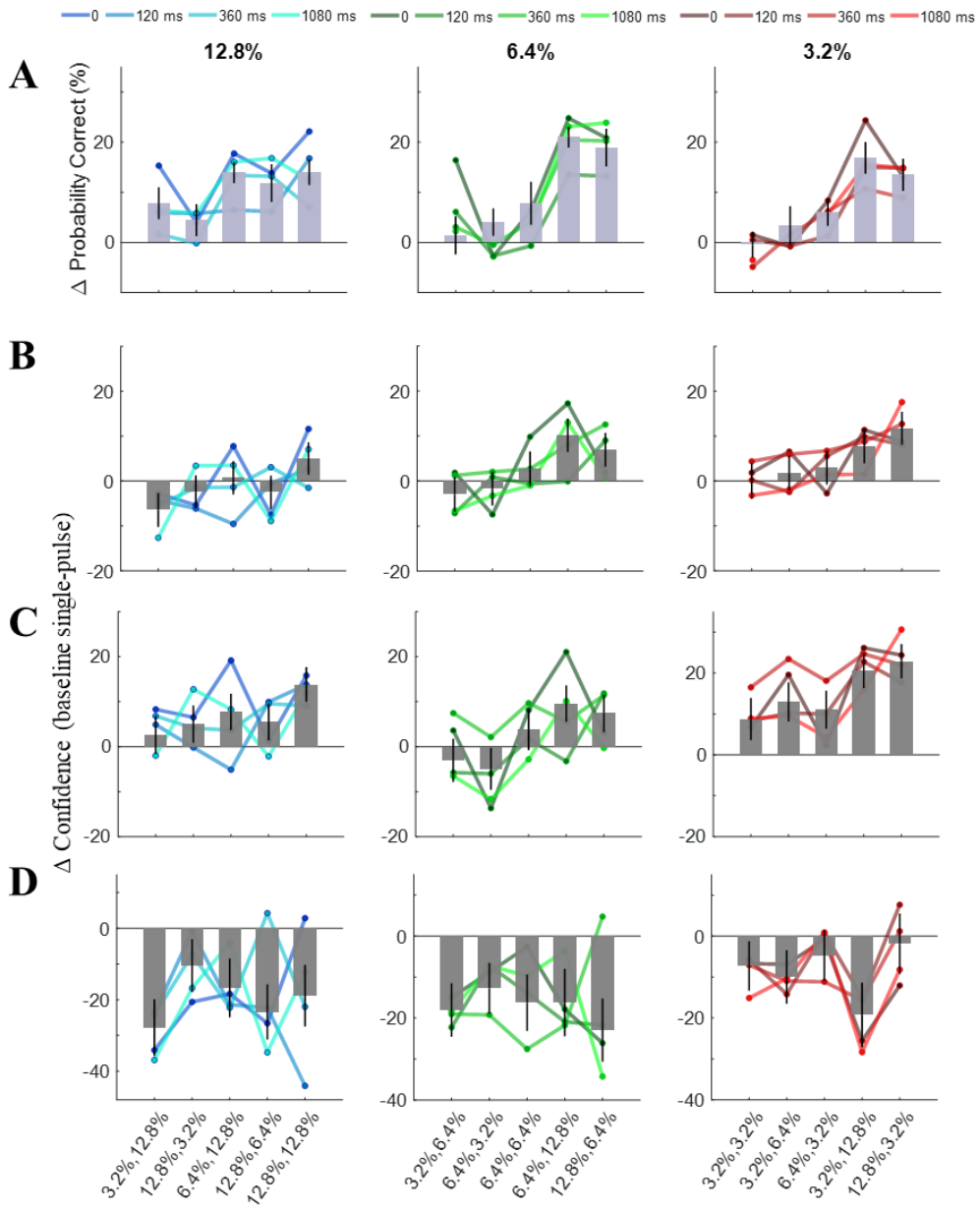
Supplementary Figure 1. Choice confidence was not depended on the sequence of motion pulses in Experiment 2. **(A)** The weak–strong pulse sequence contributed higher accuracy than the strong–weak sequence. **(B)** The weak–strong pulse sequence did not contribute higher confidence compared to the strong–weak sequence. In all panels, data are represented as group mean ± SEM. (*p<0.05).



