AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2024, Vol. 153, No. 3, 656–688
https://doi.org/10.1037/xge0001524

How Do Humans Give Confidence? A Comprehensive Comparison of Process Models of Perceptual Metacognition

Medha Shekhar and Dobromir Rahnev
School of Psychology, Georgia Institute of Technology

Humans have the metacognitive ability to assess the accuracy of their decisions via confidence judgments. Several computational models of confidence have been developed but not enough has been done to compare these models, making it difficult to adjudicate between them. Here, we compare 14 popular models of confidence that make various assumptions, such as confidence being derived from postdecisional evidence, from positive (decision-congruent) evidence, from posterior probability computations, or from a separate decision-making system for metacognitive judgments. We fit all models to three large experiments in which subjects completed a basic perceptual task with confidence ratings. In Experiments 1 and 2, the best-fitting model was the lognormal meta noise (LogN) model, which postulates that confidence is selectively corrupted by signal-dependent noise. However, in Experiment 3, the positive evidence (PE) model provided the best fits. We evaluated a new model combining the two consistently best-performing models—LogN and the weighted evidence and visibility (WEV). The resulting model, which we call logWEV, outperformed its individual counterparts and the PE model across all data sets, offering a better, more generalizable explanation for these data. Parameter and model recovery analyses showed mostly good recoverability but with important exceptions carrying implications for our ability to discriminate between models. Finally, we evaluated each model's ability to explain different patterns in the data, which led to additional insight into their performances. These results comprehensively characterize the relative adequacy of current confidence models to fit data from basic perceptual tasks and highlight the most plausible mechanisms underlying confidence generation.

Public Significance Statement

Several process models have attempted to describe the computations that underlie metacognition in humans. However, due to lack of systematic, widespread comparisons between these models, there is no consensus on what mechanisms best characterize the process of confidence generation. In this study, we tested 14 popular models of metacognition on three large data sets from basic perceptual tasks, using multiple quantitative as well as qualitative metrics. Our results highlight two mechanisms as the most plausible, generalizable features of confidence—the selective corruption of confidence by signal-dependent metacognitive noise and a heuristic strategy that uses stimulus visibility to estimate confidence. Analyzing the qualitative patterns of confidence generated by the models provides additional insights into each model's success or failure. Our results also help to establish a comprehensive framework for model comparisons that can guide future efforts.

Keywords: metacognition, confidence, perceptual decision-making, computational modeling, metacognitive noise

Supplemental materials: https://doi.org/10.1037/xge0001524.supp

As humans, we are capable of metacognitively evaluating the quality of our own decisions via confidence estimates (Metcalfe & Shimamura, 1994). A person with higher metacognitive ability has greater insight into their decisions, reliably expressing lower confidence for incorrect and higher confidence for correct decisions. Accurate metacognitive evaluations can drive learning and

This article was published Online First December 14, 2023.
Medha Shekhar https://orcid.org/0000-0002-5205-7274
Parts of this work have previously been presented at the Vision Sciences Society Annual Conference in St. Pete Beach, Florida, United States (2020), the Association for Scientific Studies of Consciousness (ASSC) Annual Conference held online (2021), and the Perceptual Metacognition Satellite to the ASSC Annual Conference in Amsterdam, the Netherlands (2022). This work was supported by the National Institutes of Health (Award R01MH119189) and the Office of Naval Research (Award N00014-20-1-2622). The authors declared no competing interests.

Medha Shekhar served as lead for formal analysis, investigation, methodology, and writing–original draft. Dobromir Rahnev served as lead for funding acquisition, project administration, and supervision. Medha Shekhar and Dobromir Rahnev contributed equally to conceptualization and writing–review and editing.

Correspondence concerning this article should be addressed to Medha Shekhar, School of Psychology, Georgia Institute of Technology, 654 Cherry Street NW, Atlanta, GA 30332, United States. Email: medha@gatech.edu

information seeking in ways that maximize our chances of achieving successful outcomes (Desender et al., 2018; Fleming, Dolan, & Frith, 2012; Koriat, 2006; Nelson & Narens, 1990; Shimamura, 2000; Yeung & Summerfield, 2012). Thus, insight into the processes underlying metacognition is important for both understanding and improving people's decision-making.

One of the most important avenues for understanding the computational mechanisms of metacognition involves the development and comparison of process models of metacognition (Fleming & Daw, 2017; Jang et al., 2012; Maniscalco & Lau, 2016; Pleskac & Busemeyer, 2010; Shekhar & Rahnev, 2021b). These models precisely specify a set of internal computations that generate the observed choice and confidence judgments, and thus offer mechanistic insight into the processes underlying metacognition. Among many other issues, process models of confidence generation allow us to determine the internal computations that transform an incoming sensory signal into a choice and a confidence rating, as well as how the processes that generate choice and confidence are related to each other. Yet, while many different models of metacognition have been developed in recent years, less attention has been given to comparing and arbitrating between the full range of existing models. Here, we set to perform a comprehensive comparison of existing models of metacognition that could then allow us to gain greater insight into the processes underlying metacognition.

Models of Metacognition

The last decade has witnessed a large proliferation of models of metacognition designed to explain subjects' choice and confidence judgments (J. W. Bang et al., 2019; Fleming & Daw, 2017; Jang et al., 2012; Maniscalco et al., 2016; Maniscalco & Lau, 2016; Shekhar & Rahnev, 2021b). We performed an extensive search for published process models of metacognition that can be tested by fitting them to choice and confidence data. We only selected models that are currently in use (e.g., we did not consider threshold models developed in the first half of the 20th century that are now largely abandoned). In the end, this process resulted in selecting 14 models of metacognition (Table 1). These models represent a wide range of hypotheses about the mechanisms underlying metacognition and make diverse assumptions about the architecture of information processing and computations that govern choice and confidence judgments. Below, we provide a brief overview of the 14 models; detailed descriptions are available in the Method.

Perhaps the most prominent among all 14 models is signal detection theory (SDT model; Green & Swets, 1966). In fact, 13 of the 14 models selected here are built in some way on top of SDT. The classic SDT model assumes the presence of Gaussian noise in the internal perceptual representations. Both the decision and confidence are made by placing noiseless criteria on the internal activations.

Three different models have proposed relatively simple extensions of SDT where confidence ratings are made less informative. The most popular among these (J. W. Bang et al., 2019; Maniscalco & Lau, 2016; Shekhar & Rahnev, 2018, 2021a, 2021b) assumes the existence of Gaussian metacognitive noise (Gauss model) that corrupts either the signal or criteria for confidence. One prominent variant of the Gauss model postulates that, in addition to the Gaussian metacognitive noise, the signal on which the choice is made decays before the confidence rating can be given (Decay model; Maniscalco & Lau, 2016), thus leading to two sources of corruption. Finally, a more recent model also postulates the existence of metacognitive noise but assumes that this noise follows a lognormal distribution (LogN model; Shekhar & Rahnev, 2021b). Increasing the mean of a lognormal distribution also increases its variance, thus making the confidence criteria far from the decision criterion noisier. Because of this property, we refer to the metacognitive noise in the LogN model as being signal-dependent.

Another six models have proposed more substantive extensions of SDT. The first one assumes the presence of additional, postdecisional processing that adds more information before confidence is given (Post-Dec model; Barrett et al., 2013). The second model proposes that confidence is based on a weighted sum of evidence from SDT and stimulus visibility (weighted evidence andvisibility [WEV]) model; Rausch et al., 2018, 2020). Finally, the last two models from this group assume that confidence is based only on choice-congruent (positive) evidence and ignores choice-incongruent (negative) evidence (Maniscalco et al., 2016; Zylberberg et al., 2012). A complete disregard of choice-incongruent evidence results in a pure positive evidence (PE) model, while a partial disregard of choice-incongruent evidence leads to a more flexible version of the PE model (PE-Flex model). A very popular model—the Bayesian confidence hypothesis (BCH)—proposes that confidence is computed as the posterior probability of a correct choice (BCH model; Sanders et al., 2016). Finally, a recent model proposes that confidence represents an observer's estimate of uncertainty in the decision variable but that the metacognitive system only has access to a noisy estimate of this uncertainty (confidence as a noisy decision reliability estimate [CASANDRE] model; Boundy-Singer et al., 2023).

The next three models propose larger departures from SDT by postulating the existence of two different pathways or decision-making systems (in comparison to all models above which are based on a single evidence stream). The most popular among these is the dual channel model (DC model), which assumes that high- and low-confidence decisions are based on separate information processing pathways called "conscious" and "unconscious" channels (del Cul et al., 2009; Maniscalco & Lau, 2016). The other two models both assume that the signals for the primary decision and confidence arise from correlated but separate Gaussian random variables. In the stochastic detection and retrieval model (SDRM; Jang et al., 2012), confidence is given by placing criteria directly on the confidence signal. In the second-order confidence (SOC model; Fleming & Daw, 2017), confidence is given based on a Bayesian computation of the probability of being correct, similar to the BCH model. However, instead of directly computing this probability from the decision variable, confidence is estimated by a second-order inference of the possible states of the decision variable via the observed confidence variable.

Finally, one model originally designed to fit response times (RT) data—the two-stage dynamic signal detection (2DSD model; Pleskac & Busemeyer, 2010)—has a structure that allows it to be fit only to choice and confidence data, and is therefore included here too. Conceptually, 2DSD is related to the Post-Dec model in that it assumes that noisy confidence evidence is accumulated after the decision was made.

Previous Work Comparing Existing Models of Metacognition

Despite the abundance of process models of metacognition, relatively few studies have compared a wide range of existing models with most model comparison work instead focusing on comparing

Table 1
Process Models of Confidence Generation

Model	Full name	Key reference	Description
SDT	Signal detection theory	Green and Swets (1966)	Confidence and choice are based on the same evidence
Gauss	Gaussian meta noise	Maniscalco and Lau (2016)	Confidence is selectively corrupted by Gaussian metacognitive noise
LogN	Lognormal meta noise	Shekhar and Rahnev (2021b)	Confidence is selectively corrupted by lognormal (signal-dependent) metacognitive noise
Decay	Noisy decay	Maniscalco and Lau (2016)	Signal for confidence undergoes Decay; corruption by Gaussian meta noise
Post-Dec	Postdecisional SDT	Barrett et al. (2013)	Additional sample of noisy evidence is added to the signal for confidence
2DSD	Two-stage dynamic signal detection	Pleskac and Busemeyer (2010)	Confidence and choice are based on two sequential stages of evidence accumulation
PE	Positive evidence	Zylberberg et al. (2012), Maniscalco et al. (2016)	Confidence is based only on decision-congruent evidence
PE-Flex	Flexible positive evidence	—	Confidence gives more weight to decision-congruent evidence than optimal
WEV	Weighted evidence and visibility	Rausch et al. (2018)	Signal for confidence is a weighted sum of evidence and stimulus visibility
DC	Dual channel	del cul et al. (2009), Maniscalco and Lau (2016)	Choices associated with low and high confidence (conscious experience) are generated by independent pathways
SDRM	Stochastic detection and retrieval model	Jang et al. (2012)	Choice and confidence based on distinct but correlated evidence samples
BCH	Bayesian confidence hypothesis	Hangya et al. (2016)	Confidence is computed as the posterior probability of being correct
SOC	Second order confidence	Fleming and Daw (2017)	Confidence is generated as posterior probability of correct choice based on higher order inference
CASANDRE	Confidence as a noisy decision reliability estimate	Boundy-Singer et al. (2023)	Confidence depends on a noisy estimate of the reliability of the decision

Note. Details on the 12 models of confidence generation examined here. SDT = signal detection theory model; Gauss = Gaussian meta noise model; LogN = lognormal meta noise model; Decay = noisy decay model; Post-Dec = postdecisional SDT model; 2DSD = two-stage dynamic signal detection model; PE = positive evidence bias model; PE-Flex = flexible positive evidence bias model; WEV = weighted evidence and visibility model; DC = dual channel model; SDRM = stochastic detection and retrieval model; BCH = Bayesian confidence hypothesis model; SOC = second-order confidence model; CASANDRE = confidence as a noisy decision reliability estimate model.

several variants of the same model or only a few different models. The earliest model comparison of a wide range of models was performed by Maniscalco and Lau (2016) who fit several versions of three broad classes of models (single channel, DC, and hierarchical) to visibility data from a metacontrast masking experiment. Their models included SDT, Gauss, Decay, and several variants of the DC model with the Decay model performing the best. Two studies by Rausch and colleagues used confidence data from a masking task and fit the SDT, Gauss, Post-Dec, Decay, PE, DC, and WEV models (Rausch et al., 2018, 2020), finding that the WEV model provided the best fits. However, none of these studies included the newly developed LogN, CASANDRE, and SOC models, or other older models (BCH, SDRM, and 2DSD). In fact, to the best of our knowledge, despite its popularity, the SOC model has never even been fit to real data. Furthermore, none of the studies above performed parameter recovery, which is important for characterizing the identifiability of models (i.e., whether different sets of parameters can produce the same data; Wilson & Collins, 2019), and some of the studies did not perform model recovery. Therefore, building on the previous research, here we sought to conduct systematic model comparison using an even wider range of models and perform both parameter and model recovery analyses.

The Current Approach

An endeavor to conduct a comprehensive model comparison necessarily involves many choices on the part of the authors. Some of the most important choices include what type of task manipulations to use, what type of data to fit, and what variations of existing models to include. We briefly discuss the choices we made in each of these domains, as well as the advantages and disadvantages of these choices.

Task Manipulations

A critical step for any model comparison is choosing the types of task manipulations to include. One possible approach is to include manipulations that produce nontrivial qualitative results, which can provide strong falsification for models that cannot explain these results. This is the approach that many studies to date have taken (Maniscalco & Lau, 2016; Rausch et al., 2018, 2020). A weakness of this approach is that such manipulations may impact the auxiliary assumptions of the models. For example, it is not well understood whether the backward masking designs used in several previous studies affect the variability of the internal distributions. While it is possible for model comparisons to be extended such that different combinations of auxiliary assumptions are tested for each model, in practice this results in a very large set of models and is rarely done. Here, we instead chose to examine data with no manipulations at all (Experiment 1), as well as data with very simple manipulations of difficulty via stimulus contrast (Experiment 2) and motion coherence (Experiment 3). While this approach does not produce nontrivial qualitative results, it obviates the need to test a large number of auxiliary assumptions.

Type of Data to Fit

Another critical choice an experimenter needs to make concerns what type of data to fit. Most models to date operate only on choice

and confidence data. However, several models have been developed to jointly explain choice, confidence, and reaction time (Kiani et al., 2014; Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2013; Vickers, 1979). Beyond these three measures, it would be desirable for a model to explain as much observable data as possible, including pupillometry (e.g., pupil dilation), physiological measures (e.g., heart rate and skin conductance), and brain measures (e.g., electroencephalography and functional magnetic resonance imaging activity across sensory and higher-order areas) since these measures are also related to confidence. In practice, however, no model can explain every observable piece of data and it is unclear whether it would be advantageous to even try to build such a model. While explaining more observable data is certainly an asset for a model, it adds an important layer of complexity and can shift the focus away from the most basic phenomena. Here we chose to maximally narrow our focus and fit models to choice and confidence data only. It is our expectation that the obtained results will inspire and constrain any model that is designed to fit a wider range of data. However, it should be noted that our decision to exclude RT results in the exclusion of most models that are based on the drift-diffusion modeling framework (except 2DSD). Accordingly, a majority of the remaining models we include here are based on SDT, since SDT remains one of the most popular frameworks for modeling two-choice tasks, particularly when RTs are excluded.

The 2DSD model was originally developed to simultaneously explain choice, confidence, and RT. However, in the current study, we use the model to generate only choice and confidence, while suppressing RT. We acknowledge that constraining the model in this way might result in the model being underutilized. However, given that we cannot fit any of the other models to RT, it becomes necessary to suppress RTs in the 2DSD model in order to fit it in our study.

Model Variants

Finally, for most models in the literature, it is possible to identify several model variants. For example, SDRM has three sources of noise (variability in the decision criterion, variability in the confidence criteria, and imperfect correlation between the choice and confidence variable) and one could test model variants where between zero and three of these noise sources are included (Jang et al., 2012). Similarly, concepts such as the existence of two channels (Maniscalco & Lau, 2016) can be implemented in several different ways. To keep the model comparison manageable, here we chose to only fit the main version of each model. We only made one exception for the PE model where, consistent with previous findings (Maniscalco et al., 2016), we included a version where choice-incongruent evidence is partially considered (PE-Flex model).

Summary of Study Design and Results

We performed the most comprehensive investigation to date on the ability of existing models of metacognition to fit choice and confidence data from simple perceptual experiments. We fit all 14 models to three data sets that are substantially larger than in previous work (Experiment 1: 59 subjects, 3,500 trials/subject; Experiment 2: 20 subjects, 2,800 trials/subject; Experiment 3: 45 subjects, 1,600 trials/subject) and therefore allow for stronger conclusions. All experiments featured simple perceptual tasks with Experiment 1 having a single difficulty level, Experiment 2 having three difficulty levels obtained by manipulating the contrast of a Gabor patch stimulus, and Experiment 3 having eight difficulty levels obtained from manipulating motion coherence in a random dot motion stimulus. Model recovery analyses showed that most models (except SDRM) can be reliably discriminated from each other, but the recoverability decreases substantially when only a single difficulty level is used. On the other hand, parameter recovery analyses showed that most models are uniquely identifiable, though three models (Decay, SDRM, and SOC) showed poor recovery of at least one model parameter. Furthermore, we explored the extent to which each model can capture qualitative patterns and found that most models failed to capture at least one qualitative effect. The only exception was the WEV model that flexibly captured all the observed qualitative patterns across the three experiments. Critically, the best-fitting model for Experiments 1 and 2 was the LogN model, but the PE model was found to provide the best fit for Experiment 3. Additionally, we tested a new model that combines features from two of the most consistently performing models (LogN and WEV), and found that it outperformed the PE model as well as its individual counterparts, suggesting an alternative mechanism for confidence that can be generalized across data sets. These findings are consistent with many recent findings on the neural mechanisms of confidence generation involving the prefrontal cortex (Allen et al., 2017; D. Bang & Fleming, 2018; Fleming et al., 2010; Fleming, Huijgen, & Dolan, 2012; Morales et al., 2018; Rounis et al., 2010; Shekhar & Rahnev, 2018; Wokke et al., 2017; Yeon et al., 2020) and constrain hypotheses about the neural architecture supporting metacognition.

Method

Data for Modeling

We fit the models to data from three experiments that contained a large number of trials and involved a perceptual discrimination with confidence ratings. All experiments have been previously reported: Experiment 1 as Experiment 2 in Haddara and Rahnev (2022), Experiment 2 as Experiment 4 in Shekhar and Rahnev (2021b), and Experiment 3 as Experiment 1 in Orchard et al. (2022). Data from all experiments have been made available through the Confidence Database (Rahnev et al., 2020). This study was not preregistered. In all experiments, subjects reported normal or corrected-to-normal vision and received monetary compensation for their participation. The studies were approved by the local Institutional Review Board. All study details for the experiments can be found in the original publications; below, we briefly discuss the basic experimental designs.

Experiment 1

A total of 75 subjects completed the experiment over the course of 7 days but 15 of them were excluded from the original publication (Haddara & Rahnev, 2022) based on preregistered criteria. In the current study, we excluded one additional subject because their metacognitive score lay outside the plausible range of values (M-Ratio = −0.5; a description of the measure M-Ratio is provided under the Analysis section). Subjects were recruited online via Amazon's Mechanical Turk and each completed a total of 3,500 trials (500 trials per day; 20 blocks of 25 trials each). The first six blocks of trials (150 trials) on Day 1 were part of a staircasing procedure to adjust the task difficulty for the rest of the experiment and were therefore excluded from the main analyses, resulting in 3,350 trials per subject.

Each trial began with subjects fixating on a small white cross at the center of the screen followed by presentation of a 7 × 7 grid containing the letters X and O for 500 ms (Figure 1a). After the stimulus offset, subjects indicated which of the two letters appeared more frequently in the grid and then rated their confidence on a scale from 1 to 4 (1 = *low confidence*, 4 = *high confidence*). Subjects' responses were untimed and collected via key presses. The average accuracy across all 7 days of the experiment was 75%. The experiment was designed using jsPsych 5.0.3.

Experiment 2

A total of 20 subjects participated in this experiment. Each subject came for three sessions held on separate days and completed a total of 2,800 trials. Each trial began with subjects fixating on a small white dot at the center of the screen for 500 ms followed by presentation of the stimulus for 100 ms (Figure 1b). The stimulus was a Gabor patch (diameter = 3°) oriented either to the left (counterclockwise) or right (clockwise) of the vertical by 45°. The Gabor patches were superimposed on a noisy background. The response screen appeared after the stimulus offset and remained till the subjects made a response. Subjects' task was to indicate the direction of the tilt (left/right) and simultaneously rate their confidence using a continuous confidence scale (ranging from 50% correct to 100% correct for each type of response) via a single mouse click. Because all models fit here require a discrete set of confidence criteria to generate ratings, we transformed the continuous confidence scale into a 6-point scale, using five equidistant criteria placed between the lowest (50) and highest (100) possible ratings. This procedure is equivalent to what was used in the original publication (Shekhar & Rahnev, 2021b).

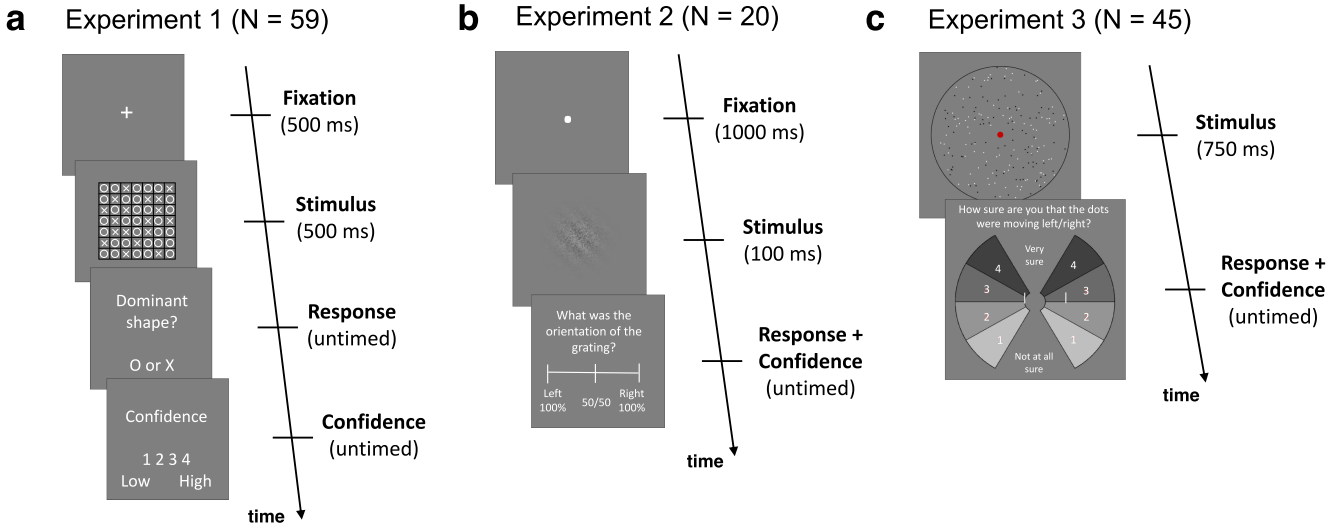
Three interleaved contrast values of 4.5%, 6%, and 8% were used. The three levels of contrast yielded three increasing levels of accuracy (Contrast 1: $M = 67\%$, $SD = 2.7\%$; Contrast 2: $M = 77\%$, $SD = 3.7\%$; Contrast 3: $M = 89\%$, $SD = 3.6\%$). In addition, to incentivize the veridical use of the continuous confidence scale, participants were awarded points based on how closely their confidence reports matched their accuracy (Fleming et al., 2016). At the end of the three sessions, participants' cumulative scores were calculated and they were rewarded a bonus based on their performance.

