Simultaneous modeling of choice, confidence, and response time in visual perception

Sebastian Hellmann, Michael Zehetleitner, and Manuel Rausch
Professur für Allgemeine Psychologie II
Katholische Universität Eichstätt-Ingolstadt

20 November 2022

Author Note

Sebastian Hellmann https://orcid.org/0000-0001-6006-5103, Michael Zehetleitner https://orcid.org/0000-0003-3363-2680 (michael.zehetleitner@ku.de), Manuel Rausch https://orcid.org/0000-0002-5805-5544 (manuel.rausch@ku.de)

Model specification and modeling analyses were preregistered at the Open Science Framework website (https://osf.io/mtr4j) before accessing the data.

Data and code is available for download at https://github.com/SeHellmann/SeqSamplingConfidenceModels

This work was in part funded by the Deutsche Forschungsgemeinschaft (grant numbers ZE887/8-1 to MZ and RA2988/3-1 to MR) and by the Profor+ funding of the Katholische Universität Eichstätt-Ingolstadt (grant number 2613 to MR). We are grateful to Carolina Ehret, Florian Sprang and Veit Zehentmaier for help with data collection. A non-reviewed version of this manuscript has been uploaded at the Open Science Framework website (https://osf.io/9jfqr). Results of this study were in parts presented at the 43rd European Conference on Visual Perception (2021), at the 64th Conference of Experimental Psychologists (2022), and at the 25th annual meeting of the Association for the Scientific Study of Consciousness (2022).

Correspondence concerning this article should be addressed to Sebastian Hellmann, KU Eichstätt-Ingolstadt, Ostenstraße 25, 85072 Eichstätt, Germany. E-mail: sebastian.hellmann@ku.de

Abstract

How can choice, confidence, and response times be modeled simultaneously? Here, we propose the new dynamical weighted evidence and visibility model (dynWEV), an extension of the drift diffusion model of decision making, to account for choices, reaction times, and confidence simultaneously. The decision process in a binary perceptual task is described as a Wiener process accumulating sensory evidence about the choice options bounded by two constant thresholds. To account for confidence judgments, we assume a period of postdecisional accumulation of sensory evidence and parallel accumulation of information about the reliability of the present stimulus. We examined model fits in two experiments, a motion discrimination task with random dot kinematograms and a post-masked orientation discrimination task. A comparison between the dynamical weighted evidence and visibility model, two-stage dynamical signal detection theory, and several versions of race models of decision making showed that only dynWEV produced acceptable fits of choices, confidence, and reaction time. This finding suggests that confidence judgments depend not only on choice evidence but also on a parallel estimate of stimulus discriminability and postdecisional accumulation of evidence.

Keywords: cognitive modeling, confidence, decision making, drift diffusion model, sequential sampling models

Simultaneous modeling of choice, confidence, and response time in visual perception

One central aspect of cognitive psychology is the study of decision making. In most situations, the decision maker can not be entirely sure about whether the choice they made was correct or the best possible choice. The resulting degree of belief in a correct decision is referred to as confidence and its evaluation is an essential metacognitive ability (Pouget et al., 2016).

Modeling in the sense of describing cognitive processes formally and mathematically offers many benefits compared to natural language descriptions. Computational models can serve as a link between behavioral psychology and neuroscience (e.g. Kiani et al., 2014; Philiastides et al., 2006; Rausch et al., 2020; van den Berg et al., 2016; Zylberberg et al., 2012). Many cognitive models contain parameters that can be interpreted psychologically and be compared across subjects (Ratcliff et al., 2001). In addition, mathematical models offer the possibility to generate precise quantitative predictions that can be tested against empirical data. In addition, simulations can reveal new qualitative patterns that can be used for falsification or as starting point for further research questions (Palminteri et al., 2017).

Static Models of Confidence

A large number of mathematical models of decision making and confidence proposed in recent years were based on Signal Detection Theory (SDT, Aitchison et al., 2015; Mamassian & de Gardelle, 2021; Maniscalco & Lau, 2016; Maniscalco et al., 2016; Rausch et al., 2018; Shekhar & Rahnev, 2021; Zawadzka et al., 2017). SDT is a general framework for modeling stimulus and observer properties in various decision tasks (Macmillan & Creelman, 2005). According to SDT, a decision arises from comparing a normally distributed sample of evidence against a decision criterion. The parameters of the normal distribution depend on the perceptual abilities of the observer and the nature of the stimulus that has to be categorized. Several generalizations of SDT were proposed to additionally model confidence judgments. These generalizations of SDT were often designed to explain different specific empirical patterns. For example, models with normal (Maniscalco & Lau, 2016) or log-normal noise in the confidence judgment (Shekhar & Rahnev, 2021) or the heuristic response congruent model (Maniscalco et al., 2016) explain why metacognitive accuracy is sometimes lower than expected from decision accuracy. Other models were designed to be sufficiently flexible to account for both metacognitive accuracy that is lower than expected from decision accuracy as well as metacognitive accuracy that is higher than expected from decision accuracy (Mamassian & de Gardelle, 2021). Finally, some models implement the behavior of an ideal observer (Aitchison et al., 2015). Most of these static confidence models predict a specific qualitative pattern of interaction between stimulus discriminability, correctness and confidence, namely that confidence increases with stimulus discriminability for correct responses while it decreases for higher

stimulus discriminability in incorrect responses (Rausch et al., 2018, 2020). This pattern is referred to as folded X-pattern (Kepecs & Mainen, 2012) and was observed in several experiments with humans and animals (Desender et al., 2021; Kepecs et al., 2008; Lak et al., 2017; Moran et al., 2015; Pleskac & Busemeyer, 2010; Sanders et al., 2016). In other experiments, however, the data was characterized by a qualitatively different pattern: Confidence increased with stimulus discriminability also in wrong responses (Kiani et al., 2014; Rausch et al., 2018, 2020, 2021; van den Berg et al., 2016) resulting in what was referred to as a double increase pattern (Rausch & Zehetleitner, 2019). This means that all models that can only produce the folded X-pattern are unsuited as general models of confidence in perceptual decisions. The recently proposed weighted evidence and visibility model (WEV) can explain both the folded X-pattern and the double increase pattern by assuming that observers incorporate visibility of the stimulus into confidence judgments. The WEV model showed a superior fit to data from a masked orientation discrimination task compared to other static models (Rausch et al., 2018, 2020, 2021).

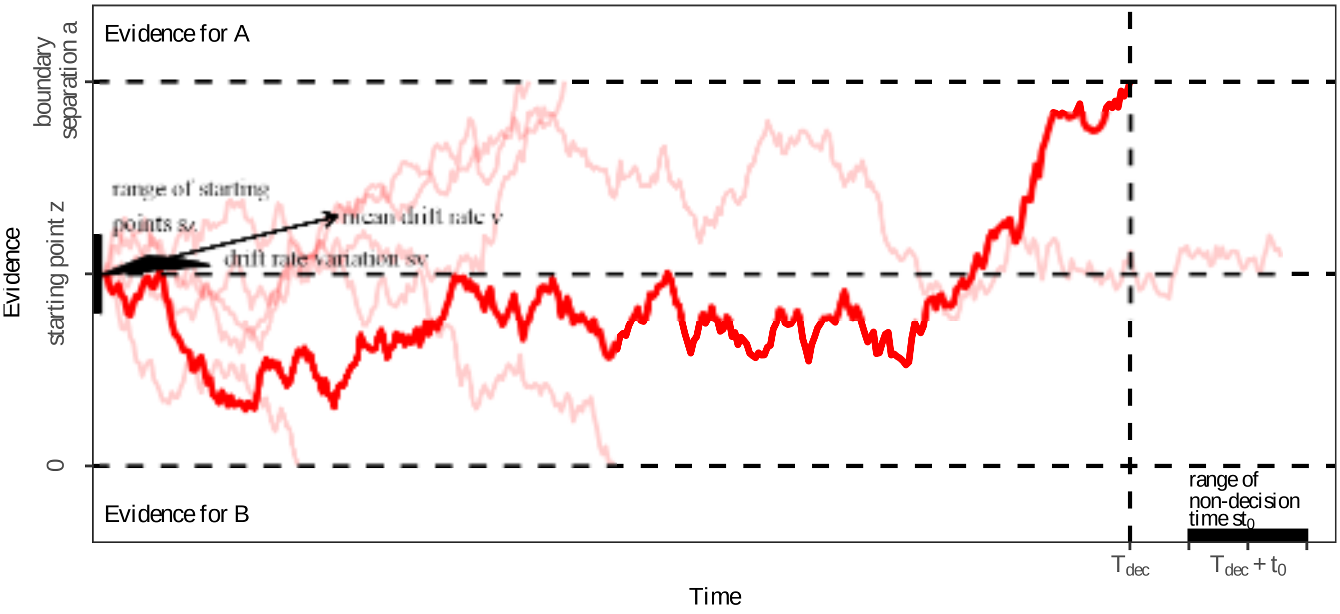
However, all static models share a substantial drawback. They can only account for response proportions and confidence ratings but not response times. The study of the relationship between confidence and response times has a long history (Vickers et al., 1985) and reaction times are closely linked to task difficulty and confidence in many decision tasks (e.g. Kiani et al., 2014; Rahnev et al., 2020). In contrast to static models, dynamical models can predict reaction time distributions and response probabilities by assuming the accumulation of evidence over time (Ratcliff & Smith, 2004).

Sequential Sampling Models of Confidence

Dynamical decision models follow the idea that the representation of evidence is not constant but is accumulated over the time course of a decision (Ratcliff, 1978; Usher & McClelland, 2001). More specifically, many sequential sampling models took the idea from SDT of normally distributed evidence samples (Ratcliff & Smith, 2004). Instead of a single observation these models assume an accumulation of evidence over time, which can be described as sequentially adding normally distributed samples to a decision variable. Ultimately, a decision is triggered when the decision variable crosses a certain threshold. This behavior is described in the limit by a continuous Gaussian process by reducing time step size. Although a Gaussian process is the most common choice, some dynamical models propose a different process, for example, Poisson counter models (LaBerge, 1994). Based on the idea of stochastic integration of evidence, several confidence models have been proposed, for instance the RTCON model (Ratcliff & Starns, 2009, 2013), two-stage signal detection theory (Pleskac & Busemeyer, 2010),the bounded accumulation model by Kiani et al. (2014), or the leaky evidence accumulation models by Pereira et al. (2021). Sequential sampling models of decision making provide explanations for various empirical patterns such as the correlation between discriminability and reaction time or the speed-accuracy trade-off (Bogacz

Figure 1

Example for a drift diffusion process



Note. The process starts at a location that is uniformly distributed around z with a range of sz. The drift rate is normally distributed around with standard deviation s . The sign of depends on the stimulus identity, which is A in this case. The Wiener process evolves with a diffusion coefficient set to 1 until it hits either the lower (0) or upper (a) boundary, at which time point a decision for the respective alternative is initiated. The observable reaction time is the decision time t plus a uniformly distributed non-decision time component, which is uniformly distributed with minimum t_0 and range st_0.

et al., 2006; Pleskac & Busemeyer, 2010; Vickers et al., 1985).

Sequential sampling models of decision making can be classified into (1) one dimensional diffusion models and (2) multidimensional race models.

The first class of sequential sampling models assumes one process that evolves in a diffusion-like manner with either a positive or negative drift depending on the stimulus. Each direction on the dimension represents support for one of two stimulus alternatives. There are two boundaries, one above and the other one below the starting point of the process, and when the process hits one of them for the first time, a decision is triggered for the corresponding alternative. Diffusion models thus apply only to binary choices. The different models in this class vary with respect to the stochastic process included (Wiener or Ornstein-Uhlenbeck process) or whether the boundaries are time-constant or collapsing, which is a widespread assumption in value-based decisions (Milosavljevic et al., 2010; Ratcliff et al., 2016; Tajima

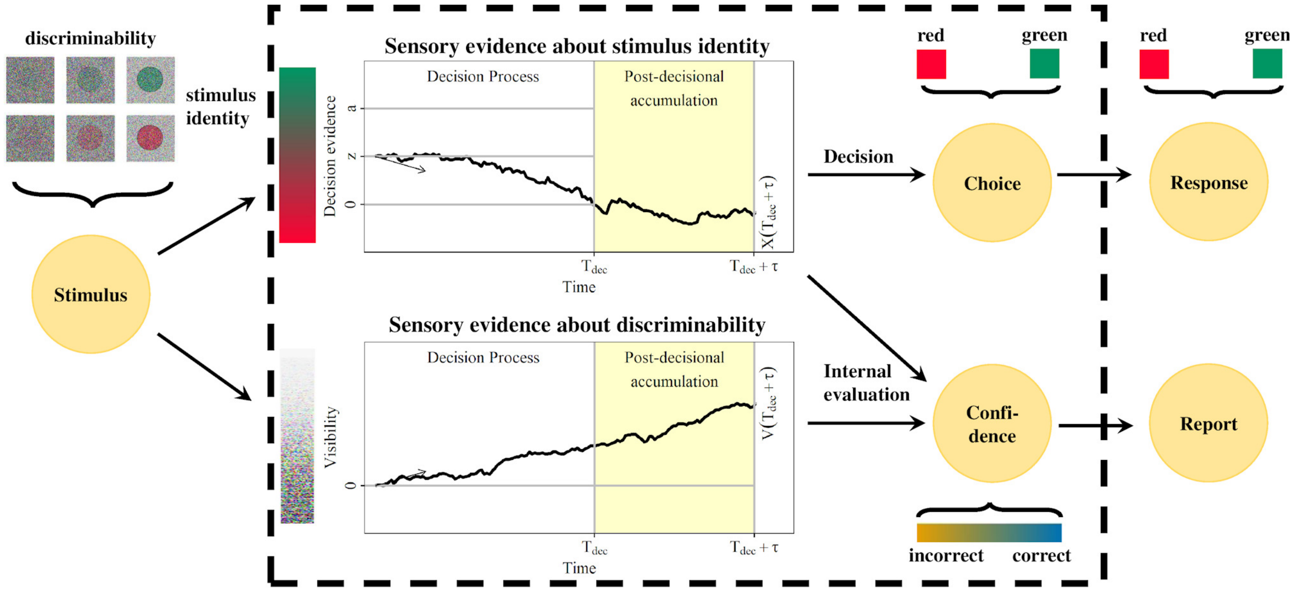
et al., 2016; Zhang et al., 2014). The drift diffusion model is characterized by a Wiener process with drift depending on stimulus discriminability and time-constant boundaries (Fig. 1). This approach is justified by the interpretation of evidence accumulation as updating a log-likelihood ratio test statistic and of the decision process as the continuous limit of a sequential probability-ratio test that implements an optimal policy for a speed-accuracy trade-off (Bogacz et al., 2006; Wald & Wolfowitz, 1948; Wald, 1947). In addition, the drift diffusion model includes a non-decision time component of reaction times that represents time spent on encoding and producing the motor response (Ratcliff & McKoon, 2008). Moreover, inter-trial variations in drift rate, starting point, and non-decision time were added to deal with empirically observed patterns such as fast or slow errors depending on task difficulty (Ratcliff & Rouder, 1998; Starns et al., 2012). Figure 1 visualizes the model and all its parameters.