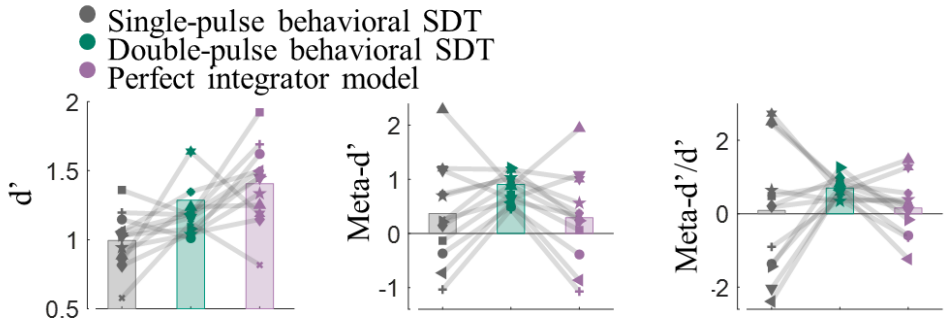
Supplementary Figure 2. The Interplay between confidence, accuracy, and interpulse intervals in double-pulse trials in Experiment 2. **(A)** Choice accuracy for double-pulse trials grouping in all possible interval conditions. **(B)** Pooling data across all time intervals calculated the confidence of double-pulse trials. Each data point addresses pooled data from indicated sequence pulse and its reverse order (e.g., 12.8– 3.2% and 3.2 –12.8%).

1199

This is a provisional file, not the final typeset article



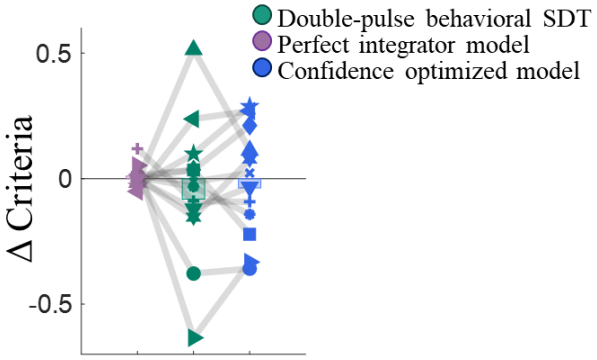
Supplementary Figure 3. Variation of accuracy or confidence in double-pulse trials baselined by the corresponding coherence (3.2%, 6.4% and 12.8% for each column) in Experiment 2. **(A)** Considering all the trials, accuracy improved in combination with almost all pulses compared to the baseline. **(B)** Considering all the trials, confidence improved in combination with stronger pulses, whereas the confidence in sequence with a weaker pulse either decreased or remained constant. **(C)** In correct-choice trials, the increasing effect of stronger pulses is more obvious and the confidence even slightly improved in combination with weaker pulses compared to the corresponding baseline. **(D)** Interestingly, in incorrect trials, the confidence decreased in every condition. The colored line representing matching data for each of four possible gaps. In the bar graph, the data are represented as group mean ± SEM.



Supplementary Figure 4. Comparison of models and human behavior in Experiment 2. Stimulus sensitivity (d'), metacognitive sensitivity ($Meta\text{-}d'$), and metacognitive efficiency ($Meta\text{-}d'/d'$) estimated for single-pulse trials, double-pulse trials and the perfect integrator models.

