

# Psychological Review

## Modeling Speed-Accuracy Trade-Offs in the Stopping Rule for Confidence Judgments

Stef Herregods, Pierre Le Denmat, Luc Vermeylen, and Kobe Desender

Online First Publication, December 8, 2025. <https://dx.doi.org/10.1037/rev0000603>

### CITATION

Herregods, S., Le Denmat, P., Vermeylen, L., & Desender, K. (2025). Modeling speed-accuracy trade-offs in the stopping rule for confidence judgments. *Psychological Review*. Advance online publication. <https://dx.doi.org/10.1037/rev0000603>

# Modeling Speed–Accuracy Trade-Offs in the Stopping Rule for Confidence Judgments

Stef Herregods, Pierre Le Denmat, Luc Vermeulen, and Kobe Desender  
Research Unit Brain and Cognition, Leuven Brain Institute, Katholieke Universiteit Leuven

Making a decision and reporting your confidence in the accuracy of that decision are thought to be driven by the same mechanism: the accumulation of evidence. It is well known that choices and reaction times are well accounted for by a computational model assuming noisy accumulation of evidence until crossing a decision boundary (e.g., the drift diffusion model). Decision confidence can be derived from the amount of evidence following postdecision evidence accumulation. Currently, the stopping rule for postdecision evidence accumulation is underspecified. In the current work, we quantitatively and qualitatively compare the ability of five prominent models of confidence couched within evidence accumulation to account for this stopping rule. We collected data for two experiments in which participants were instructed to make fast or accurate decisions and to give fast or carefully considered confidence judgments. We then compared the different models in their ability to capture the speed–accuracy effects on confidence. Both at the quantitative and the qualitative level, the data were best accounted for by our newly proposed flexible confidence boundary model, in which postdecision accumulation terminates once it reaches one of two opposing slowly collapsing confidence boundaries. Inspection of the parameters of this model revealed that instructing participants to make fast versus accurate decisions influenced the height of the decision boundaries, while instructing participants to make fast versus careful confidence judgments influenced the height of the confidence boundaries. Our data show that the stopping rule for confidence judgments can be well described as an accumulation-to-bound process, and that the height of these confidence boundaries is under strategic control.


**Keywords:** confidence, decision making, drift diffusion model, computational modeling

**Supplemental materials:** <https://doi.org/10.1037/rev0000603.supp>

Human decision making is accompanied by a sense of confidence. Humans often report high confidence when they make correct decisions and low confidence when they make incorrect decisions (Fleming et al., 2010). Understanding the computational underpinnings of decision confidence is of high importance, given that humans use decision confidence to adapt subsequent behavior (Desender et al., 2018, 2019; Folke et al., 2016). In recent work, identifying the computational underpinnings of decision confidence has been identified as an important common goal for the field of metacognition (Rahnev et al., 2022). Given that decision confidence reflects an evaluation of the accuracy of a decision, computational accounts of decision confidence usually depart from decision-making models and aim to explain the computation of confidence within these models.

In many decision-making scenarios, human observers face the challenging task of making accurate decisions based on noisy evidence. Many theories of decision making assume that people solve this challenge by accumulating multiple pieces of evidence. Accumulation-to-bound models specifically propose that evidence is accumulated sequentially until the accumulated evidence reaches a predefined decision boundary. Once the decision boundary is reached, the model makes a choice (for review, see Gold & Shadlen, 2007; Figure 1A). Within the drift diffusion model (DDM), evidence accumulates toward one of two opposing decision boundaries, with the additional assumption that evidence for both choice options is perfectly anticorrelated (Ratcliff & McKoon, 2008). In its most basic implementation, the DDM explains the dynamics of decision making using only three main parameters: a drift rate, corresponding to the

Peter D. Kvam served as action editor.

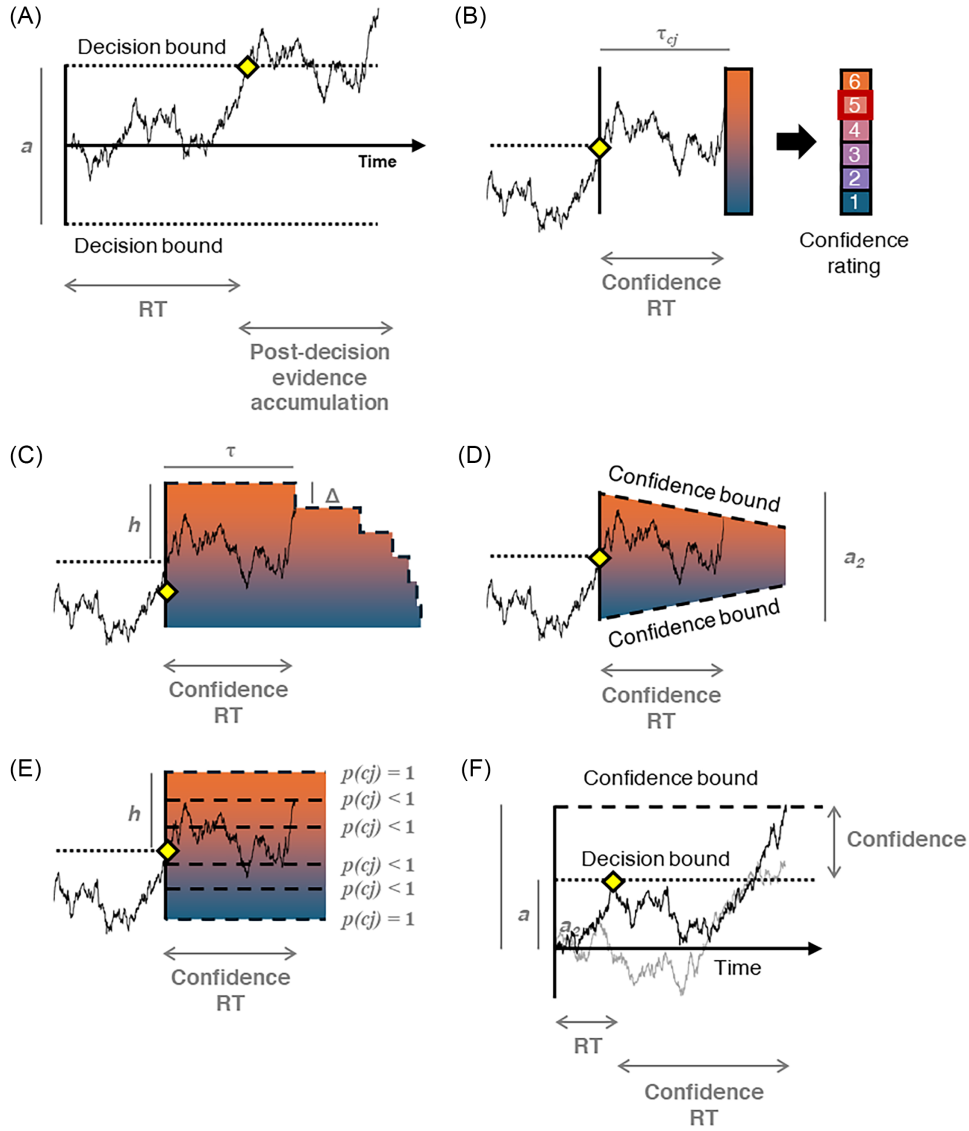
Kobe Desender  <https://orcid.org/0000-0002-5462-4260>

All hypotheses, sample sizes, exclusion criteria for participants, analyzed variables, the experimental design, and planned analyses were preregistered on the Open Science Framework registries (Herregods & Desender, 2021; <https://doi.org/10.17605/OSF.IO/Z2UCM> and <https://osf.io/vyh4k/overview>). This research was supported by a project grant from the Research Foundation Flanders, Belgium (FWO-Vlaanderen No. G0B0521N); a Central Europe Leuven Strategic Alliance grant from the Katholieke Universiteit Leuven (Central Europe Leuven Strategic Alliance/21/010); and an FWO postdoctoral fellowship to Luc Vermeulen from the Research Foundation Flanders,

Belgium (FWO-Vlaanderen No. 1242924N).

Stef Herregods played a lead role in data curation, formal analysis, investigation, and methodology and an equal role in writing–original draft. Pierre Le Denmat played a supporting role in formal analysis, supervision, and writing–review and editing. Luc Vermeulen played a supporting role in methodology, software, supervision, and writing–review and editing. Kobe Desender played a lead role in conceptualization, funding acquisition, and supervision, a supporting role in investigation and software, and an equal role in writing–original draft.

Correspondence concerning this article should be addressed to Kobe Desender, Research Unit Brain and Cognition, Leuven Brain Institute, Katholieke Universiteit Leuven, Tiensestraat 102, 3000 Leuven, Belgium. Email: [kobe.desender@kuleuven.be](mailto:kobe.desender@kuleuven.be)

**Figure 1***Competing Models Implementing Postdecision Confidence Computation*

**Note.** (A) According to the DDM, evidence accumulates until reaching a decision boundary at which point a decision is made, here indicated by the yellow diamond. Boundary separation ( $a$ ) can vary, changing the speed-accuracy trade-off. Confidence is often modeled by allowing the accumulation process to continue after boundary crossing; however, the exact nature of the stopping rule for postdecision processing is unclear. (B) The 2DSD model proposes that postdecision evidence accumulates for a fixed time period ( $\tau$ ) after making the decision. The final state of the postdecision accumulator then informs the confidence rating. (C) The CCB model assumes that the accumulation of postdecision evidence terminates once it reaches a slowly collapsing confidence boundary, with absolute height ( $h$ ), collapse time ( $\tau$ ), and collapse height ( $\Delta$ ) as free parameters. (D) In the newly proposed FCB model, postdecision evidence accumulates until reaching one of two slowly collapsing confidence boundaries, with confidence boundary separation  $a_2$  and confidence urgency  $u_2$ . (E) The optional stopping model assumes postdecision evidence accumulation with an evidence threshold for each level of confidence judgment, with evidence level  $c_{jx}$  and probability of absorbing  $p_{c_{jx}}$  as free parameters. The outermost thresholds are fully absorbing. (F) The race model proposed by Van Zandt and Maldonado-Molina (2004) assumes separate evidence accumulators for both decision options. Confidence is quantified as the difference in evidence once one or both accumulators reach the confidence boundary, with boundary height  $a_2$ . DDM = drift diffusion model; 2DSD = two-stage dynamic signal detection theory model; CCB = collapsing confidence boundary model; FCB = flexible confidence boundary; RT = reaction time. See the online article for the color version of this figure.

strength of the evidence accumulation process; a decision boundary, indicating the degree of evidence required before a decision is made; and nondecision time, capturing nondecision-related components. This simple tenet has proven to be a powerful framework that can account for a realm of behavioral and neurophysiological data. For example, accumulation-to-bound signals such as described by the DDM have been observed in human (Donner et al., 2009; O'Connell et al., 2012) and primate (Gold & Shadlen, 2007) neurophysiology. Most prominently, the DDM can explain the trade-off between speed and accuracy that characterizes all forms of speeded decision making (Bogacz, Wagenmakers, et al., 2010; Bogacz et al., 2006). When participants are instructed to make speeded versus accurate decisions, the DDM explains these data by changing the height of the decision boundary (although the selectivity of this effect has been debated; Rafiei & Rahnev, 2021). Decreasing the decision boundary effectively lowers the required level of evidence before reaching it, promoting fast responses at the expense of accuracy. Given that participants are able to change the decision boundary based on instructions (amongst many other manipulations), it is believed that the height of the decision boundary is under voluntary strategic control (Balci et al., 2011; Bogacz, Hu, et al., 2010).