Stimuli were generated using the Psychophysics Toolbox (Brainard, 1997) in MATLAB (MathWorks) and presented on a computer monitor (21.5-in. display, 1,920 × 1,080 pixel resolution, 60 Hz refresh rate). Subjects were seated in a dim room and positioned 60 cm away from the screen.

Experiment 3

Forty-five subjects participated in this experiment and completed a total of 1,600 trials within a single session. Each trial began with the presentation of the stimulus for 750 ms followed by the presentation of a response screen. The stimulus consisted of a random dot kinematogram—a circular aperture containing 200 moving dots (100 black, 100 white; diameter = 0.07) and a red central fixation mark (diameter = 0.35°). Each dot moved in a direction that was randomly chosen from a wrapped Gaussian distribution with a mean of 22° either to the left or right of the vertical. Subjects were required to simultaneously indicate the direction of motion (left/right) and their

Figure 1
Schematic of the Tasks in Experiments 1–3



Note. (a) Task in Experiment 1. Each trial began with fixation for 500 ms and was followed by the presentation of a grid containing the letters "X" and "O." Subjects had to first indicate which of these two letters appeared more frequently in the grid and then rate their confidence in their response on a 4-point scale. (b) Task in Experiment 2. Each trial began with fixation for 1 s followed by the presentation of a noisy Gabor patch tilted 45° either to the left or right of the vertical for 100 ms. Subjects were instructed to indicate the tilt of the Gabor patch while simultaneously rating their confidence on a continuous scale from 50 to 100. (c) Task in Experiment 3. Each trial began with presentation of the stimulus for 750 ms. The stimulus consisted of 200 black and white dots moving within a circular aperture. Subjects simultaneously indicated the net direction of dot motion (left/right) and their confidence on a 4-point scale by clicking on a wedge in the response screen. Figure adapted from "Internal Noise Measures in Coarse and Fine Motion Direction Discrimination Tasks and the Correlation With Autism Traits," by E. R. Orchard, S. C. Dakin, and J. J. A. van Boxtel, 2022, *Journal of Vision*, *22*(10), Article 19 (https://doi.org/10.1167/jov.22.10.19). CC BY 4.0. See the online article for the color version of this figure.

confidence on a 4-point scale by a button press. The response screen contained a confidence wheel (Figure 1c) divided into eight wedges. The four wedges on the left/right indicated leftward/rightward motion, respectively, and were numbered 1–4 to indicate the level of confidence for that response.

The trials consisted of eight interleaved difficulty levels. Difficulty was manipulated by varying the degree of randomness (coherence) in dot motion. Specifically, each dot moved in a direction drawn randomly from a Gaussian distribution with $M \pm 22°$ and an *SD* of 0°, 35°, 45°, 60°, 70°, 80°, 90°, or 100°. When the standard deviation was 0°, all the dots moved in the same direction. The eight levels of coherence yielded eight decreasing levels of accuracy (98.8%, 96.4%, 93.4%, 82.19%, 74%, 66.4%, 61%, and 56.9%).

Model Details

We fit 14 process models:

1. SDT model
2. Gaussian meta noise (Gauss) model
3. Lognormal meta noise (LogN) model
4. Noisy decay (Decay) model
5. Postdecisional SDT (Post-Dec) model
6. WEV model
7. PE bias model
8. PE-Flex bias model
9. 2DSD model
10. BCH model
11. Confidence as a noisy decision reliability estimate (CASA NDRE) model
12. DC model
13. SDRM
14. SOC model

Models 1–11 postulate the existence of a single system for the generation of choice and confidence (single process models), while Models 12–14 assume distinct decision-making systems/pathways for generating decisions (dual process models). Below, we give details about each of the 14 models.

SDT Model

We implement the classic SDT model described by Green and Swets (1966) under the assumption that confidence criteria are placed on the evidence axis. SDT assumes that each trial generates noisy sensory evidence, r_{sens}, drawn from a Gaussian distribution such that, $r_{sens} = N\left(-\frac{\mu_{sens}}{2}, \sigma^2_{sens}\right)$, when the stimulus belongs to the first category, S_1, and $r_{sens} = N\left(\frac{\mu_{sens}}{2}, \sigma^2_{sens}\right)$, when the stimulus belongs to the second category, S_2. Here, μ_{sens} is the distance between the evidence distributions corresponding to the two stimulus categories and σ_{sens} is the standard deviation of each distribution. The primary decision about the identity of the stimulus is generated by comparing r_{sens} with a decision criterion, c_0 such that $r_{sens} < c_0$ leads to a response "S_1" and $r_{sens} \geq 0$ leads to a response "S_2."

Confidence decisions are generated by a set of confidence criteria, $[c_{-n}, c_{-n+1}, \ldots, c_{-1}, c_1, \ldots, c_{n-1}, c_n]$, where n is the number of ratings on the confidence scale. The criteria c_i are monotonically increasing with $c_{-n} = -\infty$ and $c_n = \infty$. When the primary response is "S_2," confidence is generated using the criteria $[c_0, c_1 \ldots, c_n]$ such that r_{conf} falling within the interval $[c_i, c_{i+1})$ results in a confidence of $i + 1$. When the primary response is "S_1," confidence is generated using the criteria $[c_{-n}, c_{-n+1}, \ldots, c_0]$ such that r_{conf} falling within the interval $[c_i, c_{i+1})$ results in a confidence of $-i$.

Critically, SDT assumes that the stimulus and confidence decisions are based on the same sensory evidence, such that the signal for confidence $r_{conf} = r_{sens}$ and there is no additional noise is generating confidence.

Gaussian Meta Noise Model (Gauss)

The Gauss model is equivalent to SDT with the exception of the additional assumption that the sensory evidence underlying confidence, r_{conf}, is selectively corrupted by additional "metacognitive noise" such that, $r_{conf} = N(r_{sens}, \sigma^2_{meta})$, where σ_{meta} is the amount of metacognitive noise added to the sensory signal (J. W. Bang et al., 2019; Maniscalco & Lau, 2016; Shekhar & Rahnev, 2018, 2021b). The model we implement here is the version that we have used previously in Shekhar and Rahnev (2018). Note that another interpretation of the same model is that the metacognitive noise is added to the confidence criteria rather than the sensory evidence; the two interpretations result in equivalent models (Shekhar & Rahnev, 2021b).

The Gauss model can sometimes generate apparently nonsensical scenarios where metacognitive noise causes r_{sens} and r_{conf} to lie on opposite sides of the decision criterion c_0. For example, $r_{sens} < c_0$ (and thus the Type-1 response is "S_1") but $r_{conf} > c_0$ (which corresponds Type-1 response "S_2"). When such "cross-over" scenarios arise, the Gaussian metanoise model is constrained to generate a confidence of 1.

LogN Model

The LogN model is equivalent to SDT with the exception of the additional assumption that metacognitive noise drawn from a lognormal distribution is added to the confidence criteria (first implemented in Shekhar & Rahnev, 2021b). The confidence criteria, c_i, thus follow the following lognormal probability distribution, $g_{lognormal}$:

$$c_i \sim g_{lognormal}(x \mid \mu_i, \sigma^2_{meta}) = \begin{cases} \frac{1}{(x - c_0)\sqrt{2\pi\sigma^2_{meta}}} e^{-\frac{(\ln(x-c_0)-\mu_i)^2}{2\sigma^2_{meta}}}, x \in (c_0, \infty) & \text{if } i > 0 \\ \frac{1}{(c_0 - x)\sqrt{2\pi\sigma^2_{meta}}} e^{-\frac{(\ln(c_0-x)+\mu_i)^2}{2\sigma^2_{meta}}}, x \in (-\infty, c_0) & \text{if } i < 0, \end{cases} \quad (1)$$

where μ_i and σ^2_{meta} are the mean and variance of the Gaussian random variable obtained by taking log of c_i and $i = -n + 1, \ldots, -1, 1, \ldots, n - 1$. The parameters μ_i are constrained so that $\mu_{-n+1} \leq \cdots \leq \mu_{-1}$ and $\mu_1 \leq \cdots \leq \mu_{n-1}$. The confidence criteria, c_i, are generated as perfectly correlated random variables ensuring that the criteria, $[c_{-n}, c_{-n+1}, \ldots, c_{-1}, c_1, \ldots, c_{n-1}, c_n]$, are strictly increasing and do no cross over each other. Additionally, all confidence criteria, c_i, are bounded on one end by c_0 such that, there they do not cross over the decision criterion (as in the Gauss model). The variance of each confidence criterion, c_i, is given by $(e^{\sigma^2_{meta}} - 1) \times e^{2\mu_i + \sigma^2_{meta}}$ when $i > 0$ and by $(e^{\sigma^2_{meta}} - 1) \times e^{-2\mu_i + \sigma^2_{meta}}$ when $i < 0$, implying that variance of the confidence criteria scales with their distance from the decision criterion.

We note that it is possible to alternatively model metacognitive noise as noise in the confidence signal, rather than noise in the confidence criteria. Here, we chose to model metacognitive noise as trial-by-trial variability in the confidence criteria because we believe that criterion noise is the more likely source of metacognitive noise. Mechanistically, we can explain criterion noise as arising from observers' being unable to hold stable criteria across trials (random criterion jitter) or from using other unmodeled sources of information when computing confidence. On the other hand, the assumption that metacognitive noise arises from the addition of noise to the confidence signal does not offer such an immediate mechanistic explanation for its source. Additionally, it assumes the existence of a second-order representation of the confidence signal that is corrupted by noise of unknown origin.

Noisy Decay Model (Decay)

The Decay model is equivalent to Gauss model (it assumes the presence of Gaussian metacognitive noise) but additionally postulates that the sensory signal underlying the perceptual decision, r_{sens}, undergoes a process of decay before reaching the stage of confidence generation. We follow the implementation that is described in Maniscalco and Lau (2016). Therefore, the sensory signal for confidence is given by the formula $r_{conf} = N(r_{sens} \times p_{decay}, \sigma^2_{meta})$, where $p_{decay} \in [0, 1]$, represents the proportion of the signal that is retained for confidence and σ_{meta} is the metacognitive noise.

Postdecisional SDT Model (Post-Dec)

The Post-Dec model is equivalent to SDT with the exception of the additional assumption that the signal underlying confidence, r_{conf}, contains the original signal for the choice, r_{sens}, as well as an additional sample of postdecisional evidence. Therefore,

$$r_{conf} \sim \begin{cases} N\left(r_{sens} - \frac{\delta_{post} \times \mu_{stim}}{2}, \delta_{post}\right), & \text{if stimulus} = S_1 \\ N\left(r_{sens} + \frac{\delta_{post} \times \mu_{stim}}{2}, \delta_{post}\right), & \text{if stimulus} = S_2 \end{cases}, \tag{2}$$

where δ_{post} controls the amount of postdecisional evidence added to r_{sens}. The Post-Dec model we describe here was first implemented by Rausch et al. (2020).

WEV Model

The WEV model was first developed by Rausch et al. (2018). Here, we use the version of their model that is described in Rausch et al. (2020). The WEV model is equivalent to the Gauss model (it assumes the presence of Gaussian metacognitive noise) but postulates that confidence is also influenced by stimulus visibility along with the original sensory signal (Rausch et al., 2018). Therefore, the signal underlying confidence is computed as a weighted average of the sensory signal (r_{sens}) and a visibility signal such that:

Equation 3 (see below)

where w_{vis} refers to the weight given to stimulus visibility, $\bar{\mu}_{stim}$ refers to the mean stimulus signal across all experimental conditions, and σ_{meta} refers to the Gaussian metacognitive noise in the confidence sample. To obtain a measure of stimulus visibility, the stimulus strength on the current trial is compared to the average stimulus signal across all experimental conditions. Due to the need for averaging signal strengths across experimental conditions, the WEV model requires that the data must consist of at least two conditions with varying difficulty levels (otherwise, the model becomes equivalent to the Gauss model). Therefore, the WEV model cannot be fit to Experiment 1 which consists of a single difficulty level and we only fit this model to Experiments 2 and 3.

PE Model

The PE model assumes that confidence ratings are based only on the evidence that favors the chosen response. We use a version of this model that has been previously described by Maniscalco et al. (2016). The PE model considers separately the internal evidence for each of the two choices, r_{S_1} and r_{S_2}, such that:

$$\begin{aligned} r_{s_1} &\sim N\left(\frac{\mu_{stim}}{2} - c_0, \sigma^2_{sens}\right) \& r_{s_2} \sim N(c_0, \sigma^2_{sens}), \quad \text{if stimulus} = S_1 \\ r_{s_1} &\sim N(-c_0, \sigma^2_{sens}) \& r_{s_2} \sim N\left(\frac{\mu_{stim}}{2} + c_0, \sigma^2_{sens}\right), \quad \text{if stimulus} = S_2 \end{aligned}, \tag{4}$$

where c_0 reflects the observer's inherent bias to favor one response over another. Perceptual decisions are generated based on comparisons between r_{s_1} and r_{s_2} such that $r_{s_1} > r_{s_2}$ results in a response of "S_1" and $r_{s_1} \leq r_{s_2}$ results in a response of "S_2."

To generate the signal for confidence, the model only considers evidence in favor of the response that has been made, such that the sensory signal underlying confidence is given as: $r_{conf} = r_{congruent}$, where $r_{congruent} = r_{S_1}$ when stimulus response is "S_1" and $r_{congruent} = r_{S_2}$ when stimulus response is "S_2." The confidence decision is generated by defining a distinct set of confidence criteria for each response type—$[c_{i_0}, c_{i_1}, c_{i_2} \ldots c_{i_{n-1}}, c_{i_n}]$, where n is the number of ratings on the confidence scale and $i = \{1, 2\}$ refers to the class of stimulus responses ("S_1" or "S_2"). The criteria c_{ij} are monotonically increasing with $c_{i_0} = -\infty$ and $c_{i_n} = \infty$ when $i = 2$ and monotonically decreasing with $c_{i_0} = \infty$ and $c_{i_n} = -\infty$ when $i = 1$. Confidence responses are generated such that r_{conf} falling within the interval $[c_{i_j}, c_{i_{j+1}})$ results in a confidence of $j + 1$.

$$r_{conf} = \begin{cases} N\left((1 - w_{vis}) \times r_{sens} - w_{vis} \times \frac{(\mu_{stim} - \bar{\mu}_{stim})}{2}, \sigma^2_{meta}\right), & \text{if response} = S_1 \\ N\left((1 - w_{vis}) \times r_{sens} + w_{vis} \times \frac{(\mu_{stim} - \bar{\mu}_{stim})}{2}, \sigma^2_{meta}\right), & \text{if response} = S_2 \end{cases}, \tag{3}$$

PE-Flex Model

We also implemented a more PE-Flex model in which the confidence decision selectively overweighs evidence in favor of the response that has been made. The model was inspired by previous findings that the PE bias can be ameliorated via feedback and perhaps other factors (Maniscalco et al., 2016). However, the version of the PE model we implement here has never been used before. The PE-Flex model is identical to the PE model except that the confidence signal, r_{conf}, is a weighted difference of decision-congruent and incongruent evidence such that: $r_{\text{conf}} = w_{pe} \times r_{\text{congruent}} - (1 - w_{pe}) \times r_{\text{incongruent}}$, where $w_{pe} \in [.51, 1]$ controls how much weight is given to decision-congruent evidence (setting w_{pe} to 1 makes this model identical to the PE model).

2DSD Model

The original 2DSD model generates choice and confidence along with reaction times via a diffusion process (Pleskac & Busemeyer, 2010). In this model, the choice and RT are generated as in Ratcliff's diffusion model (Ratcliff & McKoon, 2008), whereas confidence is generated by running the diffusion process for an additional period after the choice has been made. However, since it is difficult to compare models fit to different sets of data (one that include and do not include RT), we suppress the RT generation in the 2DSD model while preserving the way the model generates choice and confidence.

Choice probabilities that arise from a diffusion process have been previously derived by Ratcliff (1978) as:

$$p(R_1|S_1) = \frac{e^{\left(\frac{4\delta\theta}{\sigma^2}\right)} - e^{\left(\frac{2\delta(\theta - z)}{\sigma^2}\right)}}{e^{\left(\frac{4\delta\theta}{\sigma^2}\right)} - 1}$$
$$p(R_2|S_2) = \frac{e^{\left(\frac{-4\delta\theta}{\sigma^2}\right)} - e^{\left(\frac{-2\delta(\theta + z)}{\sigma^2}\right)}}{e^{\left(\frac{-4\delta\theta}{\sigma^2}\right)} - 1}, \quad (5)$$

where $p(R_i|S_j)$ corresponds to the probability of responding "R_i" when stimulus S_j is presented, δ is the diffusion rate which controls the rate at which the sensory signal gets incorporated into the decision, z is the starting point of the diffusion process (a positive value of z results in a bias toward "S_2" responses), θ is the choice threshold which controls the amount of evidence required to make a choice, and σ is the drift coefficient that controls the noise in the diffusion process. We can then calculate the probability of making incorrect choices simply as: $p(R_2|S_1) = 1 - p(R_1|S_1)$ and $p(R_1|S_2) = 1 - p(R_2|S_2)$.

2DSD assumes that postdecisional evidence accumulation continues after the perceptual choice and that confidence is derived from the evidence that is collected during the second stage of diffusion. Pleskac and Busemeyer (2010) show that due to the noisy accumulation process, at the end of the second stage of diffusion, the final accumulated evidence for confidence (r_{conf}) follows a Gaussian distribution, such that when stimulus S_1 is presented:

$$r_{\text{conf}} \sim \begin{cases} N(-\tau\delta - \theta,\ \sigma^2\tau), & \text{if response} = S_1 \\ N(-\tau\delta + \theta,\ \sigma^2\tau), & \text{if response} = S_2 \end{cases}, \quad (6)$$

where τ is the strength of the postdecisional signal. Similarly, when the stimulus S_2 is presented:

$$r_{\text{conf}} \sim \begin{cases} N(\tau\delta - \theta,\ \sigma^2\tau), & \text{if response} = S_1 \\ N(\tau\delta + \theta,\ \sigma^2\tau), & \text{if response} = S_2 \end{cases}. \quad (7)$$

Since choice and confidence in this model are generated from distinct processes, the confidence criteria are not constrained by the decision criterion. Therefore, the set of criteria that generate confidence for "S_1" responses and the criteria that generate confidence for "S_2" responses form distinct sets, each extending from $-\infty$ to ∞. For each response type, we define the criteria as $[c_{i_0},\ c_{i_1},\ c_{i_2} \ldots c_{i_{n-1}},\ c_{i_n}]$, where n is the number of ratings on the confidence scale and $i = \{1, 2\}$ refers to the class of stimulus responses, S_i. The criteria c_{ij} are monotonically increasing with $c_{i_0} = -\infty$ and $c_{i_n} = \infty$ when $i = 2$ and monotonically decreasing with $c_{i_0} = \infty$ and $c_{i_n} = -\infty$ when $i = 1$. Confidence responses are generated such that r_{conf} falling within the interval $[c_{i_j},\ c_{i_{j+1}})$ results in a confidence of $j + 1$.

BCH Model

A Bayesian observer computes the posterior probability associated with each stimulus class and chooses the stimulus class that maximizes the posterior. However, in all our experiments both stimulus classes are equally likely to occur (i.e., priors are equal). Therefore, the posterior probability of a choice depends only the likelihood of the observed evidence under each stimulus class and the primary decision depends only on comparison of likelihoods, and choices are generated identical to the SDT model.

The posterior probability of each choice is computed as:

$$P(S_i|r_{\text{sens}}) = \frac{1}{e^{-\frac{2r_{\text{sens}}\mu_i}{\sigma^2_{\text{sens}}}} + 1}. \quad (8)$$

Assuming Gaussian evidence distributions where $r_{\text{sens}} \sim N(\mu_i, \sigma^2_{\text{sens}})$, $i = [1, 2]$, and $\mu_i = -\frac{\mu_{\text{stim}}}{2}$ for stimuli belonging to S_1, and $\mu_i = \frac{\mu_{\text{stim}}}{2}$ for stimuli belonging to S_2. The observer generates choices by comparing the posterior odds to a decision criterion, such that "S_1" responses are generated when the posterior odds, $\frac{P(S_1|r)}{P(S_2|r)} > t_0$ and "S_2" responses are generated otherwise, where $t_0 \in [0, 1]$. In the BCH model, confidence is defined as the posterior probability of a correct choice. Therefore, confidence is generated by comparing the maximum posterior estimate: $\max\{P(S_1|r), P(S_2|r)\}$ to a set of confidence criteria defined in the posterior probability space, $[t_{-n}, t_{-n+1}, \ldots, t_{-1}, t, \ldots, t_{n-1}, t_n]$, where n is the number of ratings on the confidence scale. The criteria are monotonically increasing with $t_{-n} = 0$ and $t_n = 1$. When the primary response is "S_2," confidence is generated using the criteria $[t_0, t_1, \ldots, t_n]$, such that $P(S_2|r)$ falling within the interval $[t_i, t_{i+1})$ results in a confidence of $i + 1$. When the primary response is "S_1," confidence is generated using the criteria $[t_{-n}, t_{-n+1}, \ldots, t_0]$, such that $P(S_1|r)$ falling within the interval $[t_i, t_{i+1})$ results in a confidence of $-i$. It should be noted that the BCH model becomes discriminable from the SDT model only in the presence of more than one task condition and under the assumption that confidence criteria remain fixed across the different task conditions. While the SDT model assumes that confidence criteria are fixed in the evidence space, the BCH model assumes that criteria are fixed in the posterior probability space—leading to differences in how different task conditions interact with confidence. Therefore, the BCH was not fit to Experiment 1 which consists of a single difficulty level and we only fit this model to Experiments 2 and 3.