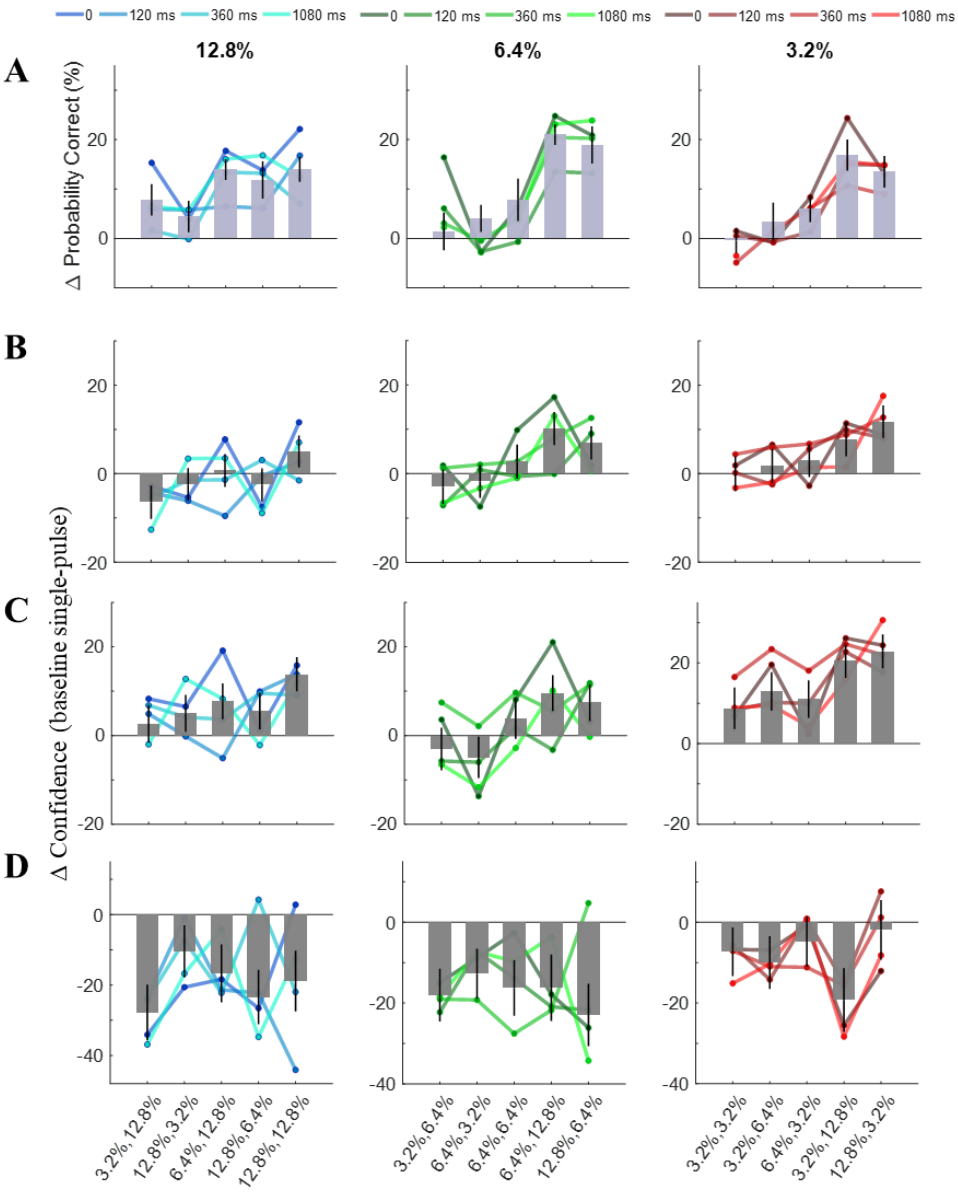
A univariate ANOVA showed that d' between models fitted to double or single-pulse trials and the perfect integrator model significantly differed (F(2,33) = 9.99; $p = 0.41 \times 10^{-4}$). Also, a univariate ANOVA showed that $Meta\text{-}d'$ between models fitted to double-pulse or single-pulse trials and the perfect integrator model partially differed (F(2,33) = 1.04; $p = 0.09$). We also computed metacognitive efficiency ($Meta\text{-}d'/d'$), A univariate ANOVA revealed an insignificant difference on all three models (F(2,33) = 2.50; $p = 0.10$).), We also applied the *t*-test as a post hoc procedure to compare all pairs of d', $Meta\text{-}d'$, $Meta\text{-}d'/d'$ calculated in three models (**Supplementary Table 5**).

This is a provisional file, not the final typeset article



Supplementary Figure 5. Variation of confidence criteria baselined to single-pulse trials for the perfect integrator model vs. the double-pulse trials model and the type 2 optimized model in Experiment 2.