The drift diffusion model provides an accurate explanation of reaction time distributions in perceptual as well as memory tasks (Ratcliff, 1978; Ratcliff et al., 2009, 2016) and in fast as well as very slow reaction time tasks (Lerche & Voss, 2019). However, the diffusion process always finishes at a constant threshold. Thus, the only signal available for confidence judgments is decision time (Ratcliff, 1978). To the best of our knowledge, no study so far has fit the classical drift diffusion model directly to empirical confidence data. Therefore, we investigated how confidence can be explained by a drift diffusion confidence model (DDConf) in which confidence is assumed to be a decreasing function of decision time. Pleskac and Busemeyer (2010) suggested a two-stage dynamical signal detection model (2DSD) by adding a postdecisional period of additional evidence accumulation to a drift diffusion. After this second stage, the final evidence is compared to a set of criteria, as in static models of confidence (see, e.g. Maniscalco & Lau, 2016; Rausch et al., 2018).

The second class of sequential sampling models is formed by race models (Gold & Shadlen, 2007). These models assume one process or dimension for each available alternative and are thus applicable to multi-alternative decisions. There is, again, a boundary on the processes, and the process that hits its boundary first determines the decision. Model variants differ in whether they assume independent or correlated noise in the diffusion of the different processes (Kiani et al., 2014; Moreno-Bote, 2010; van den Berg et al., 2016; Zylberberg et al., 2012) and can also include inhibitory interactions or decay (Usher & McClelland, 2001). According to the balance of evidence hypothesis, confidence is a function of the difference between the amount of evidence accumulated in the different processes until the time of the decision (Vickers et al., 1985). However, decision time influences confidence judgments in addition to the difference between the two accumulators (Kiani et al., 2014). Another version of a sequential sampling model with multiple accumulators, which was constructed to account for confidence judgments, is the RTCON model (Ratcliff & Starns, 2009; Starns et al., 2012). In the present study, we compare

Figure 2

Dynamical weighted evidence and visibility model



Note. The decision process follows the drift diffusion model (top path). After the initial decision time T_{dec}, there is a constant time period of length , in which the evidence process continues accumulating. In parallel, the visibility process accrues information about stimulus discriminability (bottom path). In the end, the final states of both processes are read out to calculate the confidence variable and formulate a confidence report.

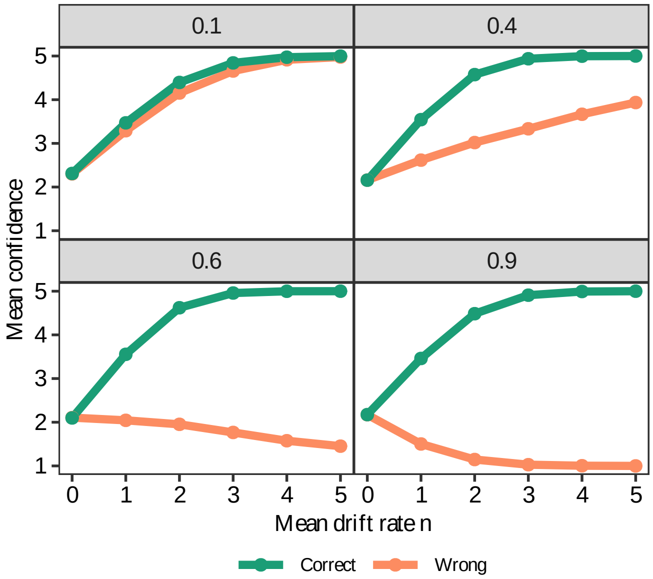
diffusion-like confidence models with a set of race models, in which processes accrue evidence for their respective alternatives with either uncorrelated (= 0) or anti-correlated (= 0.5) noise and opposite drift directions. Confidence is either computed according to the balance of evidence hypothesis or as a function of decision time and state of the losing accumulator at decision time.

Dynamical weighted evidence and visibility model

Here, we present the Dynamical weighted evidence and visibility model (dynWEV), a dynamical version of the WEV model. It is based on a drift diffusion decision mechanism with postdecisional accumulation as in 2DSD. But in addition, the dynWEV adopts the idea of WEV by assuming that there is a second source of information contributing to confidence. The core idea is that two internal variables determine confidence judgments: the evidence about the identity of the stimulus as well as the visibility of the stimulus (see Fig. 2). Visibility in the context of our model is defined as an internal representation of the discriminability of the stimulus. Visibility is beneficial for the computation of confidence because it allows the observer to put the evidence about the identity of the stimulus into context: When the stimulus is highly visible, it is reasonable to assume that the evidence about the identity of the stimulus will also be

Figure 3

Effect of the weight parameter w on mean confidence in dynWEV



Note. Simulation of the mean confidence rating in correct and incorrect decisions for varying levels of discriminability with different weights on the evidence about stimulus identity w (panels). For each level of coherence and each value of the weight parameter 10^5 trials were simulated. The value of the mean drift rate forms the x-axis. The other parameters were set as following: a = 1, z = .5, s_z = 0, s = 0.5, t_0 = 0, s_{t0} = 0, = 1, s_V = 1, $_V$ = 1, $_{11}$ = $_{11}$ = 0, $_{12}$ = $_{12}$ = 0.5, $_{13}$ = $_{13}$ = 1.0, $_{14}$ = $_{14}$ = 1.5.

accurate. Thus, a high degree of confidence is appropriate. Likewise, when the stimulus is barely visible, the evidence about the identity of the stimulus is likely also poor and even possibly misleading (Rausch et al., 2018). From a Bayesian perspective, observers should make use of all information available, that relates to the uncertainty of the decision. This means, that if internal evidence about the discriminability is available, which is independent of the evidence about stimulus identity but predictive of decision accuracy, an optimal observer needs to take this additional information into account in the computation of confidence judgments (Rausch & Zehetleitner, 2019).

Visibility in the model is also gathered dynamically over the time course of a trial in a second accumulation process (Fig. 2). To determine the degree of confidence, evidence about the identity of the stimulus and visibility are weighted and combined into one confidence variable. The weights between the two processes is expected to depend on the characteristics of the stimulation and the task: Some stimulus material may allow observers to estimate discriminability with precision. In this case, a substantial weight on visibility would be expected. In contrast, other stimulus material may leave observers without any cues

to estimate discriminability, resulting in a considerable weight on evidence about the identity. Simulations show that whenever there is a considerable weight on the visibility process, dynWEV predicts a double increase pattern as a function of stimulus discriminability and accuracy of the choice (upper panels in Fig. 3). The reason for this behavior is that the drift of the visibility accumulator and, thus, its mean final state is assumed to increase with stimulus discriminability irrespective of the choice. For small weights on visibility, the model produces the folded X-pattern (lower panels in Fig. 3). Therefore, the dynWEV model is consistent with both the folded X-pattern as well as the double increase pattern.

The precise formulation of the model can be found in the Analysis section. The postdecisional accumulation allows for changes-of-mind (Moran et al., 2015; Pleskac & Busemeyer, 2010; Resulaj et al., 2009; van den Berg et al., 2016). Particularly in incorrect decisions, the process sometimes tends to the opposite direction than the initial decision. If the postdecisional evidence strongly contradicts the initial decision, the observer may change its mind and choose the other alternative.

Rationale of the present study

The aim of the present study was to investigate which model provides the best explanation for the joint distribution of choice, reaction time, and confidence. For this purpose, we compared the fit to empirical data of the dynWEV model with a drift diffusion confidence model, 2DSD and several versions of race models, including independent and anti-correlated processes. We expected that the dynWEV would be more precise in the prediction of reaction time and confidence distributions compared to the alternative models and produce the best fit when compared in terms of the Bayesian information criterion (BIC, Schwarz, 1978) and the Akaike information criterion (AIC, Akaike, 1974). We also expected that dynWEV would be able to reproduce empirical patterns of the relationship between stimulus discriminability, confidence, and response time distribution. The reason is that only the dynWEV model includes a separate accumulation process for visibility, giving the model sufficient flexibility to account for the relationship between choice, reaction time, and confidence.

The race models were included as possible alternatives to dynWEV as they were previously found to be able to produce a double increase pattern of confidence as well (Kiani et al., 2014). On the other hand, 2DSD may only predict a folded X-pattern (Desender et al., 2021). With respect to the relationship of confidence and stimulus discriminability, particularly for small inter-trial variation of drift rate, 2DSD is similar to dual-channel like static models of confidence, where confidence is based on an independent or at least partially independent evidence sample (Mamassian & de Gardelle, 2021; Moran et al., 2015).

To compare the performance of the models in explaining empirical data, we analyzed data from two visual discrimination tasks with confidence judgments for which the double increase pattern has been observed previously (Kiani et al., 2014; Rausch et al., 2018): An orientation discrimination task with

masked sinusoidal gratings and a motion discrimination task with random dot stimuli. Previous studies showed that the double increase pattern is a greater challenge to account for in cognitive modeling (Rausch et al., 2018).

Method

We analyzed data from two visual discrimination tasks with confidence judgments. Both experiments involved a within-subject manipulation of stimulus discriminability. The first experiment was a masked orientation discrimination tasks where stimulus discriminability was manipulated by varying stimulus-onset-asynchrony. The second experiment was a motion direction discrimination task with coherence as manipulation of discriminability.

Participant recruitment

Participants were recruited using a derivative of the Online Recruitment System for Economic Experiments (Greiner, 2015) at the Catholic University Eichstätt-Ingolstadt. Participation was compensated either with 8€ per hour or with course credits (for undergraduate students). Before the experiment, participants were informed about the possibility of leaving the experiment without any negative consequences. They also provided written informed consent for participating in the experiment. They reported normal or corrected-to-normal vision, no history of neuropsychological or psychiatric disorders, and also not being on psychoactive medication. All participants were naive to the hypotheses of the study. The study protocol was approved by the Ethics Committee of the Katholische Universtität Eichstätt-Ingolstadt.

Apparatus

All experiments were performed in a darkened room on a Display++ LCD monitor (Cambridge Research Systems, UK) with a screen diagonal of 81.3 cm, set at a resolution of 1,920 × 1,080 pixels and a refresh rate of 120 Hz. The participants were seated at an approximate distance of 60 cm from the monitor. The experiments were conducted with PsychoPy (Peirce, 2007, 2009) on a Fujitsu ESPRIMO P756/E90+ desktop computer with Windows 8.1. Participants used a Cyborg V1 joystick (Cyborg Gaming, UK) for the response.

Experiment 1: Masked orientation discrimination

Participants. We collected data from 16 participants (15 female, 1 male) aged between 18 and 28 ($M = 20.4$ $SD = 2.4$) over three sessions.

Stimuli. The target stimulus was a square sinusoidal grating with a size of 3 × 3 and one cycle per degree (maximal luminance: 64 cd/m^2; minimal luminance: 21 cd/m^2) presented in front of a gray background (44 cd/m^2). The orientation was randomly set, either horizontal or vertical. The mask was a

checkerboard pattern (size: 4 × 4 , five rows and columns) with black (0 cd/m^2) and white (88 cd/m^2) boxes.

Trial Structure and Design. Figure 4A shows the sequence of events in one trial of Experiment 1. All trials started with a white fixation cross for one second in the center of the screen followed by the target stimulus with random, horizontal or vertical orientation. After a variable stimulus-onset-asynchrony (SOA) a mask replaced the target. Five levels of SOA were used: 8.3, 16.7, 33.3, 66.7, and 133.3 ms. After the mask, which was visible for 500 ms, two scales were presented. The one on the upper part on the screen was labeled vertical, the one on the lower part horizontal. The left end of both scales additionally had the label unsure while the right end was labeled sure. Participants had to indicate the perceived orientation of the grating together with their confidence using a joystick that moved a mark to the scale of their choice and the position representing their degree of confidence. The mark became only visible when the joystick was moved 50% of the distance towards one of the two choice options so participants were not biased from the starting point of the index. The participant confirmed their response by pulling the trigger of the joystick. Response times were measured as time from stimulus onset until the participant pulled the trigger. If the choice was incorrect the word Error! was presented for one second. Participants were instructed to report the orientation of the grating and their confidence as accurately as possible without time pressure to keep consistency with previous studies (Rausch et al., 2018, 2020). Instructing participants to report as accurately as possible was also intended to ensure high quality of confidence reports.

In each session the participant performed one training block and nine experimental blocks with 60 trials each. In the experimental block, each possible SOA appeared 12 times in a random order. Only the data from these blocks are used for analysis. With three sessions per participant, this results in a total of 1620 trials per participant. Each session took between 45 and 50 minutes.

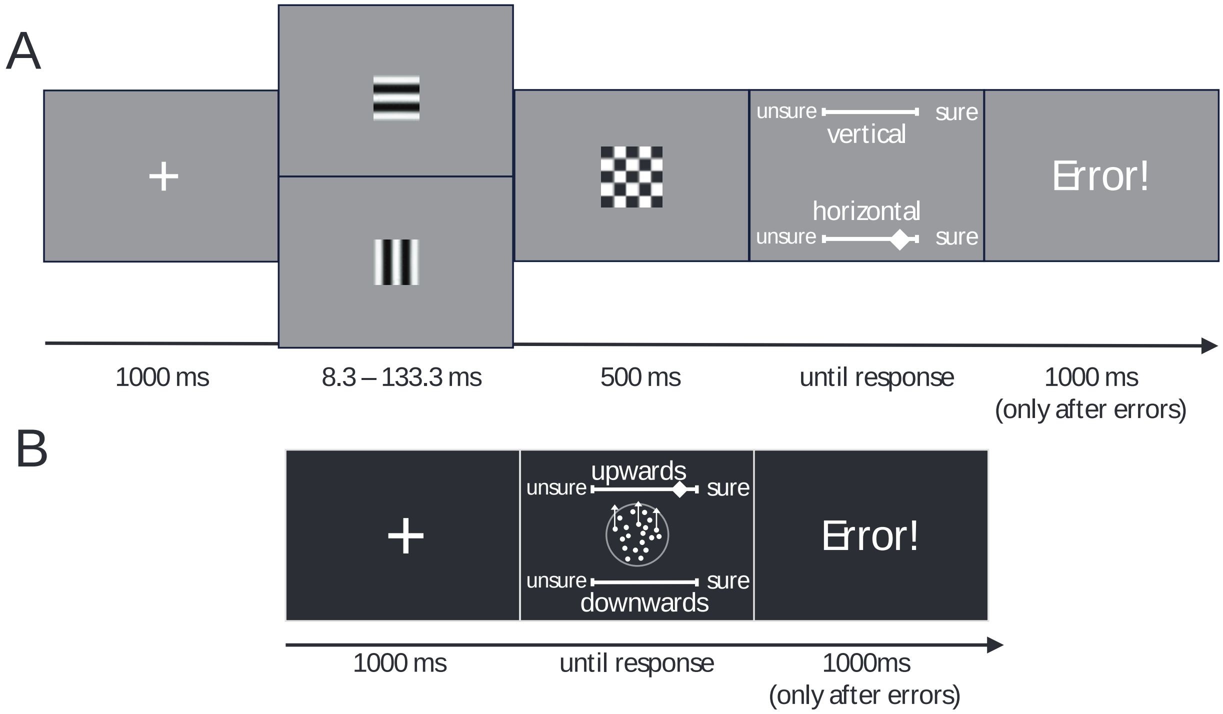
Experiment 2: Motion direction discrimination

Participants. We collected the data for the second experiment in two different time periods, separated by several months. The final sample consists of 42 participants (17 male, 25 female) of age 19 to 54 ($M = 23.3$, $SD = 6.0$). In the first data collection period, 30 participants each took part in one session of the experiment. Four participants conducted more trials because of a technical issue in the beginning of the experiment. All trials were used for analysis for these participants. Data from six additional participants, who aborted the experiment before finishing all blocks, was not considered in the analyses. In the second data collection period, eight participants ran through three sessions, two completed two sessions and two participants completed one session.