Given the success of the DDM in explaining decision making, several attempts have been made to explain decision confidence within this model. Capitalizing on the notion that confidence is typically queried after a decision has been made, Pleskac and Busemeyer (2010) put forward the two-stage dynamic signal detection theory model (2DSD), which proposes that the process of evidence accumulation does not terminate once a decision boundary has been crossed, but rather there is continued accumulation of (postdecision) evidence, which determines decision confidence. If additional postdecision evidence confirms the initial decision, the model will produce a high confidence response. If additional postdecision evidence contradicts the initial decision, the model produces low confidence or even changes its mind about the initial decision (Resulaj et al., 2009; van den Berg et al., 2016). Given that postdecision evidence is most likely to contradict initial decisions when these were incorrect, this account can explain why confidence is usually higher for correct than for incorrect decisions (Moran et al., 2015; Pleskac & Busemeyer, 2010), and why confidence better tracks accuracy when participants take more time to report confidence (Yu et al., 2015).

Previous modeling work has shown that the 2DSD model can jointly explain choices, reaction times, and decision confidence (see also Calder-Travis et al., 2024; Hellmann et al., 2023; van den Berg et al., 2016; Zylberberg et al., 2016). Strikingly, much less attention has been devoted toward the speed with which confidence reports are provided. This is remarkable, given that confidence reaction times (RTs) are highly informative about the underlying computations (Moran et al., 2015). As a consequence, models such as the 2DSD propose a very simplistic stopping rule for the postdecision evidence accumulation process (see also Desender et al., 2022; Hellmann et al., 2023; Yu et al., 2015) or do not contain an explicit stopping rule at all (Balsdon et al., 2020; Pereira et al., 2021, 2022). Within the 2DSD model, an additional parameter is included, which controls the duration of the postdecision processing time (i.e., the time between the choice and the confidence report). Thus, in the 2DSD, the stopping rule for confidence judgments is to stop accumulating postdecision evidence once a certain amount of time (potentially with some variability around it) has passed (Figure 1B). Although the time-based

stopping rule proposed by the 2DSD model generally provides a good fit to confidence reports, to the best of our knowledge this model has not been used to jointly account for confidence and confidence RTs. Thus, it remains unclear whether the time-based stopping rule can account for confidence RTs, which typically show the same right-skewed distributions as choice RTs. Moreover, to the best of our knowledge, a comprehensive model fitting exercise providing a qualitative and quantitative comparison with competing models to unravel more appropriate mechanisms has not yet been performed. Therefore, in the current work, we set out to do exactly this and compare the ability of 2DSD to several other candidate models in explaining the stopping rule for confidence judgments.

Apart from the 2DSD model, the first competing model that will be considered is the collapsing confidence boundary model (CCB) put forward by Moran et al. (2015). Following the idea that confidence RTs might provide reliable information about the stopping rule of confidence, Moran et al. proposed a model with postdecision accumulation until reaching a single slowly CCB (Figure 1C). The authors showed that the CCB model was able to account for a variety of empirical patterns involving confidence RTs and confidence judgments, which could not be accounted for by the 2DSD. The work from Moran et al. thus suggests a (collapsing) confidence boundary as the stopping rule for confidence judgments.

A second competing model that will be considered has been proposed by Van Zandt and Maldonado-Molina (2004) in an effort to explain response reversals in recognition memory. In short, these authors propose a race between two fully independent evidence accumulators toward a decision boundary (determining the choice and a corresponding RT) and subsequently toward a confidence boundary (determining the confidence RT). The level of confidence is then quantified as the distance between both accumulators at the moment of reaching the confidence boundary (Figure 1F). This model is similar in spirit to the CCB put forward by Moran et al. (2015), except that it posits the existence of a flat confidence boundary (as opposed to a CCB), that it quantifies confidence as the balance-of-evidence between both accumulators (see also Vickers, 1979), and that it proposes a race between two fully independent accumulators (as opposed to having full inhibition between both accumulators, as in the DDM).

A third competing model we consider is the optional stopping model (Pleskac & Busemeyer, 2010). According to this model, participants set multiple evidence-based thresholds (i.e., multiple confidence boundaries), and once the postdecision accumulated evidence reaches one of these thresholds, the accumulation process stops with a certain probability. Thus, upon reaching a specific threshold, the participant may report the level of confidence associated with that threshold, or evidence accumulation may continue. The outermost thresholds are absorbing, always leading to a confidence judgment when crossed (Figure 1E).

Finally, we propose a fourth competing model that extends the CCB approach such that it allows for more flexibility in capturing various associations between confidence and confidence RTs. Specifically, due to its architecture, the CCB always predicts that longer confidence RTs are associated with lower confidence (see Figure 1C). Such a pattern is quite common in experiments where participants report their confidence on a scale running from uncertain (i.e., 50% correct) to certain (i.e., 100% correct). However, in speeded decision-making tasks, it is well known that participants can sometimes detect themselves making an error (Hester & Garavan, 2005) and even provide graded reports about these (Boldt & Yeung, 2015).



In such experiments where participants are asked to report their confidence on a scale that ranges from certainly wrong (i.e., 0% correct) to certainly correct (i.e., 100% correct), sometimes the relation between confidence RTs and confidence shows an inverted U-shape (i.e., similar to bow effects observed in absolute judgment tasks; Hollands & Dyre, 2000; Kvam et al., 2023). To account for such a pattern, we here propose a model in which both choice and confidence are implemented as a process of continuous accumulation until one of two opposing boundaries is reached. For convenience, we refer to this model as the flexible confidence boundary (FCB) model (Figure 1D).

In the current work, we compared the ability of these five models to accurately describe the stopping rule for confidence judgments by fitting them to data from two experiments (one with a binary and one with a 6-choice confidence report). To provide a strong experimental manipulation of the stopping rule, in both experiments, we manipulated the level of caution that participants used to provide their confidence reports. Specifically, just like how individuals can modulate their choice boundaries according to speed–accuracy trade-off (SAT) instructions (Ratcliff & McKoon, 2008), we similarly instructed participants to either provide very fast confidence judgments or think carefully about their level of confidence. Remarkably, although there are numerous studies that have investigated SATs in choice formation (for review, see Bogacz, Wagenmakers, et al., 2010), to the best of our knowledge it has yet to be investigated whether similar trade-offs can be observed in confidence formation, and if so, which of the above-mentioned computational models of confidence can best account for such trade-offs.

## Experiment 1

### Methods and Materials

#### *Preregistration and Code*

All hypotheses, sample sizes, exclusion criteria for participants, analyzed variables, the experimental design, and planned analyses were preregistered on the Open Science Framework registries (Herregods & Desender, 2021; <https://doi.org/10.17605/OSF.IO/Z2UCM>), unless specified as exploratory. Additionally, all code and data are made publicly available on GitHub (<https://github.com/StefHerregods/ConfidenceBounds>).

#### *Participants*

We decided a priori to test a minimum of 40 viable participants, in line with previous SAT research (Desender et al., 2022). Participant recruitment continued until this sample size was met after applying exclusion criteria. In total, 51 participants took part in Experiment 1 in return for course credit. From the total data set, one participant gave the same confidence rating in more than 95% of the trials, and 10 participants required too many training trials or did not complete the experiment in time. Data from these participants were excluded from further analyses. The final data set comprised 40 participants (36 female), with a mean age of 18.0 ( $SD = 0.6$ , range = 17–19). All participants had normal or corrected-to-normal vision and signed informed consent before their participation. The experiment was approved by the local ethics committee.

#### *Stimuli and Apparatus*

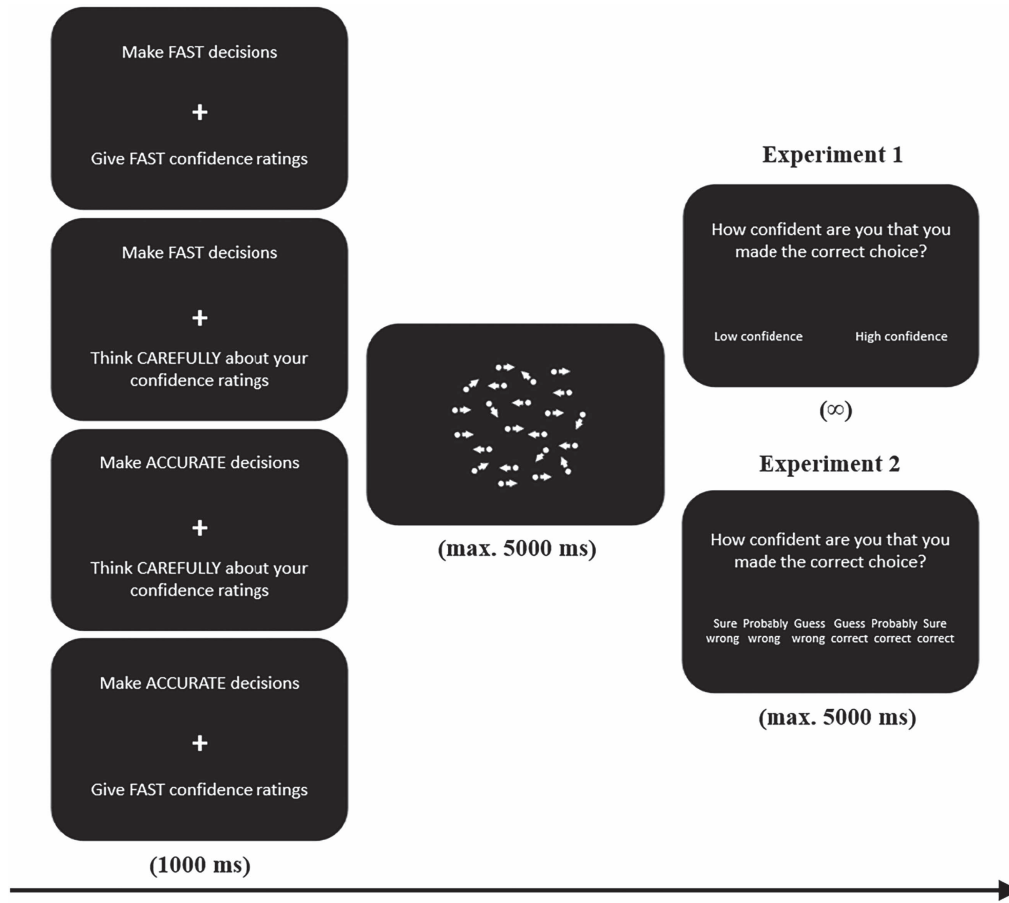
The experiment was programed using Python v3.6.6 and PsychoPy (Peirce et al., 2019). Participants completed the experiment on 24-inch liquid-crystal display screens using an AZERTY keyboard, with blue stickers indicating buttons used for confidence judgments and red stickers indicating decision-making buttons.