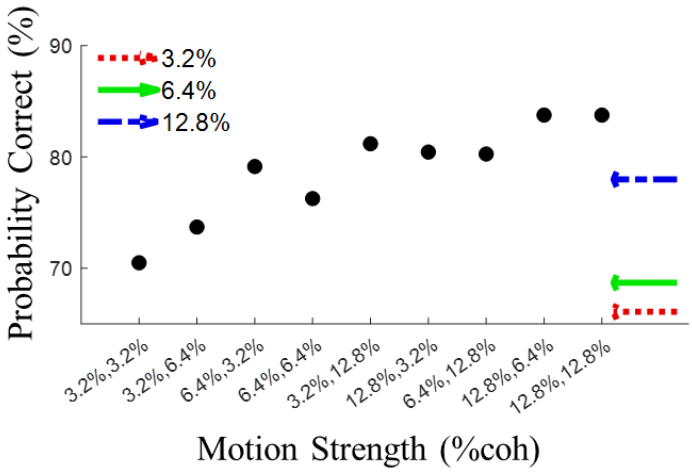
1210



Supplementary Figure 6. Variation of accuracy or confidence in double-pulse trials baselined by the corresponding coherence (3.2%, 6.4%, and 12.8% for each column) for the different interpulse intervals in Experiment 1. **(A)** Considering all the trials, accuracy improved in combination with almost all pulses compared to the baseline. **(B)** Considering all the trials, confidence improved in combination with stronger pulses, whereas the confidence in sequence with a weaker pulse either decreased or remained constant. **(C)** In correct-choice trials, the increasing effect of stronger pulses is more significant and the confidence even slightly improved in combination with weaker pulses compared to the corresponding baseline. **(D)** Interestingly, in incorrect trials, the confidence decreased in every condition—the colored line representing matching data for each of four possible gaps. In the bar graph, the data are represented as group mean ± SEM.

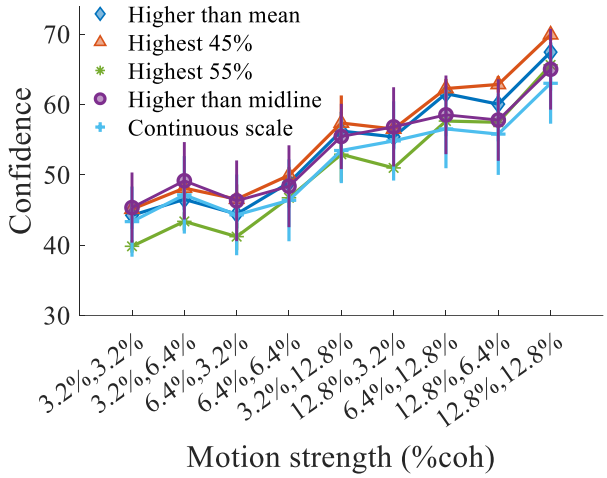
1211

This is a provisional file, not the final typeset article



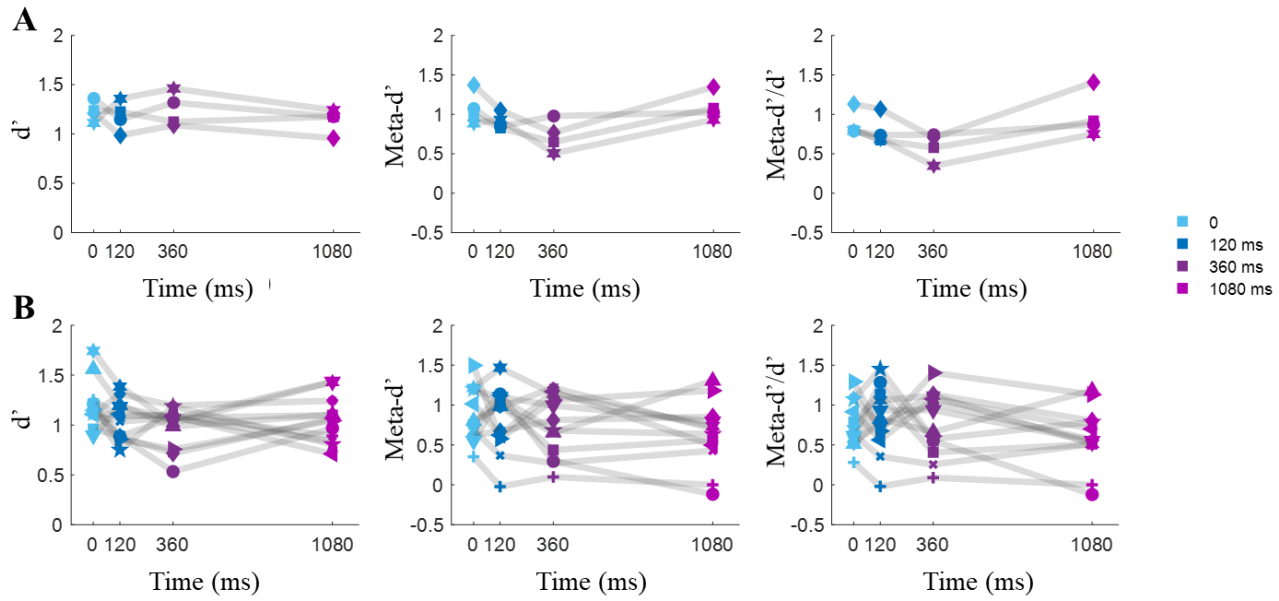
Supplementary Figure 7. Accuracy in single and double pulse(s) trials when the same number of trials from each coherence with the same confidence range is selected randomly.