Stimuli. In this experiment, the target stimulus were white dots (111 cd/m^2) moving within a circular patch of 5 diameter presented on a black background (0 cd/m^2). In each frame, there were 262

Figure 4

Sequence of events in one trial



Note. Experiment 1 (A) and Experiment 2 (B).

dots with a size of two pixels. The noise dots were drawn at a random location in each frame, the target dots were moving either up or down at a constant speed of 5per second. If they reached the border of the stimulus region or after 100 frames, the target points were reinitialized at random locations. The proportion of target points, i.e. the motion coherence, was varied randomly from trial to trial in five levels: 1.6% 3.2% 6.4% 12.8% and 25.6%.

Trial Structure and Design. The sequence of events in one trial of Experiment 2 is shown in Figure 4B. All trials started with a white fixation cross presented for one second in the center of the screen followed by the target. The target stimulus was presented on the screen until a response was given. Motion coherence was varied between trials. Together with the target stimulus, two horizontal scales were presented above and below the stimulus, labeled upwards and downwards respectively. The left end of both scales was labeled unsure while the right end was labeled sure. Participants indicated their perceived motion direction and confidence simultaneously with a joystick by moving a mark to the scale corresponding to their choice and position on the scale corresponding to the degree of their confidence and confirmed their response by pulling the trigger of the joystick. Response times were measured as time

between stimulus onset and button press of the joystick. If the choice was incorrect the word Error! was presented for one second after the trial. Participants were instructed to report the direction of motion and their confidence as accurately as possible without time pressure to encourage precise confidence reports and to keep consistency with a previous study (Kiani et al., 2014).

Each session took about 45 minutes and consisted of one training block and eight experimental blocks. In each experimental block, each coherence level and motion direction combination was presented eight times in random order, resulting in 80 trials per block. Thus, depending on the number of sessions, participants completed 640 to 1920 trials.

Analysis

The free software for statistical computing R was used for all analyses (R Core Team, 2021).

Data exclusion and preprocessing

Participants were excluded if their overall accuracy was not above chance. Specifically, if their accuracy was below 50% or if the Bayes factor for the comparison against 50% in a binomial model assuming a logistic prior with a scale factor of 0.5 on the log-odds was less than 3. For the computations we used the function proportionBF from the BayesFactor package in R (Morey et al., 2018). Moreover, if the confidence ratings were equal in at least 90% of the trials the participant was excluded. However, none of the participants met the exclusion criteria so we analyzed the full sample.

In addition, we excluded trials in which the participantsŠ reaction time was smaller than 300 ms or grater than the mean plus four times the standard deviation of the participantsŠ individual reaction time distribution. These criteria were previously used by Pleskac and Busemeyer (2010) and similar exclusion is common in the literature (e.g. Ratcliff & Smith, 2004). If the minimal number of trials for each stimulus identity and discriminability condition was below 20 after the trial level exclusion we intened to drop the participant from further analyses. However, no participant was excluded from analysis due to fewer observations. Overall, on average 0.69% and 0.88% of trials were eliminated in Experiment 1 and 2, respectively.

All mathematical models were constructed to predict discrete confidence judgments. Therefore, the confidence reports on the analogue scale were binned to a five level discrete variable with breaks at 20%, 40%, 60%, and 80% of the continuous scale length, the same number of levels as had been used in previous studies (Rausch et al., 2018, 2021).

Mathematical model formulation

Drift diffusion model. First, we present the drift diffusion model for decision making, which forms the basis for the drift diffusion confidence model, the two-stage dynamical signal detection model, as

well as the dynamical weighted evidence and visibility model.

The drift diffusion model assumes that evidence is accumulated as a Wiener process with constant drift. There are two time constant decision thresholds, each corresponds to one choice option of the binary task. At the time the accumulation process reaches one of the thresholds for the first time, the decision for the respective alternative is triggered. We implemented the drift diffusion based confidence model consistent with the standard formulation of the drift diffusion decision model (Ratcliff et al., 2016) using the ddiffusion function from the R package rtdists (Singmann et al., 2020). This means, we included also the inter-trial variability of drift rate, starting point and non-decision time for the motor response in the decision task. These additional parameters are known to increase the fit to real reaction time distributions considerably as without them response time distributions of correct and wrong decisions would be identical (Ratcliff et al., 2007). The process describing the evidence accumulation is

$$X(t) = x_0 + W(t)$$

with starting point x_0, which is drawn from a uniform distribution Unif [z $\frac{s_z}{2}$ + $\frac{s_z}{2}$], and a Wiener process W with diffusion constant 1 and drift rate µ. The diffusion constant is set to 1 because it acts as a scaling factor in the model. The drift rate varies across trials according to a normal distribution with mean drift rate and standard deviation s , which is denoted as drift rate variation. The mean drift rate depends on stimulus category and discriminability. For binary decision tasks stimulus category may be represented by S { 1 1} and determines the sign of . The magnitude of depends on stimulus discriminability and therefore varied between experimental conditions, while the other parameters were kept constant. Thus, there was one parameter $_i$ for each of the five levels of stimulus discriminability. In a trial with discriminability level i, the mean drift rate is therefore equal to S $_i$. The decision time is T_{dec} := min{t|X (t) / [0 a]} with response R = 1, if X (T_{dec}) a and R = 1, if X (T_{dec}) 0.

Drift diffusion confidence model. According to the drift diffusion confidence model (DDConf) the process generating a perceptual decision is identical to the drift diffusion model. In this model the decision time may serve as a signal for difficulty because the expected decision time decreases with discriminability. This means that $\frac{1}{T_{dec}}$ may serve as a proxy for signal strength. Indeed, in a Bayesian decision framework, the optimal confidence is a function of $\frac{a}{T_{dec}}$ (Moreno-Bote, 2010). As we use discretized confidence ratings, we assume a set of thresholds to which the decision time is directly compared to. This comparison leads to a confidence rating of C = i, if T_{dec} [$_R$ $_R$ $_1$), with $_1$ $_1$ = 0 and $_1$ $_1$ = . In accordance with standard models from signal detection theory, we allow the confidence thresholds to vary for the two choice options. Finally, in accordance with previous studies the non-decision time component is assumed to vary uniformly between trials,

$T_{err} \sim Unif[t_0, t_0 + s_{t_0}]$. The observable response time is the sum of the decision time T_{dec} and the non-decision time T_{err}. It should be noted that according to DDConf the overlap of response time distributions between different levels of confidence is only due to the variation in the non-decision time component. All in all, for five different levels of discriminability, the model has following parameters $\theta = (z, s_z, a, \nu_1, \dots, \nu_5, s_\nu, t_0, s_{t_0}, \theta_{-1,1}, \dots, \theta_{-1,4}, \theta_{1,1}, \dots, \theta_{1,4})$. An overview of all parameters can be found in Table 1.

Two-stage dynamical signal detection model. The 2-stage dynamical signal detection model (2DSD) also assumes a drift diffusion-based decision as the previously described model. In contrast to the DDConf, 2DSD assumes that the process is not killed at decision time but continues to accumulate evidence for a fixed time period τ after the decision. At the end of this postdecisional period the state of the process is read out and compared to a set of criteria to form a discrete confidence judgment. Using the same notation as in the previous section the confidence variable is mathematically defined as $c := X(T_{dec} + \tau)$. Similarly to DDConf, the confidence variable is compared to the set of criteria, depending on the decision, such that $C = i$, if $c \in [\theta_{R,i-1}, \theta_{R,i})$, with $\theta_{-1,0} = \theta_{1,5} = \infty$ and $\theta_{-1,5} = \theta_{1,0} = -\infty$. As in DDConf, we also included a varying non-decision time component in the 2DSD model. Because decision and confidence judgments are reported simultaneously in the present study, we assume that all processes, encoding, decision, postdecisional accumulation, and response production happen sequentially and the observable response time is the sum of decision time, postdecisional accumulation period plus the non-judgment component, i.e. $RT = T_{dec} + \tau + T_{err}$. All in all, 2DSD has one additional parameter compared to DDConf. An overview of all parameters can be found in Table 1.

Dynamical weighted evidence and visibility model. In the present study, we propose a new sequential sampling model that extends the static WEV model (Rausch et al., 2018, 2020) to take reaction times into account. The dynamical weighted evidence and visibility model (dynWEV) includes not only evidence about stimulus identity but also the visibility of the stimulus as an internal measure of the discriminability in the computation of confidence. The dynamical WEV model combines this idea with 2DSD (Pleskac & Busemeyer, 2010). This means that for the accumulation of decision-relevant evidence, we assume a drift diffusion process until a decision is made and a fixed time period of postdecisional evidence accumulation in line with the 2DSD model. However, we also assume that there is a second process accumulating information about the discriminability of the stimulus. This process is denoted visibility process as its the dynamical equivalence of visibility in the WEV model (Rausch et al., 2018). We propose this process is again an independent Wiener process $V(t)$ with drift μ_V and diffusion constant s_V^2. The state of the visibility process is also read out at time $T_{dec} + \tau$. The internal confidence variable is a weighted sum of the form

Table 1

List and short description of all parameters fitted for the different models.

Parameter	Description	Models using the parameter
$_i$	mean drift rates for drift diffusion (diffusion-based models) or correct accumulation process (race models), $i = 1, \ldots, 5$ (one parameter per stimulus discriminability)	all
t_0	minimal non-decision time	all
s_{t_0}	range of uniformly distribution for non-decision time	all
$_{R,k}$	set of confidence criteria, $R = _{1,k}$, $k = 1, \ldots, 4$ (confidence is discretized into five steps)	all
s	variation in drift rate of the decision process	DDConf, 2DSD, dynWEV, dynVis
a	distance between upper and lower decision boundary for decision process	DDConf, 2DSD, dynWEV, dynVis
z	mean starting point of decision process	DDConf, 2DSD, dynWEV, dynVis
s_z	range of uniform distribution for starting point in decision process	DDConf, 2DSD, dynWEV, dynVis
	length of inter-rating period	2DSD, dynWEV
w	weight on decision evidence for confidence variable	dynWEV
s_V	variability in visibility process	dynWEV, dynVis
$_V$	variation in drift rate of visibility process	dynWEV, dynVis
A, B	thresholds for the two accumulation processes	IRM, PCRM, IRMt, PCRMt
w_X, w_{RT} and w_{Int}	weights on loosing accumulator, decision time and interaction for the confidence variable	IRMt, PCRMt

$$c = wRX(T_{dec} + \quad) + (1 \quad w)V(T_{dec} + \quad)$$

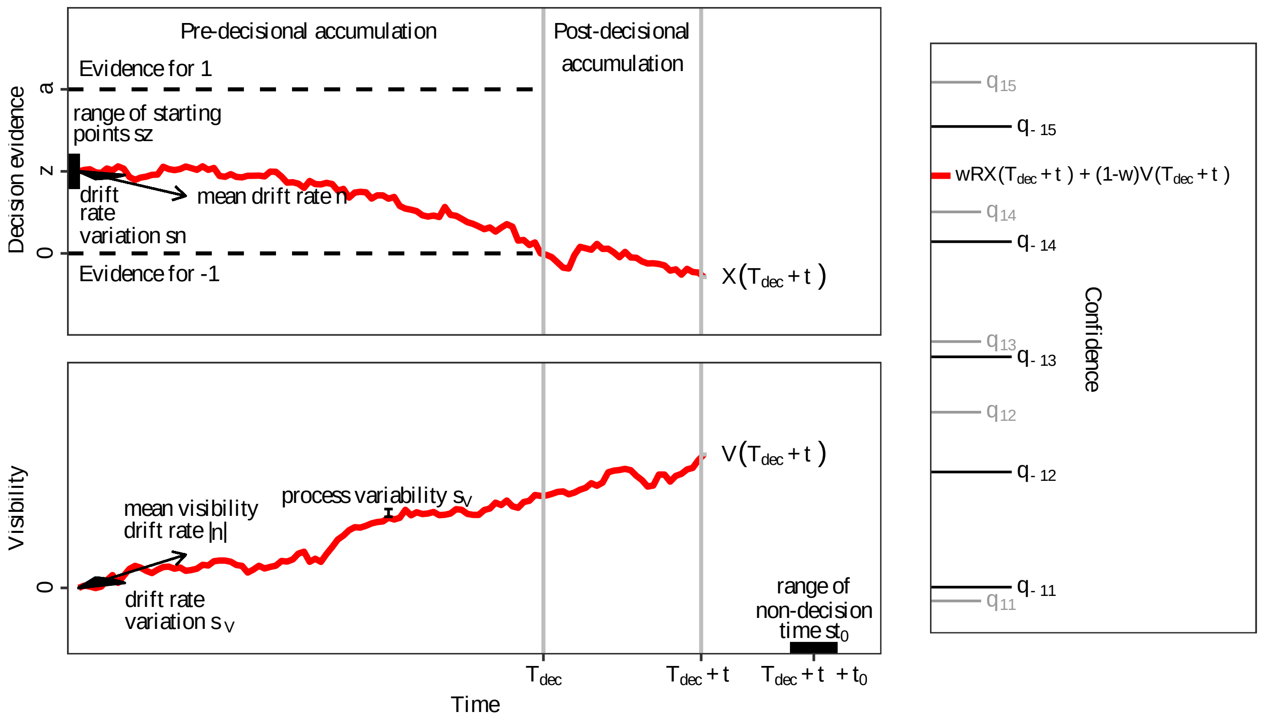
with the weight on the decision evidence w ranging between 0 and 1. This parameter represents the degree to which confidence judgments are based on the internal strength of decision relevant (X) or decision-irrelevant (V) stimulus features. Thus, if $w = 1$, dynWEV is equivalent to 2DSD, as the visibility process has no influence on confidence. Note, that the decision evidence is multiplied by R, because for a lower decision $R = \quad 1$, a lower value of X would support the decision while a high value would contradict the decision. The opposite holds for a upper decision $R = 1$. The final categorization of the confidence variable is equivalent to 2DSD. For the present study, we assume that drift rate of the visibility process varies with stimulus discriminability in the way that stimuli that are easier to discriminate are also perceived as more visible. The average drift in the visibility process μ_V is therefore not a separate parameter but assumed to be equal to the absolute value of the average drift of the decision process $| \quad |$, such that visibility is independent of stimulus identity, which is represented by the sign of the mean drift rate in the decision process. In addition, we include drift rate variation in the visibility process $_V$ similar to the drift rate variation in the decision process. For the purpose of the present study, we assume that the drift is normally distributed with mean $| \quad |$ and standard deviation $_V$ and that the drift variation in the visibility process is independent of the drift variation in the decision process (see also Discussion). As in 2DSD there is a uniformly varying non-decision time component T_{err} and the observed response time is assumed to be $T_{dec} + \quad + T_{err}$ as all required processes are assumed to occur sequentially before the overt response. Therefore, dynWEV includes three more parameters, $(w\ s_V\ _V)$, in addition to the parameters of 2DSD. An overview of all parameters can be found in Table 1. In addition, Figure 5 shows an illustration of the dynWEV model with all parameters.