CASANDRE Model

Generation of the primary choice in the CASANDRE model follows standard SDT assumptions. For confidence, the model assumes an additional stage of processing based on the observer's estimate of the reliability of their choices (Boundy-Singer et al., 2023). Therefore, in CASANDRE, the confidence variable represents choice reliability, r_{conf}, as $r_{conf} = \frac{|r_{sens} - c_0|}{\hat{\sigma}_{sens}}$. This definition of choice reliability is based on the reasoning that stronger evidence samples (r_{sens}) and lower sensory uncertainty (σ_{sens}) lead to more consistent choices. Critically, the model assumes that the observer does not have direct access to the level of sensory uncertainty and relies on a noisy estimate, $\hat{\sigma}_{sens}$, modeled as a random variable drawn from a lognormal distribution with a mean equal to the true value of sensory uncertainty (σ_{sens}):

$$\hat{\sigma}_{sens} \sim g_{lognormal}(x \mid \sigma_{sens}, \sigma^2_{meta}) = \frac{1}{x\sqrt{2\pi\sigma^2_{meta}}}\, e^{-\frac{(\ln(x)-\sigma_{sens})^2}{2\sigma^2_{meta}}}, \quad (9)$$

where σ_{meta}, termed as "meta-uncertainty" captures the uncertainty in their estimate of sensory uncertainty.

Since choice and confidence are generated from distinct processes, the confidence criteria are not constrained by the decision criterion. Therefore, we define two sets of confidence criteria to generate confidence for "S_1" and "S_2" responses, with each set extending from 0 to ∞. For each response type, we define the criteria as $[c_{i_0},\ c_{i_1},\ c_{i_2} \ldots c_{i_{n-1}}, c_{i_n}]$, where n is the number of ratings on the confidence scale and $i = \{1, 2\}$ refers to the class of stimulus responses, S_i. The criteria c_{ij} are monotonically increasing with $c_{i_0} = 0$ and $c_{i_n} = \infty$. Confidence responses are generated such that r_{conf} falling within the interval $[c_{i_j},\ c_{i_{j+1}})$ results in a confidence of $j + 1$.

DC Model

DC models were originally developed to explain the phenomenon of blindsight where observers can perform a visual task above chance level without being consciously aware of the stimuli. Such models assume that perceptual decisions associated with low confidence (assumed to be equivalent to no conscious experience) and high confidence (assumed to involve conscious experience) are processed by independent channels. Specifically, the "conscious" channel processes decisions associated with high confidence, whereas the "unconscious" channel processes low confidence decisions. The model therefore has a peculiar structure where on each trial it is first necessary to determine the predicted confidence before it can be determined which channel should be used for the primary decision. We implemented a version of the DC model that generates decisions associated with low conscious experience (i.e., low confidence) by combining inputs from both the "conscious" and "unconscious" channels. The model version has been previously shown to have the best performance among all variants within this class of models (Maniscalco & Lau, 2016).

According to the DC model, internal evidence for the conscious and unconscious channels, $r_{conscious}$ and $r_{unconscious}$, are generated independently from Gaussian distributions, such that:

$$r_{conscious} \sim \begin{cases} N\left(-\frac{\mu_{conscious}}{2}, \sigma^2_{sens}\right), & \text{if stimulus} = S_1 \\ N\left(\frac{\mu_{conscious}}{2}, \sigma^2_{sens}\right), & \text{if stimulus} = S_2 \end{cases} \quad (10)$$

and

$$r_{unconscious} \sim \begin{cases} N\left(-\frac{\mu_{unconscious}}{2}, \sigma^2_{sens}\right), & \text{if stimulus} = S_1 \\ N\left(\frac{\mu_{unconscious}}{2}, \sigma^2_{sens}\right), & \text{if stimulus} = S_2 \end{cases}, \quad (11)$$

where $\mu_{conscious}$ and $\mu_{unconscious}$ are the means of the evidence distributions for the conscious and unconscious channels and $\mu_{unconscious} = w_{unconscious} \times \mu_{conscious}$. The parameter $w_{unconscious} \in [0, 1]$ controls the fraction of evidence present in the conscious channel that is available to the unconscious channel.

The evidence coming from these two channels is then combined into a weighted sum such that $r_{conscious+unconscious} = r_{conscious} \times w_{combined} + r_{unconscious} \times (1 - w_{combined})$, where the weight given to evidence from the conscious channel is $w_{combined} = \frac{\mu_{conscious}}{\mu_{conscious} + \mu_{unconscious}}$.

Confidence decisions are first generated from $r_{conscious}$ by setting the confidence criteria, $[c_{-n}, c_{-n+1}, \ldots, c_{-1}, c_1, \ldots, c_{n-1}, c_n]$, on the decision axis for the conscious channel, where n is the number of ratings on the confidence scale. The criteria c_i are monotonically increasing with $c_{-n} = -\infty$ and $c_n = \infty$. When the primary response is "S_2," confidence responses greater than 1 (conscious decisions) are generated using the criteria $[c_1, \ldots, c_n]$ such that $r_{conscious}$ falling within the interval $[c_i, c_{i+1})$ results in a confidence of $i + 1$. When the primary response is "S_1," confidence responses greater than 1 are generated using the criteria $[c_{-n}, c_{-n+1}, \ldots, c_{-1}]$, such that $r_{conscious}$ falling within the interval $[c_i,\ c_{i+1})$ results in a confidence of $-i$. Finally, confidence responses of 1 (unconscious decisions) are given when $c_{-1} \leq r_{conscious\ +\ unconscious} \leq c_1$.

The stimulus judgments for "unconscious" decisions (decisions associated with a confidence of 1), are based on comparisons of $r_{conscious\ +\ unconscious\ with\ the\ decision\ criterion\ c0}$. If confidence is larger than 1, stimulus decisions are based on the response indicated by the confidence criteria for the conscious channel. For instance, if $r_{conscious} < c_{-1}$, the stimulus is classified as S_1, whereas if $c_1 < r_{conscious}$, the stimulus is classified as S_2. However, if there is a conflict between stimulus classifications indicated by the conscious and combined channels, the stimulus decision provided by the conscious channel prevails, but confidence for that decision is set to 1. For example, if $r_{conscious\ +\ unconscious} > c_0$ (the combined channel indicates a response of S_2) but $r_{conscious} < c_{-1}$ (the conscious channel indicates a response of S_1), the stimulus would be classified as S_1 (in accordance with the conscious channel) but confidence would be set to 1.

SDRM

SDRM assumes that the evidence variables for the primary decision and confidence, r_{sens} and r_{conf}, are generated from bivariate Gaussian distributions such that:

$$r_{sens} \sim \begin{cases} N\left(-\frac{\mu_{sens}}{2}, \sigma^2_{sens}\right), & \text{if stimulus} = S_1 \\ N\left(\frac{\mu_{sens}}{2}, \sigma^2_{sens}\right), & \text{if stimulus} = S_2 \end{cases} \quad (12)$$

and

$$r_{conf} \sim \begin{cases} N\left(-\frac{\mu_{sens}}{2}, \sigma^2_{sens}\right), & \text{if stimulus} = S_1 \\ N\left(\frac{\mu_{sens}}{2}, \sigma^2_{sens}\right), & \text{if stimulus} = S_2 \end{cases}. \quad (13)$$

Importantly, the two evidence samples are correlated and can be described by the following covariance structure:

$$\begin{pmatrix} r_{\text{sens}} \\ r_{\text{conf}} \end{pmatrix} \sim \begin{cases} N(-\boldsymbol{\mu}, \Sigma), & \text{if stimulus} = S_1 \\ N(\boldsymbol{\mu}, \Sigma), & \text{if stimulus} = S_2 \end{cases}, \tag{14}$$

where $\boldsymbol{\mu} = \left[\frac{\mu_{\text{sens}}}{2}, \frac{\mu_{\text{sens}}}{2}\right]$ and the covariance matrix $\Sigma = \begin{bmatrix} \sigma^2_{\text{sens}} & \rho\sigma^2_{\text{sens}} \\ \rho\sigma^2_{\text{sens}} & \sigma^2_{\text{sens}} \end{bmatrix}$ is such that σ^2_{sens} specifies the variance in the sensory evidence samples for both choice (r_{sens}) and confidence (r_{conf}) and $\rho \in [0, 1]$ controls the correlation between them.

Stimulus decisions are generated by comparing r_{sens} with a noisy decision criterion drawn from a Gaussian distribution, $C_0 \sim N(c_0, \sigma^2_{\text{dec}})$, where c_0 is the mean criterion location and σ^2_{dec} quantifies the trial-by-trial variability of the decision criterion. Note that SDRM is the only model considered here that includes noise in the decision criterion. To generate confidence, a distinct set of confidence criteria, $[c_{i_0}, c_{i_1}, c_{i_2} \ldots c_{i_{n-1}}, c_{i_n}]$, are defined for each response type where n is the number of ratings on the confidence scale and $i = \{1, 2\}$ refers to the class of stimulus responses ("S_1" or "S_2"). The criteria c_{ij} are monotonically increasing with $c_{i_0} = -\infty$ and $c_{i_n} = \infty$ when $i = 2$, and monotonically decreasing with $c_{i_0} = \infty$ and $c_{i_n} = -\infty$ when $i = 1$. Confidence responses are generated such that r_{conf} falling within the interval $[c_{i_j}, c_{i_{j+1}})$ results in a confidence of $j + 1$. The model assumes that confidence criteria on each trial are drawn from Gaussian distributions such that:

$$c_{i_j} \sim g_{\text{Gauss}}(x \mid \mu_{i_j}, \sigma^2_{\text{meta}}) = \frac{1}{\sqrt{2\pi\sigma^2_{\text{conf}}}} e^{-\frac{(x-\mu_{i_j})^2}{2\sigma^2_{\text{conf}}}}, \quad x \in (-\infty, \infty), \tag{15}$$

where μ_{i_j} and σ^2_{conf} are the mean and variance of the Gaussian distribution controlling criterion variability with $i = \{1, 2\}$ referring to the response class and $j = \{1, \ldots, n\}$ indexing the criterion. To maintain the order of the criteria, the parameters μ_{i_j} are constrained so that $\mu_{i_1} \leq \mu_{i_2} \cdots \leq \mu_{i_{n-1}}$.

SOC Model

SOC assumes that the sensory evidence samples for the perceptual and metacognitive systems are generated from a bivariate Gaussian distribution (Fleming & Daw, 2017), such that:

$$r_{\text{sens}} \sim \begin{cases} N\left(-\frac{\mu_{\text{sens}}}{2}, \sigma^2_{\text{sens}}\right), & \text{if stimulus} = S_1 \\ N\left(\frac{\mu_{\text{sens}}}{2}, \sigma^2_{\text{sens}}\right), & \text{if stimulus} = S_2 \end{cases} \tag{16}$$

and

$$r_{\text{conf}} \sim \begin{cases} N\left(-\frac{\mu_{\text{sens}}}{2}, \sigma^2_{\text{conf}}\right), & \text{if stimulus} = S_1 \\ N\left(\frac{\mu_{\text{sens}}}{2}, \sigma^2_{\text{conf}}\right), & \text{if stimulus} = S_2 \end{cases}. \tag{17}$$

These evidence samples obey the following covariance structure:

$$\begin{pmatrix} r_{\text{sens}} \\ r_{\text{conf}} \end{pmatrix} \sim \begin{cases} N(-\boldsymbol{\mu}, \Sigma), & \text{if stimulus} = S_1 \\ N(\boldsymbol{\mu}, \Sigma), & \text{if stimulus} = S_2 \end{cases}, \tag{18}$$

where $\boldsymbol{\mu} = \left[\frac{\mu_{\text{sens}}}{2}, \frac{\mu_{\text{sens}}}{2}\right]$, the covariance matrix is $\Sigma = \begin{bmatrix} \sigma^2_{\text{sens}} & \rho\sigma_{\text{sens}}\sigma_{\text{conf}} \\ \rho\sigma_{\text{sens}}\sigma_{\text{conf}} & \sigma^2_{\text{conf}} \end{bmatrix}$, σ^2_{sens} and σ^2_{conf} specify the variance in the sensory evidence samples for choice (r_{sens}) and confidence (r_{conf}), and the parameter $\rho \in [0, 1]$ controls the correlation between the two samples.

According to SOC, the perceptual system generates a binary response (R) by comparing r_{sens} to a decision criterion, c_0, such that $r_{\text{sens}} > c_0$ leads to the response "S_2" and $r_{\text{sens}} \leq c_0$ leads to the response "S_1." The metacognitive system computes confidence as the posterior probability that the perceptual decision is correct (i.e., $R = S$ where S is the true stimulus identity). This posterior probability computation is considered a "second-order" computation because the metacognitive system uses r_{conf} to infer the probability distribution across all possible decision variables, r_{sens}, that could have generated the observed response, R. This type of inference is possible because the metacognitive system is assumed to have knowledge of the covariance structure, Σ, that governs the relationship between the two decision variables, r_{sens} and r_{conf}. A measure of confidence can then be computed by marginalizing across this probability distribution. Therefore, the probability of a given response, R, being correct, is computed as:

$$p(R = S|r_{\text{conf}}, R, \Sigma) = \begin{cases} p(S_1|r_{\text{conf}}, R, \Sigma), & \text{if } R = S_1 \\ 1 - p(S_1|r_{\text{conf}}, R, \Sigma), & \text{if } R = S_2 \end{cases}, \tag{19}$$

where $p(S|r_{\text{conf}}, R, \Sigma)$ is the posterior probability of stimulus S, conditional on the confidence variable r_{conf}, the observed response, R, and the covariance structure, Σ, describing the association between the confidence and decision variables.

For a detailed breakdown of the Bayesian computation of $p(S|r_{\text{conf}}, R, \Sigma)$, see Fleming and Daw (2017). Briefly, the posterior probability of a correct response can be computed as:

$$p(S|r_{\text{conf}}, R, \Sigma) = \frac{p(R|r_{\text{conf}}, S, \Sigma)p(S|r_{\text{conf}}, \Sigma)}{\sum_S p(R|r_{\text{conf}}, S, \Sigma)p(S|r_{\text{conf}}, \Sigma)}, \tag{20}$$

where $p(R|r_{\text{conf}}, S, \Sigma)$ refers to the likelihood of the observed response given the stimulus S, and $p(S|r_{\text{conf}}, \Sigma)$ refers to the prior probability of the stimulus, both probabilities being conditional on r_{conf} and Σ.

In this model, confidence is generated in terms of posterior probability, which lies within the interval [0, 1]. To transform these probability values into discrete confidence ratings, two sets of confidence criteria are defined: $c_{i_0}, c_{i_1}, c_{i_2} \ldots c_{i_{n-1}}, c_{i_n}$, where n is the number of ratings on the confidence scale and $i = \{1, 2\}$ refers to the class of stimulus responses (S_1 or S_2). The criteria c_{ij} are monotonically increasing with $c_{i_0} = -\infty$ and $c_{i_n} = \infty$ when $i = 2$, and monotonically decreasing with $c_{i_0} = \infty$ and $c_{i_n} = -\infty$ when $i = 1$. Confidence responses are generated such that r_{conf} falling within the interval $[c_{i_j}, c_{i_{j+1}})$ results in a confidence of $j + 1$.

Model Parameters

General Parameters

For most models the primary decision about the identity of the stimulus was generated using the same two parameters: μ_{stim}, which controls the strength of the sensory signal, and c_0, which

controls the location of the decision criterion. The only exception is for 2DSD where these two parameters are substituted with equivalent parameters from the diffusion model—δ (drift rate) and z (starting point). Without loss of generality, the sensory noise, σ_{sens}, was set to 1 for all models except CASANDRE and was thus not considered a free parameter. For CASANDRE, we instead fixed the stimulus strength parameter, μ_{stim}, to 1 and allowed σ_{sens} to vary because confidence in this model is derived directly from values of sensory uncertainty, rather than stimulus strength. It is standard practice to model changes in stimulus strength as changes in the mean of stimulus distributions (keeping their variance fixed). In fact, several of the models we use in these studies have been fit to data under the same assumptions (Maniscalco et al., 2016; Maniscalco & Lau, 2016; Rausch et al., 2018, 2020). However, we note that it is possible to alternatively model these effects as changes in the variance of stimulus distributions (keeping the means fixed) or as changes in both the mean and variance of stimulus distributions. Future studies should explore the effects of these assumptions on model performances. In Experiment 1, which only featured a single difficulty level, we used a single free parameter for μ_{stim}, whereas in Experiments 2 and 3, which featured three and eight stimulus levels, respectively, we fit μ_{stim} separately to each contrast/coherence level. For CASANDRE, we fit σ_{sens} separately to each contrast/coherence level. It should also be noted that for SDRM, which alone assumes a noisy decision criterion, the parameter c_0 refers to the Gaussian distribution from which the trial-by-trial values are sampled (for all other models, the decision criterion is fixed at c_0). Finally, a set of $2 \times (n\text{-}1)$ confidence criteria were specified to generate the confidence ratings where n refers to the number of ratings on the confidence scale ($n = 4$ for Experiments 1 and 3, and $n = 6$ for Experiment 2). The confidence criteria were assumed to be the same across the different contrasts in Experiments 2 and 3 to provide maximum constraints to the models. Therefore, all models shared equivalent eight (Experiment 1), 14 (Experiment 2), or 15 (Experiment 3) general parameters.

Model-Specific Parameters

Each model makes unique assumptions about how metacognitive processes operate to transform the sensory signal into a confidence rating, often requiring additional parameters (Table 2). We note that the confidence generation parameters for three models—Decay, PE-Flex, and DC—vary with the level of stimulus contrast. For Decay and DC, the decision to allow these parameters to vary with stimulus contrast levels was made in order to remain consistent with their original instantiations (Maniscalco & Lau, 2016). For the PE-Flex model, we allowed the confidence-specific parameter to vary with stimulus contrast because this made the model perform significantly better for Experiment 2 (Supplementary Results in the online supplemental materials). Therefore, we present the more flexible version of the PE-Flex model here.

Model Fitting

Model fitting was based on a maximum likelihood estimation (MLE) procedure that searched for the set of parameters that maximize the log-likelihood associated with the full probability distribution of responses. The log-likelihood was computed using the following formula:

$$\text{Log} - \text{likelihood} = \sum_{i,j,k} \log(p_{ijk}) \times n_{ijk}, \tag{21}$$

where p_{ijk} and n_{ijk} are the response probability and number of trials, respectively, associated with the stimulus class $i = \{1, 2\}$, confidence response $j = \{-4,-3,-2,-1,1,2,3,4\}$ for Experiments 1 and 3, and $j = \{-6, -5.. -1, 1, \ldots 6\}$ for Experiment 2 (where negative confidence responses correspond to S_1 responses), and task difficulty level, $k = 1$ for Experiment 1; $k = \{1,2,3\}$ for Experiment 2, and $k = \{1,2,3,4,5,6,7,8\}$ for Experiment 3. The parameter search was conducted using the Bayesian Adaptive Direct Search (BADS) toolbox, Version 1.0.5 (Acerbi & Ma, 2017).

The MLE procedure requires us to compute the response probabilities associated with each type of confidence response for a given set of model parameters. For SDT, Gauss, LogN, Decay, 2DSD, BCH, CASANDRE, WEV, and SDRM, we derived the expression for computing the response probabilities from the model's assumptions and estimated these probabilities either numerically or analytically (see Supplementary Methods in the online supplemental materials). However, for Post-Dec, PE, PE-Flex, DC and SOC, we were unable to derive analytical expressions for computing response probabilities. Therefore, we simulated these models using 1 million trials and estimated response probabilities as the proportion of trials that were associated with each type of confidence response. The very large number of trials used minimized the increased noisiness for stimulation-based fitting and we did not observe a penalty for these models in our model recovery analyses. To obtain more reliable estimates of model fits for each model, we ran the fitting procedure five times for each of these models and chose the fits that were associated with the highest log-likelihood values. We generally observed similar model and parameter recovery for models for which analytical expressions were or were not available, thus suggesting that our fitting procedure did not unduly disadvantage the models without analytical expressions.

Model Selection

The ultimate goal of model comparisons is to select the model which is most likely to have generated the observed data. In order to quantify the plausibility of each model, we evaluated how closely the models fit the observed data using the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Both AIC and BIC measure the goodness-of-fits generated by a model, while including penalties for the number of free parameters. BIC applies a harsher penalty than AIC, thus tending to favor simpler models.

AIC and BIC were computed using standard formulas: $\text{AIC} = -2\log L + 2k$ and $\text{BIC} = -2\log L + k\log(n)$, where k refers to the total number of free parameters of the model and n refers to the number of trials in the data. Lower values of AIC and BIC indicate better quality of fits. To assess whether AIC differences between models are significant, we obtained bootstrapped 95% confidence intervals (CIs) on the summed AIC differences between models by subsampling the individual AIC differences with replacement and summing across the subsamples. The bootstrapped intervals were computed from 100,000 data samples using the MATLAB function bootci which uses a bias corrected and accelerated percentile method. Confidence intervals that do not contain zero are indicative

Table 2
Model-Specific Parameters Associated With Each of the 14 Models

Model	Additional parameters	Total number of parameters		
		Expt 1	Expt 2	Expt 3
SDT	None	8	14	15
Gauss	σ_{meta} (meta noise)	9	15	16
LogN	σ_{meta} (meta noise)	9	15	16
Decay	σ_{meta} (meta noise), ρ_{decay} (proportion of sensory signal available for confidence)	10	18	24
Post-Dec	δ_{post} (postdecisional signal)	9	15	16
2DSD	τ (postdecisional signal)	9	15	16
PE	None	8	14	15
PE-Flex	$\boldsymbol{w}_{pe}$ (weight given to PE)	9	17	23
WEV	w_{vis} (weight given to stimulus visibility), σ_{meta} (meta noise)	—	16	17
BCH	None	—	14	15
CASANDRE	σ_{meta} (meta uncertainty)	9	15	16
DC	$\boldsymbol{w}_{unconscious}$ (proportion of the available signal accessed by the unconscious channel)	9	17	23
SDRM	σ_{dec} (noise in decision criterion), σ_{conf} (noise in confidence criteria), ρ (correlation between evidence samples for primary choice and confidence)	11	17	18
SOC	σ_{conf} (noise in confidence criteria), ρ (correlation between evidence samples for primary choice and confidence)	10	16	17

Note. All models have an equivalent set of general parameters that describe how the primary decision is generated. However, the models differ in their assumptions about how confidence is generated, thus requiring a unique set of parameters to describe confidence decisions. The parameters indicated in bold text are dependent on the stimulus contrast/coherence levels (i.e., there is a separate free parameter for each contrast/coherence). Therefore, for Experiments 2/3, which feature three/eight difficulty levels, each of these parameters increase the total number of free parameters by three/eight. The WEV model was not fit to Experiment 1 because in the presence of a single difficulty level, the model becomes equivalent to the Gauss model. Expt = experiment; SDT = signal detection theory model; Gauss = Gaussian meta noise model; LogN = lognormal meta noise model; Decay = noisy decay model; Post-Dec = postdecisional SDT model; 2DSD = two-stage dynamic signal detection model; PE = positive evidence bias model; PE-Flex = flexible positive evidence bias model; WEV = weighted evidence and visibility model; BCH = Bayesian confidence hypothesis model; CASANDRE = confidence as a noisy decision reliability estimate model; DC = dual channel model; SDRM = stochastic detection and retrieval model; SOC = second-order confidence model.

of a significant difference in the AIC values of the models being compared. The same procedure was also used for BIC values.