#### *Procedure*

Each experimental trial started with the display of a white fixation cross on a black background for 1 s (see Figure 2). Instructions regarding the speed–accuracy regime were shown above and below the fixation cross for decision-making and confidence judgments, respectively. Depending on the block, the instructions were to either “make fast decisions” or “make accurate decisions” and to “give fast confidence ratings” or “think carefully about your confidence ratings” for choices and confidence reports, respectively. For convenience, we will refer to both types of instructions as choice SAT and confidence SAT, respectively. Next, a dynamic random dot motion stimulus was presented until participants gave a response. If participants did not provide a response within 5 s, the message “Too slow, please respond faster” was shown on the screen. Motion coherence was controlled by the proportion of dots consistently moving toward the left versus the right side of the screen. During the main experiment, three levels of coherence were used (.1, .2, and .4). Participants were instructed to press the “c” or “n” key with the thumbs of their left and right hands to indicate whether they thought dots were moving toward the left or the right, respectively. If participants responded within 5 s, they were subsequently asked about their level of confidence. The text “How confident are you that you made the correct choice?” appeared on top of the screen, and participants pressed the “e” or the “u” key with their index fingers, mapped to high and low confidence, respectively (mapping counterbalanced across participants). Confidence judgments were transformed to numeric values, with “low confidence” as zero and “high confidence” as one.

The experiment started with three practice blocks (24 trials each). In Block 1, participants only made random dot motion decisions with a coherence of .5 for all trials. During this block, they received feedback about choice accuracy after each trial. Participants repeated Block 1 until achieving an average accuracy of 85% or more. Block 2 was identical except that the same three coherence levels as in the main phase were used (.1, .2, and .4). Participants repeated Block 2 until achieving an average accuracy of 60% or more. In Block 3, participants no longer received trial-by-trial feedback but instead were asked about their level of confidence after each trial. Afterward, participants took part in 12 blocks of 60 trials each. In each block, there was an equal number of coherent left and right dot motion trials and an equal occurrence of the three coherence levels. Finally, each block had specific instructions about the speed–accuracy regime for decision-making and confidence judgments. These instructions appeared both before each block and at the start of each trial (i.e., during the fixation cross). Speed–accuracy regime instructions were constant within a block but switched after each block. Each combination of instructions appeared three times, and the order of appearance was counterbalanced across participants using a Latin square. After each block, participants received feedback about their average accuracy, average RT, and average confidence RT of the preceding block.

**Figure 2**  
*Example of an Experimental Trial*



*Note.* During presentation of the fixation cross, participants received specific instructions regarding the speed–accuracy regime for choices (above fixation) and confidence (below fixation). These instructions were constant within a block, but switched each block. Next, participants made binary choices about random dot motion and afterward indicated their level of confidence on a 2-point scale (Experiment 1) or a 6-point scale (Experiment 2). Duration of the presentation of each screen is indicated between brackets. Max. = maximum.

## Statistical Analyses

RTs and confidence RTs on correct trials, accuracy, and confidence judgments on correct trials were analyzed using mixed effects models. All models included at least a random intercept per participant and all manipulations (choice SAT, confidence SAT, and coherence) and their interactions as fixed effects, unless otherwise specified. These models were then extended with random slopes in order of the biggest increase in Bayesian information criterion (BIC), until the addition of random slopes led to a nonsignificant increase in likelihood or until the random effects structure was too complex to be supported by the data (leading to an unstable fit). We used the lmer and glmer functions of the lme4 package (Bates et al., 2015) to fit the linear and generalized linear mixed models, respectively, in R (R Core Team, 2021). The calculation of  $p$  values is based on chi-square estimations using the Wald test from the car-package (Fox & Weinberg, 2019). Due to violations of the assumptions of normally distributed residuals and homoscedasticity, all RTs and confidence RTs were log-transformed and mean-centered. Furthermore, evidence

accumulation model performance was compared across three different model specifications using BIC values computed as explained in Solway and Botvinick (2015). Finally, the influence of the speed–accuracy manipulations on the estimated model parameters of the best performing model was examined using repeated measures analyses of variance (ANOVAs) and follow-up paired  $t$  tests, as implemented in the rstatix package (Kassambara, 2023).

## Model Specification

For the 2DSD, CCB, and FCB<sub>simple</sub> models, we simulated noisy evidence accumulation using a random walk approximation of the drift diffusion process (Tuerlinckx et al., 2001). A random walk process started at  $z \times a$ , with  $z$  being an unbiased starting point of .5, and continued to accumulate until the accumulated evidence reached 0 or  $a$  (corresponding to the height of the decision boundaries). At each time step  $\tau$ , the accumulated evidence was updated with  $\Delta$ , with the update rule shown in Equation (1):

$$\Delta = v \times \tau + \sigma \times \sqrt{\tau} \times N(0, 1), \quad (1)$$

with  $v$  indicating the drift rate,  $N$  indicating the standard normal distribution,  $\tau$  indicating precision, which was set to .001 in all simulations, and  $\sigma$  indicating within-trial noise, which was fixed to 1. Choice and RT were quantified at the moment of boundary crossing. We used an accuracy coding scheme, such that crossing the upper decision boundary equals a correct choice and crossing the lower decision boundary an incorrect choice. An additional time  $ter$  was added to predicted RTs to capture nondecision-related processes. After the accumulated evidence reached 0 or  $a$ , evidence continued to accumulate at each time step  $\tau$  with displacement  $\Delta_{post}$ , with the postdecision update rule shown in Equation (2):

$$\Delta_{post} = v_2 \times \tau + \sigma \sqrt{\tau} \times N(0, 1), \quad (2)$$

with  $v_2$  corresponding to the drift rate governing postdecisional processing. Allowing dissociations between drift rate and postdecisional drift rate is necessary to account for differences in metacognitive accuracy between participants (Desender et al., 2022). Postdecisional accumulation continued until a certain criterion was reached, and this stopping rule differed across the models we investigated.

For the first model, based on the 2DSD model of Pleskac and Bussemeyer (2010), evidence continues to accumulate postdecision for a variable interjudgment time, normally distributed with mean  $\tau_{ej}$  and standard deviation  $SD_{\tau}$ . Confidence judgments are then determined by the crossing of confidence criterion  $c + a$  for correct decision trials or  $-c$  for incorrect decision trials. More specifically, higher evidence than the decision criterion toward the chosen decision leads to a “high confidence” response, and less evidence to a “low confidence” response. Note that in the original 2DSD model proposed by Pleskac and Bussemeyer, across-trial variability in drift rate and starting point variability were also modeled, but we opted not to implement these parameters such that all the different models considered here are similar in their decision process and only differ in the postdecision stopping mechanisms.

In the second model, based on the CCB model of Moran et al. (2015), postdecision evidence accumulates until it reaches a single “stepwise” CCB (i.e., an evidence-based criterion) with the starting height being determined by parameter  $h$ . When the accumulator reaches the confidence boundary at height  $h$ , the model gives a high confidence response. After the collapse time  $\tau_{cp}$ , the criterion height is lowered with collapse height  $\Delta$ . In the variant with multiple discrete confidence levels, predicted model confidence lowers with each collapse of the confidence boundary. Note that in the current experiment, with only two levels of confidence, the model simplifies to a single step such that the confidence boundary only takes two heights (i.e.,  $h$  corresponding to high confidence and  $h - \infty$  corresponding to low confidence).

Third, for the FCB<sub>simple</sub> model, two confidence boundaries that both slowly collapse demarcate the area of evidence accumulation. The height of the confidence boundaries is given by Equation (3):

$$\begin{aligned} \text{if}(\text{choice} = a) \text{ upper confidence boundary} &= a + a_2 - u \times t_2 \\ \text{if}(\text{choice} = a) \text{ lower confidence boundary} &= a - a_2 + u \times t_2 \\ \text{if}(\text{choice} = 0) \text{ upper confidence boundary} &= 0 - a_2 + u \times t_2 \\ \text{if}(\text{choice} = 0) \text{ lower confidence boundary} &= 0 + a_2 - u \times t_2, \end{aligned} \quad (3)$$

with  $a_2$  corresponding to the height of the confidence boundaries,  $u$  indicating the amount of (linear) urgency, and  $t_2$  indicating

postdecision time. Given that Experiment 1 only has two levels of confidence (high vs. low), confidence here fully coincides with the boundary that was reached (i.e., high vs. low confidence when reaching the upper vs. lower confidence boundary; see Figure 1D).

Finally, the Van Zandt model was implemented as two independent evidence accumulators that race toward a single decision boundary and then continue to race toward a single confidence boundary with heights  $a$  and  $a_2$ , respectively. The accumulators were defined as specified for the other models but started at zero and had a separate drift rate for the correct choice accumulator  $v_{\text{correct}}$  and the error choice accumulator  $v_{\text{error}}$ . RTs and decision accuracy were defined by the first accumulator to cross the decision boundary. Afterward, evidence accumulation continued for both accumulators with the same drift rate until one of them crossed the confidence boundary, at which point a confidence judgment was made. If the evidence at this point was higher for the first accumulator that reached the decision boundary, a high confidence response was given. Otherwise, the simulation ended with a low confidence response.

Notably, in all models, an additional time  $ter_2$  was added to predicted confidence RTs to capture non-confidence-related processes (e.g., pressing a confidence button). In contrast to  $ter$ , which is by definition always positive, we also allowed  $ter_2$  to take negative values to account for the possibility that postdecision evidence accumulation already starts before an overt response has been made (e.g., during the motor execution of the first response). An overview of the free parameters used in all models can be found in the Supplemental Materials.

Note that we did not implement the optional stopping model of Pleskac and Bussemeyer (2010) to fit the data of Experiment 1. This model assumes that the outermost confidence thresholds are fully absorbing, that is, the probability of stopping when reaching these thresholds is one, so in an experiment with only two levels of confidence, there is no optional stopping in this model, leading to an unfair representation of that model variant.

## Parameter Estimation and Model Fit

We estimated best fitting parameters separately for each participant and within each condition by jointly minimizing three error functions based on quantile and proportion optimization of the RT and confidence RT distributions and of the confidence judgment proportions. This fitting approach allowed us to evaluate both whether the parameters of interest (i.e., choice and confidence boundaries) varied between conditions and to test whether other parameters (drift rates and nondecision times) did not. For completeness, Supplemental Figure S10 shows the fit of the winning model (FCB<sub>simple</sub>) when jointly fitting the data from the different SAT conditions, with all parameters shared across SAT conditions except for the choice and confidence boundaries and confidence urgency, which were allowed to vary as a function of SAT condition. Quantiles were computed separately for correct and error trials in both observed and simulated data for two types of quantiles, (a) decision RT and (b) confidence RT (RT<sub>conf</sub>), and the confidence judgment proportions. The resulting error functions are shown in Equations 4–6:

$$\text{MSE}_{\text{RT}} = \frac{\sum (\text{oRT}_{i,q} - \text{sRT}_{i,q})^2}{\text{sRT}_{i,q}}, \quad (4)$$

$$\text{MSE}_{\text{RTconf}} = \frac{\sum (\text{oRTconf}_{i,q} - \text{sRTconf}_{i,q})^2}{\text{sRTconf}_{i,q}}, \quad (5)$$

$$\text{MSE}_{\text{confidence proportions}} = \frac{\sum (\text{oPROPconf}_{i,cj} - \text{sPROPconf}_{i,cj})^2}{\text{sPROPconf}_{i,cj}}, \quad (6)$$

with *oRT* and *sRT* referring to observed and simulated RT proportions, and *oRTconf* and *sRTconf* to observed and simulated confidence RT proportions, across multiple quantiles (*q*) (.1, .3, .5, .7, and .9), for correct and error trials (*i*) separately. Similarly, *oPROPconf* and *sPROPconf* refer to the observed and simulated confidence judgment proportions across all levels of confidence (*cj*) (0, 1), also for correct and error trials (*i*) separately. We minimized the sum of the error functions given by Equations (4)–(6) using differential evolution optimization to optimize all parameters. The evolutionary algorithm was operationalized by means of the DEoptim package, with the number of iterations set to 500 and the number of simulated trials set to 2000 (Mullen et al., 2011). Model fitting was done separately per participant. To assess model fits, we simulated choices, RTs, confidence judgments, and confidence RTs from the estimated parameters.