1212



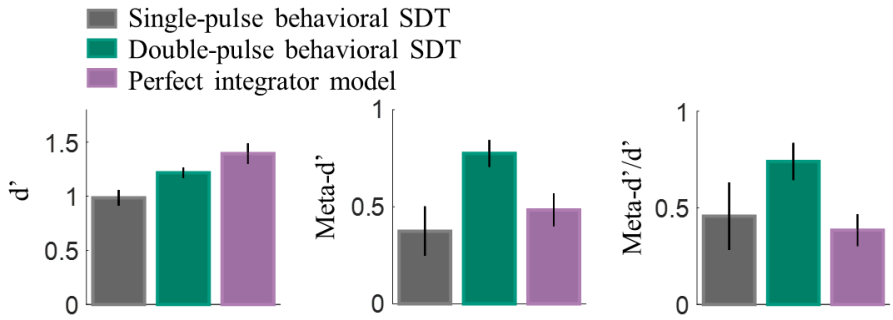
Supplementary Figure 8. Graded confidence and four methods of confidence categorization. A univariant ANOVA showed that graded confidence and probability of high confidence calculated by four different approaches for each participant in double-pulse trials did not significantly differ ($F(4,145) = 7.99$; $p = 0.41$). Four paired-samples *t*-tests between the probability of high confidence calculated by different categorization methods and graded confidence showed no difference (all *ps* > 0.10).

This is a provisional file, not the final typeset article



Supplementary Figure 9. Stimulus sensitivity (d'), metacognitive sensitivity ($Meta\text{-}d'$) and, metacognitive efficiency ($Meta\text{-}d'/d'$) estimated for double-pulse trials were independent of the interpulse interval in **(A)** Experiment 1 and **(B)** Experiment 2. Each dot represents the data of each participant.