Model selection based on AIC/BIC comparisons rely on the fixed-effects assumption that all subjects generate behavior using a single model, with the implication that the winning model explains the behavior of all subjects. However, an alternative approach is to treat models as random effects such that the generative model can vary between subjects. Within this approach, each model is assigned a certain probability of being the generative model (model frequency; p_{model}) and we select the model that is has the highest frequency in the population (note that this procedure is similar to a random-effects statistical test rather than a count of how many subjects are described best by each model). In addition to model frequencies, random-effects analyses also allow us to compute exceedance probabilities, p_{exc}, which measures the higher-order probability that a given model is the most frequent model in the set of models being compared. We used the Variational Bayes Analysis toolbox (Daunizeau et al., 2014) to estimate the model frequencies (p_{model}) and their protected exceedance probabilities (p_{exc}).

Model and Parameter Recovery

Model recovery is a crucial prerequisite for model selection as it validates our ability to practically discriminate between models using the data we have. The goal of model recovery analyses is to assess whether models can be uniquely identified from the data they generate. For this purpose, we first generate data from a chosen model and then fit all the models to these data. The chosen model is deemed to be recoverable when it can be reliably identified as the best-fitting model for the data that it has generated. If a model that generated the data is liable to be confused with other models, it implies that these models cannot be distinguished from each other and therefore, comparisons between these models may not be valid.

For these analyses, we simulated 50 data sets (treated as separate, synthetic subjects) from each model by uniformly sampling their parameters across their plausible range of values. For each model and parameter, we determined these plausible ranges based on the minimum and maximum values observed from model fits. To be able to practically evaluate model recovery, models were simulated using the same number of trials that were originally contained in each of the two experiments. All the models were fit to each simulated data set. We then computed model frequencies (p_{model}) for each data set to determine the probability that the model that generated the data set is assigned as the generative model for those data. SDRM was excluded from model recovery analyses in Experiment 3 because the time cost for fitting the model made recovery unfeasible. Specifically, SDRM takes on average 24–36 hr to fit a single data set and model recovery for each experiment requires fitting the model to 50 data sets generated by each of the 14 models requiring us to fit SDRM to a total of 700 data sets.

Parameter recovery, on the other hand, can tell us whether a model's parameters can be uniquely identified. Inability to recover a model's parameters may suggest that there is redundancy in the model's description and indicates that we may be unable to identify the unique contribution of that parameter to the process we are trying to model.

To assess, parameter recovery for each model, we simulated a synthetic experiment (separately for each experiment) by randomly sampling each of its parameters along the observed range of values and matching the number of trials from the corresponding experiment. Then, we fit the model to the simulated data and correlated the parameters recovered from the fits to the true values of those parameters.

Qualitative Analyses

In addition to quantifying each model's ability to fit the data, we also wanted to obtain qualitative insight into why models succeeded or failed in capturing these data. Therefore, we tested models on their ability to fit observed behavior. Specifically, we assessed how closely they can fit individual differences in several measures including metacognitive ability, z-transformed receiver operating characteristic curve (zROC) shapes (Shekhar & Rahnev, 2021b), the folded-X pattern (FXP; Hangya et al., 2016), and task performance. Task performance was quantified using the SDT-derived measure for stimulus sensitivity, d'. The measure is computed as $d' = \varphi^{-1}(\text{HR}) - \varphi^{-1}(\text{FAR})$ where HR and FAR refer to the observed hit rate and false alarm rates, respectively, and φ^{-1} is the inverse of the cumulative standard normal distribution that transforms cumulative probabilities into z-scores. Metacognitive ability was quantified using the measure M-Ratio computed as the ratio between metacognitive sensitivity meta-d' and their task sensitivity d' (Maniscalco & Lau, 2012). The zROC functions were the standard plots of the relationship between an observer's z-transformed hit rate (zHR) and z-transformed false alarm rate (zFAR) for different locations of the classification criterion. Finally, the FXP refers simply to plotting average confidence for correct and incorrect trials as a function of stimulus difficulty. The name of the pattern comes from the fact that as the task becomes easier (d' increases), confidence for correct responses increases but confidence for incorrect responses decreases. We computed these measures for each individual subject while also generating the corresponding model fits for that subject by simulating each model with its best-fitting parameters. To avoid any confound arising from trial counts, we simulated models using the same number of trials as contained in the actual experiments.

Likelihood-Ratio (LR) Test

Some of the models that we include in this study are nested within others and differ only by the addition of a single parameter. For these pairs of models, we performed nested comparisons to assess the usefulness of the additional parameter by conducting LR tests. The LR test allows us to determine whether the improvement in model fits resulting from the additional of a parameter are large enough to consider the parameter useful for explaining the data. For each subject within a comparison, we computed the LR test statistic as $\text{LR} = 2(\hat{l} - \hat{l}_0)$ where $\hat{l}$ and $\hat{l}_0$ refer to the maximum log-likelihood estimates of the full and reduced models, respectively. We then assessed the significance of the LR value at $\alpha = .05$. If the p value associated with the LR was smaller than .05, we rejected the reduced model in favor of the full model, concluding that the additional parameter was necessary to explain the data for that individual subject.

Transparency and Openness

All data, as well as code for analysis and model fitting, are available at https://osf.io/g8f9x.

Results

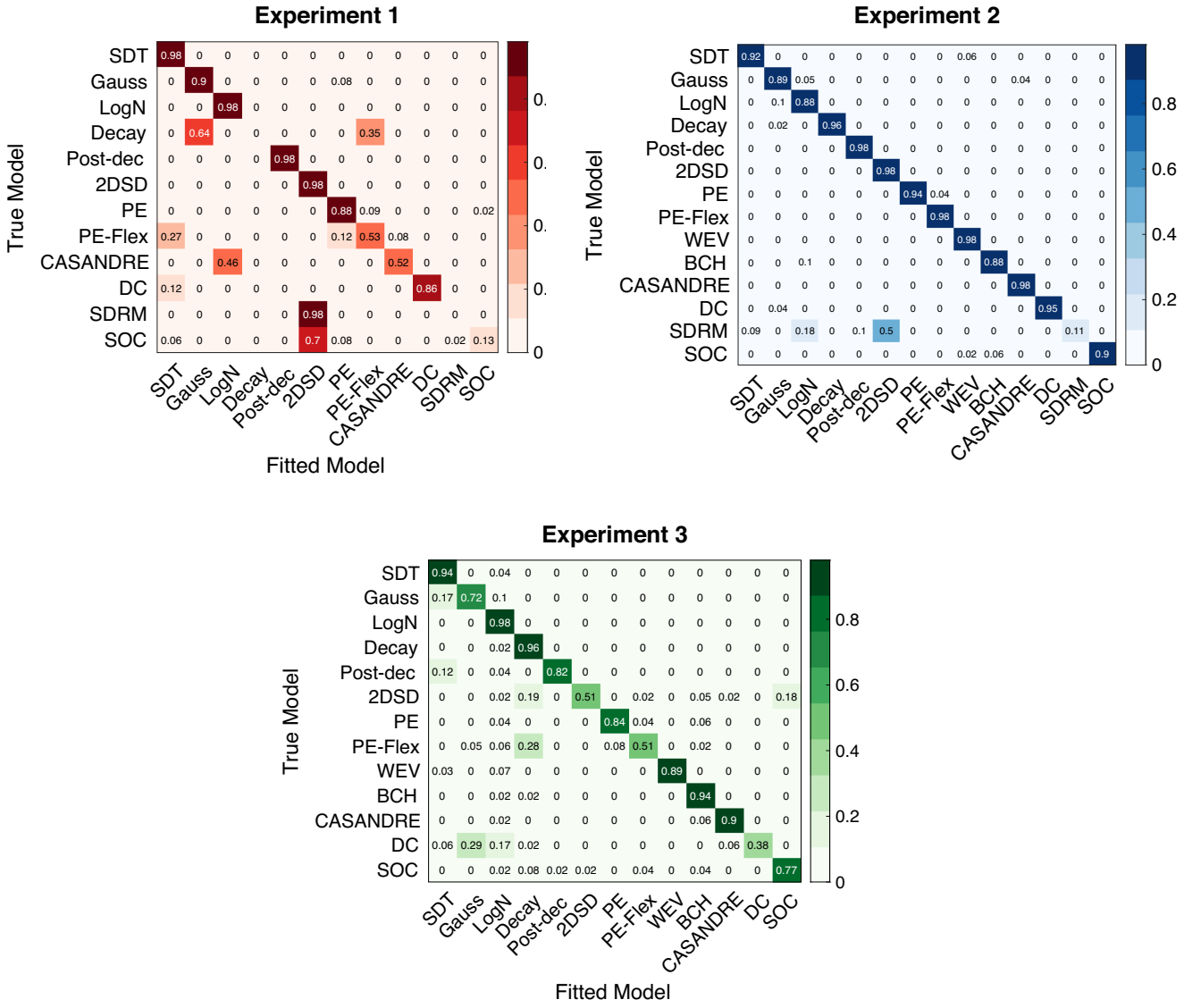
We tested competing theories of the mechanisms of confidence generation by fitting 14 process models to three large data sets that feature simple perceptual discrimination tasks and confidence ratings (Figure 1). We used a variety of different criteria for judging the quality of the fit—including both AIC and BIC scores, as well as both fixed- and random-effects modeling. Furthermore, to gain deeper understanding into the underlying mechanisms, we examined several aspects of each model including model recovery, parameter recovery, and ability to fit qualitative behavioral effects.

Model Recovery

Before performing model comparisons, it is critical to validate our ability to discriminate between models using the data that we have. Therefore, we performed model recovery analyses by quantifying the probability that a model that generated a given data set would be correctly identified as the generative model for that data set. To compute this probability for each model, we fit all the models to 50 synthetic data sets generated by that model and computed the model frequencies (p_{model}) associated with each model. A high p_{model} value for a given model indicates good recoverability for that model. We computed p_{model} values using both AIC and BIC scores. We found that AIC-derived model frequencies yielded reasonably high recovery probabilities for most models. Overall recovery was highest for Experiment 2 which featured multiple difficulty levels as well as a large number of trials per difficulty level (Figure 2). Specifically, the average p_{model} value for generating models was .65 in Experiment 1, .88 in Experiment 2, and .77 in Experiment 3, with 26 out of 39 values above .85. Conversely, BIC-derived model frequencies were heavily biased toward the SDT and PE models (since these models have the lowest number of parameters; Figure S1 in the online supplemental materials) and thus resulted in lower p_{model} values for all experiments (.40 for Experiment 1, .75 for Experiment 2, and .66 for Experiment 3). These results suggest that, in the context of the current models and data, AIC provides a more appropriate measure for model comparisons and therefore we use it as the primary comparison measure in the rest of the article.

Examining the exact pattern of model confusion revealed in Figure 2 leads to several important insights. First, seven models (SDT, Gauss, LogN, Post-Dec, PE, WEV, and BCH) show consistently high p_{model} values ($>.72$) across all experiments, suggesting that we can confidently arbitrate between these models using the data sets that we have. Note that the WEV and BCH models were only fit to Experiments 2 and 3 because the models require data sets that contain at least two different difficulty levels. Second, three other models (Decay, CASANDRE, and SOC) feature low to medium recoverability for Experiment 1 (p_{model} between 0 and .52) but high recoverability for Experiments 2 and 3 (p_{model} was between .77 and .98), suggesting that these models only become distinguishable from the rest in the presence of multiple difficulty levels. Third, there are four models that show inconsistent recovery across the three experiments. While PE-Flex shows high recoverability for Experiment 2 ($p_{\text{model}} = .98$), its recovery probability drops for both Experiments 1 ($p_{\text{model}} = .53$) and 3 ($p_{\text{model}} = .51$), suggesting that the model requires multiple difficulty levels as well as a large number of total trials to become discriminable from others models.

Figure 2
Model Recovery Analyses



Note. For each of the 50 simulated data sets from each model, we computed AIC-derived model frequencies (p_{model}) across all models. The values along the diagonal represent the probability that the model that generated the data is assigned as the generative model for that data set. Seven models—SDT, Gauss, LogN, Post-Dec, PE, WEV, and BCH—show consistently high model frequencies (>0.72) across all experiments. Two of the models—Decay and SOC—and show inconsistent recovery (being recoverable only for Experiments 2 and 3) whereas SDRM is always confused with at least one other model in each experiment. We observe better overall recovery for Experiment 2 compared to Experiments 1 and 3, suggesting discriminability of models is maximized both by having multiple difficulty levels within an experiment as well as including a high number of trials per condition. Note that the WEV and BCH models were only fit to data from Experiments 2 and 3 because in the presence of a single difficulty level (as in Experiment 1), they become equivalent to the Gauss and SDT models, respectively. Also note that SDRM was not included in model recovery analysis for Experiment 3 because of the heavy time cost associated with fitting SDRM to 700 data sets. AIC = Akaike information criterion; SDT = signal detection theory model; Gauss = Gaussian meta noise model; LogN = lognormal meta noise model; Decay = noisy decay model; Post-Dec = postdecisional SDT model; 2DSD = two-stage dynamic signal detection model; PE = positive evidence bias model; PE-Flex = flexible version of the PE model; CASANDRE = confidence as a noisy decision reliability estimate model; DC = dual channel model; SDRM = stochastic detection and retrieval model; SOC = second-order confidence model; WEV = weighted evidence and visibility model; BCH = Bayesian confidence hypothesis model. See the online article for the color version of this figure.

On the other hand, 2DSD, and DC show good recovery for Experiments 1 and 2 ($p_{\text{model}} > .86$), but relatively poor recovery for Experiment 3 ($p_{\text{model}} < .51$), implying that recovery of these models is more sensitive to the total number of trials. Finally, the SDRM model produced low recoverability values in both experiments ($p_{\text{model}} \leq .11$) with SDRM-generated data fit best by the 2DSD model. This result implies that SDRM may not be practically distinguishable from 2DSD in the current data sets. Note that SDRM

was excluded from model recovery in Experiment 3 because of the heavy time cost associated with fitting the model to 700 data sets (see Method for more details).

In general, Experiment 2, which consists of multiple interleaved difficulty levels as well as a large number of total trials, affords the best recoverability for models. The improvement in recovery for Experiment 2 over Experiments 1 and 3 suggest that the recoverability of models is constrained by both the number task conditions in the experiment as well as the total number of trials (Experiment 2 had 2,800 trials per subject, whereas Experiment 3 had 1,600 trials per subject). While including a greater number of difficulty levels in an experiment is beneficial for model recovery, having fewer trials can lead to worse recovery in spite of this benefit.

Parameter Recovery

Beyond model recovery, we also assessed how well it is possible to recover the parameters of each generating model. This parameter recovery analysis revealed that, for most models, all parameters can be uniquely identified with very high fidelity (Supplementary Results and Figures S2 and S3 in the online supplemental materials), at least for designs with trial numbers as high as in the current experiments. Nevertheless, there were several notable exceptions such as the decay parameter in the Decay model (for Experiment 1), the decision noise parameter in the SDRM model (for Experiments 1 and 2), and the correlation parameters in both the SDRM (for Experiment 1) and SOC models (for all experiments). All three of these models feature at least two sources of noise that can at least partially mimic each other and therefore recovering all of their parameters may require more complex manipulations. In the context of basic tasks like the ones used here, stable parameter recovery for these models may require using reduced versions of the models where some sources of noise are ignored. Finally, as with the model recovery analyses, we found an overall better parameter recoverability in Experiment 2, suggesting that both the presence of multiple difficulty levels and a large number of total trials are important for accurate parameter recovery.

Model Comparisons

Having examined both model and parameter recovery for all models, we turned to the central question of the article, namely how well did the different models fit the empirical data. Overall, we found that the LogN model provided the best fits for Experiments 1 and 2. However, for Experiment 3, which featured eight difficulty levels, the PE model emerged as the best-fitting model.

For Experiments 1 and 2, LogN had lower AIC scores (indicating better fit) than all other models (Figure 3 and Tables S1 and S2 in the online supplemental materials). These AIC differences favoring LogN were significant when assessed with 95% bootstrapped confidence intervals for all pairs of models except for the WEV model in Experiment 2 (95% CI [−32.33, 517.34]). In fact, only SDRM in Experiment 2 was within 100 AIC points of the LogN model showing the substantial advantage of LogN over all competing models. In Experiment 1, the next best-fitting models were CASANDRE (124.7 AIC points worse), SOC (218.56 AIC points worse), SDRM (223.86 AIC points worse), PE-Flex (294.71 AIC points worse), and 2DSD (374.09 AIC points worse), with the remaining models having substantially higher AIC values (over 1,000 AIC points worse). In Experiment 2, the next best model was SDRM (73.71 AIC points worse), followed by CASANDRE (119.3 AIC points worse) and WEV (162.55 AIC points worse), with other models producing much higher AIC values (over 225 AIC points worse). Examining the subject-level data from both experiments (Figure S4 in the online supplemental materials) shows that no other model fit more individual subjects better than LogN.

Model fitting for Experiments 1 and 2 yielded relatively consistent results, with LogN, CASANDRE, and SDRM models being in the top four for both experiments. However, for Experiment 3, we observed a shift in the pattern of model performances. The PE model (which was previously ranked in the bottom three for Experiments 1 and 2) was found to be the best-performing model. The next best-fitting model was WEV (268.07 points worse), with all other models producing substantially worse AIC scores (over 1,063 points worse). The AIC scores favoring PE were significant for all pairs of models when assessed with 95% bootstrapped confidence intervals, except for the WEV model (95% CI [−30.03, 803.49]).

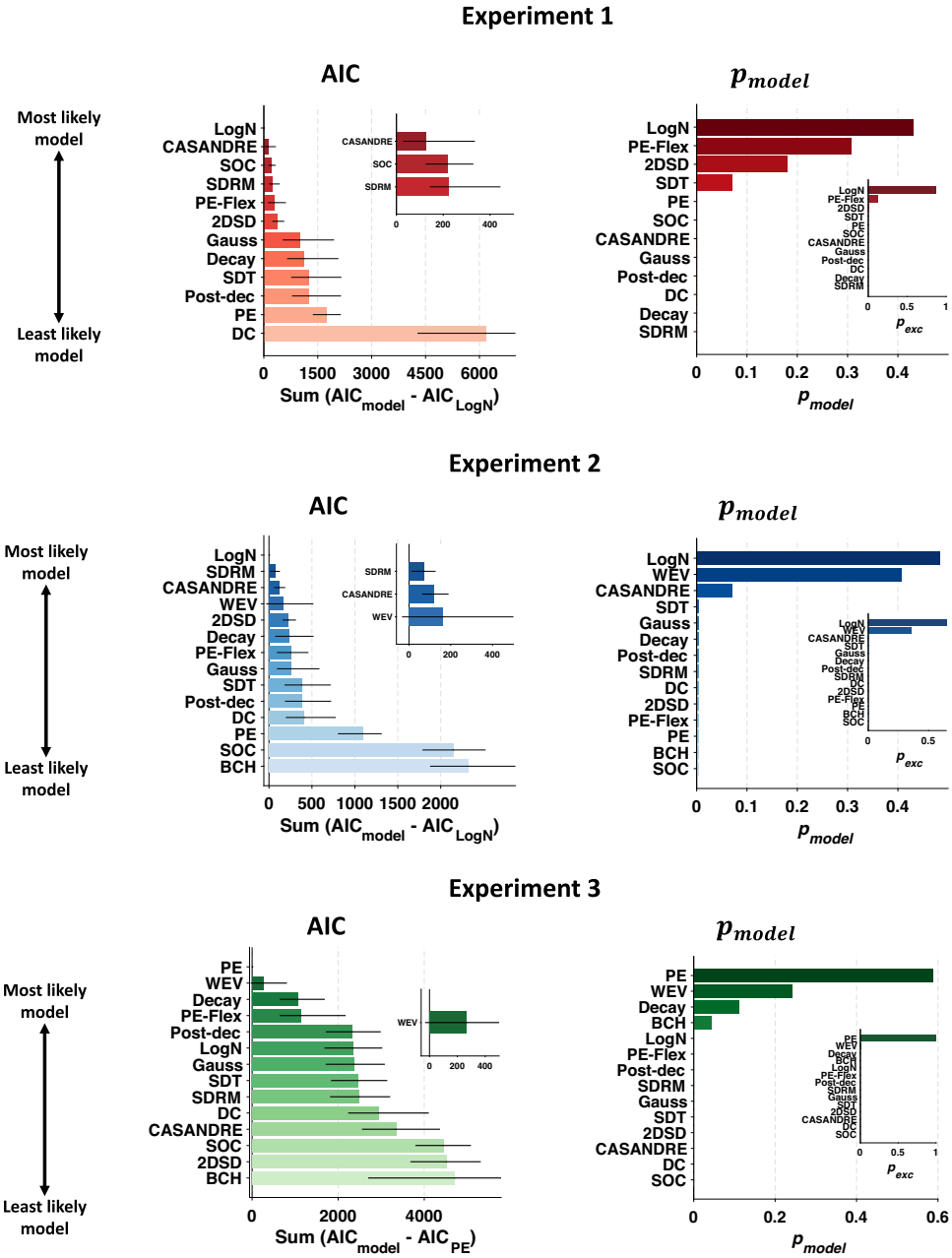
Directly comparing AIC values constitutes a fixed-effects analysis, which tacitly assumes no variability in the population. Because this may not be a reasonable assumption (i.e., it is possible that different subjects are described best by different models), we also performed a random-effects analysis, which assumes that the true model can vary between subjects. Specifically, we computed model frequencies (p_{model}) that reflect the probability of a given model being the generative model in our data set. Note p_{model} is a model-based measure (Daunizeau et al., 2014) and is not the same as computing the proportion of subjects for which a given model yields the lowest AIC/BIC values. The results corroborated the conclusions of our fixed-effects analyses. For Experiments 1 and 2, the LogN model yielded the highest model frequencies (Experiment 1: $p_{model} = .43$; Experiment 2: $p_{model} = .48$). The only other models in Experiment 1 with model frequencies higher than .01 were PE-Flex ($p_{model} = .31$), 2DSD ($p_{model} = .18$), and SDT ($p_{model} = .07$). For Experiment 2, there were two models with model frequencies higher than .01: WEV ($p_{model} = .40$) and CASANDRE ($p_{model} = .07$). Finally, for Experiment 3, the PE model was assigned the highest model frequency ($p_{model} = .59$), with three other models having model frequencies higher than .01: WEV ($p_{model} = .24$), Decay ($p_{model} = .11$), and BCH ($p_{model} = .05$).