Finally, to compare fit across models, we computed BIC values for each fit separately, according to Equation (7):

$$\text{BIC} = n_{\text{par}} \times \ln(n_{\text{trials}}) + n_{\text{trials}} \times \ln\left(\frac{\text{MSE}}{n_{\text{trials}}}\right), \quad (7)$$

with  $n_{\text{par}}$  and  $n_{\text{trials}}$  referring to the number of free parameters and the number of observed trials, respectively. MSE (Mean Squared Error) indicates the sum of  $\text{MSE}_{\text{choice RT}}$ ,  $\text{MSE}_{\text{confidence RT}}$ , and  $\text{MSE}_{\text{confidence proportions}}$ .

## Results

### Behavioral Analysis

Trials with RTs below .2 s were excluded from the data set (00.40%) (Moran et al., 2015). In addition, confidence RTs slower than 5 s were excluded (00.10%; note that choice RTs slower than 5 s were excluded by design). Next, we report a set of analyses testing how RTs, confidence RTs, accuracy, and confidence judgments were influenced by motion coherence (3 levels: .1, .2, and .4), choice SAT (2 levels: fast vs. accurate), and confidence SAT (2 levels: fast vs. careful). For a summary of these results, see Supplemental Table S6.

For RTs on correct trials (shown in Figure 3A), as expected, we found a significant effect of choice SAT instructions,  $\chi^2(1) = 68.87, p < .001$ , but not of confidence SAT instructions,  $\chi^2(1) = 0.56, p = .455$ . Choice RTs were shorter when participants were instructed to respond fast ( $M = 0.91$  s,  $SD = 0.51$ ) versus accurate ( $M = 1.33$  s,  $SD = 0.79$ ). Also, the main effect of motion coherence was significant,  $\chi^2(2) = 687.25, p < .001$ , reflecting shorter RTs with increasing motion coherence. Additionally, we found a significant interaction between the choice SAT and confidence SAT,  $\chi^2(1) = 15.35, p < .001$ , reflecting that the choice SAT effect was more expressed when participants were instructed to provide accurate versus careful confidence ratings. There was also a significant interaction between choice SAT and coherence,  $\chi^2(1) = 43.24, p < .001$ , reflecting that the choice SAT effect was slightly larger for low coherence trials. All other

effects were not significant,  $ps > .525$ . For accuracy, we likewise found a significant effect of the choice SAT,  $\chi^2(1) = 8.25, p = .004$ , and coherence,  $\chi^2(2) = 165.58, p < .001$ , but not of the confidence SAT,  $\chi^2(1) = 0.63, p = .429$ . As shown in Figure 3B, participants responded more correctly when instructed to be accurate ( $M = 72.28\%$ ,  $SD = 0.45$ ) compared to when instructed to be fast ( $M = 70.78\%$ ,  $SD = 0.45$ ), and accuracy increased with motion coherence. All other effects were not significant,  $ps > .071$ .

For confidence RTs on correct trials, we found significant effects of the confidence SAT instructions,  $\chi^2(1) = 77.06, p < .001$ , and coherence,  $\chi^2(2) = 14.29, p = .001$ . As expected, choice SAT instructions did not influence confidence RTs,  $\chi^2(1) = 0.42, p = .518$ . As can be seen in Figure 3C, confidence RTs were faster when participants were instructed to make fast ( $M = 0.30$  s,  $SD = 0.30$ ) versus careful ( $M = 0.73$  s,  $SD = 0.66$ ) confidence judgments. Additionally, we found a significant interaction between choice SAT and confidence SAT,  $\chi^2(1) = 6.28, p = .012$ , reflecting a small spillover from choice SAT into confidence RTs (mostly visible in the “accurate” condition). All other effects were not significant,  $ps > .464$ . Finally, for confidence judgments (see Figure 3D), we observed a significant main effect of coherence,  $\chi^2(2) = 120.71, p < .001$ , reflecting that confidence increased with the proportion of motion coherence. There were no significant main effects of choice SAT,  $\chi^2(1) = 1.23, p = .267$ , nor confidence SAT,  $\chi^2(1) = 0.27, p = .606$ . There was only a small but significant interaction between choice SAT and confidence SAT,  $\chi^2(1) = 4.89, p = .027$ , reflecting that participants more often reported high confidence for fast ( $M = .63, SD = 0.48$ ) than for accurate ( $M = .60, SD = 0.49$ ) choices in the fast confidence condition, whereas there was a smaller difference in the careful confidence condition ( $M = .61, SD = 0.49$  vs.  $M = .60, SD = 0.49$ , respectively). Finally, there was an interaction between choice SAT and coherence,  $\chi^2(2) = 9.97, p = .007$ , reflecting that the relation between confidence and coherence was slightly stronger in the accurate compared to the fast choice condition. All other effects were not significant,  $ps > .160$ .

Given that there was no effect of the SAT manipulations on average confidence, we additionally examined whether there was a difference in confidence resolution (i.e., the relation between confidence and accuracy). To do so, we ran an exploratory type II receiver-operating characteristic curve analysis separately for each condition (ignoring coherence). A 2-way ANOVA on these estimates showed a main effect of confidence SAT,  $F(1,39) = 15.42, p < .001$ , but not from choice SAT,  $p = .599$ , nor was there an interaction,  $p = .491$ . As can be seen in Figure 4A, the relation between confidence and accuracy (expressed in area under the curve [AUC] units) was higher when participants were instructed to make deliberate versus fast confidence ratings. Thus, although confidence did not strongly change on average, there was clear evidence that the precision with which participants distinguished correct from incorrect decisions was improved when deliberately computing confidence.

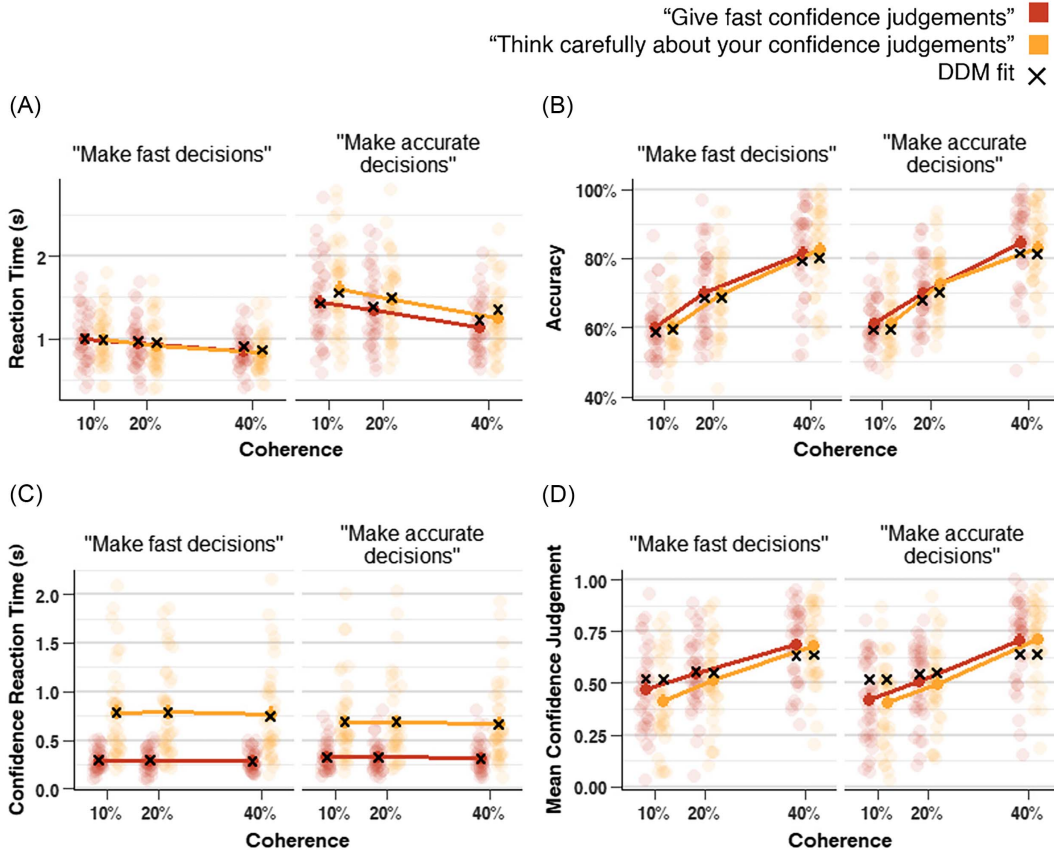
### Modeling the Stopping Rule for (SATs in) Confidence

Next, we compared the ability of four prominent models of confidence to explain the stopping rule for postdecision evidence accumulation. Our modeling framework departed from the classical DDM, a popular evidence accumulation model that accounts well for choices and RTs in perceptual decisions (Ratcliff & McKoon, 2008). To also account for confidence within the DDM, we allow the evidence to



**Figure 3**

*The Influence of Choice SAT and Confidence SAT on Reaction Times (A), Accuracy (B), Confidence RTs (C), and Confidence (D) for Experiment 1*



*Note.* As expected, when participants were instructed to make fast versus accurate choices, this led to fast versus slow choice RTs (A) and, to a lesser extent, to less and more accurate choices (B), respectively. When participants were instructed to make fast versus deliberate confidence judgments, this led to fast versus slow confidence RTs (C). There was no main effect on confidence judgments (D). Error bars correspond to the standard error of the means (SEM), transparent dots indicate means of individual participants, and black crosses indicate fits from the FCB<sub>simple</sub> model. DDM = drift diffusion model; FCB = flexible confidence boundary; RT = reaction time; SAT = speed-accuracy trade-off. See the online article for the color version of this figure.

accumulate after it has reached a threshold (postdecision evidence accumulation; Pleskac & Bussemeyer, 2010). Critically, we tested different possibilities as to how this process of postdecision evidence accumulation should stop to account for the observed data. In the 2DSD, postdecision evidence accumulation terminates after a fixed amount of time has passed. In the CCB model, postdecision evidence accumulation terminates after it reaches a slowly collapsing confidence boundary. In the FCB<sub>simple</sub> model, postdecision evidence accumulation terminates once it reaches one of two slowly collapsing confidence boundaries. Finally, according to the model proposed in (Van Zandt & Maldonado-Molina, 2004), confidence is determined by the difference in evidence between two independent racing accumulators once one of them reaches a (flat) confidence boundary.