Dynamical visibility model. We also included the dynamical visibility model (dynVis), a restricted version of the dynWEV model without postdecisional accumulation, to test whether postdecisional accumulation is really necessary to explain confidence. In this restricted version, is set to 0. Because $= 0$ implies that there is no variation in the final state of the evidence process any more, because it is always at the decision threshold at decision time, we can directly set $w = 1$ because confidence is only explained by variation in the visibility process. DynVis has two additional parameters compared to DDConf: $(s_V\ _V)$. An overview of all parameters can be found in Table 1.

Race Models. In contrast to the drift diffusion model, race models (RM) assume two accumulation processes, one for each decision alternative. In the present study, following Moreno-Bote (2010), we model the two processes as Wiener processes, each starting at 0 and having upper boundaries A and B respectively. Evidence is thus described as two-dimensional Gaussian process $X(t) = (X_1(t)\ X_2(t))$

Figure 5

Illustration of the parameters of the dynamical weighted evidence and visibility model



Note. Left columns shows the accumulation processes for decision evidence (top) and evidence about visibility (bottom). The decision process follows a drift diffusion model with parameters a (boundary separation), z (mean starting point), sz (starting point variation), (mean drift rate), and s (drift rate variation). The visibility process evolves in parallel with mean drift set to the absolute value of the mean drift rate of the decision process (| |) and an independent drift rate variability $_V$. The visibility process has an additional parameter s_V for the process variability. A decision R is triggered as soon as the decision process reaches the lower or upper threshold, in this case R = 1. This time point is denoted T_{dec}. In this illustration the decision was correct as the threshold and direction of mean drift rate correspond. After the decision both processes continue evolving for a fixed duration . Then, a weighted sum of the final states is computed depending on the weight parameter w (see right panel). The resulting confidence variable is compared against a set of criteria depending on the initial choice. Observable response times are T_{dec} + plus the non-decision time component t_0 with variation st_0.

with constant drift μ = (μ $_1$ μ$_2$) and covariance matrix = ^{2} ^{1} $_1$. Because is just a scaling parameter in this model, we set it to 1. The parameter represents the correlation of the two processes. The sign and magnitude of drift rates is determined by the stimulus category and discriminability, respectively, similar to the mean drift rate in drift diffusion-based models. Precisely, μ = (S $_i$ S $_i$) for stimulus category S { 1 1} and discriminability level i {1 5}. This means that the first accumulator indicates

evidence in favor of stimulus category 1 and the second for category -1. The time of decision is defined as $T_{dec} := \min\{t \mid X_1(t) > A \lor X_2(t) > B\}$ and the response is 1, if $X_1(T_{dec}) > A$ and -1, if $X_2(T_{dec}) > B$. We also assume a uniformly distributed non-decision component of the reaction time, like in dynWEV and 2DSD.

For the generation of confidence, we took two possible models into consideration. According to the Balance of Evidence (BoE) hypothesis, confidence is a monotone function of the state of the loosing accumulator (Vickers et al., 1985). This follows the idea that the difference in the final amount of evidence is a clue for perceptual ambiguity. In easy decisions, evidence should exclusively support the chosen alternative while an ambiguous stimulus produces similar amounts of evidence for all possible choices. In the race model, the state of the winning accumulator is fixed at the threshold, so the difference indicating the BoE is completely determined by the state of the loosing accumulator. Yet, previous studies provided evidence that subjects additionally take the reaction time into account (Kiani et al., 2014). This behavior seems also plausible, since low intensity stimuli lead not only to a low precision but also to long periods of accumulation until a threshold is met. In addition, Moreno-Bote (2010) showed that in a Bayesian framework, the posterior of a correct decision, say for alternative 1, is a function of the final state of the loosing accumulator divided by the square root of decision time, $X_2(T_{dec})/\sqrt{T_{dec}}$. Therefore, we included the race model with a time-dependent confidence variable by combining the two indicators and the decision time as additional indicator in a linear way, such that

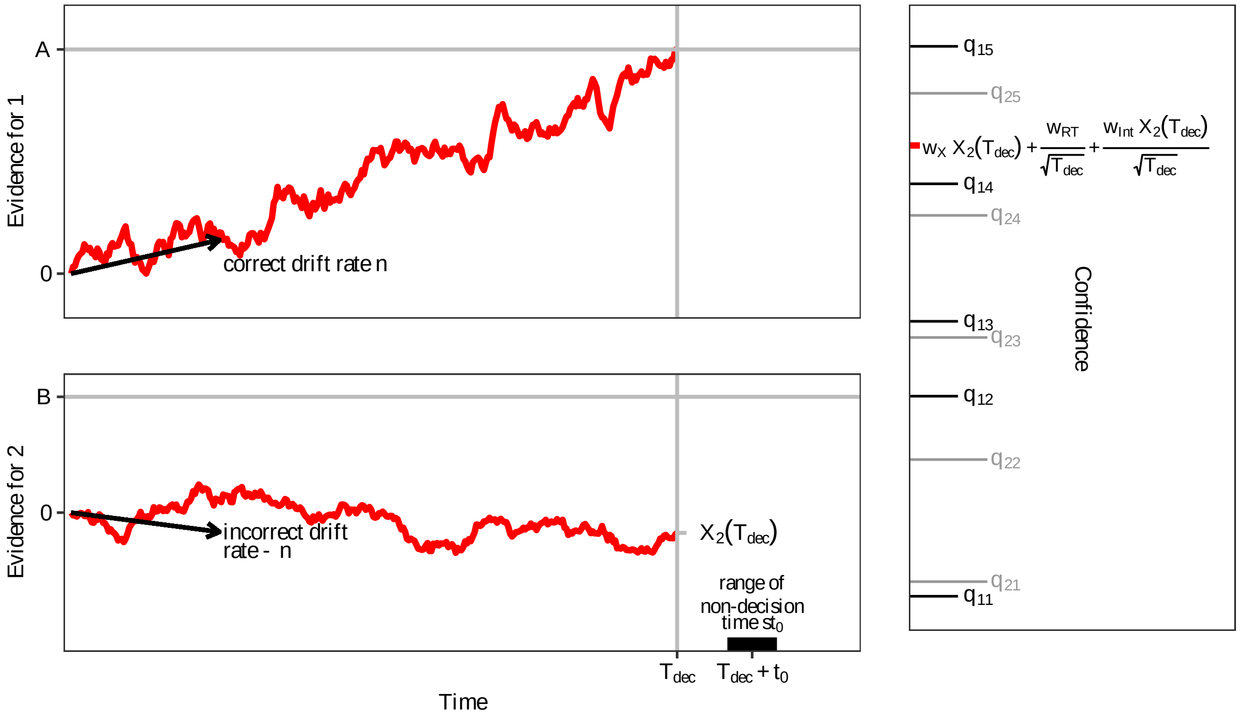
$$c := w_X X_2(T_{dec}) + w_{RT} \frac{1}{\sqrt{T_{dec}}} + w_{Int} X_2(T_{dec}) \frac{1}{\sqrt{T_{dec}}}.$$

The parameters w_X, w_{RT}, and w_{Int} are weight parameters indication how strong the confidence variable is influenced by the BoE, decision time, or the optimal ratio. Thus, the model may represent and individual trade-off between these indicators for choice accuracy. Similar to the other models, the confidence variable is compared against a set of criteria. For the model comparison, we use two different fixed values for ρ as analytical solutions were available only for $\rho = 0$ and $\rho = -.5$. We refer to the model characterized by $\rho = 0$ as independent race model (IRM), and for $\rho = -.5$ as partially anti-correlated race model (PCRM). Both are implemented representing the Balance of Evidence hypothesis (i.e. $w_X = 1, w_{RT} = w_{Int} = 0$) and with a time-dependent confidence variable ($w_X, w_{RT}, w_{Int} \in \mathbb{R}_+$), leading to four different variations of the race models. We denote models implementing the Balance of Evidence hypothesis as IRM and PCRM and the models with time-dependent confidence variable as IRMt and PCRMt, respectively. For models with time-dependent confidence variable, we set $w_X + w_{RT} + w_{Int} = 1$. This prevents an arbitrary scaling of the process parameters, coefficients and confidence thresholds and is similar to fixing one of the coefficients to 1 (but delivers a more intuitive interpretation of the coefficients as weights, similar to the weight in

dynWEV). An overview of all parameters can be found in Table 1 and an illustration of the race models with all parameters involved is shown in Figure 6.

Figure 6

Illustration of the parameters of the race models model of confidence



Note. Race models assume two separate processes, one for each alternative (left column). The process accumulating evidence for the correct choice (here top) has a positive drift rate. The other process (bottom) has the same absolute but negative drift rate. Each process has an upper threshold (A and B). When one of the processes reaches its threshold a decision is triggered. In this illustration the decision was correct. At decision time T_{dec} the state of the losing accumulator (here X_2) is used as measure for the Balance of Evidence. Balance of Evidence as well as decision time are combined in a internal confidence variable depending on the weight parameters w_X, w_{RT}, and w_{Int}. This confidence variable is compared against a set of thresholds depending on the choice to generate discrete confidence judgments. This illustration depicts a race model with time-dependent confidence measure. Race models with Balance of Evidence based confidence share the same architecture with only the weight parameters fixed to $w_X = 1$ and $w_{RT} = w_{Int} = 0$. Race models may differ whether the diffusion noise is independent (IRM) or anti-correlated (PCRM), which is not represented in this illustration.

Parameters and Fitting Procedure

The various models share many parameters. Table 1 shows an overview and a short description of all parameters. Both experiments had five steps of stimulus discriminability (i.e. SOA in Exp. 1 and

confidence in Exp. 2) and two opposite categories (i.e. horizontal vs. vertical orientation in Exp. 1 and upwards vs. downwards motion in Exp. 2). Therefore, we modeled the stimulus categories as $S \in \{-1, 1\}$ and fitted for each level of discriminability an intensity parameter $\nu_i \geq 0, i = 1, ..., 5$. For the diffusion-based models (DDConf, 2DSD, dynWEV, and dynVis), the mean drift rate in a specific trial is accordingly set to $\nu = S\nu_i$. In RMs, the drift rate is set to $\mu = (S\nu_i, -S\nu_i)$. The choices were denoted similarly as $R \in \{-1, 1\}$, where the upper bound in diffusion-based models and the first accumulator in RMs drive decision $R = 1$ and the lower bound and second accumulator $R = -1$, respectively. Moreover, the parameters for the non-decision time t_0 and s_{t_0} are shared across models. As all models were formulated to produce discrete confidence outcomes, we discretized the judgments in the empirical data to get five confidence levels. Confidence thresholds to separate between these levels, $\theta_{R,k}, R \in \{-1, 1\}, k \in \{1, ..., 4\}$, may vary between the decision response options and are present in all models. In summary, there are 15 parameters common to all models. DDConf contains four additional parameters: s_ν, a, z, s_z (19 in total). 2DSD includes the parameter τ in addition to those from the DDConf leading to 20 parameters in total. DynWEV requires the parameters τ, w, σ_V, and s_V in addition to DDConf leading to 23 parameters. DynVis only requires the parameters σ_V and s_V in addition to the DDConf model and includes 20 parameters in total. IRM and PCRM require only A and B, which are equivalent to a and z in 2DSD and dynWEV, in addition to the common parameters (17 in total), and IRMt and PCRMt require further w_X, w_{RT}, and w_{Int} (19 effective parameters in total, because we fixed the sum of weight parameters to 1).

The parameters were fitted separately for each participant using a maximum likelihood procedure assuming independence across trials. This method uses the full information available in the data but has the drawback of being sensitive to outliers (Ratcliff & Tuerlinckx, 2002), which is why we excluded excessively slow and fast responses (see above). Formulas for the likelihood functions of all models are included in the Supplemental Material. For model m, a set of parameters θ and data vectors, consisting of stimulus category S, stimulus discriminability Q, observed decisions R, reaction times T and confidence ratings C (the dependent variables), we used the negative log-likelihood

$$L_m(\theta) = -\sum_{n=1}^{N} \log(P_m(R_n, T_n, C_n \mid \theta, S_n, Q_n))$$

as loss function, where $n = 1, ..., N$ are the different trials. The derivation of probability densities is provided in the Supplementary Material. The minimization procedure started with a broad grid search in which the likelihood is computed for different parameter constellations. After the grid search, the five best parameter sets were used as initial values for a optimization algorithm. We used the BOBYQA algorithm for box constrained optimization implemented in the bobyqa function of the minqa package (Bates et al., 2015; Powell, 2009). Details for the settings of this routine can be found in the code. We restarted the

optimization four times, using the previously found result as initial value for the next iteration to prevent the algorithm from getting stuck in a local minimum.

Model comparison

For a quantitative comparison of the fits of the models, we used the Bayesian information criterion (BIC) and Akaike information criterion (AIC). The BIC is derived using LaplaceŠs method for approximating the marginal likelihood in a Bayesian context. Both criteria take the likelihood of the data as well as the number of parameters into account to avoid overfitting due to unnecessary freedom. Thus, BIC and AIC implement a trade off of parsimony and model fit (Schwarz, 1978). They are defined by

$$BIC_m = 2L_m(\tilde{}) + k\log(N)$$

and

$$AIC_m = 2L_m(\tilde{}) + 2k$$

where k is the number of parameters in the respective model and $\tilde{}$ is the maximum likelihood parameter estimation. A low BIC or AIC indicate a better model as the data is captured without introducing unnecessary complexity in a model.

We computed BIC and AIC for each participant and model. To compare the quantitative fit of the dynWEV to that of the other models, we performed Bayesian paired t-tests for the mean difference information criteria assuming a standard Cauchy distribution with scale parameter 1 as prior distribution for the standardized effect size (Rouder et al., 2009). For this purpose we use the function ttestBF from the BayesFactor package in R. Bayes factors were interpreted with respect to their statistical evidence according to established guidelines (Lee & Wagenmakers, 2014). Reported 95% equal-tailed CIs were generated using 10^6 samples from the posterior distribution using the same prior as for the Bayes factors.