It should be noted that the fixed-effects (Figure 3; left column) and random-effects (Figure 3; right column) analyses produced some important differences (e.g., SDRM was ranked high in fixed- but low in random-effects analyses, whereas the opposite was true for PE-Flex). While the exact sources of these discrepancies are unclear, these results underscore the benefit of performing different types of analyses and only placing confidence in results that are consistent across analyses.

In addition to model frequencies, the random-effects analyses also allow us to compute protected exceedance probability (p_{exc}), which measures the second-order probability that a model is the most likely generative model across all subjects. Given that the LogN model had the highest model frequencies in Experiments 1 and 2, it is not surprising that it had the highest exceedance probability (Experiment 1: $p_{exc} = .87$; Experiment 2: $p_{exc} = .77$). The only other models with nonzero exceedance probabilities were PE-Flex in Experiment 1 ($p_{exc} = .13$) and WEV in Experiment 2 ($p_{exc} = .23$). For Experiment 3, the PE model had the highest

Figure 3

Model Fitting Results



Note. The plots show comparisons based on AIC values for (a) Experiment 1, (b) Experiment 2, and (c) Experiment 3. The LogN model significantly outperforms all other models for Experiments 1 and 2, but the PE model is the best-fitting model for Experiment 3. (Left) Summed differences in AIC values between each model and the winning model. Higher positive values indicate worse fits. Error bars represent bootstrapped 95% confidence intervals. Insets present a zoomed-in version of the same data from the main figure. (Right) Random-effects analysis of model frequencies. The LogN model is identified as the most likely generative model in the population based on model frequencies and exceedance probabilities (insets) for Experiments 1 and 2. For Experiment 3, the PE model is the most likely generative model in the population. AIC = Akaike information criterion; PE = positive evidence bias model; WEV = weighted evidence and visibility model; Decay = noisy decay model; PE-Flex = flexible version of the PE model; Post-Dec = postdecisional SDT model; LogN = lognormal meta noise model; Gauss = Gaussian meta noise model; SDT = signal detection theory model; SDRM = stochastic detection and retrieval model; DC = dual channel model; CASANDRE = confidence as a noisy decision reliability estimate model; SOC = second-order confidence model; 2DSD = two-stage dynamic signal detection model; BCH = Bayesian confidence hypothesis model; p_{model} = model frequencies; p_{exc} = protected exceedance probabilities. See the online article for the color version of this figure.

exceedance probability ($p_{exc} = 0.996$) with no other model having an exceedance probability greater than .005.

The analyses above were performed on AIC values because the use of AIC resulted in more robust model recovery (see above). Nevertheless, to assess the robustness of our model selection results, we repeated all analyses using BIC and BIC-derived model frequencies (Supplementary Results, Figure S5, and Tables S3 and S4 in the online supplemental materials). Fixed-effects BIC analyses once again favored the LogN model over all others for Experiments 1 and 2 (Figure S6 in the online supplemental materials). However, random-effects analyses favored SDT and PE over LogN. These results are not surprising since the model recovery analyses demonstrated that BIC is heavily biased toward SDT and PE as the models with fewest parameters (Figure S1 in the online supplemental materials). For Experiment 3, fixed-effects BIC as well as BIC-derived random-effects analyses favored the PE model over all others, with the WEV model being the next best-performing model.

There was some disagreement between the three experiments, the two types of analyses (fixed- and random-effects), and two goodness-of-fit measures (AIC and BIC). Therefore, to obtain an overall picture from all model selection analyses, we pooled together all results by assigning ranks to models based on their performance for each measure. For instance, a rank of 1 was assigned to the model that provided the best fit and a rank of 12 (for Experiment 1) or 14 (for Experiments 2 and 3) for the model that provided the worst fit according to that measure. We then computed the average rank of each model across all the four analyses—AIC-derived fixed-effects, AIC-derived random-effects, BIC-derived fixed-effects, and BIC-derived random-effects. We found that the LogN performed best overall with an average rank of 1.25 for Experiment 1 and 1.5 for Experiment 2 (where 1 is the highest mean rank a model can achieve; Figure 4). The next two best-performing models for Experiment 1 were PE-Flex (average rank = 3.75) and 2DSD (average rank = 4.25), whereas the next two best-performing models for Experiment 2 were CASANDRE (average rank = 3.25) and WEV (average rank = 3.5). For Experiment 3, PE was the best-performing model with an average rank of 1, followed by the WEV model (average rank = 2) and the LogN model (average rank = 5.75).

Averaging the ranks across the three experiments still favored the LogN model (average rank = 2.83), with the three other models substantially behind (SDT: average rank = 5.25; CASANDRE: average rank = 6.16; PE: average rank = 6.16). However, the WEV model, which was fit only to Experiments 2 and 3, obtained an average rank of 2.75 across the two experiments, outperforming the LogN model (average rank across Experiments 2 and 3 = 3.5). Taken together, these results imply that model performance is not fully generalizable across different stimuli and tasks, possibly because none of the models fully describes the underlying confidence mechanisms. Later, we explore a new model that combines mechanisms from two existing models to investigate if a single model can provide good fits across all three experiments.

Qualitative Fits to Different Patterns in the Data

While quantitative model comparisons help us select the model that best describes the data, they do not reveal the reasons behind why models fail or succeed in providing a good fit. Therefore, to gain more qualitative insight into the performance of the models, we tested how closely each model can explain different qualitative aspects of the data. Some of these qualitative patterns have been proposed to directly index different underlying computations, but these proposals have not been tested by examining whether a wide variety of models can explain them. Specifically, here we focused on four aspects of the data: the observed metacognitive ability (quantified with M-Ratio), the shape of the zROC functions, the confidence for both correct and error trials (known to typically form a FXP), and the stimulus sensitivity (d'). Beyond testing the ability of these qualitative patterns to index specific computations, these analyses might offer clues for the reasons behind poor fits to the data for a given model.

Explaining Observed Metacognitive Ability

The first qualitative pattern in the data that we examined is the observed metacognitive ability (quantified using the measure M-Ratio). For each model in the three experiments, we plotted the observed value of M-Ratio against the value obtained from the best fit for that model (Figure 5). To quantify the strength of the correspondence between empirical observations and model fits, we examined the regression slope between the two. A slope of 1 corresponds to an excellent overall qualitative fit in the group (though it is still possible to observe deviations for individual subject), whereas a slope of 0 indicates no relationship between the model's fits and observations. For Experiments 1 and 2, we found that the models that consistently provided the best quantitative fits (had the lowest AIC scores) for both experiments (LogN, CASANDRE, and SDRM) yielded moderate to high slopes for both experiments, ranging between .46 and 1 (all slopes significantly different from 0, $p < .05$). On the other hand, the models that provided the worst AIC fits across both experiments (Post-Dec, PE, and DC) were always observed to have slopes close to 0 (none of the slopes were significantly different from 0, $p > .05$). In contrast, for Experiment 3, M-Ratio is less reliable in distinguishing between models that perform well and models that perform poorly. The individual model fits to M-Ratio also become less precise overall, with none of the slopes exceeding 0.2. The best-fitting PE model has the second lowest slope of 0.02.

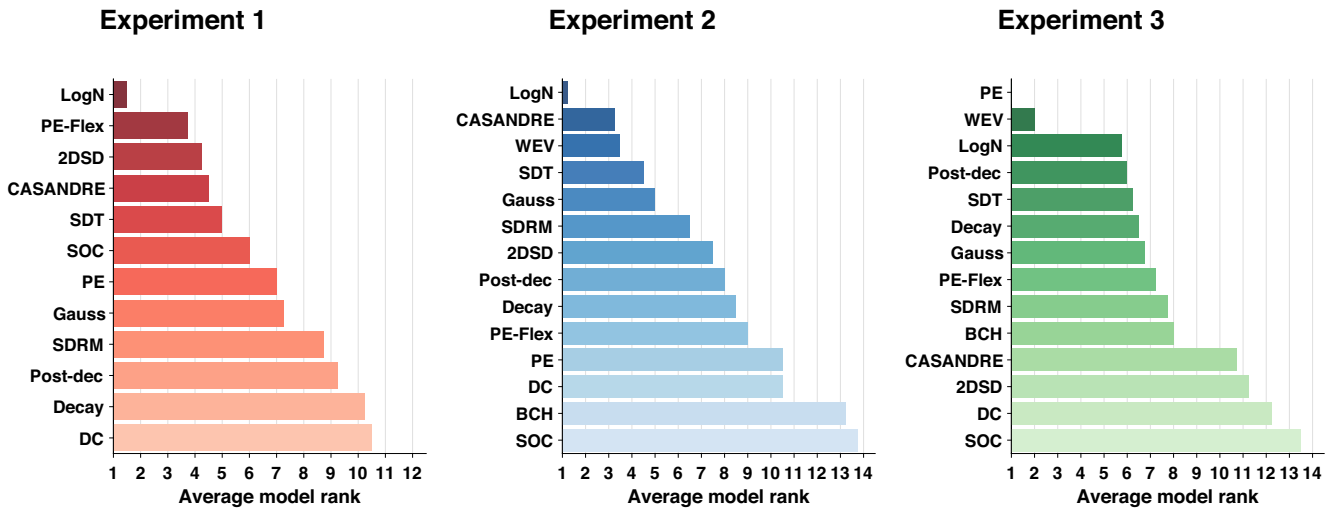
For Experiments 1 and 2, however, examining the errors of each model reveals several effects of interest. Both the SDT and Post-Dec models always generate M-Ratio values close to 1, which is unsurprising because both models lack built-in mechanisms to allow them to capture metacognitive inefficiency. One previously unappreciated effect is that the PE model always generates M-Ratio values around 0.5. This phenomenon likely emerges from the fact that the confidence ratings in this model ignore half of the available evidence (i.e., the evidence for the decision-incongruent choice). Nevertheless, it is important to note that other versions of the PE model that include both metacognitive noise and postdecisional evidence accumulation can still generate a range of M-Ratio values (Maniscalco et al., 2021). Finally, while the Gauss, Decay, and DC models are able to generate a range of different M-Ratio values, their fits tend to often deviate from the observed value. These results suggest that their respective mechanisms for inducing metacognitive inefficiency—Gaussian metacognitive noise and the presence of "unconscious" processing channel—are unlikely to capture well the underlying mechanisms of confidence generation.

Explaining Observed zROC Functions

In our previous work, we showed that zROC functions for perceptual decision-making tasks show a characteristic downward curvature which is indicative of a decrease in metacognitive sensitivity

Figure 4

Average Model Ranks Across Quantitative Measures of Comparison



Note. We ranked all models according to their performance for each of the four measures—AIC-derived fixed-effects, AIC-derived random-effects, BIC-derived fixed-effects, and BIC-derived random-effects—and then averaged these ranks. For Experiments 1 and 2, the LogN model exhibited the highest average rank (where 1 is the maximum average rank that can be achieved by a model). For Experiment 3, the PE model had the highest average rank across the different measures. However, the performance of the PE model was inconsistent across the three experiments. In contrast, the WEV model showed consistently good performance for both Experiments 2 and 3. AIC = Akaike information criterion; BIC = Bayesian information criterion; LogN = lognormal meta noise model; PE-Flex = flexible version of the PE model; 2DSD = two-stage dynamic signal detection model; CASANDRE = confidence as a noisy decision reliability estimate model; SDT = signal detection theory model; SOC = second-order confidence model; PE = positive evidence bias model; Gauss = Gaussian meta noise model; SDRM = stochastic detection and retrieval model; Post-Dec = postdecisional SDT model; Decay = noisy decay model; DC = dual channel model; WEV = weighted evidence and visibility model; BCH = Bayesian confidence hypothesis model. See the online article for the color version of this figure.

for higher confidence criteria (Shekhar & Rahnev, 2021b). This pattern served as the inspiration for developing the LogN model but we did not examine if other models can also capture it. Perhaps surprisingly, we found that, when averaged across subjects, all models are able to capture the basic downward curvature reasonably well in Experiment 1 (Figure 6, top). This is true even for the SDT model that is known to theoretically result in straight zROC functions. The likely reason for the downward curvature observed for SDT and other models here is that low trial counts for the extreme criteria result in a misestimation of d' (an issue extensively examined in Shekhar & Rahnev, 2021b). The two worst-fitting models in Experiment 2 (BCH and SOC) show large deviations from the observed data but still capture the overall downward shapes (Figure 6, middle). In Experiment 3, which has eight difficulty levels, the zROC curves show much larger deviations from the data for all models. However, the downward curvature is still apparent for all models. While the issues with misestimation are not insurmountable, removing this confound requires excluding a large proportion of the data and is somewhat ad hoc (Shekhar & Rahnev, 2021b). As such, when considered in isolation, the qualitative shape of zROC functions may not be a reliable criterion for evaluating a model's performance or revealing the underlying mechanisms of confidence.

Explaining Observed Confidence for Correct and Incorrect Trials (FXP)

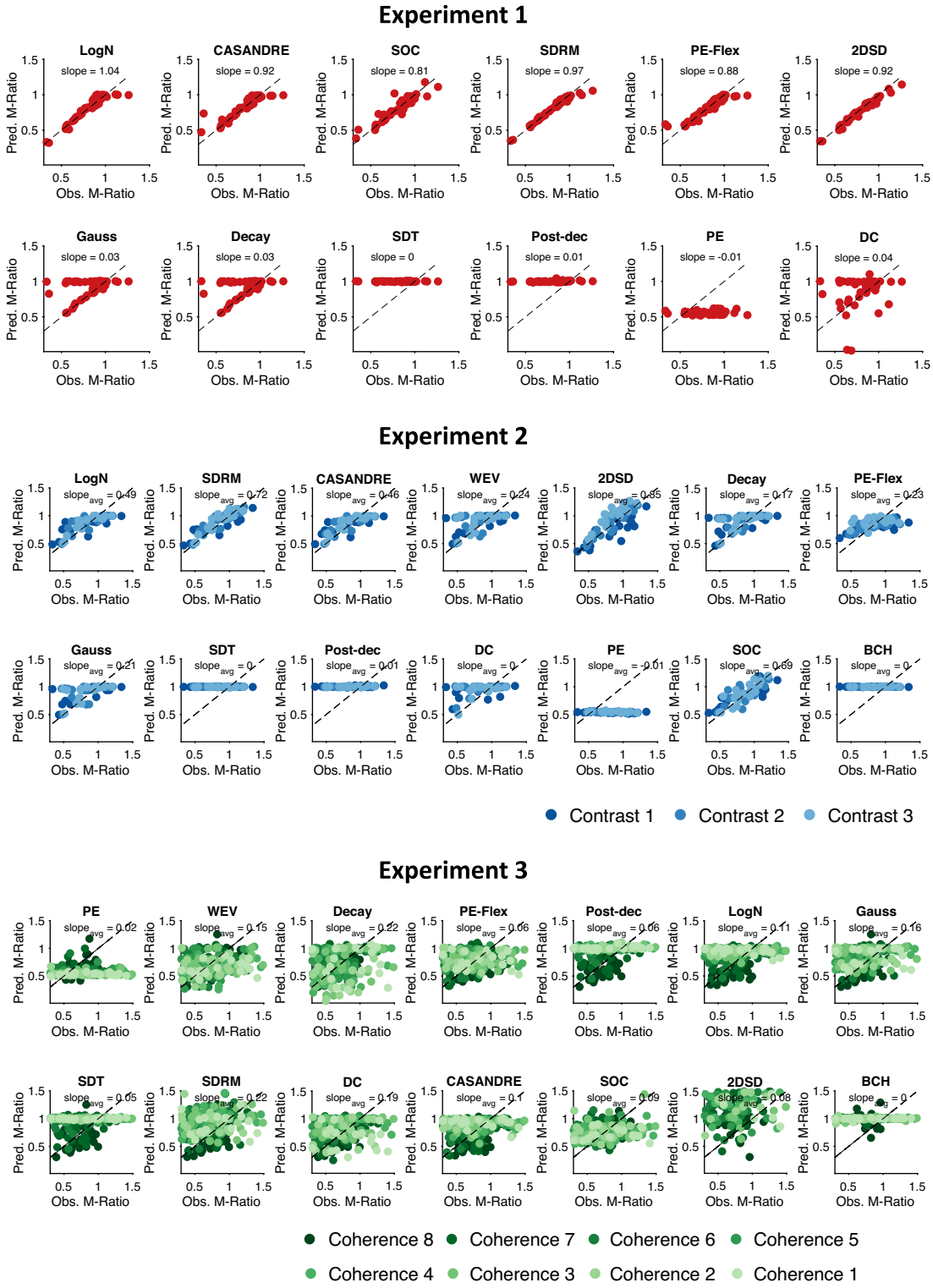
One of the well-known signatures of confidence is that it increases with stimulus discriminability for correct responses and decreases with stimulus discriminability for incorrect responses (Hangya et al., 2016; Sanders et al., 2016). When these two opposing relationships are plotted together, they resemble a folded-X shape. Since we need more than one stimulus discriminability (task difficulty) level to construct the FXP this pattern was only analyzed for Experiment 2.

Critically, we did not consistently observe the FXP in the data from these experiments. For Experiment 2, confidence for incorrect trials decreased with increasing contrast levels according to the FXP. However, for Experiment 3, we observed a different pattern where confidence for both correct and incorrect trials increased with increasing stimulus discriminability.

We find that all models except PE, SOC, and BCH generate the FXP, which can be observed in their fits to both Experiments 2 and 3 (Figure 7). For the PE, SOC, and BCH models, on the other hand, we observe an increase in confidence for incorrect responses with increasing stimulus discriminability. This observation emerges naturally from the PE model's computations which ignore evidence incongruent with the choice, resulting in confidence being directly proportional to the stimulus strength, regardless of the accuracy of the chosen response. On the other hand, for the Bayesian models (BCH and SOC), this result appears to contradict the notion that the FXP is a signature of Bayesian confidence (Hangya et al., 2016). However, there is no real contradiction because the emergence of the FXP is conditional on certain experimental conditions and model assumptions (see Discussion). Lastly, the WEV model is the only model that appears to flexibly fit both the patterns observed in Experiments 2 and 3, owing to its flexible computation of the confidence signal by adjusting the weight given to evidence over stimulus strength.

Figure 5

Model Fits for Individual Metacognitive Ability



Note. Model-generated M-Ratio values for each subject are plotted against their observed values. The slopes of the regression lines were obtained by regressing the fitted against observed values. For Experiments 2 and 3, the slopes were computed separately for each difficulty level and the average is reported. The models are arranged according to their AIC scores, with the best-performing models at the top left and the worst at the bottom right. For Experiments 1 and 2, except for the SOC model, models with higher AIC

(*figure continue*)

In Experiment 2, the SOC, PE, and BCH models which cannot fit the FXP were observed to be the worst-performing models. However, in Experiment 3, the PE and WEV models' ability to explain the observed pattern of confidence for correct and incorrect responses possibly underlies their high-performance relative to all the other models. These results suggest that whenever empirical violations of the FXP are observed (Rausch et al., 2020), this specific qualitative pattern can serve as an important indicator of model performance. However, the FXP does not fully explain performance for all models. Specifically, even though SOC and BCH are able fit the pattern of confidence observed in Experiment 3, they are ranked at the bottom.

Explaining Observed Stimulus Sensitivity

The final qualitative pattern that we examined is whether models can appropriately capture the observed stimulus sensitivity (d'). Importantly, all models have free parameters that specifically correspond to stimulus sensitivity (in Experiments 2 and 3, where we had multiple difficulty levels, all models feature a separate stimulus sensitivity-related free parameter for each difficulty level). Therefore, not surprisingly, most models were able to fit the observed d' values very well in both Experiments 1 and 2 (Figure 8). However, one notable exception was the poor fits by BCH and SOC in Experiments 2 and 3). This failure may explain why BCH and SOC provided the worst quantitative fits in those experiments, in spite of their success in fitting the observed pattern of confidence for correct and incorrect choices for Experiment 3. We speculate for the possible reasons behind this deviation in the Discussion.

Examining the Deviations for All Four Qualitative Patterns

To summarize the results from the four qualitative patterns in the data (M-Ratio, zROC functions, FXP, and d'), we calculated the mean absolute error in each model's fits with respect to the observed data for each subject and averaged the error across subjects (Figure 9). In general, we find that the average errors in all four patterns correlate with the ranking of each model according to AIC fits, though the relationship is imperfect for each of the four patterns. The positive relationship between these errors and model performance also becomes noisier as the complexity of the task increases from Experiment 1 to Experiment 3, implying that for complex experimental designs, qualitative analyses may become less reliable in accounting for model performances.

Examining the Plausibility of Individual Model Parameters

Some of the models that we have used in this study feature completely different architectures than any other model (e.g., 2DSD, SDRM, and SOC). However, some pairs of models differ only in a single parameter. Formally, the simpler model (the one with fewer parameters) is said to be "nested" within the more complex model. Here, we examine all pairs of nested models to evaluate whether the parameters added by the more complex models enhance their ability to explain the data. To do so, we perform a LR test and examine if there is statistically significant evidence for the additional parameters. Note that the LR should never be smaller than 0 because the more complex model should always be able to fit the data at least as well as the simpler model. However, 111 of the 685 LR values we observed were negative (16%), indicating that the more complex model likely became stuck in a local minimum. Even though such cases are easy to correct by simply using the fit from the simpler model, we have kept them here for purposes of full transparency (correcting them would not change any of the conclusions we draw below).