### Model Recovery

After fitting the models described above to the data, separately for each participant and condition, we first performed model recovery to assess whether it is indeed possible to distinguish these models

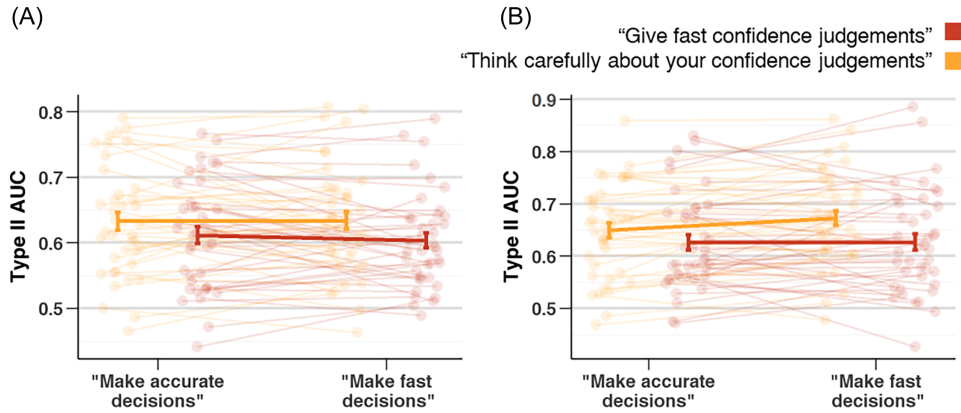
based on the empirical data. We first simulated data from each model using the parameters estimated from the experimental data and then refitted all models to the simulated data sets using the same fitting procedure as for the original data. As shown in Table 1, on average, each model yielded the lowest BIC when fit by the generating model. When looking at individual data sets, we also found that most data sets were best fitted by their generating model, with one notable exception. For the CCB model, the majority of individual data sets were best fit (i.e., had the lowest BIC) by the FCB<sub>simple</sub> model, indicating potential misidentification at the individual level despite correct average-level recovery. Note that this is not surprising, given that in a task with only two levels of confidence, both models can behave very similarly.

### Model Comparison

Next, we summarized model averages of MSE and BIC values in Table 2. As shown, the FCB<sub>simple</sub> model returns the lowest mean BIC ( $\Delta\text{BICs} > 80$ ) and was the best fitting model for most of the

**Figure 4**

Confidence Resolution in (A) Experiment 1 and (B) Experiment 2, Expressed as Type II AUC



*Note.* Although confidence SAT instructions did not have a clear effect on average confidence, we did observe a clear effect on confidence resolution, which was not the case for choice SATs. Same conventions as in Figure 3. AUC = area under the curve; SAT = speed-accuracy trade-off. See the online article for the color version of this figure.

participants (71.25%). Post hoc pairwise contrasts, accounting for participants and different conditions, showed that the FCB<sub>simple</sub> model produced significantly lower BIC values than the CCB model,  $t(636) = -92.2, p < .001$ , the Van Zandt race model,  $t(636) = -80.0, p = .002$ , and the 2DSD model,  $t(636) = -162.9, p < .001$ . This advantage was primarily driven by the FCB<sub>simple</sub> model's better fit to both confidence RTs and confidence proportions. Notably, although 2DSD performs well in capturing confidence proportions (as shown by the confidence RT MSE), it completely fails to capture the confidence RT patterns found in the empirical data. The CCB model and the Van Zandt model do a relatively better job compared to the 2DSD in this respect, suggesting that the stopping rule of confidence judgments is best explained by an accumulation-to-bound mechanism and not uniquely by a time-based stopping criterion. Finally, note that

the CCB and the Van Zandt model perform worse relative to the FCB<sub>simple</sub> model, mostly because the CCB model underperformed both in terms of confidence RT and confidence proportions, and the Van Zandt model captured the observed confidence RTs well but performed worse in terms of confidence proportions.

To get better insights into the strengths and weaknesses of each model, we also visualized qualitative signatures regarding confidence RTs and confidence resolution of each model type (Palminteri et al., 2017). As shown in Figure 5A, the race model based on Van Zandt and Maldonado-Molina (2004) failed to capture the accuracy versus confidence RT relation observed in the behavioral data. Furthermore, the CCB model could not account for the observed mean and variance in confidence as a function of confidence RT (Figure 5B and 5C). More specifically, the CCB model predicts that fast confidence responses always lead to a high confidence judgment, and vice versa. Therefore, the model's limited flexibility makes it difficult to simultaneously capture both confidence RTs and confidence judgment proportions. Last, as shown in Figure 5D, the Van Zandt race model predicts higher confidence resolution than observed.

Finally, we visualized the fit of the winning FCB model. Figure 5E and 5F shows that, on average, the predicted confidence RT distributions align well with the observed distributions. Furthermore, the observed patterns of mean RT, confidence RT, accuracy, and confidence judgments are captured by the model (see Figure 3). A more in-depth analysis of the model fit can be found in Supplemental Figure S1.

**Table 1**  
*Model Recovery for Experiment 1*

Generating model	Fitted model			
	FCB <sub>simple</sub>	CCB	Van Zandt	2DSD
Relative BIC				
FCB <sub>simple</sub>	<b>0</b>	837.68	778.18	595.16
CCB	65.20	<b>0</b>	51.25	137.13
Van Zandt	539.62	993.80	<b>0</b>	1120.28
2DSD	248.36	1386.08	944.94	<b>0</b>
No. of data sets best fitted				
FCB <sub>simple</sub>	<b>150</b>	1	2	7
CCB	<b>105</b>	50	5	0
Van Zandt	0	0	<b>160</b>	0
2DSD	12	1	6	<b>141</b>

*Note.* The table includes (a) BICs of the model recovery and (b) the number of data sets best fitted by each model type for every generating model. BIC values are relative with respect to the best fitting model for each generating model. Lowest BIC values are indicated in bold. The model type fitting most data sets best for each generating model is indicated in bold. FCB = flexible confidence boundary; CCB = collapsing confidence boundary model; 2DSD = two-stage dynamic signal detection theory model; BIC = Bayesian information criterion.

### Model Parameters

Having established that the FCB<sub>simple</sub> model provides a good fit to the experimental data and outperforms the competing models under consideration, we next turn toward the estimated parameters. We hypothesized that SAT instructions for choices selectively affected choice boundaries, leaving confidence boundaries unaffected. Likewise, we expected SAT instructions about confidence to selectively affect confidence boundaries, leaving choice boundaries unaffected. These observations would support the hypothesis that indeed the stopping rule for both choices and choice confidence are



**Table 2**  
*Experiment 1 Model Comparison*

Model type	Log RT MSE	Log confidence RT MSE	Log confidence proportions MSE	Free parameters	Mean BIC	Best fit (%)
2DSD	0.47	1.70	-1.26	10	-448.17	13.13
CCB	0.37	1.40	-0.69	9	-518.80	10
FCB <sub>simple</sub>	<b>0.52</b>	<b>0.32</b>	<b>-1.05</b>	<b>9</b>	<b>-611.05</b>	<b>71.25</b>
Van Zandt	0.54	0.67	0.74	10	-531.01	5.63

*Note.* Results from a model comparison show that the stopping rule for confidence is best accounted for by an accumulation-to-bound mechanism, and specifically by a model including postdecision evidence accumulation that terminates once it reaches one of two opposing slowly collapsing confidence boundaries (i.e., the FCB<sub>simple</sub> model, indicated in bold). Best fit shows the number of participants for which the corresponding model provides the lowest BIC. RT = reaction time; FCB = flexible confidence boundary; CCB = collapsing confidence boundary model; 2DSD = two-stage dynamic signal detection theory model; BIC = Bayesian information criterion; MSE = mean squared error.

under strategic control. To investigate the interpretability of the model parameters, we first performed a parameter recovery analysis. All parameters recovered well ( $rs > .81$ ), with the exception of the confidence boundary urgency parameter ( $u$ ), which was therefore excluded from the analyses. For ease of reading, below we only report the analyses of a priori interest; the entire set of analyses can be found in the Supplemental Figures S3 and S4.

We used a repeated measures ANOVA to examine the influence of choice SAT (fast vs. accurate), confidence SAT (fast vs. careful), and their interaction on estimated decision boundaries. As expected, we found a strong and significant effect of the choice SAT,  $F(1, 39) = 70.36, p < .001, \eta_p^2 = .64$ , but no effect of confidence SAT,  $F(1, 39) = 1.50, p = .229, \eta_p^2 = .04$ , nor an interaction between both,  $F(1, 39) = 1.98, p = .167, \eta_p^2 = .05$ . As can be seen in Figure 6A, when participants were asked to make fast decisions, the separation between both choice boundaries was smaller ( $M = 1.55, SD = 0.37$ ) compared to when they were asked to make accurate decisions ( $M = 1.99, SD = 0.43$ ).

The same analysis on the estimated confidence boundary separation revealed a significant main effect of confidence SAT,  $F(1, 39) = 26.15, p < .001, \eta_p^2 = .40$ , but also of choice SAT,  $F(1, 39) = 8.17, p = .007, \eta_p^2 = .17$ , and the interaction between both types of instruction,  $F(1, 39) = 6.22, p = .017, \eta_p^2 = .14$ . Follow-up paired  $t$  tests showed that the effect of the confidence RT instructions on confidence boundary separation was significant both when decision SAT was to be accurate,  $t(39) = -4.31, p < .001$ , and when decision SAT was to be fast,  $t(39) = -5.14, p < .001$ . As shown in Figure 6B, confidence boundary separation was smaller when participants were asked to make fast confidence judgments ( $M = 1.06, SD = 0.25$ ) than when participants were asked to make accurate confidence judgments ( $M = 2.13, SD = 1.46$ ). In sum, choice boundaries were modulated by confidence SAT instructions but also by decision SAT instructions and the interaction between both.

Finally, as a sanity check, we confirmed that estimated drift rates scaled with motion coherence using a repeated measures ANOVA,  $F(1.28, 49.73) = 173.25, p < .001, \eta_p^2 = .82$ . The other parameters were not allowed to vary by the instruction conditions; their mean estimates can be found in Table 3.

## Interim Summary

In Experiment 1, participants were instructed to make fast or accurate decisions and to make fast or careful confidence judgments,

depending on the block they were in. At the behavioral level, we observed that participants were indeed able to selectively speed up choices or confidence judgments when instructed to do so. Next, we fitted four different postdecisional evidence accumulation models to explain choices, reaction times, confidence, and confidence RTs. Comparison of model fit showed that a DDM with postdecision evidence accumulation until reaching one of two slowly collapsing confidence boundaries, that is, the newly proposed FCB model, explained the data best. Notably, according to the FCB model, the mechanism underlying SATs was a modulation of the decision boundary for choices and a modulation of the confidence boundary for confidence. Thus, these findings suggest that the stopping rule for confidence judgments is best explained as an accumulation-to-bound mechanism, and that just like the decision boundary for choices, these confidence boundaries are under voluntary strategic control.

One limitation of Experiment 1 is that participants were only allowed to give binary confidence ratings (high or low). This design choice made for an easy modeling approach because it allows to directly map high and low confidence onto the upper and lower confidence boundaries in the FCB model, respectively. It is well known, however, that humans can provide more fine-grained estimates of their performance. Thus, this begs the question of whether the FCB model is also the preferred model in a task with a fine-grained confidence scale. To this end, in Experiment 2, we replicated Experiment 1, but now using a more fine-grained 6-choice confidence scale.