1215



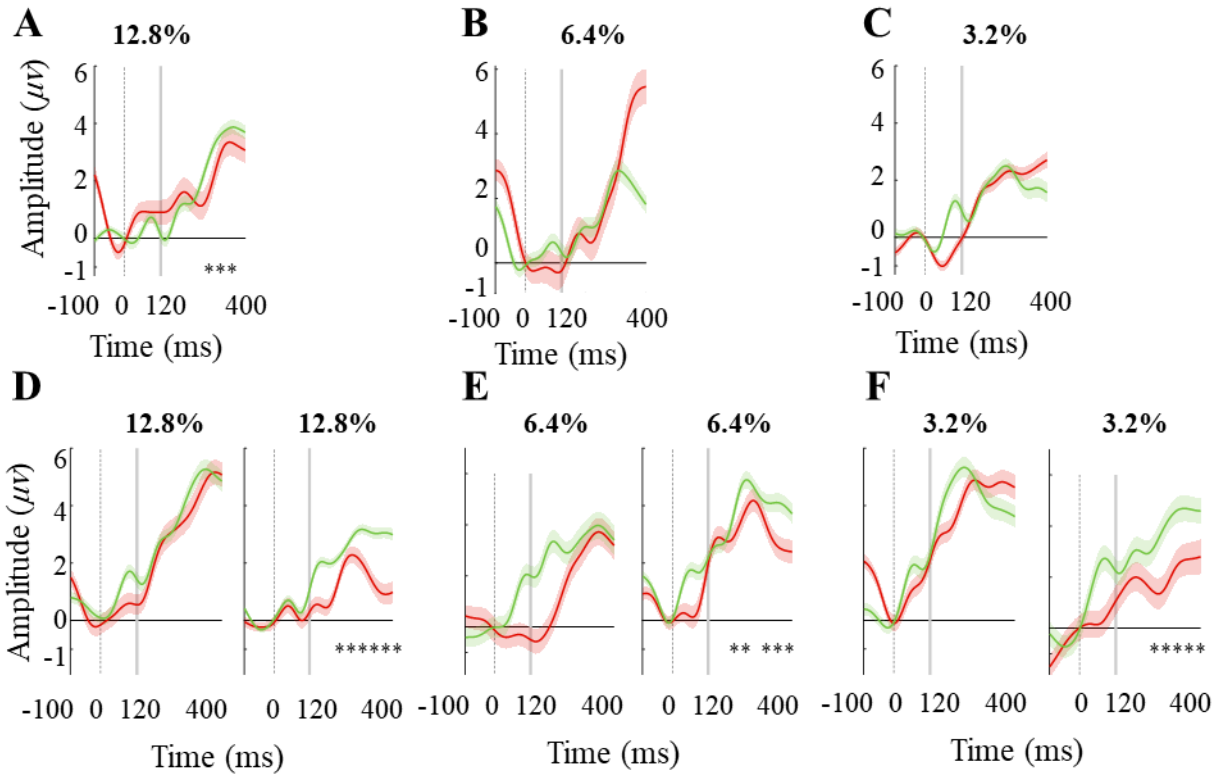
Supplementary Figure 10. Comparison of models and human behavior considering the same numbers of trials in Experiment 1. Stimulus sensitivity (d'), metacognitive sensitivity ($Meta-d'$), and metacognitive efficiency ($Meta-d'/d'$) estimated for single-pulse trials, double-pulse trials, and the perfect integrator model.

We compared d', $Meta-d'$ and, $Meta-d'/d'$ of fitted models to single/double-pulse trials and simulated data by the perfect integrator model, following up with three Dunn pair tests. A Kruskal-Wallis test showed that d' between models fitted to double/single-pulse trials and the perfect integrator model did not significantly differ ($H(3) = 3.23$; $p = 0.20$). We also applied the Dunn test as a post hoc procedure to compare all pairs of d' from three models. No d' in models significantly differed from others (all $ps > 0.21$).

Also, a Kruskal-Wallis test showed that $Meta-d'$ between models fit to double/single-pulse trials and the perfect integrator model significantly differed ($H(3) = 6.96$; $p = 0.03$). Post-hoc Dunn was used to compare all pairs of $Meta-d'$ from three models. The difference of $Meta-d'$ of single-pulse and double-pulse trials was significant ($p = 0.03$, $CI = [-12.58, -0.41]$). However, the difference of $Meta-d'$ was insignificant for single-pulse trials and the perfect integrator model ($p = 0.87$, $CI = [-7.83, 4.33]$) and for double-pulse trials and the perfect integrator model ($p = 0.17$, $CI = [4.75, 10.83]$).