Metacognitive Noise

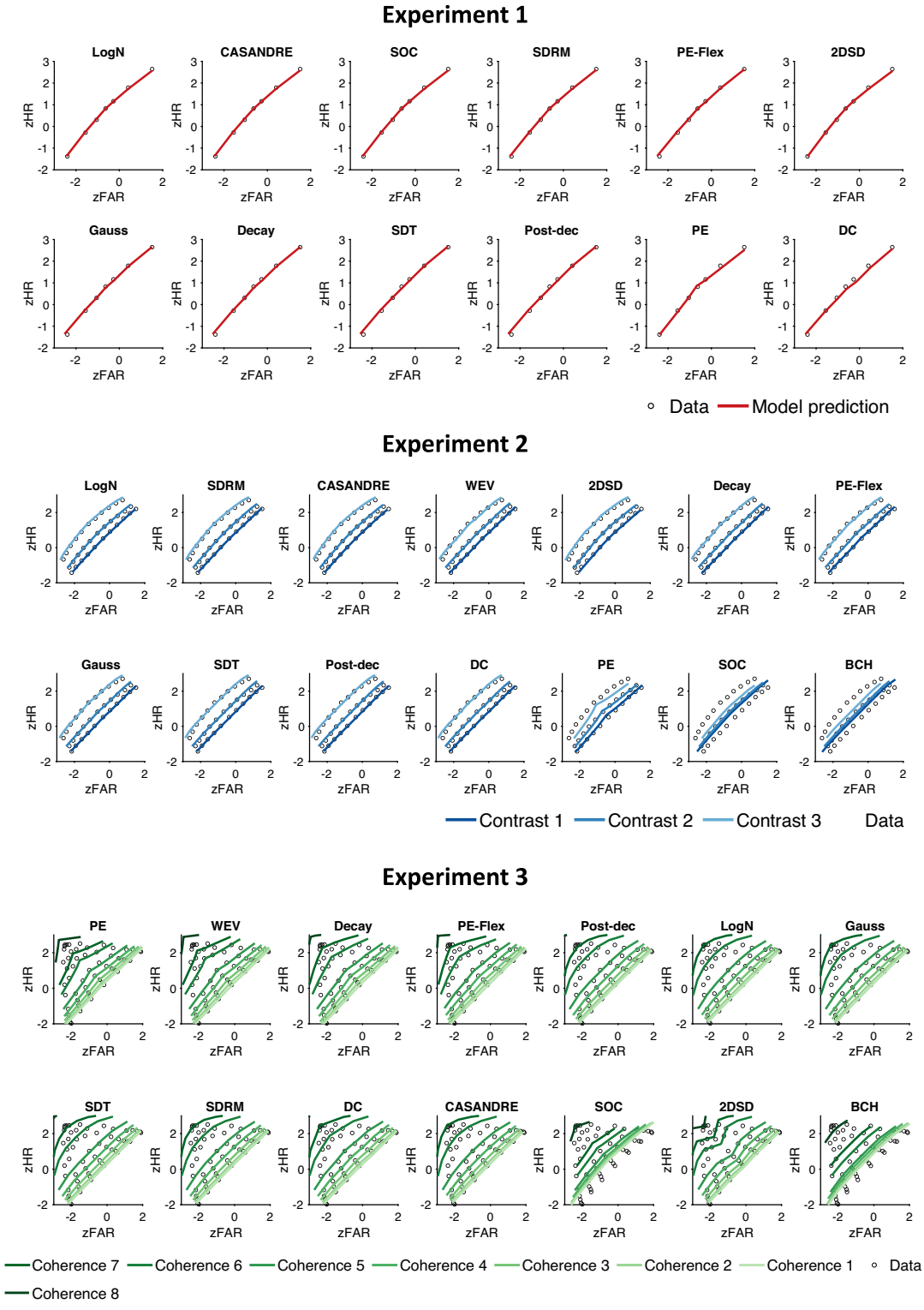
Two pairs of models differ only by whether they include metacognitive noise as a model parameter: Gauss is equivalent to SDT with Gaussian metacognitive noise, whereas LogN is equivalent to SDT with lognormal metacognitive noise The LR test rejected the simpler SDT model in favor of the Gauss model for 15 out of 59 subjects in Experiment 1, for 8 of the 20 subjects in Experiment 2, and for 14 out of 45 subjects in Experiment 3 (Figure 10a). On the other hand, the test rejected the SDT model in favor of the LogN model for 34 of the 59 subjects in Experiment 1, 11 of the 20 subjects in Experiment 2, and 22 out of 45 subjects in Experiment 3 (Figure 10b). These results mimic our previous studies where we also found the Gauss and LogN models to be an improvement over SDT (J. W. Bang et al., 2019; Shekhar & Rahnev, 2021b). These results suggest that metacognitive noise is necessary to explain confidence ratings for a significant proportion of subjects (on average 54% of subjects across experiments). For the remaining subjects, we find that SDT is sufficient to explain behavior. Our previous work (Shekhar & Rahnev, 2021b) showed that individual differences in metacognitive ability can impact the performance of models. Specifically, the model evidence for LogN increases as subjects' metacognitive scores decrease—with model evidence favoring SDT for subjects who have close to ideal M-Ratio. These findings are in line with our previous work and suggest that the assumption of

Figure 5 (*continued*)

scores are generally able to more reliably capture individual variations in metacognitive ability (as indicated by higher values of slopes for the models in the top row for both experiments). For Experiment 3, the model fits for M-Ratio become less precise and the distinction in M-Ratio fits between models in the top and bottom rows becomes less clear. AIC = Akaike information criterion; LogN = lognormal meta noise model; CASANDRE = confidence as a noisy decision reliability estimate model; SOC = second-order confidence model; SDRM = stochastic detection and retrieval model; PE-Flex = flexible version of the PE model; 2DSD = two-stage dynamic signal detection model; Gauss = Gaussian meta noise model; Decay = noisy decay model; SDT = signal detection theory model; Post-Dec = postdecisional SDT model; PE = positive evidence bias model; DC = dual channel model; WEV = weighted evidence and visibility model; BCH = Bayesian confidence hypothesis model; Pred. = predicted; Obs. = observed; avg = average. See the online article for the color version of this figure.

Figure 6

Model Fits for zROC Functions Averaged Across Subjects



Note. zROC functions plot the zHRs against zFARs. zROC functions are plotted separately for each task condition in Experiments 2 and 3. As can be seen from the plots, most models are able to fit the general shape of the zROC functions with relatively high precision for Experiments 1 and 2. However, for Experiment 3, which has eight difficulty levels, the fits become much less precise. The models are arranged according to their AIC scores for each experiment. zROC = z-transformed

(*figure continue*)

lognormal metacognitive noise may be necessary only for a subset of subjects with imperfect metacognitive scores.

Postdecisional Evidence Accumulation for Confidence

SDT and Post-Dec differ only in that the Post-Dec model includes a parameter for postdecisional accumulation of evidence related to the confidence rating. The LR test, however, failed to reject the SDT model for 55 out of 59 subjects in Experiment 1 and for all 20 subjects in Experiment 2 (Figure 10c). For Experiment 3, however, the Post-Dec model was favored over the SDT model in 23 out of 45 subjects. Based on these experiments, evidence for postdecisional evidence accumulation remains inconsistent. We also tested whether Post-Dec was favored particularly for subjects who show hypermetacognitive efficiency (i.e., M-Ratio > 1). We found no evidence for Post-Dec among these subjects in Experiments 1 and 2: the LR test could not reject SDT over Post-Dec for all five such subjects in Experiment 1 and all seven such subjects in Experiment 2. In Experiment 3, the LR test favored Post-Dec over SDT for 15 out of 27 subjects. These results suggest that, in the absence of model-based evidence, observations of hypermetacognitive efficiency (M-Ratio > 1) should not automatically be interpreted as support for postdecisional evidence accumulation.

Decay of the Confidence Signal

The Gauss and Decay models differ only in that the Decay model includes a parameter that allows the sensory signal to undergo decay before confidence is given. The LR test could not reject the simpler Gauss model for all 59 subjects in Experiment 1, but the evidence in Experiment 2 was mixed with the Decay model being favored for 12 out of the 20 subjects (Figure 10d). However, for Experiment 3, the Decay model was clearly favored for 44 out of 45 subjects. These results suggest that the explanatory power provided by the signal decay parameter may depend on qualitative features of the data, such that the decay parameter may prove useful for certain stimulus and experimental designs.

Partial PE Bias in Confidence

The PE and PE-Flex models differ only in that the PE-Flex model includes a parameter that allows for flexible, partial (instead of complete) neglect of decision-incongruent evidence. The LR test clearly favored the PE-Flex model for Experiments 1 and 2: it was preferred for 45 out of 59 subjects in Experiment 1 and for 18 out of 20 subjects in Experiment 2 (Figure 10e). However, for Experiment 3, the PE model was favored over the PE-Flex model for 29 out of 45 subjects. These analyses provide mixed evidence about whether the presumed PE bias in confidence is complete or partial.

Stimulus Visibility Effect on Confidence

The Gauss and WEV models differ only in that the WEV model includes a parameter that allows for the direct influence of task performance (named "stimulus visibility") in confidence ratings. The LR test rejected the simpler Gauss model for 10 out of 20 subjects in Experiment 2 and for 38 out of 45 subjects in Experiment 3 (note that WEV could not be fit in Experiment 1; Figure 10f). These results suggest that the addition of weighted visibility indeed benefits model fits, particularly for experimental designs such as in Experiment 3. We further explore the conditions that favor the visibility parameter in the Discussion.

Combining LogN and WEV: The logWEV Model

Our results showed that model ranks were relatively consistent across Experiments 1 and 2 but were drastically different for Experiment 3 (Figure 4). Specifically, models such as LogN, CASANDRE, and SDRM—which always featured in the top five models across Experiments 1 and 2—showed a steep drop in performance for Experiment 3 (LogN ranked sixth, CASANDRE ranked 11th, and SDRM ranked ninth). On the other hand, the PE model which was always found in the bottom three for Experiments 1 and 2, provided the best fits for Experiment 3. Of all these models, however, the WEV was the only model that performed consistently well across both Experiments 2 and 3 (ranked fourth in Experiment 2 and second in Experiment 3) and was able to fit all the observed qualitative patterns in the data across experiments.

At the outset, these results suggest that in Experiment 3, the assumption of LogN alone fails to explain confidence. Although the LogN model was the best model in Experiments 1 and 2, for Experiments 2 and 3 the WEV model either performed equally well or outperformed LogN. However, the WEV model reduces to the Gauss model for Experiment 1 (with a single task level), which performs significantly worse than LogN (ranking seven out of 12). Finally, WEV was the only model that could explain the qualitative patterns of confidence across all experiments. Based on these observations, we wanted to test whether a model that incorporates features from both these successful models could allow for a model that performs consistently well across all the experimental paradigms tested here and fits all the qualitative features across different data sets.

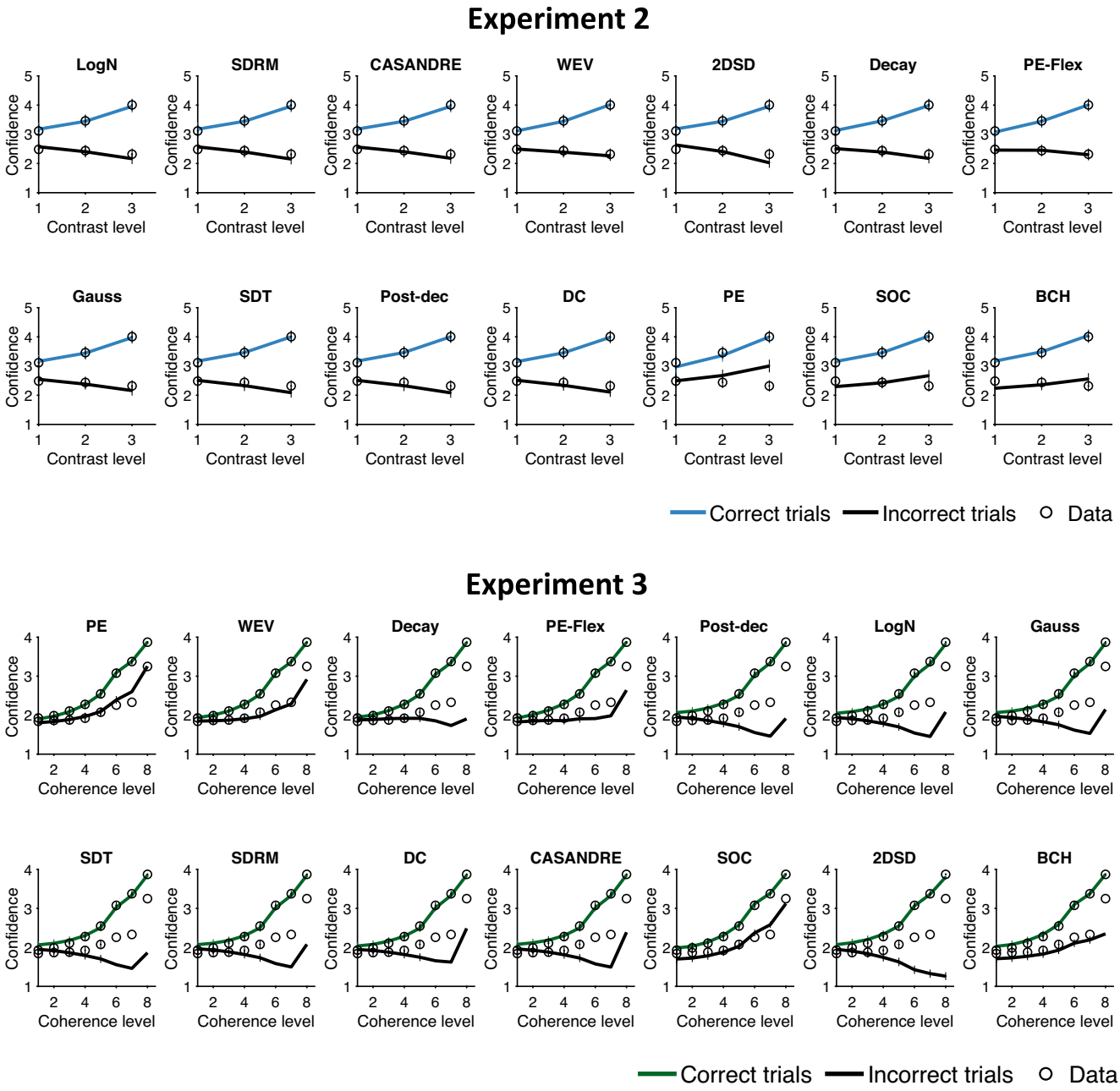
We developed the logWEV model by modifying the WEV model such that the confidence variable (which is a weighted sum of evidence and stimulus visibility) is corrupted by lognormal rather than Gaussian noise (see Supplementary Methods in the online supplemental materials for more details). We fit the logWEV model to Experiments 2 and 3, and compared its performance to the models that were previously ranked in the top five for each study (Figure 11). In terms of AIC

Figure 6 (*continued*)

receiver operating characteristics; zHR = *z*-transformed hit rate; zFAR = *z*-transformed false alarm rate; AIC = Akaike information criterion; LogN = lognormal meta noise model; CASANDRE = confidence as a noisy decision reliability estimate model; SOC = second-order confidence model; SDRM = stochastic detection and retrieval model; PE-Flex = flexible version of the PE model; 2DSD = two-stage dynamic signal detection model; Gauss = Gaussian meta noise model; Decay = noisy decay model; SDT = signal detection theory model; Post-Dec = postdecisional SDT model; PE = positive evidence bias model; DC = dual channel model; WEV = weighted evidence and visibility model; BCH = Bayesian confidence hypothesis model. See the online article for the color version of this figure.

Figure 7

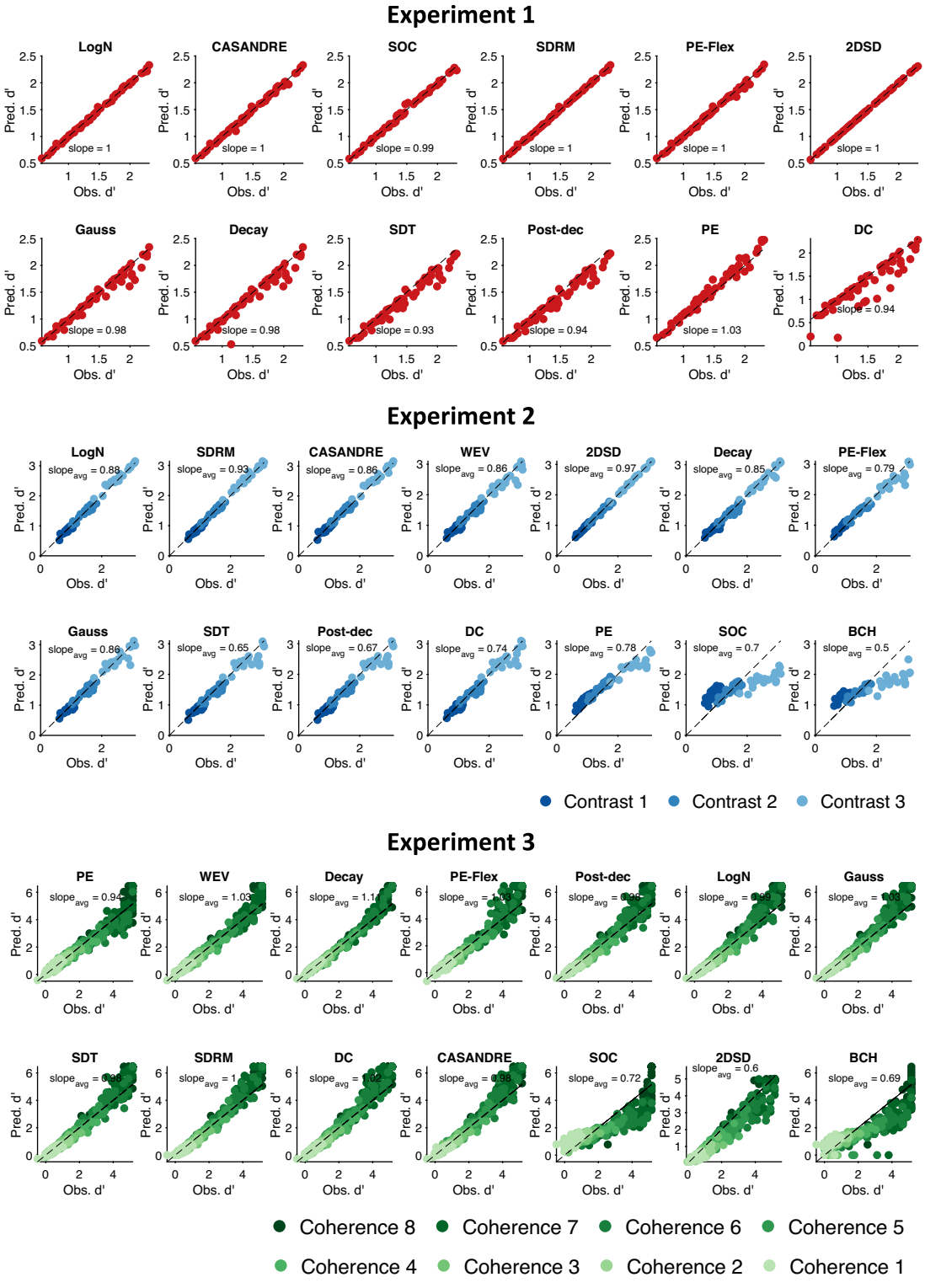
Model Fits for the FXP Averaged Across Subjects for Experiment 2



Note. The FXP is considered a classic signature of confidence that can be used to infer the underlying computations (Sanders et al., 2016). In Experiment 2, all but two models were able to capture the qualitative pattern well. Only the two models with the lowest AIC scores, PE and SOC, showed substantial deviations in that, for both models, confidence increased with stimulus discriminability for incorrect responses. The models are arranged according to their AIC scores. In Experiment 3, we did not observe the FXP. Rather, confidence for both correct and incorrect choices increased with stimulus discriminability. Most models—except PE, WEV, BCH, and SOC—failed to reproduce this pattern. Error bars show *SEM*. FXP = folded-X pattern; AIC = Akaike information criterion; LogN = lognormal meta noise model; SDRM = stochastic detection and retrieval model; CASANDRE = confidence as a noisy decision reliability estimate model; WEV = weighted evidence and visibility model; 2DSD = two-stage dynamic signal detection model; Decay = noisy decay model; PE-Flex = flexible version of the PE model; Gauss = Gaussian meta noise model; SDT = signal detection theory model; Post-Dec = postdecisional SDT model; DC = dual channel model; PE = positive evidence bias model; SOC = second-order confidence model; BCH = Bayesian confidence hypothesis model. See the online article for the color version of this figure.

Figure 8

Model Fits for Individual First-Order Task Performance (d′)



Note. Model-generated *d′* values for each subject are plotted against their observed values. For Experiments 2 and 3, the slopes are computed separately for each of the difficulty levels and the average is reported. All models are generally able to capture individual *d′* levels well with the notable exception of the SOC and BCH models in Experiment 2 and 3. The models are arranged according to their AIC scores. AIC = Akaike information criterion; LogN = lognormal meta noise model; CASANDRE = confidence as a noisy

(*figure continue*)

scores, the logWEV model was the second best-fitting model for Experiment 2 although its scores were not significantly different from the winning LogN model (AIC difference = 68.44, 95% CI [−75.01, 332.07]). For Experiment 3, the logWEV model was the best-fitting model with AIC differences being significant for all pairs of models (PE = 215.45, 95% CI [80.45, 412.41]; WEV = 483.53, 95% CI [234.46, 985.56]; Decay and PE-Flex > 1,279 AIC points). We also performed random-effects analyses with the AIC scores to obtain model frequencies and exceedance probabilities. For Experiment 2, the logWEV model had the third highest model frequency and exceedance probability after the LogN and WEV models ($p_{model} = .19$; $p_{exc} = .05$). For Experiment 3, the logWEV model had the second highest model frequency and exceedance probability ($p_{model} = .38$; $p_{exc} = .42$), although these scores were only marginally lower than the winning PE model ($p_{model} = .41$; $p_{exc} = .58$). Looking at overall model rankings across different measures (AIC/BIC) and analyses (fixed/random effects), the logWEV model ranked second (average rank = 2.75) for both Experiments 2 and 3. The logWEV model's overall rankings were a significant improvement over the WEV model in both experiments (average rank of 4.75 for Experiment 2 and 4 for Experiment 3) and the LogN model in Experiment 3 (average rank for Experiment 2 = 1.5; average rank for Experiment 3 = 7.75).

Finally, we assessed the logWEV model's ability to fit qualitative patterns in the data—namely d' and M-Ratio scores for individual subjects, zROC functions, and the FXP. As seen in Figure 11, the logWEV model provided close fits to the observed data across both experiments. As noted previously, the poor performance of most models including LogN in Experiment 3 could be attributed to their inability to fit the observation that confidence increased with stimulus discriminability for correct as well as incorrect choices. Since the LogWEV model incorporates the stimulus visibility effect, it was able to fit this pattern for Experiment 3. Together, these results suggest that logWEV model may offer a promising combination of the LogN and WEV models that generalizes well across different experimental designs and is able to reproduce all their observed qualitative features.