## Experiment 2

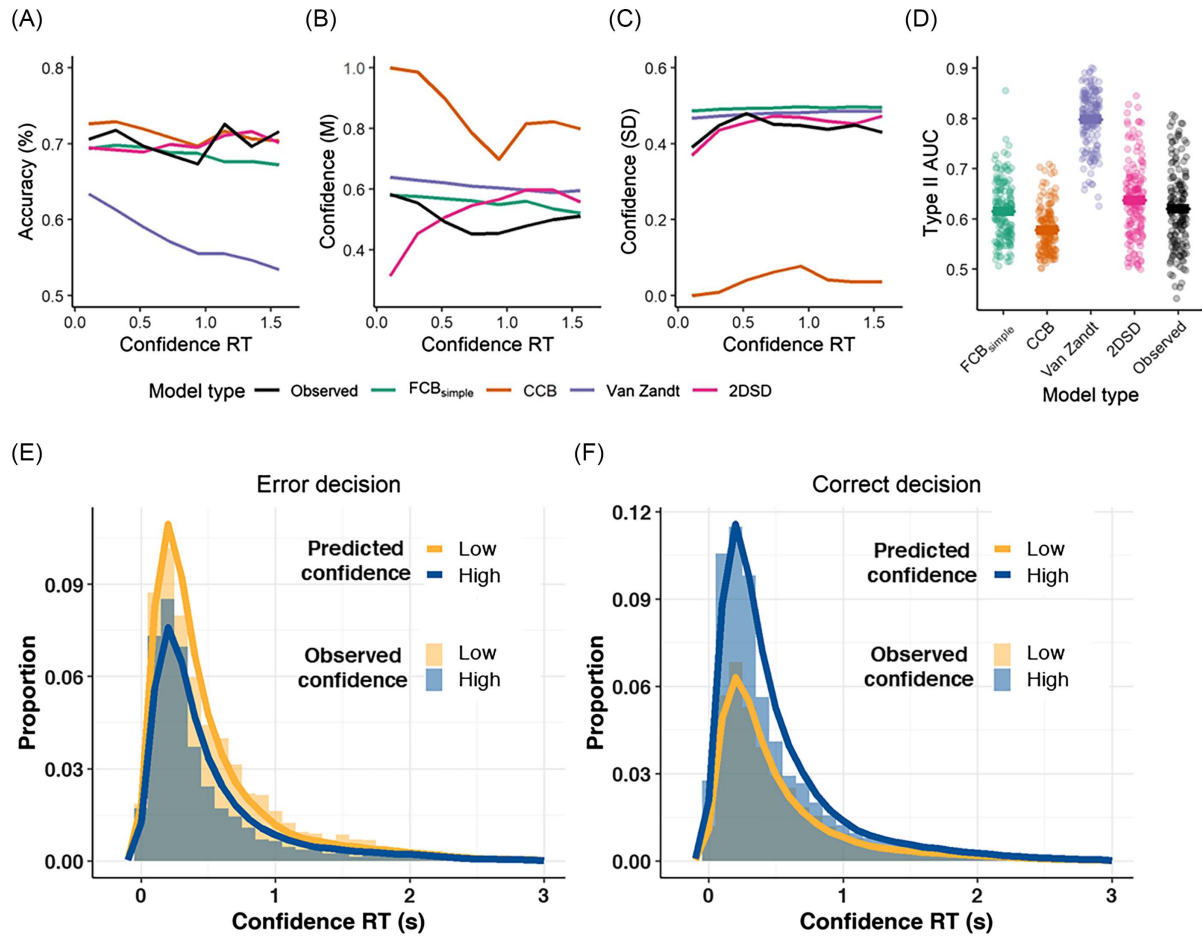
### Methods and Materials

#### Preregistration and Code

The preregistration of this experiment can be found on the Open Science Framework registries (Herregods & Desender, 2021; <https://doi.org/10.17605/OSF.IO/VYH4K>), and all code and data can be found on GitHub (<https://github.com/StefHerregods/ConfidenceBounds>).

#### Participants

A total of 54 participants participated in Experiment 2. Requirements and recruitment were identical to Experiment 1,

**Figure 5***Qualitative Model Fits of Experiment 1*

*Note.* (A)–(D) Accuracy (A), average confidence (B), and variability in confidence (C) as a function of confidence RT, and type II ROC (D) shown for behavioral data, as well as with fitted model predictions. For visualization purposes, confidence RT was divided into 30 equal-sized bins, after excluding the lowest and highest 5% of observed confidence RTs. Note that line graphs do not show individual model patterns but rather show aggregated simulations across participants after binning into 30 equal-sized bins. As a consequence, for example, the seemingly negative-going slope of the CCB model in panel B actually reflects the average of many individuals showing a step-function from high (1) to low (0) confidence at a specific confidence RT. See Supplemental Figure S7 for a version of these plots based on average parameter estimates. (D) Each dot corresponds to data from a single participant. (E)–(F) Predicted and observed confidence RTs separately for high and low confidence trials after an incorrect (E) and a correct (F) decision by the FCB<sub>simple</sub> model. FCB = flexible confidence boundary; CCB = collapsing confidence boundary model; 2DSD = two-stage dynamic signal detection theory model; RT = reaction time; ROC = receiver-operating characteristic curve; AUC = area under the curve. See the online article for the color version of this figure.

with the additional criterion of not having participated in Experiment 1. Data from six participants were removed for not having an accuracy above chance level (as assessed by a binomial test), and four participants for requiring more than seven training blocks. Finally, four participants did not finish the experiment in time. The final sample comprised 40 participants (33 female), with a mean age of 18.5 ( $SD = 1.3$ , range = 17–24).

### Stimuli and Apparatus

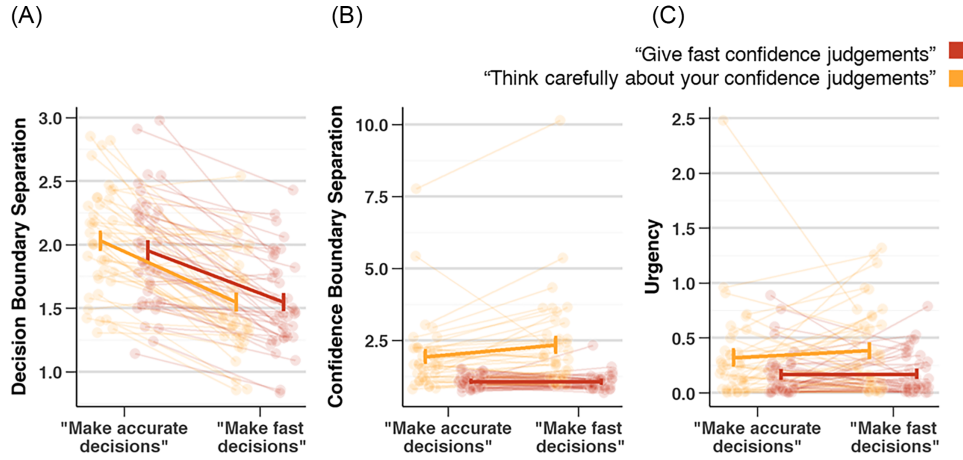
Experiment 2 used the same apparatus and stimuli as in Experiment 1.

### Procedure

The experiment was identical to Experiment 1, except for the following two exceptions: First, instead of a binary confidence rating, participants could choose between six options: “sure wrong,” “probably wrong,” “guess wrong,” “guess correct,” “probably correct,” and “sure correct,” using the “1,” “2,” “3,” “8,” “9,” and “0” keys on top of the keyboard. These six options were mapped onto a 1–6 confidence scale, with the direction of the scale (low-to-high vs. high-to-low) counterbalanced across participants. Second, a time limit of 5 s was imposed on indicating confidence judgments, equal to the time limit during decision making. If a participant did not respond

**Figure 6**

*Influence of Choice SAT and Confidence SAT on Decision Boundaries and Confidence Boundaries in Experiment 1*



*Note.* Instructing participants to make fast versus accurate choices influenced estimated decision boundaries and confidence boundary separation (A) and (B), but not confidence boundary urgency (C). Instructing participants to provide fast versus careful confidence ratings influenced estimated confidence boundary separation (B) and urgency (C), but did not affect decision boundary separation (A). Same conventions as in Figure 3. SAT = speed–accuracy trade-off. See the online article for the color version of this figure.

within this limit, they were instructed to respond faster in future trials with the following text: “Too slow. ... Please respond faster.”

### Model Specification and Fit

Given that participants indicated confidence on a 6-point scale, we adapted the models such that they could produce confidence on a

similar scale. The 2DSD model now included five confidence criteria ( $c_1$  through  $c_5$ ; in comparison to only one confidence criterion in Experiment 1), which mapped postdecision evidence onto a 6-point confidence scale. The CCB model, based on Moran et al. (2015), continued the “staircase-like” pattern corresponding to a discrete collapse of a single confidence boundary. As implemented in Moran et al., the collapse time parameter  $\tau_{cj}$  was halved for each

**Table 3**

*Mean (Standard Deviations) Estimates for the Parameters of the Best Fitting FCB Models*

Parameter	AA	AF	FA	FF
<b>Experiment 1</b>				
$a$	2.03 (0.42)	1.95 (0.44)	1.55 (0.39)	1.55 (0.36)
$v_1$ (coherence = 0.1)	0.19 (0.13)	0.20 (0.10)	0.25 (0.19)	0.23 (0.17)
$v_2$ (coherence = 0.2)	0.43 (0.20)	0.39 (0.21)	0.51 (0.31)	0.50 (0.30)
$v_3$ (coherence = 0.4)	0.82 (0.42)	0.84 (0.39)	1.02 (0.51)	0.96 (0.48)
$ter$	0.47 (0.24)	0.41 (0.16)	0.34 (0.10)	0.36 (0.13)
$ter_2$	0.07 (0.25)	0.05 (0.07)	0.01 (0.31)	0.05 (0.08)
$v_{ratio}$ ( $= v_2/v_1$ )	0.85 (0.43)	1.17 (0.69)	0.64 (0.34)	1.01 (0.60)
$a_2$	1.93 (1.26)	1.07 (0.22)	2.33 (1.63)	1.05 (0.28)
$u$	0.32 (0.43)	0.17 (0.21)	0.38 (0.38)	0.17 (0.19)
<b>Experiment 2</b>				
$a$	2.17 (0.44)	1.90 (0.39)	1.60 (0.36)	1.46 (0.31)
$v_1$ (coherence = 0.1)	0.22 (0.13)	0.24 (0.16)	0.25 (0.18)	0.27 (0.17)
$v_2$ (coherence = 0.1)	0.42 (0.22)	0.40 (0.27)	0.52 (0.28)	0.50 (0.31)
$v_3$ (coherence = 0.1)	0.89 (0.50)	0.90 (0.49)	1.23 (0.65)	1.07 (0.59)
$Ter$	0.47 (0.18)	0.45 (0.15)	0.38 (0.12)	0.39 (0.10)
$ter_2$	−0.22 (0.28)	−0.14 (0.22)	−0.25 (0.32)	−0.10 (0.18)
$v_{ratio}$ ( $= v_2/v_1$ )	0.56 (0.29)	0.59 (0.32)	0.49 (0.21)	0.57 (0.28)
$a_2$	5.34 (1.58)	4.27 (1.51)	5.37 (2.04)	3.81 (1.21)
$z_2$	0.62 (0.11)	0.66 (0.10)	0.60 (0.10)	0.62 (0.11)
$u_{upper}$	1.93 (0.96)	2.71 (1.28)	1.65 (1.03)	2.52 (1.12)
$u_{lower}$	0.62 (0.70)	0.58 (0.72)	0.62 (0.66)	0.46 (0.61)