Model identification analysis

Because of the possibility that the best fitting model is not the generative model that underlies the data, we conducted a model mimicry analysis. Previous studies of sequential sampling models showed that there are parameter regions where there is a high level of model mimicry when models are fitted only to accuracy and reaction time data (Bogacz et al., 2006; Bose et al., 2020). Because of the high complexity of the models used in this study, it is not possible to find parameter regions a priori where specific models are equivalent, besides the trivial exception that dynWEV is equivalent to 2DSD if the weight on the visibility accumulator is 0, i.e. w = 1 and dynVis is a special case of dynWEV. Therefore, we relied on simulating artificial data, fitting and comparing the resulting information criteria to examine whether the generative

model also achieves the best quantitative model fit. We generated simulations from the overall second best fitting model using the fitted parameters and the same number of trials as in the empirical data. With this method we replicated the situation of the experiments as accurately as possible. Then, we fitted the best and second-best fitting model to the simulations and compared which model is preferred by comparing the BIC.

Parameter recovery analysis

To access whether our experimental procedure and fitting methods allowed us to robustly recover the parameters of the dynWEV model, we conducted a parameter recovery study. For this purpose, we simulated one artificial data set per participant using the parameter sets obtained from the model fitted to empirical data. The number of observations for each simulated data set was equal to the number of trials for the respective participant in the actual experiments. We then fitted the dynWEV model again individually to the simulated data sets and compared the true underlying parameters with the fitted parameters.

Transparency and openness

Model specification for 2DSD, dynWEV, and the race models and analyses for all experiments were preregistered at the Open Science Framework website (https://osf.io/mtr4j, Hellmann & Rausch, 2022). The parameter recovery study was conducted after the first submission of the manuscript. Experiment files, raw data, and code are publicly available via GitHub at https://github.com/SeHellmann/SeqSamplingConfidenceModels.

Results

Experiment 1: Masked orientation discrimination

All participants were included in the analysis. The maximum proportion of trials excluded for each participant was 1.05%.

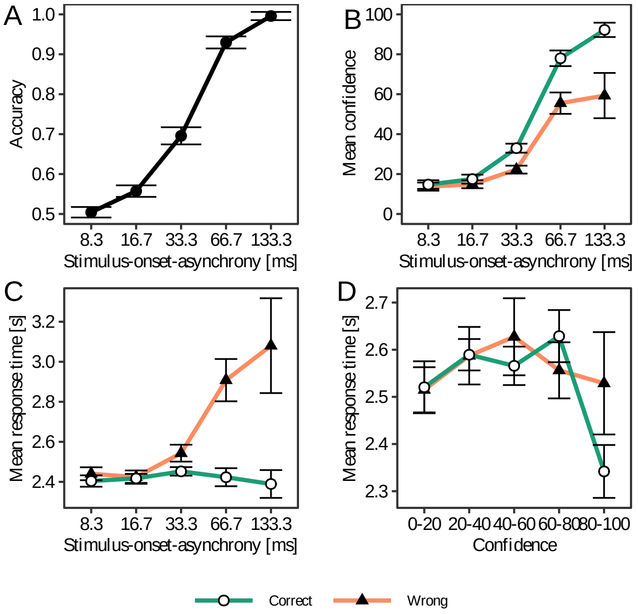
Behavioral results

Accuracy did not vary between sessions $BF = .11$. Accuracy was at chance level at the minimal SOA of 8.3 ms ($M = 50.4\%$, $SD = 2.9\%$) and close-to-perfect ($M = 99.6\%$, $SD = .6\%$) at the maximum SOA of 133.3 ms (Fig. 7A). Mean confidence increased in correct responses from 14.7% ($SD = 8.63\%$) of the visual scale for the shortest SOA to 92.2% ($SD = 2.39\%$) for the highest SOA. Similarly, confidence increased also in error trials from the hardest condition ($M = 13.7\%$, $SD = 8.13\%$) to the easiest condition ($M = 59.3\%$, $SD = 3.08\%$, Fig. 7B). Response times also increased in incorrect responses from low discriminability ($M = 2.44$ s, $SD = 0.13$ s) to high discriminability ($M = 3.08$ s, $SD = 0.63$ s). In spite of that, reaction times varied only slightly for correct trials ranging from 2.40 s ($SD = 0.11$ s) in the lowest

SOA to 2.39 s (SD = 0.28 s) in the highest SOA (Fig. 7C). There was only a weak relationship between confidence and response times with a mean Gamma correlation across participants of .05 (SD = .27, see Fig. 7D).

Figure 7

Descriptive results from Experiment 1



Note. Mean accuracy (A), confidence ratings (in % of visual scale, B), and response times (C) for different levels of stimulus onset asynchrony (SOA) and mean response times for different levels of confidence (in % of visual scale, D). Error bars represent within-subject standard errors.

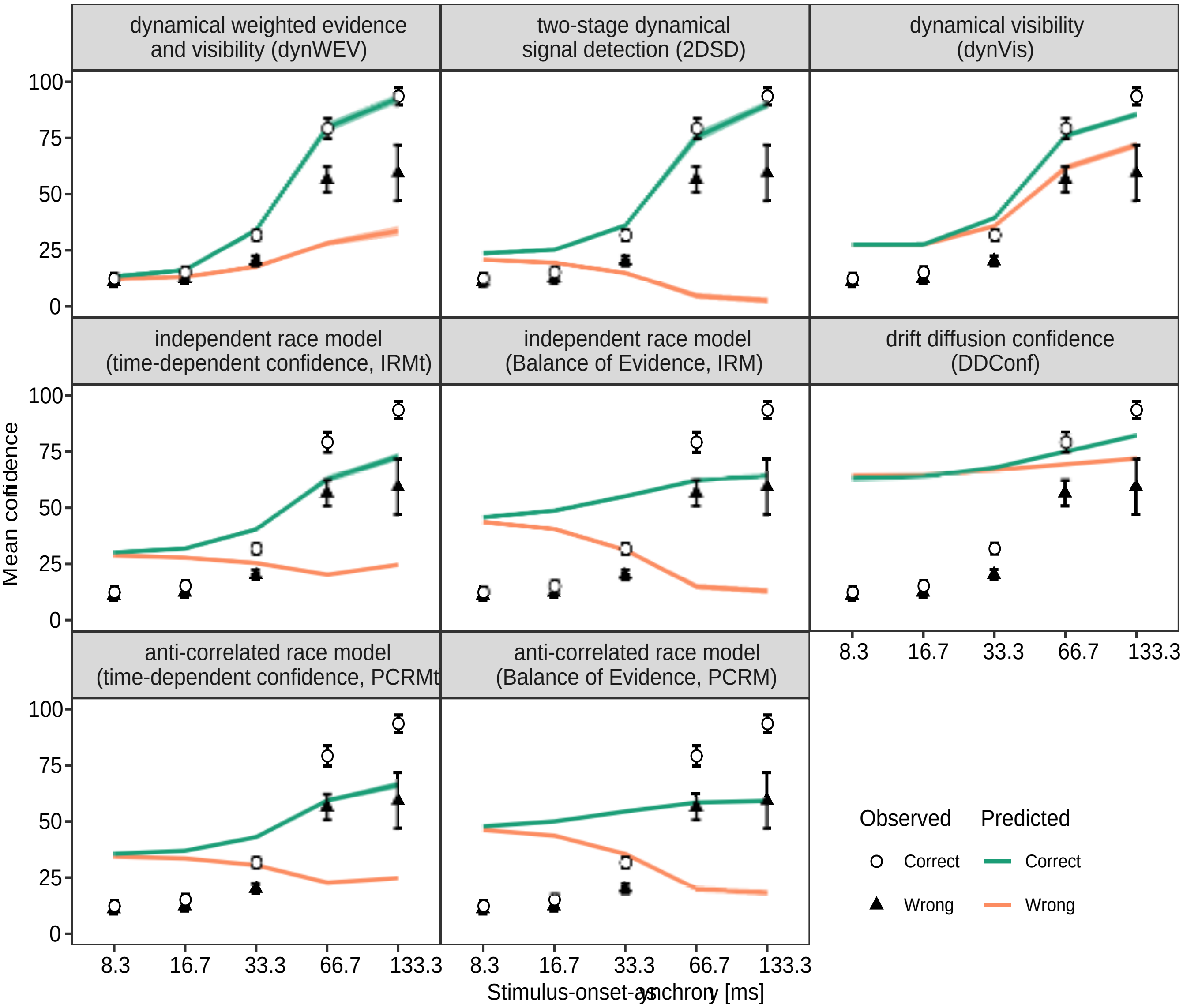
Model results

Summary statistics of the fitted parameters for all models can be found in Supplementary Table 3. There was a strong correlation between the parameter , assessing the postdecisional accumulation time in the dynWEV and 2DSD models, and metacognitive sensitivity, i.e. the degree to which confidence judgments differentiated between correct and incorrect trials (see Supplementary Fig. 1).

Visual inspection of empirical data and model predictions. Most models fitted accuracy for the different levels of SOA well (Suppl. Fig. 2). The DDConf model showed the most prominent deviations from empirical data, predicting a too flat curve and higher accuracy for lower levels of SOA than observed. The other diffusion based models, i.e. 2DSD, dynWEV, and dynVis slightly underestimated accuracy for easier conditions. Figure 8 shows the observed and fitted patterns of mean confidence judgments for different discriminability levels and correctness of the response in Experiment 1. The

Figure 8

Observed mean confidence vs. confidence predicted by model fits for Experiment 1



Note. Empirical (points) and fitted (lines) mean confidence ratings as function of stimulus onset asynchrony for correct (green, circles) and incorrect (orange, triangles) decisions. Error bars around points and shaded areas around lines represent within-subject standard errors. Ribbons around lines are barely visible because of small intervals.

DDConf model missed the changes in confidence with SOA almost completely. As expected, 2DSD and the race models without time dependent confidence (IRM and PCRM) were not able to account for the double increase pattern. Indeed, only the dynWEV and dynVis models were able to reproduce the increase of confidence with discriminability in incorrect trials. Although the prediction of the dynWEV model was

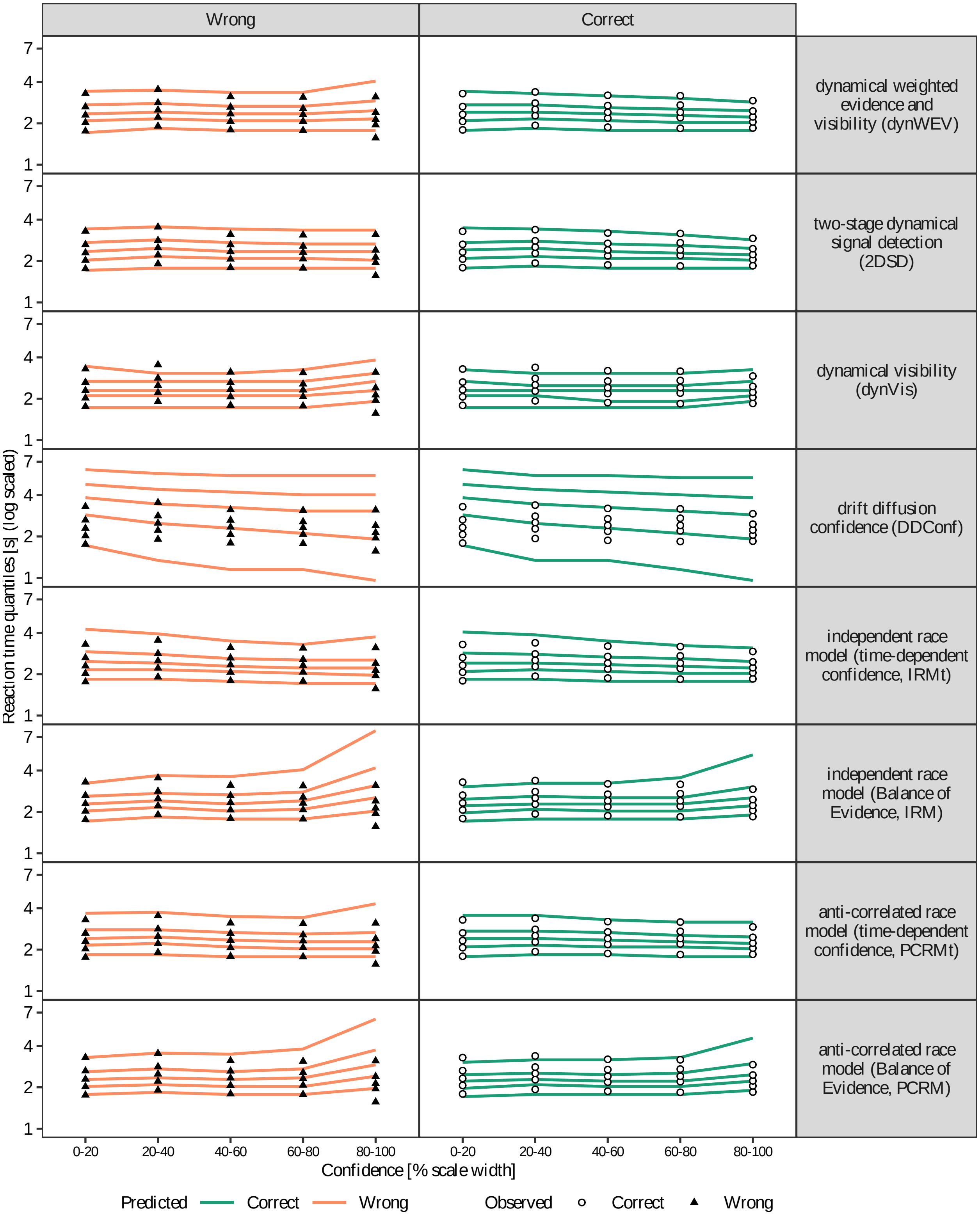
accurate in general, it underestimated the increase in confidence in incorrect trials. It should be noted that the number of errors for the two highest SOA levels was only 1.5% (see Figure 7A). Therefore, the likelihood as objective function put a smaller weight on these points, which is why predictions may be less accurate in this region. The dynVis model on the other hand overestimated the confidence for the difficult conditions, where many trials were available for both correct and wrong decisions. The race models with time dependent confidence (IRMt and PCRMt) both produced constant confidence in incorrect decisions and on the other hand underestimated the steepness of the correct response curve. The deviation of race models from the empirical confidence reports is particularly visible in the joint distribution of confidence and correctness (Suppl. Fig. 4). Supplementary Figure 4 also shows that all models except for dynWEV failed to account for the data in the same way: For difficult stimuli, the probability of high confidence reports was systematically overestimated and the prevalence of low confidence was underestimated whereas the opposite is true for the easiest two conditions.

To visualize reaction times as a function of confidence, Figure 9 shows log-transformed response time quantiles for different levels of confidence. The most striking discrepancies between empirical and predicted RT distributions is apparent for DDConf, which produced very broad distributions. These broad distributions are explained by the high values in variation of the non-decision time component necessary for accounting for the strong overlap of response time distributions between confidence ratings. The best fit to reaction time was achieved by 2DSD. The BoE race models did not account well for the reaction time at high confidence responses, while all the other models struggled with the reaction at high confidence specifically in incorrect trials. However, high confidence errors represent only 1.2% of observations.