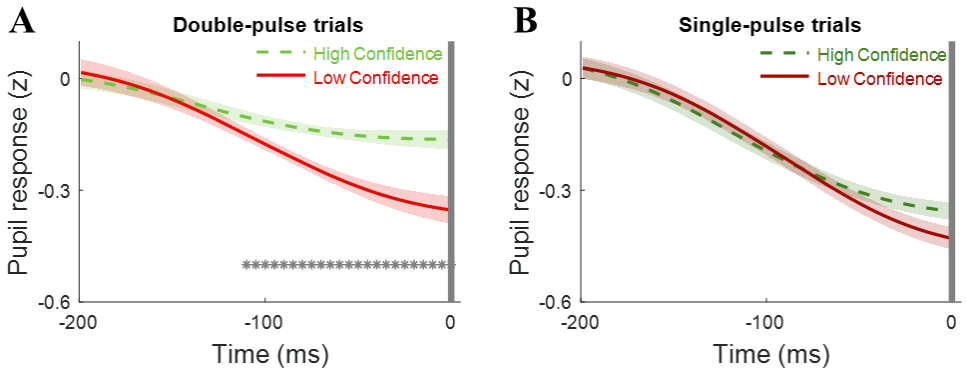
We also computed metacognitive efficiency ($Meta-d'/d'$). A Kruskal-Wallis test revealed a significant difference on all three models ($H(3) = 7.42$, $p = 0.02$). Metacognitive efficiency in double-pulse and the perfect integrator model differed significantly ($p = 0.04$, $CI = [0.16\ 12.33]$) while in double-pulse and single-pulse models, partially differed ($p = 0.07$, $CI = [-11.83, 0.33]$). The difference of single-pulse and the perfect integrator model was not significant ($p = 0.99$, $CI = [-5.58\ 6.58]$).

This is a provisional file, not the final typeset article



Supplementary Figure 11. ERPs in single and double-pulse trials in correct trials. **(A) (B) (C)** ERPs of correct single-pulse trials shows an insignificant difference in weaker motion strength in high and low confidence level trials. **(D) (E) (F)** ERPs of correct trials in the two levels of confidence are distinct after the stimulus onset. The shading region around the mean indicates SEM. * indicate p<.05 from a *t*-test of the difference between the two-time.

1236



Supplementary Figure 12. Standardized pupil response across time-window aligned to the feedback in correct trials, high confidence trials (green) vs. low confidence trials (red) in: **(A)** double-pulse trials, **(B)** single-pulse trials. The shading region around the mean indicates SEM. * indicate $p < .05$.

1237

This is a provisional file, not the final typeset article

14 Supplementary Appendix 2: Signal detection theory models

In the binary decision, the observer must discriminate between stimuli labeled S_2 or labeled S_1. Each stimulus presentation generates a value on an internal decision axis (**Figure 1B**), corresponding to the evidence in favor of S_1 or S_2. Evidence generated by each stimulus class is normally distributed across the decision axis, and the distance between these distributions in standard deviation units (d') which measures how well the observer can discriminate S_1 from S_2 (Maniscalco & Lau, 2012). The observer sets a decision criterion cr, such that all signals exceeding cr are labeled S_2 and all those failing to exceed cr are labeled S_1. The observer also sets criteria $cr_{2,"S1"}$ and $cr_{2,"S2"}$ to determine confidence ratings around the decision criterion cr. These two thresholds must be well-ordered so that $\mathrm{cr}_{2,"S1"} < cr < \mathrm{cr}_{2,"S2"}$ (**Figure 1B**). When a S_2 response is made, a confident S_2 response requires the evidence also to have surpassed the $cr_{2,"S2"}$ threshold (Maniscalco & Lau, 2012, 2014).

14.1 Confidence *Hit Rate* and *False Alarm Rate*

Sweeping the $\mathrm{cr}_{2,"S2"}$ criterion across the decision axis generates different values of confidence false alarm rate ($Prob(conf = "h" \mid stim \neq resp)$ and confidence hit rate ($Prob(conf = "h" \mid stim = resp)$. A summary of the observer's confidence performance is provided by hit rate (Hit Rate2) and false alarm rate (False Alarm Rate2) (Maniscalco & Lau, 2014):

$$\begin{aligned} \text{Hit Rate2} &= \text{Prob}(\text{conf} = "h" \mid \text{stim} = \text{resp}) = \frac{\text{n(high conf correct)}}{\text{n(correct)}}, \\ \text{False Alarm Rate2} &= \text{Prob}(\text{conf} = "h" \mid \text{stim} \neq \text{resp}) = \frac{\text{n(high conf incorrect)}}{\text{n(incorrect)}}, \end{aligned} \tag{S1}$$

where $n(cond)$ denotes a count of the total number of trials satisfying the condition $cond$.