*Note.* AA, AF, FA, and FF refer to speed–accuracy trade-off instructions to be accurate/cautious (A) or fast (F), with the first index referring to the decision and the second index referring to confidence. A schematic overview of the FCB<sub>full</sub> parameters of Experiment 2 can be found in Supplemental Figure S9. FCB = flexible confidence boundary.

successive criterion, and the model produced confidence on a 6-point scale. For the FCB model, we changed the implementation such that confidence no longer corresponds to the confidence boundary that was reached (as in Experiment 1). Instead, confidence now depends on the level of accumulated evidence when reaching the confidence boundary. We evenly divided the space in between the two confidence boundaries into six levels, and the model produced a level of confidence between 1 and 6 depending on the state of the accumulated evidence. Note that for Experiment 2, we implemented two variants of the FCB model. In addition to the FCB simple model used for Experiment 1, we also included FCB<sub>full</sub>, in which we separately estimated urgency for the upper ( $u_{\text{upper}}$ ) and lower ( $u_{\text{lower}}$ ) confidence boundary, and in which we allowed the starting point bias for confidence to be freely estimated ( $z_2$ ). The development of FCB<sub>full</sub> was done in order to allow the model to account for various relationships between confidence judgments and confidence RTs (discussed in more detail below). Fits of FCB<sub>full</sub> when jointly fitting the data from the different SAT conditions with all parameters shared across SAT conditions, except for the choice and confidence boundaries and confidence urgency, can be found in Supplemental Figure S11. The Van Zandt race model was adapted to a 6-point confidence scale by converting the postdecisional evidence difference between the two accumulators (evidence in favor of the chosen option—evidence in favor of the other option) into six categorical values based on the height of the confidence boundary,  $a_2$ . Specifically, the range between  $-a_2$  and  $a_2$  was evenly distributed into six categories, transforming evidence differences to confidence judgments between 1 and 6.

For Experiment 2, we also fitted a model based on the optional stopping model of Pleskac and Busemeyer (2010). In this model, postdecision evidence continued to accumulate until it reached one of six evidence thresholds, with the height of these thresholds given by  $cj_1$  to  $cj_6$  and each threshold corresponding to one of the possible confidence ratings. The probability of terminating the accumulation process and giving a confidence judgment when crossing a threshold was set to one for the outermost thresholds,  $cj_1$  and  $cj_6$ , and freely estimated for the other thresholds ( $p_{cj2}$  to  $p_{cj5}$ ).

## Results

### Behavioral Analysis: Mixed Effects Modeling

Data were analyzed in the same way as described in Experiment 1. Trials with a decision time of less than 0.2 s were excluded (0.30%). For a summary of these results, see Supplemental Table S7. A mixed effects model on choice RTs on correct trials showed a significant effect of choice SAT,  $\chi^2(1) = 57.91, p < .001$ , and coherence,  $\chi^2(2) = 956.89, p < .001$ . Unexpectedly, there was also a significant effect of confidence SAT,  $\chi^2(1) = 34.02, p < .001$ . Additionally, we found significant interactions between the choice SAT and confidence SAT,  $\chi^2(1) = 8.72, p = .003$ , between coherence and choice SAT,  $\chi^2(2) = 21.80, p < .001$ , and between coherence and confidence SAT,  $\chi^2(2) = 12.10, p = .002$ . The three-way interaction between choice SAT, confidence SAT, and coherence was not significant,  $\chi^2(2) = 0.65, p = .722$ . As can be seen in Figure 7A, choice RTs were shorter when participants were instructed to respond fast ( $M = 0.93$  s,  $SD = 0.51$ ) versus accurate ( $M = 1.33$  s,  $SD = 0.76$ ); however, the effect was not as selective as in Experiment 1, because choice RTs were also shorter when participants were instructed to provide fast ( $M = 1.07$  s,

$SD = 0.62$ ) versus careful confidence ratings ( $M = 1.20$  s,  $SD = 0.72$ ). The same analysis on accuracy likewise showed significant main effects of choice SAT,  $\chi^2(1) = 6.05, p = .014$ , confidence SAT,  $\chi^2(1) = 4.76, p = .029$ , and coherence  $\chi^2(2) = 1202.14, p < .001$  (see Figure 7B). Accuracy was lower when participants were instructed to make fast ( $M = 71\%$ ,  $SD = 0.45$ ) compared to accurate choices ( $M = 73\%$ ,  $SD = 0.44$ ), and likewise when participants were instructed to make fast ( $M = 72\%$ ,  $SD = 0.45$ ) versus careful confidence ratings ( $M = 73\%$ ,  $SD = 0.44$ ). All other effects were not significant,  $ps > .257$ .

The same analysis on confidence RTs on correct trials showed significant main effects of confidence SAT,  $\chi^2(1) = 85.62, p < .001$ , and coherence,  $\chi^2(2) = 71.12, p < .001$ . Unexpectedly, there was also a significant main effect of choice SAT,  $\chi^2(1) = 9.64, p = .002$ . Finally, the interaction between the confidence SAT and coherence was significant,  $\chi^2(2) = 6.26, p = .044$ . All other effects were not significant,  $ps > .161$ . As can be seen in Figure 7C, although confidence SAT clearly affected confidence RTs in the expected way, the effect was not as selective as in Experiment 1. Confidence RTs were shorter when participants were instructed to make fast ( $M = .39$  s,  $SD = 0.34$ ) versus careful ( $M = .76$  s,  $SD = 0.60$ ) confidence ratings, and counterintuitively confidence RTs were slightly longer when participants were instructed to make fast ( $M = .62$  s,  $SD = 0.56$ ) versus accurate ( $M = .53$  s,  $SD = 0.47$ ) decisions.

Finally, the same analysis was carried out on confidence for correct trials. Note that for this analysis, the three-way interaction and the interaction between the choice SAT and confidence SAT were excluded because they caused variance inflation factors higher than 10. In the final model, there was a significant main effect of coherence,  $\chi^2(2) = 100.06, p < .001$ , and the confidence SAT,  $\chi^2(1) = 4.36, p = .037$ , but not of the choice SAT,  $\chi^2(1) = 0.84, p = .359$ . As can be seen in Figure 7D, variations in confidence were mostly driven by coherence, but confidence was also slightly lower when participants were instructed to make fast ( $M = 4.50$ ,  $SD = 1.24$ ) versus careful ( $M = 4.56$ ,  $SD = 1.24$ ) confidence judgments. Finally, we found a significant interaction between the confidence SAT and coherence,  $\chi^2(2) = 22.10, p < .001$ , reflecting that the confidence SAT was more pronounced on low coherence trials. The interaction between the choice SAT and coherence was found to be not significant,  $\chi^2(2) = 1.57, p = .455$ .

Similar to Experiment 1, in a non-pre-registered analysis, we additionally looked at confidence resolution by calculating Type II AUC separately for each condition. Again, a 2-way ANOVA showed a main effect of confidence SAT,  $F(1,39) = 14.49, p < .001$ , but not from choice SAT,  $p = .066$ , nor was there an interaction,  $p = .125$ . As can be seen in Figure 4B, the relation between confidence and accuracy (expressed in AUC units) was higher when participants were instructed to make careful versus fast confidence ratings, suggesting that participants gave more accurate confidence ratings in the careful condition.

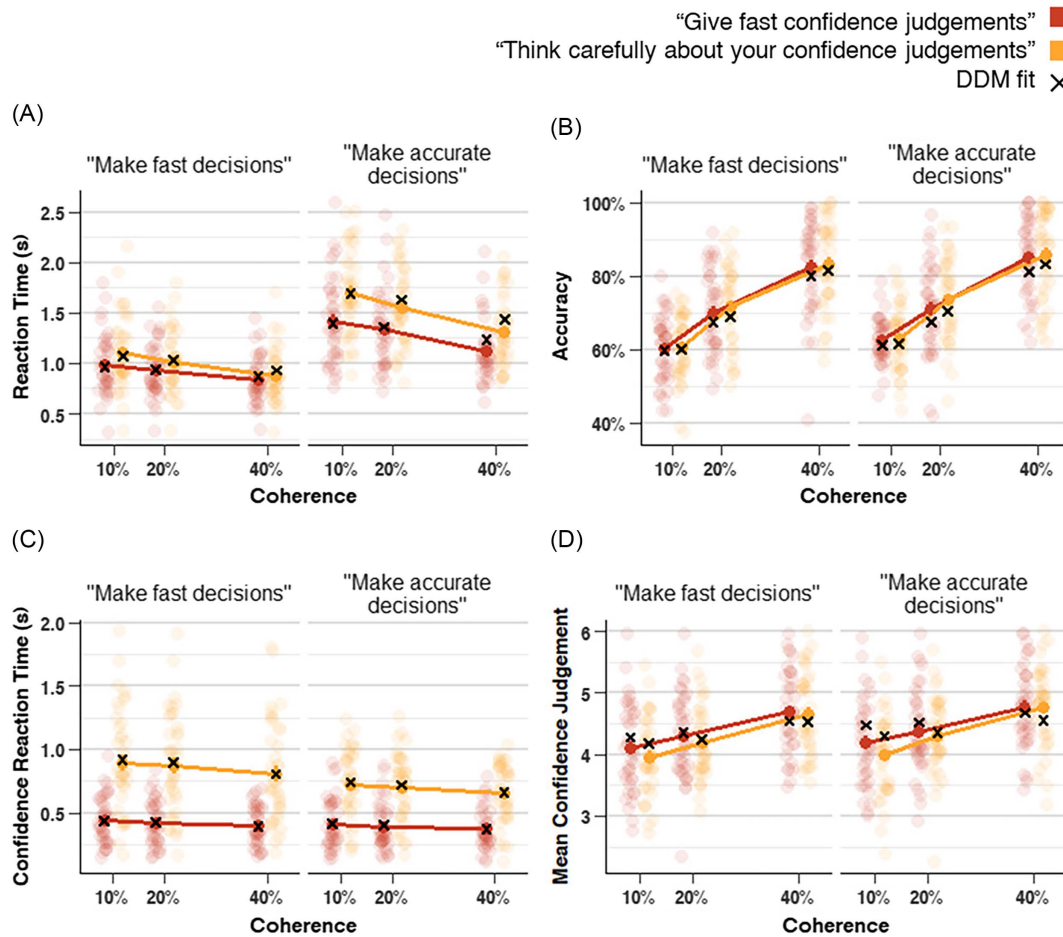
### Modeling SATs in 6-Point Scale Confidence Judgments

**Model Recovery.** We again fitted each model to the data of all 40 participants and computed BIC values. In addition to fitting the 6-choice versions of the models also used in Experiment 1 (2DSD, CCB, FCB, and Van Zandt), we also included an optional stopping variant of the 2DSD and a variant of the FCB model: FCB<sub>full</sub>, with both separate urgency parameters for the confidence boundaries and a



**Figure 7**

The Influence of Choice SAT and Confidence SAT on Reaction Times (A), Accuracy (B), Confidence RTs (C), and Confidence (D) for Experiment 2



*Note.* As expected, when participants were instructed to make fast versus accurate choices, this led to fast versus slow choice RTs (A) and, to a lesser extent, to less and more accurate choices (B), respectively. When participants were instructed to make fast versus deliberate confidence judgments, this led to fast versus slow confidence RTs (C), with less pronounced effects on confidence judgments. Same conventions as in Figure 3. DDM = drift diffusion model; RT = reaction time; SAT = speed-accuracy trade-off. See the online article for the color version of this figure.

free parameter for the postdecision starting point. For Experiment 2, model recovery showed that all models were identifiable, both when looking at mean BIC as well as when looking at the number of data sets best fitted by a specific model (see Table 4 for an overview of the model recovery). Importantly, whereas for Experiment 1, the CCB and FCB<sub>simple</sub> models, were sometimes confused when looking at individual data sets, using the six-option confidence scale from Experiment 2 allowed us to better dissociate both models. Note that when including the FCB<sub>simple</sub> model in the model confusion, it was often better fitted by the more flexible FCB<sub>full</sub> variant. Given that these two FCB variants are not dissociable, we decided not to include the FCB<sub>simple</sub> model in the model comparison of Experiment 2.