Model comparison in terms of information criteria. We compared the model fits for AIC and BIC. In terms of BIC, dynWEV was preferable over the other seven models and achieved the smallest value for all participants (see also Suppl. Fig. 6). The second best model was 2DSD (mean BIC difference to dynWEV: $M = 256$, $SD = 120$). DynVis also performed better than all race models (mean BIC difference to dynWEV: $M = 630$, $SD = 244$). Among the race models IRMt performed best ($M = 892$, $SD = 245$). The worst model was DDConf with $M = 5783$, $SD = 1717$. The BIC was smaller for the dynWEV model compared to all other models for all participants. A Bayesian t-test was conducted to compare the BIC values for the dynWEV model with the other models. For the comparison with the 2DSD the Bayes factor revealed decisive evidence in favor of dynWEV ($BF_{10} = 5.2 \times 10^4$, CI of posterior for : [1.13 2.92]). The results for the other comparisons were more extreme ($BF_{10} > 10^5$, see Supplementary Table 1 for more details). The dynWEV model outperformed the other models by an even larger margin in terms of AIC, as the BIC has a higher penalty for the parameters and therefore dynWEV gets the higher penalty in the BIC.

Figure 9

Observed response time quantiles vs. quantiles predicted by model fits for Experiment 1



Note. Empirical (triangles and points) and fitted (lines) response time quantiles (log scaled; probabilities: .1, .3, .5, .7, .9) across confidence judgments for incorrect (left column) and correct (right column) answers. Empirical quantiles were computed from the whole data set without regarding for participants. Predicted quantiles were computed after aggregating the individual response time densities.

Experiment 2

For analysis the samples from the two data collection periods were combined to one sample but model results were very similar for the two subsets. All of the participants who had finished the experiment, performed above chance overall. From the individual number of trials a maximum proportion of 1.6% was excluded because of extreme response times. All participants were included in the analyses.

Behavioral results

Accuracy ranged from 57.6% (SD = 9.0%) for minimal motion coherence to 99.3% (SD = 1.6%) for the maximal motion coherence (Fig. 10A). Again, mean confidence increased in both correct and incorrect responses from 41.6% (SD = 7.28%) and 38.7% (SD = 9.25%), respectively of the visual scale at 1.6% coherence to 93.5% (SD = 18.4%) and 87.1% (SD = 19.2%), respectively at 25.6% coherence (Fig. 10B). In the second experiment response times decreased in correct trials from a mean 3.38 s (SD = 0.44 s) in the hardest condition to 2.31 s (SD = 0.81 s) in the easiest condition. Here, mean response times decreased also for errors from 3.451 s (SD = 0.54 s) for the lowest coherence level to 2.14 s (SD = 0.51 s) for the highest coherence (Fig. 10C). There was medium to strong relationship between confidence and response times with a mean Gamma correlation across participants of -.36 (SD = .27, see Fig. 10D).

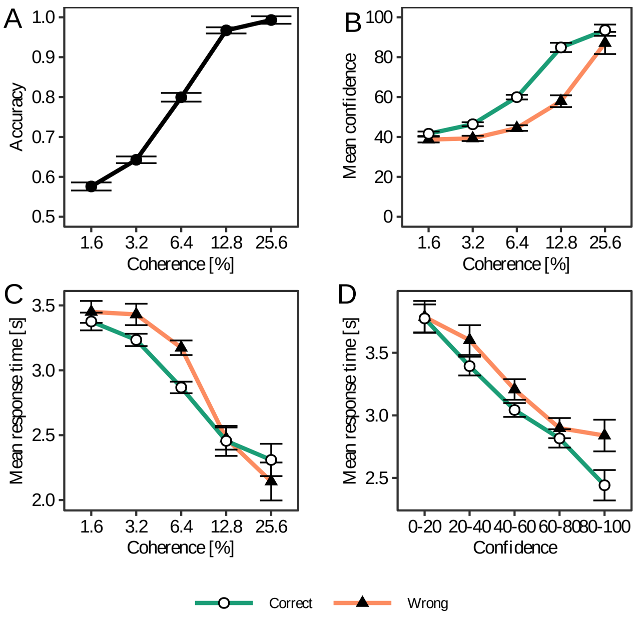
Model results

Summary statistics of the fitted parameters for all models can be found in Supplementary Table 4. There was a strong correlation between the parameter in dynWEV and 2DSD and metacognitive sensitivity (see Supplementary Fig. 1).

Visual inspection of empirical data and model predictions. In Experiment 2 all models fitted the shape of the accuracy curve as a function of coherence accurately with the exception of dynVis, which overestimated the steepness of the curve by underestimating accuracy in difficult conditions (Suppl. Fig. 3). Similarly to Experiment 1, dynWEV seemed to best approximate mean confidence judgments across conditions (Fig. 11), while 2DSD and race models based on balance of evidence missed the increase in confidence with stimulus discriminability in wrong decisions completely. DynVis and DDConf showed the double-increase pattern but overestimated confidence in difficult situation and showed almost no difference in confidence between correct and incorrect decisions. Although showing a slight

Figure 10

Descriptive results from Experiment 2



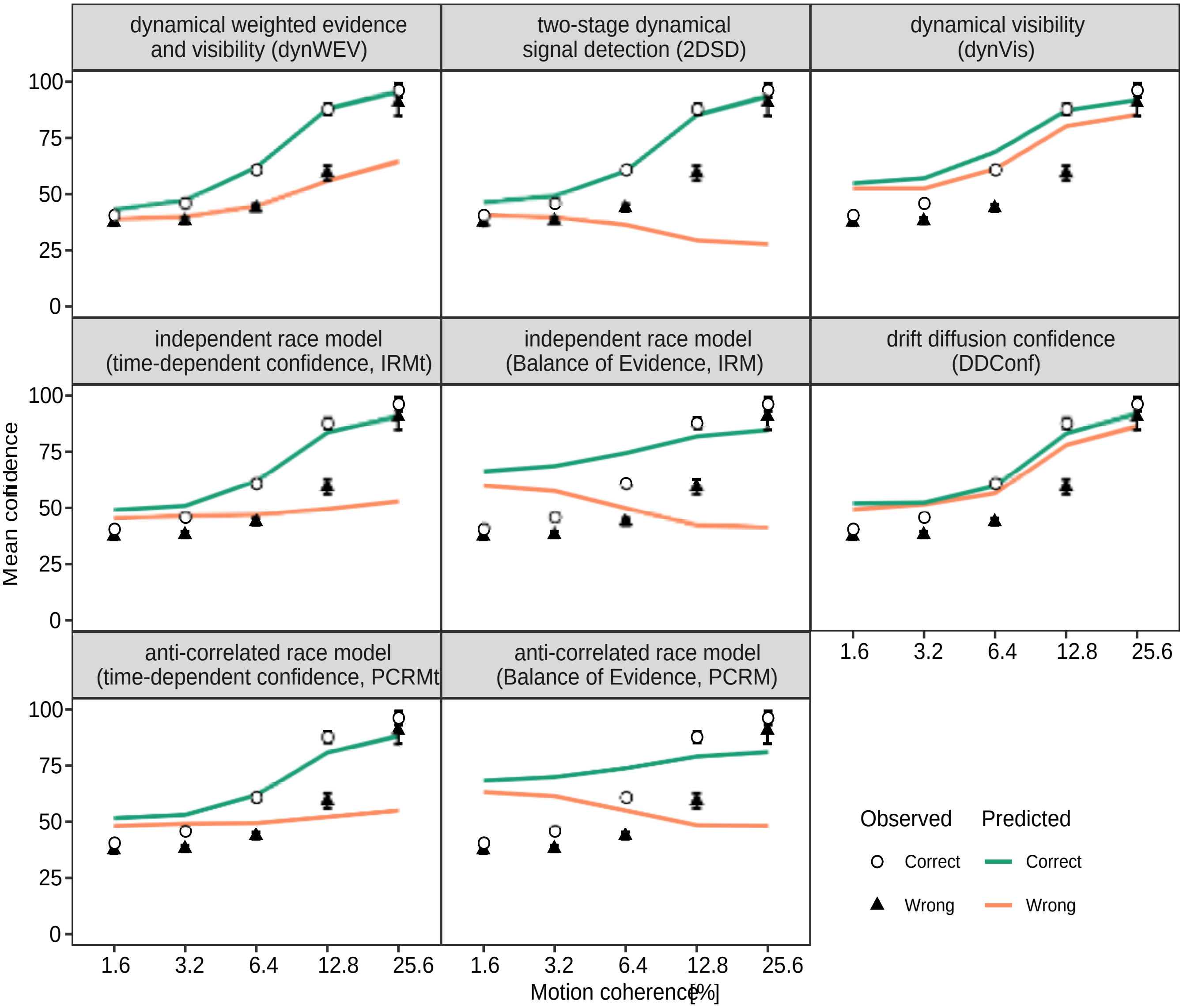
Note. Mean accuracy (A), confidence ratings (in % of visual scale, B), and response times (C) for different levels of stimulus onset asynchrony (SOA) and mean response times for different levels of confidence (in % of visual scale, D). Error bars represent within-subject standard errors.

increase, IRMt and PCRMt underestimated the slope of confidence in incorrect trials as a function of coherence. Wrong decisions in the easiest condition (25.6% coherence) form 0.1% of all observations. Therefore, mean confidence in incorrect decisions for the highest coherence did not strongly influence model fitting. With respect to the discrete empirical response distribution, the deviation was most apparent in the most extreme conditions and extreme confidence reports. All models except for dynWEV tended to underestimate the proportion of high confidence in easy conditions while for hard conditions they overestimated the proportion of high confidence trials (see Suppl. Fig. 5). For low confidence, the opposite was the case. These deviations were most pronounced in DDConf and race models. Concerning response times, dynWEV and 2DSD were more accurate in fitting response time quantiles than all other models (Fig. 12). DDConf produced again a very flat RT distribution, which does not capture the distribution of the data. Race models with time dependent confidence (IRMt and PCRMt) overestimated response times, especially in low confidence trials, but they captured the overall pattern of the relationship between response time and confidence, that is higher confidence was linked to faster decisions and particularly a shorter tail of the distribution as indexed by the upper quantiles. Time-independent confidence race

models as well as dynVis seem to have missed this pattern as decision time increases for extreme high and extreme low confidence. It can also be seen that dynWEV and 2DSD, the two models based on the drift diffusion process, tended to overestimate the tail in high confidence errors, but high confidence errors form only 2.7% of the data.

Figure 11

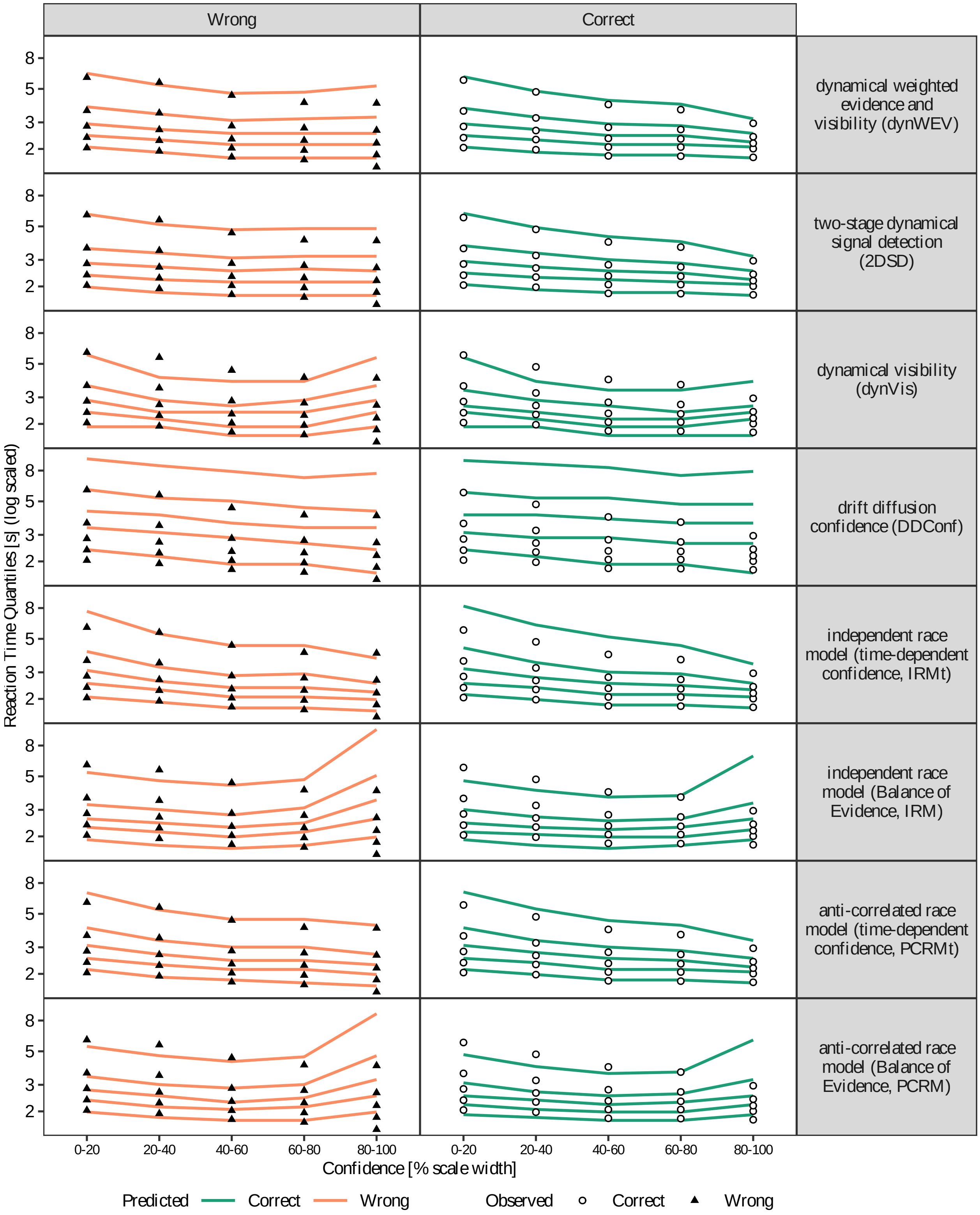
Observed mean confidence vs. confidence predicted by model fits for Experiment 2



Note. Empirical (points) and fitted (lines) mean confidence ratings as function of stimulus onset asynchrony for correct (green, circles) and incorrect (orange, triangles) decisions. Error bars around points and shaded areas around lines represent within-subject standard errors. Ribbons around lines are barely visible because of small intervals.

Figure 12

Observed response time quantiles vs. quantiles predicted by model fits for Experiment 2



Note. Empirical (triangles and points) and fitted (lines) response time quantiles (log scaled; probabilities: .1, .3, .5, .7, .9) across confidence judgments for incorrect (left column) and correct (right column) answers. Empirical quantiles were computed from the whole data set without regarding for participants. Predicted quantiles were computed after aggregating the individual response time densities.