Discussion

We comprehensively examined the ability of 14 of the most popular process models of metacognition to fit data from basic perceptual discrimination experiments from three large data sets of increasing complexity. In Experiments 1 and 2, the LogN model was robustly selected as the model that provided the best fit to the data. In Experiment 3, the PE model yielded the best fits to the data. Finally, we tested a composite model that combines two of the most consistently well-performing models—LogN and WEV—and found that the resulting logWEV model performed better than any other individual model across the three experiments. These results shed light on the most plausible mechanisms underlying confidence generation and lay a solid foundation for future computational work on confidence generation.

The Generalizability of Model Results Across Data Sets

Our results showed good agreement between Experiments 1 and 2, but Experiment 3 seemed to produce very different results compared to the first two experiments. Specifically, the models that performed best for Experiments 1 and 2 (LogN, CASANDRE, and SDRM) were some of the worst-performing models in Experiment 3, whereas the best-performing model for Experiment 3 (PE) was in the bottom three in Experiments 1 and 2. Furthermore, Experiments 2 and 3 produced qualitatively different patterns of confidence for correct and error trials: while Experiment 2 generated the classic FXP, Experiment 3 showed a clear violation of this pattern. One possible explanation for these differences is that they are caused by the differences in stimuli. Specifically, in Experiment 2, the stimulus was a Gabor patch embedded within noise in such a way that the overall contrast of the stimulus (noise + Gabor patch) was always held constant. This may have made the difficulty of the stimulus harder to perceive, thus making it harder for subjects to use heuristic computations based on low-level stimulus cues about task difficulty. However, in Experiment 3, task difficulty was manipulated by changing the variance of motion of the dots, which is more readily apparent (high variance increases the seeming overall randomness of dot motion) and thus low-level stimulus cues about task difficulty are easier to use. This explanation is consistent with our observation that the WEV model—which can flexibly adapt the weight given to such cues—is able to maintain good fits in both experiments, whereas models such as PE and LogN—which are more rigid in their assumptions about the evidence used for confidence—only perform well in one but not the other experiment. These differences highlight the need to use different experimental designs in order to capture all relevant components of confidence computations.

The LogN Model

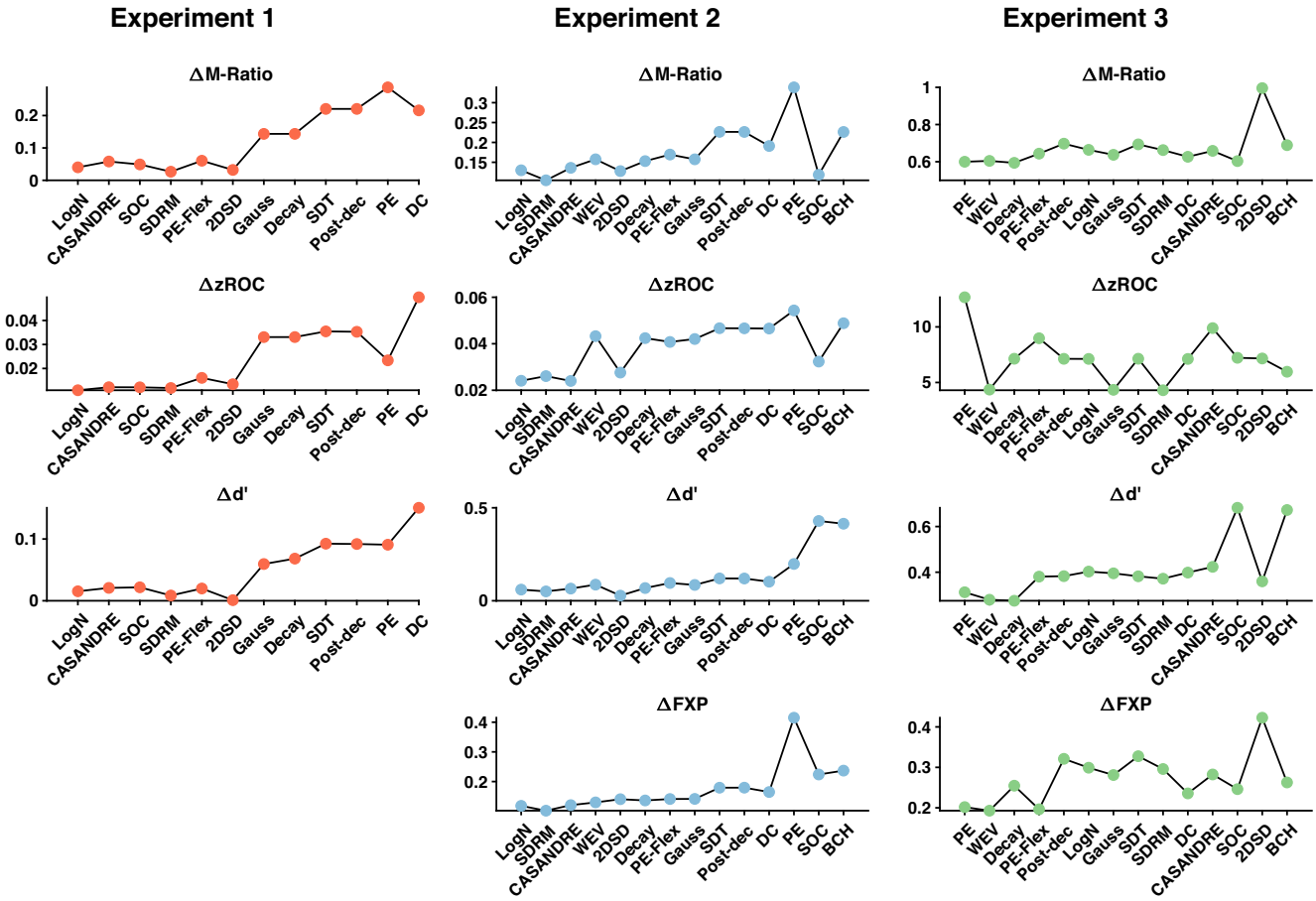
The LogN model was the best-fitting model across Experiments 1 and 2. LogN assumes that the primary decision is made identically to SDT and only diverges from SDT by introducing lognormal metacognitive noise. Our results build on previous studies in demonstrating that metacognitive noise is necessary to capture confidence ratings well (J. W. Bang et al., 2019; Maniscalco & Lau, 2016; Shekhar & Rahnev, 2021b; Xue et al., 2021). However, it is important to keep several caveats in mind. First, metacognitive noise likely occurs due to a constellation of factors such as serial dependence, stimulus variability, arousal, and so forth (Shekhar & Rahnev, 2021a) and including these factors as explicit components in a model is likely to reduce the need for this nonspecific noise term. Second, while there seems to be good evidence that metacognitive noise is likely to be signal-dependent and that confidence and

Figure 8 (*continued*)

decision reliability estimate model; SOC = second-order confidence model; SDRM = stochastic detection and retrieval model; PE-Flex = flexible version of the PE model; 2DSD = two-stage dynamic signal detection model; Gauss = Gaussian meta noise model; Decay = noisy decay model; SDT = signal detection theory model; Post-Dec = postdecisional SDT model; PE = positive evidence bias model; DC = dual channel model; WEV = weighted evidence and visibility model; BCH = Bayesian confidence hypothesis model; Pred. = predicted; Obs. = observed. See the online article for the color version of this figure.

Figure 9

Average Deviations for Each of the Qualitative Patterns



Note. We plotted the average deviation for each model for each of the four qualitative patterns we examined (M-Ratio, zROC functions, FXP, and *d'*). Models are arranged according to their AIC scores. All qualitative patterns are related to the models' AIC values but none of the relationships are perfect. zROC = z-transformed receiver operating characteristics; FXP = folded-X pattern; AIC = Akaike information criterion; LogN = lognormal meta noise model; CASANDRE = confidence as a noisy decision reliability estimate model; SOC = second-order confidence model; SDRM = stochastic detection and retrieval model; PE-Flex = flexible version of the PE model; 2DSD = two-stage dynamic signal detection model; Gauss = Gaussian meta noise model; Decay = noisy decay model; SDT = signal detection theory model; Post-Dec = postdecisional SDT model; PE = positive evidence bias model; DC = dual channel model; WEV = weighted evidence and visibility model; BCH = Bayesian confidence hypothesis model. See the online article for the color version of this figure.

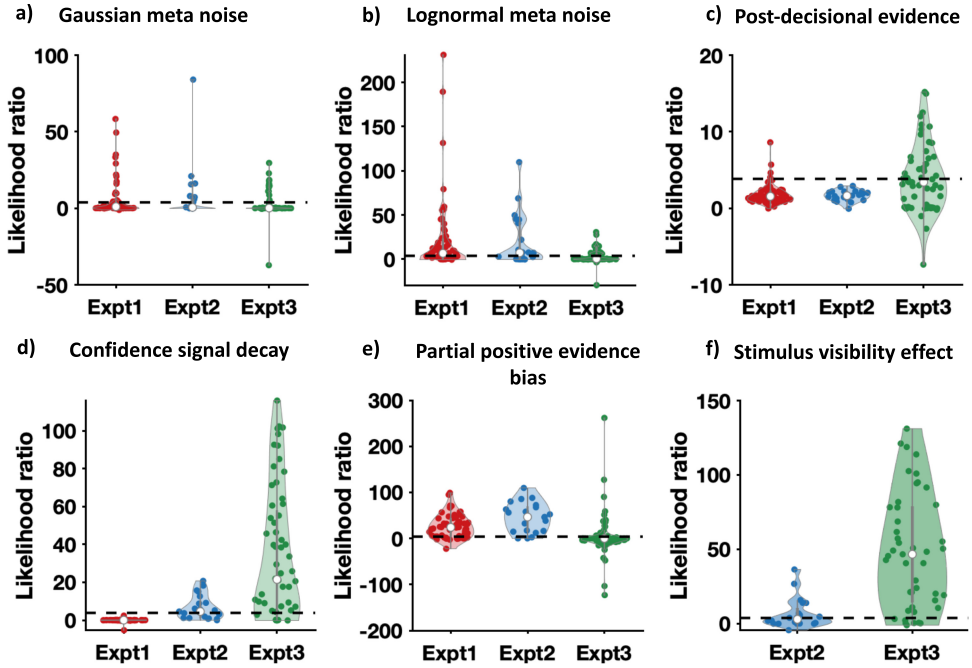
decision criteria are unlikely to cross, the exact shape of its underlying distribution may not be exactly lognormal. Third, the two experiments here used a very simple design; more complex designs would require more complex models. For example, if the stimulus manipulation changes not only the mean but also the variance of the underlying distributions, then confidence criteria may vary accordingly (Adler & Ma, 2018a; Denison et al., 2018). The LogN can easily accommodate criteria changing with the variance of the internal distributions or other external manipulations but is silent about such manipulations. Lastly, we see a drop in the LogN model's performance for Experiment 3, suggesting that the model does not fully capture the complexity of confidence computations across different tasks. Overall, we see the LogN model as a simple but solid foundation which future modeling efforts can build on. The LogN model's success also carries implications for the characteristics of noise in neural systems, although the causal factors underlying signal-dependent noise remain to be elucidated.

The WEV Model

The WEV model performed very well in our study—it was the second best-fitting model in both the experiments in which it was tested in terms of overall rank. In fact, it was the only model whose performance generalized across the data sets in Experiments 2 and 3 and it was the only model that was able to flexibly capture all the qualitative features across data sets. Furthermore, the inclusion of the "visibility" parameter led to an improvement over the simpler Gauss model (which is otherwise equivalent to WEV except for the extra parameter) in a substantial proportion of subjects (50% in Experiment 2 and 84% in Experiment 3). Therefore, the notion that confidence is influenced by a visibility heuristic should be taken seriously and investigated further. Nevertheless, we believe that more work is needed to clarify how the stimulus "visibility" used in the model generalizes across paradigms. In practice, this label fits better in the metacontrast masking designs examined by Rausch et al. (2018, 2020) where

Figure 10

Examining the Plausibility of Individual Model Parameters



Note. For each of the three experiments, we compared six pairs of nested models that only differ in the presence of a single parameter: (a) SDT versus Gauss (parameter: Gaussian metacognitive noise), (b) SDT versus LogN (parameter: lognormal metacognitive noise), (c) SDT versus Post-Dec (parameter: postdecisional evidence accumulation), (d) Gauss versus Decay (parameter: confidence signal decay), (e) PE versus PE-Flex (parameter: partial use of PE), and (f) Gauss versus WEV (parameter: stimulus visibility effect; Experiments 2 and 3 only because WEV is equivalent to Gauss in Experiment 1). The plots show the LR scores between the full (Model 2 in each pair) and reduced (Model 1 in each pair) models. For a substantial proportion of subjects (always over 50%), there is evidence in favor of either type of metacognitive noise, for partial (as opposed to complete) PE bias as well as the stimulus visibility effect. Conversely, there is little consistent evidence for postdecisional evidence accumulation or confidence signal decay across the three experiments. The dashed vertical line marks the critical LR value at the 0.05 significance level. For LR values falling above the dashed line, one can reject the reduced model in favor of the full model. Expt = experiment; SDT = signal detection theory model; Gauss = Gaussian meta noise model; LogN = lognormal meta noise model; Decay = noisy decay model; Post-Dec = postdecisional SDT model; WEV = weighted evidence and visibility model; PE = positive evidence bias model; PE-Flex = flexible positive evidence bias model; LR = likelihood ratio. See the online article for the color version of this figure.

longer stimulus-onset asynchronies lead to both higher performance and higher subjective visibility of the stimuli. In contrast, the "visibility" label fits less well in the current designs where all stimuli were easy to detect (though the two alternatives were hard to discriminate between). In the context of the tasks used in this study, we think that a "sensitivity" or "discriminability" bias may be a more appropriate characterization of this heuristic where observers are able to maintain and use an internal estimate of their own d' to inform their confidence. However, the WEV model assumes that the observer's internal estimates of stimulus discriminability are perfect (i.e., noiseless and unbiased), which is unlikely to be true. It is also an open question how such "discriminability" is computed and used in more real-world scenarios where more than two alternatives are available (Rahnev, 2020). Nevertheless, it should be noted that issues of generalizability emerge for most current models and are not specific to the assumptions made by the WEV model.

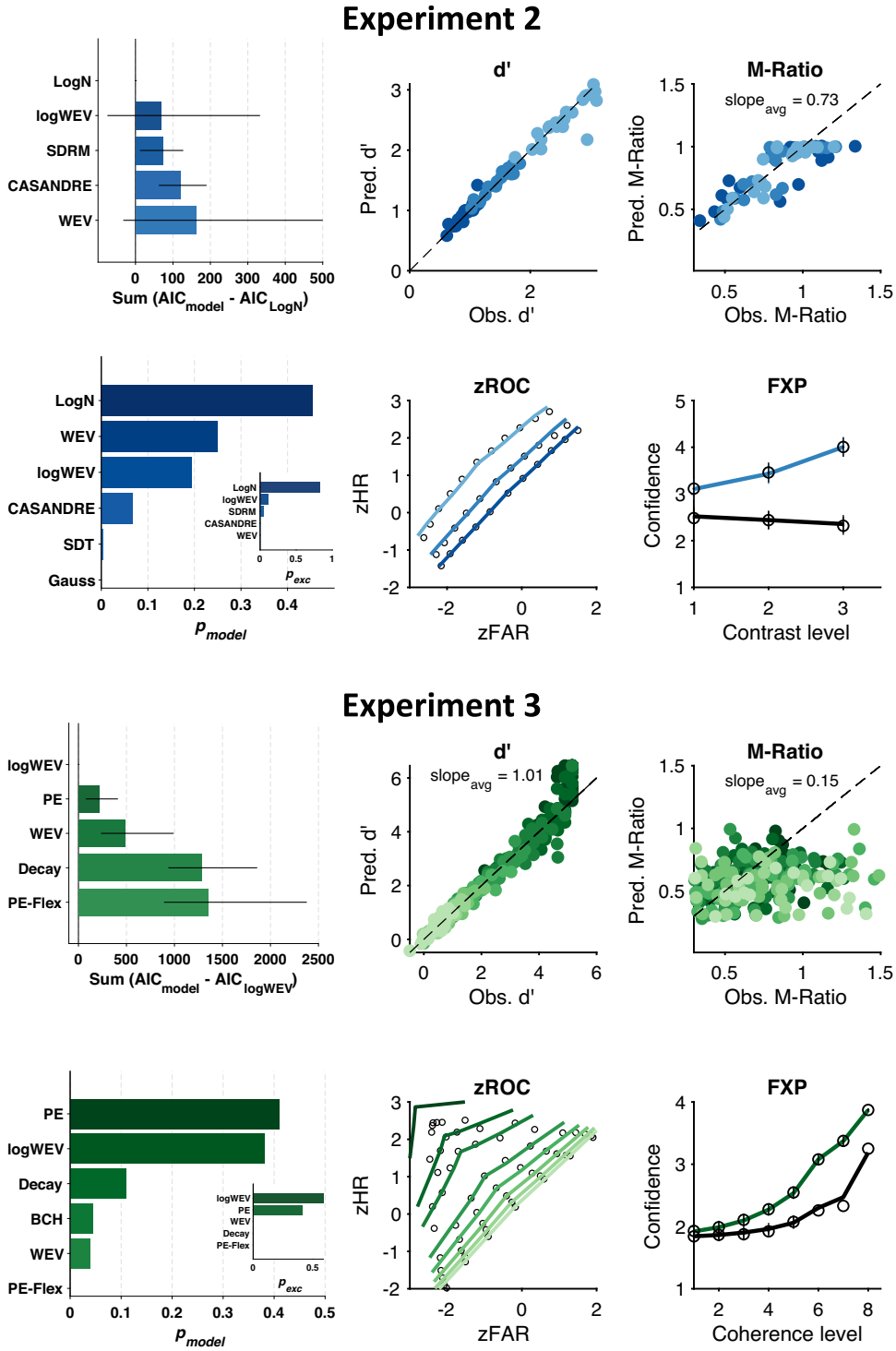
Implications for Various Proposals Related to Confidence Computation

Confidence as the Probability of Being Correct

The notion that confidence is computed as a Bayesian probability of being correct has gained substantial popularity in recent years (Hangya et al., 2016; Pouget et al., 2016; Sanders et al., 2016). Two out of the 14 models examined here—the SOC and BCH models—implemented this proposal (all of the rest implement a more traditional, distance-to-criterion computation). The results were stark: SOC had the second lowest average AIC value in Experiment 1 (where the presence of a single difficulty level obviated the difference between Bayesian and distance-to-criterion computations) but provided some of the worst fits of all models in Experiments 2 (where the presence of three difficulty levels led to a large difference between Bayesian and distance-to-criterion computations). Even the

Figure 11

Results for the logWEV Model in Experiments 2 and 3



Note. (Top) Results for Experiment 2. Summed AIC differences between the best-fitting model and the previous top five models. AIC comparisons show that the logWEV model is the second best-fitting model, although the AIC difference with the best-fitting LogN model is not significant. Random-effects analyses show that the logWEV model has the third highest model frequency after the LogN and WEV models. Qualitative analyses show that the logWEV model is able to fit individual *d'*, M-Ratio scores, zROC functions, and the FXP. (Bottom) Results for Experiment 3. The logWEV model is the best-fitting model with the lowest AIC scores (the other four models have significantly higher AIC scores).

(*figure continue*)

BCH model which was fit only to Experiments 2 and 3, consistently ranked in the bottom three for these experiments. These findings using the simple design in Experiments 2 and 3 accord well with several other recent empirical investigations with more complex designs that also cast doubt over the Bayesian computation of the probability of being correct as underlying confidence ratings (Adler & Ma, 2018a, 2018b; Bertana et al., 2021; Denison et al., 2018; Li & Ma, 2020; Locke et al., 2020).

The FXP is popularly assumed to be a signature of Bayesian confidence computations. This pattern was first demonstrated by Hangya et al. (2016) via simulations of a model that implements Bayesian confidence. However, both the Bayesian models we implement here—BCH and SOC—fail to generate this pattern. Instead, confidence in these models increases with stimulus discriminability for both correct and incorrect trials. This apparent discrepancy can be resolved when we account for the differences in model assumptions and experimental design. In their article, Hangya et al. (2016) defined stimulus discriminability as the magnitude of internal evidence for a single stimulus level. In our study, in contrast, we define stimulus discriminability based on the stimulus itself (specifically, its contrast/coherence) and externally vary this attribute. Indeed, Hangya et al. (2016) specifically noted that the emergence of the FXP relies on the assumption that the observers have no inherent knowledge of the different experimental conditions. Furthermore, in a follow-up article by the same group, Sanders et al. (2016) varied stimulus discriminability externally, but their model still assumed that observers have no ability to discriminate between different levels of the stimulus and instead assumed single underlying stimulus level. Our Bayesian models, however, do not make this assumption and allow observers to account for the different stimulus conditions, which explains the observed violation of the FXP. More importantly, the BCH model is indistinguishable from standard SDT when modeling a single stimulus level. The models can only be distinguished when they account for more than one stimulus condition.

Postdecisional Accumulation or Decay of Evidence for Confidence

One popular proposal regarding the computation underlying confidence is that it is at least partly based on a process of either accumulation or decay that occurs after the primary decision has been made (Calder-Travis et al., 2020; Pleskac & Busemeyer, 2010; Wokke et al., 2020). Three of the current models implemented different versions of this general proposal: Post-Dec and 2DSD implemented postdecisional accumulation, whereas Decay implemented postdecisional signal loss. For Experiments 1 and 2, we found no evidence that these model features lead to improved model fits for the current data sets. In fact, using a LR test to compare Post-Dec to its reduced version that lacks postdecisional accumulation (SDT) and Decay to its reduced version that lacks postdecisional signal loss (Gauss), we found evidence against the postdecisional parameters introduced by both models. Furthermore, although a small proportion of subjects displayed hypermetacognitive efficiency (M-Ratio > 1; 12 out of 79 subjects across both experiments), Post-Dec did not provide a better fit for even these subjects compared to SDT. For Experiment 3, however, evidence supported the decay and postdecisional mechanisms in a majority of the subjects. These inconsistent findings between data sets suggest that the utility of postdecisional mechanisms may depend on the stimulus set and experimental design. Further work is required to exactly characterize the conditions where such postdecisional mechanisms become useful. Of course, postdecisional processes certainly operate in cases where the stimulus continues to be presented after a decision has been made (Rollwage et al., 2018, 2020; Schulz et al., 2020) and are also likely to be important when decisions are made in a speeded manner. However, when this is not the case (such as in the current experiments), these mechanisms may not always improve fits in the context of simple perceptual discrimination tasks.