14.2 Decision *Hit Rate* and *False Alarm Rate*

In the SDT model, the decision hit rate (Hit Rate1) and the decision false alarm rate (False Alarm Rate1) are also calculated as follows:

$$\begin{aligned} \text{Hit Rate1} &= \frac{\text{n(resp=}Si,\ \text{stim=}Si\ )}{\text{n(stim = }Si)}, \ \text{i = 1,2} \\ \text{False Alarm Rate1} &= \frac{\text{n(resp=}Si,\ \text{stim=}Sj\ )}{\text{n(stim = }Sj)}, \ \text{i = 1, j = 2 or i = 2, j = 1} \end{aligned} \tag{S2}$$

where i and j represent the stimulus classification. After calculating the Hit Rate1 and False Alarm Rate1 of each participant, d' and cr are calculated as follows for each participant:

$$d' = \phi^{-1}(\text{Hit Rate1}, 0, 1) - \phi^{-1}(\text{False Alarm Rate1}, 0, 1)$$
$$cr = -0.5 * [\phi^{-1}(\text{Hit Rate1}, 0, 1) + \phi^{-1}(\text{False Alarm Rate1}, 0, 1)] \quad \text{(S3)}$$

here, ϕ^{-1} is the inverse of a function that represents a normal cumulative distribution and is calculated as follows:

$$\phi(s, \mu, \sigma) = \int_{-\infty}^{0} N(v, \mu, \sigma) dv, \quad \text{(S4)}$$

where $N(v, \mu, \sigma)$ is a Normal distribution with mean (μ) and standard deviation (σ). After the above computations, to simplify, we may consider the value of cr as zero point and move the distribution diagrams related to each option on the axis of the evidence.

By setting d', cr and two criteria $cr_{2,"S1"}$ and $cr_{2,"S2"}$ (**Figure 1B**), the probabilities of each confidence rating conditional on a given stimulus and response (Hit Rate2 and False Alarm Rate2) can be calculated theoretically according to the following equations:

$$\text{Prob}(\text{conf} = "h" | \text{stim} = S1, \text{resp} = "S1") = \text{HitRate2}_{"S1"} = \frac{\phi\left(\text{cr}_{2,"S1"}, -\frac{\text{d}'}{2}\right)}{\phi\left(\text{cr}, -\frac{\text{d}'}{2}\right)}$$

$$\text{Prob}(\text{conf} = "h" | \text{stim} = S2, \text{resp} = "S1") = \text{FalseAlarmRate2}_{"S1"} = \frac{\phi\left(\text{cr}_{2,"S1"}, \frac{\text{d}'}{2}\right)}{\phi\left(\text{cr}, \frac{\text{d}'}{2}\right)}$$

$$\text{Prob}(\text{conf} = "h" | \text{stim} = S1, \text{resp} = "S2") = \text{HitRate2}_{"S2"} = \frac{1 - \phi\left(\text{cr}_{2,"S2"}, \frac{\text{d}'}{2}\right)}{1 - \phi\left(\text{cr}, \frac{\text{d}'}{2}\right)} \quad \text{(S5)}$$

$$\text{Prob}(\text{conf} = "h" | \text{stim} = S2, \text{resp} = "S2") = \text{FalseAlarmRate2}_{"S2"} = \frac{1 - \phi\left(\text{cr}_{2,"S2"}, -\frac{\text{d}'}{2}\right)}{1 - \phi\left(\text{cr}, -\frac{\text{d}'}{2}\right)}$$

This is a provisional file, not the final typeset article

In the SDT model, there are different methods for adjusting the model with the data observed. In the method we used, d' and cr were calculated from the participants' performance (Eq. S3). Then, using maximum likelihood estimation (MLE) and Eq. S1 and S5 and by altering the value of the confidence criteria while holding d' and cr constant, a set of (Hit Rate2, False Alarm Rate2) pairs ranging between (0, 0) and (1, 1) were generated. Moreover, $Meta\text{-}d'$ was found by fitting the decision SDT model to response-specific confidence.

15 Supplementary references

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430, https://doi.org/10.1016/j.concog.2011.09.021.

Maniscalco, B., & Lau, H. (2014). Signal detection theory analysis of type 1 and type 2 data: meta-d′, response-specific meta-d′, and the unequal variance SDT model. In *The cognitive neuroscience of metacognition* (pp. 25–66). Springer, https://doi.org/10.1007/978-3-642-45190-4_3.