Model comparison in terms of information criteria. Similar to Experiment 1, we only report BIC results, here, as the comparison using AIC delivers similar results. The BIC favored dynWEV above the other models for 27 of the total 42 participants (Suppl. Fig. 7). For the other participants, in 10 cases 2DSD, in one case dynVis, and in 4 cases PCRMt delivered the best fit. In terms of the mean BIC, dynWEV obtained the best fit and 2DSD the second best fit ($M = 28$, $SD = 35$) followed by dynVis ($M = 269$, $SD = 230$). Among the race models the ones with time-dependent confidence variable, IRMt ($M = 170$, $SD = 135$) and PCRMt ($M = 191$, $SD = 179$), performed best. The Bayesian t-test comparison revealed decisive evidence for dynWEV compared to 2DSD ($BF_{10} = 3.35 \times 10^3$, posterior CI = [0.44–1.13]) with even more extreme results for the other comparisons (BF_{10} 10^5). Supplementary Table 2 shows the results for the comparison to the other models.

When analyzing the data separately for the two data collection periods Bayes factors also indicated at least strong evidence in favor of dynWEV. For the first data collection period, which included 30 participants, the Bayes factor for the comparison of dynWEV with 2DSD was 14.46 (posterior CI = [0.20–0.97]) and for the comparison with all other models at least 2.35×10^5. For the second collection period with 12 participants and different number of sessions, the comparison of dynWEV to PCRMt resulted in the lowest Bayes factor ($BF_{10} = 45.06$, posterior CI = [0.43–1.96]) still indicating very strong evidence for dynWEV. The comparison with the other models delivered again decisive evidence in favor of dynWEV ($BF_{10} > 1.14 \times 10^2$).

Model identification analysis

In both experiments dynWEV was the best and 2DSD was the second best performing model in terms of BIC and AIC. Thus, we simulated one artificial data set for each participant using the fitted parameters and number of observed trials. Then, we fitted the generative model and dynWEV to the simulated data and compared the model fit using the BIC. In this situation model mimicry is obvious as the 2DSD is a special case of dynWEV when the weight on the evidence accumulator w is equal to 1. However, for none of the 58 participants of both experiments the dynWEV performed better than 2DSD, if the data was generated by 2DSD. More precisely, in dynWEV the fitted weight parameter w was close to 1 for most participants ($M = .96$, $SD = .09$), which means that dynWEV would also prefer a single-process architecture if it is the best explanation.

A second mimicry analyses was performed to investigate whether one of the race models was falsely classified, we performed another model identification analysis with PCRMt as generative model. We chose to include PCRMt as a generative model into model identification analysis because PCRMt was the best performing model among the different flavors of the race model according to Bayes factors. For none of the 58 simulated data sets dynWEV achieved a lower BIC compared to PCRMt. Moreover, the maximum likelihood achieved by the fits was higher for PCRMt compared to dynWEV for all artificial participants, indicating that dynWEV is not able to produce the same or similar joint distributions of decisions, response times and confidence. Thus, the models seem to be identifiable by the fitting procedure in the present study.

Parameter recovery analysis

In addition to the model identification analysis we conducted a parameter recovery analysis for dynWEV to test how accurate the parameter can be estimated with the experimental data and fitting procedure. We observed a high correlation between true and observed parameters for most parameters, indicating that parameter recovery was rather robust (Suppl. Fig. 10). Among the parameters that also feature in the standard diffusion model, the lowest correlation coefficients of .7 were observed for s_z and t_0. The only two parameters for which parameter recovery was suboptimal were the variance parameters of the visibility process $_V$ (= .11) and s_V (= .27). These two parameters are closely linked in the mathematical formulae.

Discussion

The modeling analysis revealed a better fit of the dynWEV model compared to other dynamical models of decision confidence in both the masked orientation discrimination task and the random dot motion task. Specifically, only dynWEV and 2DSD, two models assuming that the choice is based on a drift diffusion process, were able to accurately fit the response time distributions as a function of different levels of confidence, while even the best performing race models using time-dependent confidence variables lacked accuracy in the explanation of response times. However, only dynWEV but not 2DSD was able to account for the increase in confidence with stimulus discriminability in both correct and incorrect decisions.

Implications of the dynWEV model

The accurate fit of the dynWEV model has some theoretical implications how human observers compute confidence judgments.

Close relationship between decision dynamics and confidence.

First, the dynWEV model demonstrates that dynamical decision models can be extended to accurately describe the distribution of choices, confidence and response times at the same time. Moreover, this study demonstrates that a single

model is able to account for experiments where the response time distributions are similar across different levels of confidence as in Experiment 1 as well as experiments where the distributions change as a function of confidence as in Experiment 2. We speculate that the strength of relationship between confidence and reaction time may vary between stimulus types used in discrimination tasks. Static stimuli lead to a weaker relationship and dynamic stimuli like RDKs induce stronger relationships (Kiani et al., 2014; Rollwage et al., 2020; van den Berg et al., 2016; Zylberberg et al., 2012). Dynamical confidence models have already been successfully applied to static stimuli (Pleskac & Busemeyer, 2010). It is important to note here that even when the relationship between confidence and response time is weak, there is a lot of information in the data to identify models by the relationship of confidence and discriminability (see Fig. 8 and 11) and between discriminability and response time (see Suppl. Fig. 8 and 9). Because confidence is related to the decision dynamics, any changes of parameters describing the decision process also affect confidence. The two best-fitting models, 2DSD and dynWEV generally provided a good fit to response times as a function of confidence by assuming that confidence is generated by a similar dynamic process as the decision itself. In spite of that, in contrast to previous studies, the winning models do not imply a direct causal connection between confidence and decision time (cf. Kiani et al., 2014). Given there is converging evidence that confidence and reaction times are related (e.g. Kiani et al., 2014; Rahnev et al., 2020), it seems problematic to model confidence ignoring reaction times as many static models of confidence do. Previous studies aiming to extend dynamical models to account for confidence often did not use the rich information provided by the joint distribution of choices, confidence, and response times to estimate parameters and examine model fit. For example, previous studies fitted only response times (Kiani et al., 2014), used aggregated data, like decision probabilities and average response times (van den Berg et al., 2016), used only the quantiles of the reaction time distribution (Pleskac & Busemeyer, 2010; Ratcliff & Starns, 2013), or fitted first the drift diffusion model parameters to the choice and response time data, and used the confidence data only to fit the confidence-specific parameters (Desender et al., 2021; Moran et al., 2015).

Postdecisional accumulation of evidence. Second, the present study supports previous findings that confidence is informed by additional evidence about the choice alternatives that is not involved in the decision process (Moran et al., 2015; Pleskac & Busemeyer, 2010; Resulaj et al., 2009; van den Berg et al., 2016), because the two best-fitting models, 2DSD and dynWEV share the assumption of a postdecisional period of evidence accumulation. A period of postdecisional accumulation is absolutely necessary in 2DSD to account for variation in confidence judgments. But also for dynWEV, the fitted parameters indicate that there is a prolonged period of postdecisional accumulation. The postdecisional period constitutes on average 67.4% of the overall response times in Experiment 1 and 46.3% in Experiment 2 (see also Suppl. Table 3 and Suppl. Table 4 for summaries of parameter fits). In addition,

dynVis, which is identical to dynWEV except it omits postdecisional accumulation, did not fit the data well. In previous studies modeling postdecisional accumulation, it was generally assumed that postdecisional accumulation of evidence occurs during the time window between the response to the task and the confidence judgments (Moran et al., 2015; Pleskac & Busemeyer, 2010; Resulaj et al., 2009). However, in the present study, participants responded to the task and reported their degree of confidence at the same time. Nevertheless, postdecisional accumulation provided a far better account of confidence than models that lacked postdecisional accumulation. Moreover, in both experiments estimates of postdecisional accumulation time were strongly correlated with metacognitive sensitivity. In line with a contribution of postdecisional accumulation, experimentally increasing the time between a perceptual decision and a confidence judgment improves metacognitive sensitivity (Moran et al., 2015; Yu et al., 2015). Moreover, some but not all electrophysiological markers of perceptual confidence are not found until after the response (Boldt & Yeung, 2015; Rausch et al., 2020). Therefore, postdecisional evidence accumulation may be necessary for the computation of confidence in perceptual decisions in humans.

Parallel estimation of stimulus discriminability. Third, the accurate fit of the dynWEV provides strong evidence that evidence about stimulus discriminability is involved in the computation of confidence (Rausch et al., 2018, 2020). Information about stimulus discriminability is at least in parts accumulated independently of decision evidence. In the masked orientation discrimination task of Experiment 1 visibility may be informed by the perceived size, form, or presentation time of the stimulus. Concerning the random dot kinematograms in Experiment 2, this means there is information available from the random dot stimulus that is predictive of stimulus discriminability but is not predictive of the choice. For example, the precision of the representation of motion orientation is predictive of stimulus discriminability i.e. coherence, but orientation is not predictive of direction of motion. Thus, the dynWEV model is consistent with probabilistic theories of perception, according to which observers take into account knowledge about the uncertainty associated with observations (Ma, 2012). A possible neural mechanism may involve posterior parietal cortex and ventral striatum, which were found to track sensory reliability independently of the choice (Bang & Fleming, 2018). Evidence about stimulus reliability seems to be an adequate explanation for the double increase pattern observed in previous studies (Kiani et al., 2014; Rausch et al., 2018, 2020). A previous study had proposed that the double increase pattern can be explained by a combination of balance-of-evidence and decision time (Kiani et al., 2014). However, in the present study no race model with time dependent evidence was able to reproduce the double increase pattern of confidence, although confidence was computed as a combination of terms including Balance of Evidence and decision time and therefore these models are in principle able to produce a double increase pattern (e.g. if confidence was solely a function of decision time, i.e. with

$w_{RT} = 1$, $w_X = 0$, and $w_{Int} = 0$). The additional constraints put on the parameters by fitting the whole joint distribution of decisions, response times, and confidence reports are likely to cause the discrepancies between empirical and predicted mean confidence (Figures 8 and 11). This again emphasizes the necessity of fitting the joint distribution of choice, response time and confidence to identify the complete cognitive architecture underlying perceptual decisions.

Specification of the dynWEV model. It should be noted that the dynWEV model presented here is only one possibility to formulate a dynamical version of a weighted evidence and visibility model of confidence. There were some choices made in the concrete formulation that were somehow arbitrary. Most importantly, we included trial-to-trial variation of the drift rate of the visibility accumulation. Variability of drift rate in the visibility process seems plausible since effects similar to the ones causing the drift variation in decision evidence accumulation should also be present in the visibility accumulation. However, although dynWEV implies that the variations of drift rates in the decision process and in the visibility process are independent, it might be assumed that these variations of drift rates in the decision and visibility process are in fact correlated. The plausibility of independent drift rate variability depends on the origin and nature of drift rate variation. In the drift diffusion model, variation in drift rates is often used to account for large reaction times in incorrect responses and to ensure a finite asymptote of accuracy when the boundary separation increases arbitrarily (Ratcliff & Rouder, 1998; Starns et al., 2012). It was previously proposed that the variability in the drift rate arises from noise in visual and memory encoding of the stimuli and the variation in accumulation from the actual comparison process, for example in a word-matching task (Ratcliff, 1981) or an orientation discrimination task (Smith et al., 2004). This interpretation would support correlated drift rates in visibility and evidence accumulation. On the other hand, there are some arguments for an independent variation. First, in an object identification task, drift rate variability in the diffusion model seems to arise from late processing of the task relevant features of the stimulus and seems to be part of the neural process of decision making (Ratcliff et al., 2009). In addition, there is evidence that different stimulus features are processed independently and in parallel in the visual system if stimulus are presented briefly (Kyllingsbaek & Bundesen, 2007), which supports our approach as the visibility process incorporates task-irrelevant and thus different features than the decision accumulator. Besides the drift rate variation there is the possibility that the diffusion noise is not independent as in our formulation but that process noise is shared between the accumulation processes. To substantiate the model specification future research may try to independently manipulate specific parameters of dynWEV through experimental conditions (Voss et al., 2004).

Limitations and open questions

Although there is strong evidence in favor of dynWEV compared to all other models presented here, there are some limitations of the present study and open questions, which should be pointed out here.

Response times in high confidence errors. First, despite of providing the best fit to response times of all models tested, the dynWEV model still overestimates response times for high confidence errors (see Fig. 9 and 12). Although these observations form a rather small proportion of observations, the pattern is apparent in both experiments. In previous studies, the pattern of response times for high confidence errors was not consistent and may be influenced by stimulus properties and task setting like instruction (Ratcliff & Starns, 2009; Starns et al., 2012). Future studies will require an extraordinary number of trials to investigate how these differences in response times for high confidence errors are explained by the properties of the decision process.

Modeling confidence response times and changes of mind. Second, for 2DSD and dynWEV, we assumed a constant duration of the postdecisional accumulation period. However, a fixed duration of postdecisional accumulation seems too restrictive. Thus, the models presented in the present study may be improved by including a mechanism that produces probabilistic postdecisional accumulation times (e.g. Moran et al., 2015). An experimental paradigm in which choice and confidence are reported subsequently would allow to explicitly model confidence response times as well as changes of mind. The modeling of confidence response time distributions may deliver further insight into the dynamics involved in the generation of a confidence judgment by postdecisional accumulation. However, to fit more sophisticated models may require using simulation-based techniques. In addition, it is not at all clear that postdecisional accumulation starts not until the initial choice response was made, such that simply separating the responses physically does not solve the research question when in time the postdecisional accumulation begins. The postdecisional accumulation period allows 2DSD and dynWEV in principle to account for changes of mind. However, the way how the models are formulated in the present study implies that participants stick to their decision during the postdecisional accumulation period and that contradicting evidence may only lead to low confidence. Under which conditions changes of mind arise is still an open research question. With respect to the dynWEV model, additional research is necessary to investigate the role of the visibility process in changes of mind. In addition, it is possible that changes of mind occur but are not observable in the present paradigm. Finally, by incorporating confidence response times and changes of mind as an additional variable in the models, a direct comparison with the race models presented in this study would not be possible. Future studies are necessary to generalize the existing race models with postdecisional accumulation.

Metacognitive Sensitivity. Third, the present study may also have implications for the large research program relying on measurements of metacognitive accuracy. A large number of different measures of metacognitive accuracy exist, some of which are model free, such as Goodman and KruskalŠs Gamma (Nelson, 1984) and the area under type-2 ROC curve (Fleming et al., 2010). However, there are also measures of metacognitive accuracy that rely on specific confidence models to disentangle between metacognitive accuracy and subjective criteria, for example meta-dŠ/dŠ (Maniscalco & Lau, 2012); $_{meta}$ (Shekhar & Rahnev, 2021), confidence efficiency (Mamassian & de Gardelle, 2021), or metacognitive noise (Guggenmos, 2022). Future studies are needed to investigate how these measures of metacognitive accuracy are related to the parameters of the dynWEV model (or other dynamical models of confidence) to see if measures of metacognitive accuracy are incomplete or even biased due to not accounting for dynamic accumulation processes.