Architectures With Different Signals for Decision and Confidence

Eleven out of the 14 models examined here (SDT, Gauss, LogN, Decay, Post-Dec, WEV, PE, PE-Flex, BCH, CASANDRE, and 2DSD) included a single evidence stream used for both the decision and confidence. However, the remaining three models (DC, SDRM, and SOC) postulated architectures where the signals for the decision and confidence are fundamentally different (though still correlated). None of these models were preferred consistently across the three experiments, thus casting doubt at the ideas that different signals underlie the primary decision and confidence judgments. The DC model specifically always ranked in the bottom three models for all experiments, strongly suggesting that the idea of different channels for high- and low-confidence decisions is not viable. One important caveat in these results is that the model recovery analyses showed very poor recoverability for SDRM in Experiments 1 and 2, and for SOC in Experiment 1 (both models were typically confused with 2DSD). However, both SDRM and SOC were almost never confused with the winning models, thus making it unlikely that the lack of good fits for SDRM and SOC is primarily due to model recoverability issues. This confusability underscores the necessity of conducting model recovery analyses in future modeling studies, especially for models that propose separate signals for decision and confidence.

Figure 11 (*continued*)

The logWEV model has the second highest model frequency and exceedance probability in the population. The logWEV model is again able to capture all the qualitative patterns of d', M-Ratio, zROC, and the FXP. AIC = Akaike information criterion; logWEV = WEV model with lognormal meta noise; PE = positive evidence bias model; WEV = weighted evidence visibility model; Decay = noisy decay model; PE-Flex = flexible version of the PE model; Pred. = predicted; Obs. = observed; avg = average; BCH = Bayesian confidence hypothesis model; WEV = weighted evidence and visibility model; p_{exc} = protected exceedance probabilities; p_{model} = model frequencies; zHR = z-transformed hit rate; zFAR = z-transformed false alarm rate; zROC = z-transformed receiver operating characteristics; FXP = folded-X pattern. See the online article for the color version of this figure.

PE Bias

One of the most prominent proposals regarding the computations underlying confidence is that they selectively ignore decision-incongruent evidence (Maniscalco et al., 2016; Peters et al., 2017; Zylberberg et al., 2012). The "positive evidence bias" was originally proposed by Zylberberg et al. (2012) based on reverse correlation analyses that suggested that decision-incongruent evidence does not influence confidence reports. However, to the best of our knowledge, these initial findings have not been replicated and other studies using reverse correlation analyses have found that confidence appears to in fact be sensitive to decision-incongruent evidence (Weise et al., 2021). Our results demonstrate that a pure PE bias (as implemented in the current PE model) is untenable in the absence of additional mechanisms. Indeed, the PE model provided some of the worst model fits for Experiments 1 and 2, and also resulted in constant M-Ratio values at 0.5 (which strongly contradicts empirical reality). The M-Ratio result had not been appreciated before but is sensible in retrospect since according to the PE model, confidence ignores half of the available information. Surprisingly, the PE model emerged as the best model for Experiment 3 and its good performance was associated with its ability to explain the observed pattern of increasing confidence for incorrect choices. However, we believe it is unlikely that this result underlies a true shift in the observers' confidence generation strategies across data sets and that a more general and flexible strategy underlies their observed behavior. Indeed, the new logWEV model—which does not include a PE bias—performed at least as well as PE for Experiment 3.

The PE-Flex model performed markedly better but still worse than models with no PE bias assumption. However, it should be noted that a recent article implemented a pure PE bias in the context of sequential sampling—where the confidence signal could accumulate both extra information and extra noise—and was able to successfully fit the observed M-Ratio values (Maniscalco et al., 2021). Therefore, to fully evaluate the notion of PE bias, it is necessary to systematically vary the different auxiliary assumptions or to devise manipulations that directly test the predictions of the PE computations.

This conclusion may appear surprising given the success of what has been dubbed "positive evidence" manipulations. In these manipulations, stimuli with high signal and high noise are matched in performance to stimuli with low signal and low noise, but the high-intensity stimuli are routinely found to produce higher confidence (Koizumi et al., 2015; Samaha et al., 2016; Zylberberg et al., 2012). Because high-intensity stimuli produce stronger evidence for both the decision-congruent and decision-incongruent choices, such confidence-accuracy dissociations can be naturally explained as a result of a PE bias. However, critically, such dissociations can also be explained in several other ways, such as by assuming that increasing both the signal and noise in a stimulus affects both the signal and variability in the internal evidence distributions. In fact, a recent study showed that high confidence for high-intensity stimuli is observed even in neural networks trained to take all evidence into account (Webb et al., 2021). Therefore, we suggest that empirical confidence-accuracy dissociations where high-intensity stimuli with matched d' produce higher confidence rating be renamed to "high-intensity–high-confidence" effect. The names "positive evidence bias" and "response-congruency effect" presuppose a mechanistic explanation rather than simply describing an empirical effect.

Confidence as an Estimate of Decision Reliability

A recent model proposed the idea that confidence is derived from observers' subjective estimates of decision reliability (CASANDRE; Boundy-Singer et al., 2023). Critically, the model assumes that an observer can maintain only a noisy representation of their own decision noise (quantified as meta-uncertainty), allowing the model to account for inefficient metacognition. The CASANDRE model performed well in Experiments 1 and 2 where it was the second and third best-performing model, respectively (according to AIC scores). However, like other models, it did not provide a good fit to Experiment 3, suggesting that model does not fully capture the complexity of confidence computations across different tasks.

One important observation regarding the CASANDRE model is that it appears somewhat related to the LogN model. This was especially the case for Experiment 1 where CASANDRE could not be reliably distinguished from LogN (50% probability of recovery) and most subjects showed nearly identical AIC scores, suggesting that the models may be similar in the presence of a single difficulty level. Nevertheless, CASANDRE performed significantly worse (though by a relatively small margin) than the LogN model in all three experiments. We note that Boundy-Singer et al. (2023) previously found that CASANDRE and LogN performed equivalently for Experiment 2, whereas we find that LogN performed better by an average of about six AIC points in that experiment. This discrepancy comes from the fact that Boundy-Singer et al. (2023) compared their CASANDRE model fits the LogN model fits from Shekhar and Rahnev (2021b) where we used a less optimal fitting algorithm. In the current study, we use the BADS algorithm (Acerbi & Ma, 2017), which is a much more efficient search algorithm compared to our previous custom-built procedure. Using BADS led to an overall improvement in the LogN model's performance (by about six AIC points on average) compared to our previous fits, but produced the same results for CASANDRE compared to Boundy-Singer et al.'s fits.

General Modeling Considerations

Model and Parameter Recovery

Model and parameter recovery have rarely been performed in the context of previous models of metacognition. While our results suggested that, within the context of our large experiments, that both model and parameter recovery for existing models is high, there were several important exceptions. First, model recovery was notably worse in Experiment 1 where only a single difficulty level was present, thus emphasizing the need for using conditions with varying difficulty. Second, model recovery was also worse in Experiment 3 compared to Experiment 2. Even though Experiment 3 contained more task conditions, it had fewer overall trials (1,600) compared to Experiment 2 (2,800), suggesting that larger trial numbers are important for reliably discriminating between models. Third, some seemingly unrelated models such as SDRM and 2DSD were confused with each other, which may suggest the presence of deeper links between the models. Fourth, several specific parameters such as correlations between the evidence for decision and confidence could not be recovered robustly, suggesting that the fitted values for such parameters need to be interpreted with extreme caution. Overall, these results underscore the importance of both model and parameter recovery analyses in future model development. Nevertheless, it is important to note that some models (e.g., SDRM) that showed relatively poor

model recovery nevertheless produced good models fits, a phenomenon that requires further focused investigation.

Qualitative Patterns in the Data

Several qualitative patterns have been proposed to directly reveal different aspects of the computations underlying confidence ratings. For example, the inverted-U shape of zROC curves has been seen as directly suggesting the existence of signal-dependent metacognitive noise (Shekhar & Rahnev, 2021b). However, we found that the zROC pattern was at least qualitatively explained by most existing models, thus questioning whether they can be used to directly infer the details of the internal computations.

Similarly, the FXP—the observation that easier stimuli lead to an increase in confidence for correct trials but a decrease in confidence for error trials—has been proposed as a signature of "statistical confidence" (Hangya et al., 2016). However, we only observed this pattern for Experiment 2. In Experiment 3, confidence increased with stimulus discriminability irrespective of whether the choice was correct or incorrect. Such violations of the FXP have also been previously reported by others (Rausch et al., 2018, 2020), suggesting that the FXP is not a reliable signature of confidence. Rather, the patterns of confidence observed for correct versus incorrect choices may be important to understand how different stimulus features or experimental paradigms interact with the confidence generation process.

A model's ability to fit individual variations in M-Ratio values as well as d' was also found to reasonably constrain model performances. These measures were particularly useful for understanding why certain models perform poorly.

Overall, qualitative patterns in the data remain important for understanding why a model may not provide a good fit but may have limited utility in and of themselves to constrain the modeling of metacognition. Furthermore, as experiments become more complex, the qualitative features of a data set become less informative in terms of explaining model performances.

Parameter Dependence on Difficulty Level

Three of the models (Decay, DC, and PE-Flex) included parameters that depend on task difficulty. While there is nothing mathematically wrong with such model specifications, it is questionable how plausible they are. The issue is that if the internal computations differ based on the difficulty level, the system should first have perfect knowledge about the difficulty level on each trial. However, most experiments (including our Experiments 2 and 3) interleave the different difficulty levels and do not inform subjects about the difficulty level used on each trial. There could, of course, be mechanisms that work to infer the condition that each trial comes from (e.g., the contrast in Experiment 2) but such an inference process is likely to be imperfect and should ideally be modeled too.

Limitations and Generality Constraints

Exploring the Full Space of Model Variants

Here we focused on fitting well-defined existing models but did not examine the full space of possible models. For example, the Decay, SDRM, and SOC models feature Gaussian metacognitive noise but may in theory be improved by the use of lognormal metacognitive noise instead. Similarly, most models can be augmented by adding postdecisional evidence accumulation, signal decay, visibility weighting, metacognitive noise, decision noise (which currently only features in SDRM), or a lapse rate parameter. In addition, models such as SDRM and Decay include multiple sources of corruption and a full investigation would test reduced versions of these models. Finally, since the beginning of our work, there have been several new models of metacognition that did not make it in our list (Guggenmos, 2022; Hu et al., 2021; Mamassian & de Gardelle, 2022). Future modeling work can build on the current results by exploring model variants with combinations of features, additional reduced model versions, as well as adding new models as they are formulated.

Using Generic Rather Than Targeted Task Data

Another limitation of the current study is that we included either no difficulty manipulation (Experiment 1) or a very simple difficulty manipulation (change in contrast or coherence levels; Experiments 2 and 3). In this sense, our data were "generic" rather than targeted toward testing any specific model. This decision was deliberate on our part as complex manipulations can lead to violations in the auxiliary assumptions made by the models. Nevertheless, this choice makes it difficult to draw conclusions about any specific proposed mechanism and instead only allows the comparison of fully formed models. Convincingly confirming or falsifying any proposed mechanism of confidence computation would thus require more targeted manipulations that specifically address predictions made by that mechanism.

Not Fitting RT or Other Types of Data

Finally, our focus in this study was maximally narrow by only considering choice and confidence data but ignoring a host of other relevant measures such as RT, brain activity, pupil dilation, and arousal. It is possible that including additional types of data would require fundamentally different types of models with only a loose connection to the existing models. Nevertheless, we hope that our results that focus only on the most basic data would still have implications for more complex models that try to capture additional types of data.

Conclusion

Using extensive model comparisons, we demonstrate the strengths and weaknesses of 14 popular models of metacognition across three large data sets. These results provide tentative support for two mechanisms—the presence of signal-dependent metacognitive noise and a selectively visibility bias that only affects confidence. Indeed, a new model based on both of these mechanisms—the logWEV model—provides the best overall performance across all three experiments. The present comprehensive assessment of models of metacognition provides a solid foundation for future efforts to build even better models of confidence generation.

References

Acerbi, L., & Ma, W. J. (2017, December 4–December 9). *Practical Bayesian optimization for model fitting with Bayesian adaptive direct search* [Conference session]. 31st Annual Conference on Neural Information Processing Systems (NIPS), Long Beach, Canada.

Adler, W. T., & Ma, W. J. (2018a). Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLoS Computational Biology*, *14*(11), Article e1006572. https://doi.org/10.1371/journal.pcbi.1006572

Adler, W. T., & Ma, W. J. (2018b). Limitations of proposed signatures of Bayesian confidence. *Neural Computation*, *30*(12), 3327–3354. https://doi.org/10.1162/neco_a_01141

Allen, M., Glen, J. C., Müllensiefen, D., Schwarzkopf, D. S., Fardo, F., Frank, D., Callaghan, M. F., & Rees, G. (2017). Metacognitive ability correlates with hippocampal and prefrontal microstructure. *NeuroImage*, *149*, 415–423. https://doi.org/10.1016/j.neuroimage.2017.02.008

Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, *115*(23), 6082–6087. https://doi.org/10.1073/pnas.1800795115

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

Barrett, A. B., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological Methods*, *18*(4), 535–552. https://doi.org/10.1037/a0033268

Bertana, A., Chetverikov, A., van Bergen, R. S., Ling, S., & Jehee, J. F. M. (2021). Dual strategies in human confidence judgments. *Journal of Vision*, *21*(5), Article 21. https://doi.org/10.1167/jov.21.5.21

Boundy-Singer, Z. M., Ziemba, C. M., & Goris, R. L. T. (2023). Confidence reflects a noisy decision reliability estimate. *Nature Human Behaviour*, *7*(1), 142–154. https://doi.org/10.1038/s41562-022-01464-x

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Calder-Travis, J. M., Charles, L., Bogacz, R., & Yeung, N. (2020 June 2). *Bayesian confidence in optimal decisions*. https://doi.org/10.31234/OSF.IO/J8SXZ

Daunizeau, J., Adam, V., & Rigoux, L. (2014). VBA: A probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Computational Biology*, *10*(1), Article e1003441. https://doi.org/10.1371/journal.pcbi.1003441

del Cul, A., Dehaene, S., Reyes, P., Bravo, E., & Slachevsky, A. (2009). Causal role of prefrontal cortex in the threshold for access to consciousness. *Brain*, *132*(9), 2531–2540. https://doi.org/10.1093/brain/awp111

Denison, R. N., Adler, W. T., Carrasco, M., & Ma, W. J. (2018). Humans incorporate attention-dependent uncertainty into perceptual decisions and confidence. *Proceedings of the National Academy of Sciences*, *115*(43), 11090–11095. https://doi.org/10.1073/pnas.1717720115

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological Science*, *29*(5), 761–778. https://doi.org/10.1177/0956797617744771

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

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: Computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1280–1286. https://doi.org/10.1098/rstb.2012.0021

Fleming, S. M., Huijgen, J., & Dolan, R. J. (2012). Prefrontal contributions to metacognition in perceptual decision making. *The Journal of Neuroscience*, *32*(18), 6117–6125. https://doi.org/10.1523/JNEUROSCI.6489-11.2012

Fleming, S. M., Massoni, S., Gajdos, T., & Vergnaud, J.-C. (2016). Metacognition about the past and future: Quantifying common and distinct influences on prospective and retrospective judgments of self-performance. *Neuroscience of Consciousness*, *2016*(1), Article niw018. https://doi.org/10.1093/nc/niw018

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, *329*(5998), 1541–1543. https://doi.org/10.1126/science.1191883

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

Guggenmos, M. (2022). Reverse engineering of metacognition. *eLife*, *11*, Article e75420. https://doi.org/10.7554/eLife.75420

Haddara, N., & Rahnev, D. (2022). The impact of feedback on perceptual decision-making and metacognition: Reduction in bias but no change in sensitivity. *Psychological Science*, *33*(2), 259–275. https://doi.org/10.1177/09567976211032887

Hangya, B., Sanders, J. L., & Kepecs, A. (2016). A mathematical framework for statistical decision confidence. *Neural Computation*, *28*(9), 1840–1858. https://doi.org/10.1162/NECO_a_00864

Hu, X., Zheng, J., Su, N., Fan, T., Yang, C., Yin, Y., Fleming, S. M., & Luo, L. (2021). A Bayesian inference model for metamemory. *Psychological Review*, *128*(5), 824–855. https://doi.org/10.1037/rev0000270

Jang, Y., Wallsten, T. S., & Huber, D. E. (2012). A stochastic detection and retrieval model for the study of metacognition. *Psychological Review*, *119*(1), 186–200. https://doi.org/10.1037/a0025960

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

Koizumi, A., Maniscalco, B., & Lau, H. (2015). Does perceptual confidence facilitate cognitive control? *Attention, Perception, & Psychophysics*, *77*(4), 1295–1306. https://doi.org/10.3758/S13414-015-0843-3

Koriat, A. (2006). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (Vol. 3, No. 2, pp. 289–325). Cambridge University Press. https://doi.org/10.1017/CBO9780511816789.012

Li, H. H., & Ma, W. J. (2020). Confidence reports in decision-making with multiple alternatives violate the Bayesian confidence hypothesis. *Nature Communications*, *11*(1), Article 2004. https://doi.org/10.1038/s41467-020-15581-6

Locke, S. M., Gaffin-Cahn, E., Hosseinizaveh, N., Mamassian, P., & Landy, M. S. (2020). Priors and payoffs in confidence judgments. *Attention, Perception, & Psychophysics*, *82*(6), 3158–3175. https://doi.org/10.3758/s13414-020-02018-x

Mamassian, P., & de Gardelle, V. (2022). Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychological Review*, *129*(5), 976–998. https://doi.org/10.1037/rev0000312

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

Maniscalco, B., Odegaard, B., Grimaldi, P., Cho, S. H., Basso, M. A., Lau, H., & Peters, M. A. K. (2021). Tuned inhibition in perceptual decision-making circuits can explain seemingly suboptimal confidence behavior. *PLOS Computational Biology*, *17*(3), Article e1008779. https://doi.org/10.1371/journal.pcbi.1008779

Maniscalco, B., Peters, M. A. K., & Lau, H. (2016). Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Attention, Perception, & Psychophysics*, *78*(3), 923–937. https://doi.org/10.3758/s13414-016-1059-x

Metcalfe, J., & Shimamura, A. P. (1994). *Metacognition: Knowing about knowing*. MIT Press. https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1738390

Morales, J., Lau, H., & Fleming, S. M. (2018). Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *The Journal of Neuroscience*, *38*(14), 3534–3546. https://doi.org/10.1523/JNEUROSCI.2360-17.2018

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and some new findings. *The Psychology of Learning and Motivation*, *26*, 125–173. https://doi.org/10.1016/S0079-7421(08)60053-5

Orchard, E. R., Dakin, S. C., & van Boxtel, J. J. A. (2022). Internal noise measures in coarse and fine motion direction discrimination tasks and the correlation with autism traits. *Journal of Vision*, *22*(10), Article 19. https://doi.org/10.1167/jov.22.10.19

Peters, M. A. K., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., Doyle, W., Kuzniecky, R., Devinsky, O., Halgren, E., & Lau, H. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour*, *1*(7), Article 0139. https://doi.org/10.1038/s41562-017-0139

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. https://doi.org/10.1037/a0019737

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

Rahnev, D. (2020). Confidence in the real world. *Trends in Cognitive Sciences*, *24*(8), 590–591. https://doi.org/10.1016/j.tics.2020.05.005

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

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review*, *120*(3), 697–719. https://doi.org/10.1037/a0033152

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

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

Rollwage, M., Dolan, R. J., & Fleming, S. M. (2018). Metacognitive failure as a feature of those holding radical beliefs. *Current Biology*, *28*(24), 4014–4021.e8. https://doi.org/10.1016/j.cub.2018.10.053

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), Article 2634. https://doi.org/10.1038/s41467-020-16278-6

Rounis, E., Maniscalco, B., Rothwell, J. C., Passingham, R. E., & Lau, H. (2010). Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. *Cognitive Neuroscience*, *1*(3), 165–175. https://doi.org/10.1080/17588921003632529

Samaha, J., Barrett, J. J., Sheldon, A. D., LaRocque, J. J., & Postle, B. R. (2016). Dissociating perceptual confidence from discrimination accuracy reveals no influence of metacognitive awareness on working memory. *Frontiers in Psychology*, *7*, Article 851. https://doi.org/10.3389/fpsyg.2016.00851

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

Schulz, L., Rollwage, M., Dolan, R. J., & Fleming, S. M. (2020). Dogmatism manifests in lowered information search under uncertainty. *Proceedings of the National Academy of Sciences*, *117*(49), 31527–31534. https://doi.org/10.1073/pnas.2009641117

Shekhar, M., & Rahnev, D. (2018). Distinguishing the roles of dorsolateral and anterior PFC in visual metacognition. *The Journal of Neuroscience*, *38*(22), 5078–5087. https://doi.org/10.1523/JNEUROSCI.3484-17.2018

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

Shekhar, M., & Rahnev, D. (2021b). The nature of metacognitive inefficiency in perceptual decision making. *Psychological Review*, *128*(1), 45–70. https://doi.org/10.1037/rev0000249

Shimamura, A. P. (2000). Toward a cognitive neuroscience of metacognition. *Consciousness and Cognition*, *9*(2), 313–323. https://doi.org/10.1006/ccog.2000.0450

Vickers, D. (1979). *Decision processes in visual perception*. Academic Press. https://doi.org/10.1016/c2013-0-11654-6

Webb, T. W., Miyoshi, K., So, T. Y., Rajananda, S., & Lau, H. (2021). *Performance-optimized neural networks as an explanatory framework for decision confidence*. bioRxiv. https://doi.org/10.1101/2021.09.28.462081

Weise, L., Forster, S. D., & Gauggel, S. (2021). Reverse-correlation reveals internal error-corrections during information-seeking. *Metacognition and Learning*, *17*(2), 321–335. https://doi.org/10.1007/s11409-021-09286-4

Wilson, R. C., & Collins, A. G. E. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*, Article e49547. https://doi.org/10.7554/eLife.49547

Wokke, M. E., Achoui, D., & Cleeremans, A. (2020). Action information contributes to metacognitive decision-making. *Scientific Reports*, *10*(1), Article 3632. https://doi.org/10.1038/s41598-020-60382-y

Wokke, M. E., Cleeremans, A., & Ridderinkhof, K. R. (2017). Sure i'm sure: Prefrontal oscillations support metacognitive monitoring of decision making. *The Journal of Neuroscience*, *37*(4), 781–789. https://doi.org/10.1523/JNEUROSCI.1612-16.2016

Xue, K., Shekhar, M., & Rahnev, D. (2021). Examining the robustness of the relationship between metacognitive efficiency and metacognitive bias. *Consciousness and Cognition*, *95*, Article 103196. https://doi.org/10.1016/j.concog.2021.103196

Yeon, J., Shekhar, M., & Rahnev, D. (2020). Overlapping and unique neural circuits are activated during perceptual decision making and confidence. *Scientific Reports*, *10*(1), Article 20761. https://doi.org/10.1038/s41598-020-77820-6

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

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

Received January 5, 2023
Revision received October 3, 2023
Accepted October 9, 2023 ■