Other sequential sampling models of confidence. In addition, we restricted the comparison of models in this study to models for which solutions to likelihood functions are available and did not include models for which we would have to approximate the likelihood by sampling. Thus, models such as RTCON (Starns et al., 2012) and the bounded accumulation model proposed by Kiani et al. (2014) were not considered in the present study. Finally, we also did not consider other decision architectures except for the drift diffusion model and the race model such as the leaky competing accumulator model (Usher & McClelland, 2001).

Multiple alternative decisions. The dynWEV model shares a drawback with all drift diffusion based models, namely it applies to binary decisions. Though the excellent fit in experimental paradigms that often use binary decisions diffusion models lack some external validity since real world decisions most often include multiple if not a continuous scale of possible alternatives. Race models may be applied to multiple alternatives in a straightforward way since every alternative has its accumulator but perform worse in fitting response time distributions. One possible explanation of the better fit of diffusion models is the inter-trial variations that make is very flexible. Indeed, most of these additional parameters are included to gather all empirical patterns of response times (Ratcliff & Rouder, 1998; Ratcliff et al., 2016). Inter-trial variability in stating point (s_z) should allow for fast errors while variation in drift rate (s) makes small errors possible. It may be possible that race models can achieve equally good fits when provided with this additional freedom. Furthermore, the visibility accumulation may be also incorporated in such a model, which would be a possible alternative candidate for the dynWEV model presented in this study.

Magnitude Sensitivity. Finally, while dynWEV is applicable to a broad range of psychophysical tasks, there is one empirical phenomenon that can not be explained by the current version of the model: magnitude sensitivity. Magnitude sensitivity describes the effect of faster decisions in decision tasks when

the stimulus magnitude (e.g. illumination or numerosity in a respective task) or value (e.g. reward in points) is increased for all available alternatives (Pirrone et al., 2021). This effect is well studied in detection tasks (Pins & Bonnet, 1996; van Maanen et al., 2012) and can be observed in various value-based and perceptual decision tasks in humans and animals (Ratcliff et al., 2018; Teodorescu et al., 2016; van Maanen et al., 2012) even if both alternatives are equal in magnitude (Kirkpatrick et al., 2021; Pirrone, Azab, et al., 2018; Pirrone, Wen, & Li, 2018). Recent studies examined different possible sequential sampling accounts for this effect. Drift diffusion models with intensity dependent noise parameters and leaky competing accumulator models both deliver explanations for the data but clear evidence in favor of a specific model is missing (Ratcliff et al., 2018; Teodorescu et al., 2016), as there is a high degree of model mimicry (Bogacz et al., 2006; Bose et al., 2020).

A similar effect as magnitude sensitivity has been reported with confidence reports. By separately manipulating signal strength and signal to noise ratios, it is possible to manipulate confidence without affecting accuracy (Koizumi et al., 2015; Odegaard et al., 2018; Samaha et al., 2016). Given that the effects of magnitude sensitivity can be explained by the dynamics of decision making, it seems necessary to investigate if the effects on confidence can be explained by these dynamics as well. The present study shows that the cognitive modeling of the joint distribution of choice, response time and confidence is a powerful tool for model comparison.

Conclusion

We proposed a sequential sampling model, the dynamical weighted evidence and visibility model, to account for the joint distribution of decisions, response times, and confidence judgments in binary perceptual decision tasks. In two different perceptual decision tasks with simultaneous confidence reports, the model fitted the empirical data better than five alternative models and captured all relevant relationships between decision, response time and confidence judgment. These observations indicate that confidence is not exclusively based on evidence utilized in the decision process, but incorporates also postdecisional evidence as well as independent evidence about the reliability of the stimulus. Moreover, we demonstrated that using the all the information available in the data in form of the full joint distribution of all dependent variables is both feasible and advantageous for model fitting and comparison.

References

Aitchison, L., Bang, D., Bahrami, B., & Latham, P. E. (2015). Doubly bayesian analysis of confidence in perceptual decision-making. PLOS Computational Biology, 11(10), Article e1004519. https://doi.org/10.1371/journal.pcbi.1004519

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

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

Bates, D., Mullen, K. M., Nash, J. C., & Varadhan, R. (2015). Minqa: Derivative-free optimization algorithms by quadratic approximation. https://cran.r-project.org/web/packages/minqa

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. Psychological review, 113(4), 700–765. https://doi.org/10.1037/0033-295X.113.4.700

Boldt, A., & Yeung, N. (2015). Shared neural markers of decision confidence and error detection. Journal of Neuroscience, 35(8), 3478–3484. https://doi.org/10.1523/JNEUROSCI.0797-14.2015

Bose, T., Pirrone, A., Reina, A., & Marshall, J. A. (2020). Comparison of magnitude-sensitive sequential sampling models in a simulation-based study. Journal of Mathematical Psychology, 94, 102298. https://doi.org/10.1016/j.jmp.2019.102298

Desender, K., Donner, T. H., & Verguts, T. (2021). Dynamic expressions of confidence within an evidence accumulation framework. Cognition, 207, Article 104522. https://doi.org/10.1016/j.cognition.2020.104522

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

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. Annual review of neuroscience, 30, 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with orsee. Journal of the Economic Science Association, 1(1), 114–125. https://doi.org/10.1007/s40881-015-0004-4

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

Hellmann, S., & Rausch, M. (2022). Modelling reaction time and confidence distributions in decision making. https://doi.org/10.17605/OSF.IO/8FZBX

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

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

Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. Neuron, 84 (6), 1329 1342. https://doi.org/10.1016/j.neuron.2014.12.015

Kirkpatrick, R. P., Turner, B. M., & Sederberg, P. B. (2021). Equal evidence perceptual tasks suggest a key role for interactive competition in decision-making. Psychological review, 128 (6), 1051 1087. https://doi.org/10.1037/rev0000284

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

Kyllingsbaek, S., & Bundesen, C. (2007). Parallel processing in a multifeature whole-report paradigm. Journal of experimental psychology. Human perception and performance, 33 (1), 64 82. https://doi.org/10.1037/0096-1523.33.1.64

LaBerge, D. (1994). Quantitative models of attention and response processes in shape identification tasks. Journal of Mathematical Psychology, 38 (2), 198 243. https://doi.org/10.1006/jmps.1994.1015

Lak, A., Nomoto, K., Keramati, M., Sakagami, M., & Kepecs, A. (2017). Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. Current biology : CB, 27 (6), 821 832. https://doi.org/10.1016/j.cub.2017.02.026

Lee, M. D., & Wagenmakers, E.-J. (2014). Bayesian cognitive modeling: A practical course. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

Lerche, V., & Voss, A. (2019). Experimental validation of the diffusion model based on a slow response time paradigm. Psychological Research, 83 (6), 1194 1209. https://doi.org/10.1007/s00426-017-0945-8

Ma, W. J. (2012). Organizing probabilistic models of perception. Trends in cognitive sciences, 16 (10), 511 518. https://doi.org/10.1016/j.tics.2012.08.010

Macmillan, N. A., & Creelman, C. D. (2005). Detection theory: A userŠs guide (2nd ed.). Lawrence Erlbaum Associates Publishers.

Mamassian, P., & de Gardelle, V. (2021). Modeling perceptual confidence and the confidence forced-choice paradigm. Psychological review, Advance online publication. 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). https://doi.org/10.1093/nc/niw002

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

Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., & Rangel, A. (2010). The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. Judgment and Decision Making, 5 (6), 437–449. https://doi.org/10.2139/ssrn.1901533

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

Moreno-Bote, R. (2010). Decision confidence and uncertainty in diffusion models with partially correlated neuronal integrators. Neural computation, 22 (7). https://doi.org/10.1162/neco.2010.12-08-930

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). Bayesfactor: Computation of bayes factors for common designs. https://cran.r-project.org/web/packages/BayesFactor

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. Psychological bulletin, 95 (1), 109–133.

Odegaard, B., Grimaldi, P., Cho, S. H., Peters, M. A. K., Lau, H., & Basso, M. A. (2018). Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. Proceedings of the National Academy of Sciences, 115 (7), E1588–E1597. https://doi.org/10.1073/pnas.1711628115

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. Trends in cognitive sciences, 21 (6), 425–433. https://doi.org/10.1016/j.tics.2017.03.011

Peirce, J. W. (2007). Psychopy—psychophysics software in python. Journal of neuroscience methods, 162 (1-2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Peirce, J. W. (2009). Generating stimuli for neuroscience using psychopy. Frontiers in neuroinformatics, 2 (10), 1–8. https://doi.org/10.3389/neuro.11.010.2008

Pereira, M., Megevand, P., Tan, M. X., Chang, W., Wang, S., Rezai, A., Seeck, M., Corniola, M., Momjian, S., Bernasconi, F., Blanke, O., & Faivre, N. (2021). Evidence accumulation relates to

perceptual consciousness and monitoring. Nature Communications, 12(1), 3261. https://doi.org/10.1038/s41467-021-23540-y

Philiastides, M. G., Ratcliff, R., & Sajda, P. (2006). Neural representation of task difficulty and decision making during perceptual categorization: A timing diagram. Journal of Neuroscience, 26(35), 8965–8975. https://doi.org/10.1523/JNEUROSCI.1655-06.2006

Pins, D., & Bonnet, C. (1996). On the relation between stimulus intensity and processing time: PiéronŠs law and choice reaction time. Perception & Psychophysics, 58(3), 390–400. https://doi.org/10.3758/BF03206815

Pirrone, A., Azab, H., Hayden, B. Y., Stafford, T., & Marshall, J. A. R. (2018). Evidence for the speed-value trade-off: Human and monkey decision making is magnitude sensitive. Decision (Washington, D.C.), 5(2), 129–142. https://doi.org/10.1037/dec0000075

Pirrone, A., Reina, A., Stafford, T., Marshall, J. A., & Gobet, F. (2021). Magnitude-sensitivity: Rethinking decision-making. Trends in Cognitive Sciences, 26(1), 66–80. https://doi.org/10.1016/j.tics.2021.10.006

Pirrone, A., Wen, W., & Li, S. (2018). Single-trial dynamics explain magnitude sensitive decision making. BMC Neuroscience, 19(1), 54. https://doi.org/10.1186/s12868-018-0457-5

Pleskac, T. J., & Busemeyer. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. Psychological review, 117(3). 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

Powell, M. J. D. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. https://www.damtp.cam.ac.uk/user/na/NA_papers/NA2009_06.pdf

R Core Team. (2021). R: A language and environment for statistical computing. https://www.R-project.org/

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. (1981). A theory of order relations in perceptual matching. Psychological review, 88(6), 552–572. https://doi.org/10.1037/0033-295X.88.6.552

Ratcliff, R., Hasegawa, Y. T., Hasegawa, R. P., Smith, P. L., & Segraves, M. A. (2007). Dual diffusion model for single-cell recording data from the superior colliculus in a brightness-discrimination task. Journal of neurophysiology, 97(2), 1756–1774. https://doi.org/10.1152/jn.00393.2006

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., Philiastides, M. G., & Sajda, P. (2009). Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. Proceedings of the National Academy of Sciences of the United States of America, 106(16), 6539–6544. https://doi.org/10.1073/pnas.0812589106

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. Psychological Science, 9(5), 347–356. https://doi.org/10.1111/1467-9280.00067

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. Psychological review, 111(2), 333–367. https://doi.org/10.1037/0033-295X.111.2.333

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. Trends in cognitive sciences, 20(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. Psychological review, 116(1), 59–83. https://doi.org/10.1037/a0014086

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

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. Psychology and Aging, 16(2), 323–341. https://doi.org/10.1037/0882-7974.16.2.323

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. Psychonomic bulletin & review, 9(3), 438–481. https://doi.org/10.3758/bf03196302

Ratcliff, R., Voskuilen, C., & Teodorescu, A. (2018). Modeling 2-alternative forced-choice tasks: Accounting for both magnitude and difference effects. Cognitive psychology, 103, 1–22. https://doi.org/10.1016/j.cogpsych.2018.02.002

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

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

Rausch, M., & Zehetleitner, M. (2019). The folded x-pattern is not necessarily a statistical signature of decision confidence. PLoS computational biology, 15(10), Article e1007456. https://doi.org/10.1371/journal.pcbi.1007456

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

Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. Nature, 461(7261), 263–266. https://doi.org/10.1038/nature08275

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

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin & Review, 16(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

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

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461–464. https://doi.org/10.1214/aos/1176344136

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

Singmann, H., Brown, S., Gretton, M., Heathcote, A., Voss, A., Voss, J., & Terry, A. (2020). Rtdists: Response time distributions.

Smith, P. L., Ratcliff, R., & Wolfgang, B. J. (2004). Attention orienting and the time course of perceptual decisions: Response time distributions with masked and unmasked displays. Vision research, 44(12), 1297–1320. https://doi.org/10.1016/j.visres.2004.01.002

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of zroc slopes with response time data and the diffusion model. Cognitive psychology, 64(1-2), 1–34. https://doi.org/10.1016/j.cogpsych.2011.10.002

Tajima, S., Drugowitsch, J., & Pouget, A. (2016). Optimal policy for value-based decision-making. Nature Communications, 7(1), Article 12400. https://doi.org/10.1038/ncomms12400

Teodorescu, A. R., Moran, R., & Usher, M. (2016). Absolutely relative or relatively absolute: Violations of value invariance in human decision making. Psychonomic Bulletin & Review, 23 (1), 22 38. https://doi.org/10.3758/s13423-015-0858-8

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. Psychological review, 108 (3), 550 592. https://doi.org/10.1037/0033-295x.108.3.550

van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. eLife, 5, Article e12192. https://doi.org/10.7554/eLife.12192

van Maanen, L., Grasman, R. P. P. P., Forstmann, B. U., & Wagenmakers, E.-J. (2012). PiéronŠs law and optimal behavior in perceptual decision-making. Frontiers in Neuroscience, 5, Article 143. https://doi.org/10.3389/fnins.2011.00143

Vickers, D., Smith, P., Burt, J., & Brown, M. (1985). Experimental paradigms emphasising state or process limitations: Ii effects on confidence. Acta Psychologica, 59 (2), 163 193. https://doi.org/10.1016/0001-6918(85)90018-6

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. Memory & Cognition, 32 (7), 1206 1220. https://doi.org/10.3758/BF03196893

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. The Annals of Mathematical Statistics, 19 (3), 326 339. https://doi.org/10.1214/aoms/1177730197

Wald, A. (1947). Sequential analysis. John Wiley.

Yu, S., Pleskac, T. J., & Zeigenfuse, M. D. (2015). Dynamics of postdecisional processing of confidence. Journal of experimental psychology. General, 144 (2), 489 510. https://doi.org/10.1037/xge0000062

Zawadzka, K., Higham, P. A., & Hanczakowski, M. (2017). Confidence in forced-choice recognition: What underlies the ratings? Journal of experimental psychology. Learning, memory, and cognition, 43 (4), 552 564. https://doi.org/10.1037/xlm0000321

Zhang, S., Lee, M. D., Vandekerckhove, J., Maris, G., & Wagenmakers, E.-J. (2014). Time-varying boundaries for diffusion models of decision making and response time. Frontiers in psychology, 5, Article 1364. https://doi.org/10.3389/fpsyg.2014.01364

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