**Model Comparison.** Results from the model comparison shown in Table 5 show that FCB<sub>full</sub> outperformed the other models in mean

BIC ( $\Delta\text{BICs} > 37$ ) and provided the best fit for the largest number of participants (45.63%). *Post hoc* pairwise contrasts, accounting for participants and different conditions, showed that the FCB<sub>full</sub> model produced significantly lower BIC values than the CCB model,  $t(636) = -121.6, p < .001$ , the Van Zandt race model,  $t(636) = -142.3, p = .002$ , and the 2DSD model,  $t(636) = -104.6, p < .001$ , but only numerically lower BIC values compared to the 2DSD optional stopping model  $t(636) = -37.4, p = .11$ .

To visually represent the qualities of different model types, Figure 8A shows the relation between confidence and confidence RTs for a representative example participant. As can be seen, while this participant showed an inverted-U association between confidence and confidence RTs, this pattern was only captured closely by the FCB<sub>full</sub> model and, to some extent, the Van Zandt model. The 2DSD model predicts similar confidence RTs for all levels of

**Table 4**  
*Model Recovery for Experiment 2*

Generating model	Fitted model				
	FCB <sub>full</sub>	CCB	Van Zandt	2DSD	Optional stopping
Relative BIC					
FCB <sub>full</sub>	<b>0</b>	147.24	516.05	151.65	192.58
CCB	72.66	<b>0</b>	176.62	128.24	270.6
Van Zandt	144.86	223.26	<b>0</b>	266.97	176.7
2DSD	110.1	176.85	420.88	<b>0</b>	181.62
Optional stopping	105.29	237.18	408.75	269.18	<b>0</b>
No. of data sets best fitted					
FCB <sub>full</sub>	<b>136</b>	10	3	9	2
CCB	6	<b>148</b>	2	3	1
Van Zandt	4	1	<b>154</b>	1	0
2DSD	23	6	5	<b>122</b>	4
Optional stopping	33	5	3	11	<b>108</b>

*Note.* Table includes (a) BICs of the model recovery and (b) number of data sets best fitted by each model type for every generating model. BIC values are relative with respect to the best fitting model for each generating model. Lowest BIC values are indicated in bold. The model type fitting most data sets best for each generating model is indicated in bold. FCB = flexible confidence boundary; CCB = collapsing confidence boundary model; 2DSD = two-stage dynamic signal detection theory model; BIC = Bayesian information criterion.

confidence because of its time-based stopping criterion, and the CCB model predicts a negative confidence RT—confidence judgment relation, driven by its staircase-shaped confidence boundaries.

Going beyond this example participant, Figure 8B–8E shows several diagnostic features both for the empirical data as well as for the fitted models. Figure 8B and 8C suggests that the 2DSD and optional stopping models did not capture the observed negative relation between confidence RTs and, respectively, accuracy and mean confidence judgments. Notably, the predictions of these models regarding the standard deviation of confidence judgments relative to confidence RTs aligned closest to the observations (Figure 8D). Namely, higher confidence RTs were related to a higher standard deviation of the confidence judgments. While the FCB<sub>full</sub> and CCB models followed this pattern, both models underestimated the standard deviation. Last, the Van Zandt model predicted the opposite pattern. Finally, as suggested by Figure 8E, confidence resolution (Type II AUC) was overestimated by the Van Zandt model and underestimated by the optional stopping model.

In summary, the best model fit to our data was FCB<sub>full</sub>, in which confidence boundaries had separate urgency parameters and the postdecision starting point could freely vary. Additionally, Figure 7 shows that the model captured the observed RTs, confidence RTs,

accuracy, and confidence judgments well. A thorough examination of the model fit is available in Supplemental Figure S2.

**Model Parameters.** Next, we investigated the influence of the SAT instructions on the parameters of the best fitting model, the FCB model with separate urgencies and a flexible postdecision starting point. A parameter recovery analysis of this model showed that all parameters recovered well, including the confidence urgency parameters (which were difficult to recover in the case of binary confidence judgments). Full results of this analysis can be found in the Supplemental Figures S5 and S6.

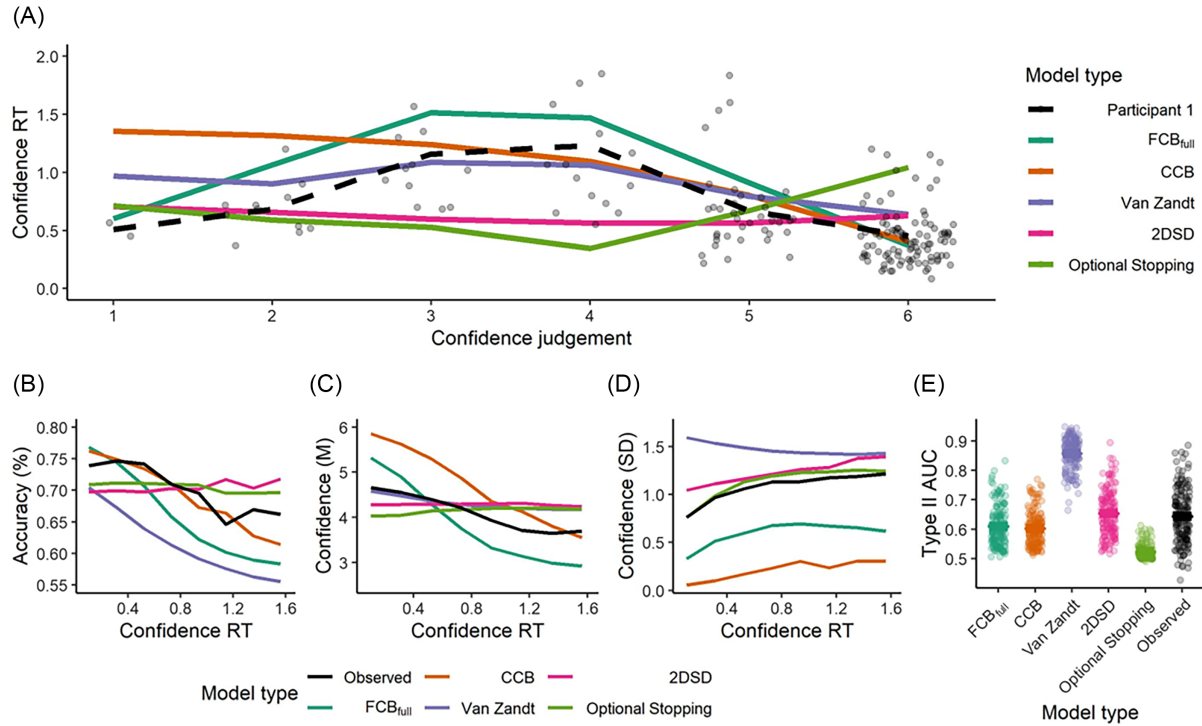
Similar to the findings of Experiment 1, we again observed that estimated decision boundaries were affected by choice SAT,  $F(1, 39) = 66.41, p < .001, \eta_p^2 = .63$ . However, we also found a significant effect of the confidence SAT,  $F(1, 39) = 31.90, p < .001, \eta_p^2 = .45$ , and an interaction between both,  $F(1, 39) = 4.82, p = .034, \eta_p^2 = .11$ . Follow-up paired  $t$  tests showed that the effect of the choice RT instructions on decision boundary separation was significant both when confidence SAT was to be accurate ( $M = 1.89, SD = 0.49$ ),  $t(39) = -7.75, p < .001$ , and when confidence SAT was to be fast ( $M = 1.68, SD = 0.42$ ),  $t(39) = -6.87, p < .001$ . As expected, choice boundaries were modulated by choice SAT instructions (Figure 9A), although the effect also seemed to scale, to a lesser extent, with confidence SAT.

**Table 5**  
*Experiment 2 Model Comparison*

Model type	Log RT MSE	Log confidence RT MSE	Log confidence proportions MSE	Free parameters	Mean BIC	Best fit (%)
2DSD	0.64	1.70	−0.52	14	−424.96	15.63
CCB	0.73	1.62	0.66	10	−408.04	6.88
FCB <sub>full</sub>	<b>0.59</b>	<b>0.77</b>	<b>0.06</b>	<b>11</b>	<b>−529.60</b>	<b>45.63</b>
Van Zandt	0.82	0.92	1.61	10	−387.31	4.38
Optional Stopping	0.76	0.64	0.11	17	−492.19	27.50

*Note.* Best fit shows the number of participants for which the corresponding model provides the lowest BIC. In line with Experiment 1, a variant of the FCB model provided best fit for the observed data, here indicated in bold. RT = reaction time; MSE = mean squared error; BIC = Bayesian information criterion; 2DSD = two-stage dynamic signal detection theory model; CCB = collapsing confidence boundary model; FCB = flexible confidence boundary.



**Figure 8***Experiment 2 Model Fit Comparison*

*Note.* (A) The relation between confidence and confidence RTs for a representative example participant (Participant 1), together with predictions of the different models. Each dot corresponds to a single observation. (B)–(E) Accuracy (B), average confidence (C), and variability in confidence (D) as a function of confidence RT, and type II ROC (E) shown for behavioral data, as well as with fitted model predictions. Note that line graphs do not show individual model patterns but rather show aggregated simulations across participants after binning into 30 equal-sized bins. As a consequence, for example, the seemingly negative-going slope of the CCB model in panel B actually reflects the average of many individuals showing a step-function from high (1) to low (0) confidence at a specific confidence RT. See Supplemental Figure S8 for a version of these plots based on average parameter estimates. FCB = flexible confidence boundary; CCB = collapsing confidence boundary model; 2DSD = two-stage dynamic signal detection theory model; RT = reaction time; ROC = receiver-operating characteristic curve. See the online article for the color version of this figure.

Second, we analyzed the confidence boundary separation. In line with findings from Experiment 1, we found a significant effect of confidence SAT,  $F(1, 39) = 58.42, p < .001, \eta_p^2 = .60$ . However, we did not find a significant effect of decision SAT,  $F(1, 39) = 1.35, p = .252, \eta_p^2 = .03$ , or the interaction between both,  $F(1, 39) = 2.14, p = .151, \eta_p^2 = .05$ . Participants increased confidence boundary separation when instructed to make careful confidence judgments ( $M = 5.36, SD = 1.81$ ) compared to when instructed to make fast confidence judgments ( $M = 4.04, SD = 1.38$ , Figure 9B). Furthermore, we found a significant effect on the starting point of evidence accumulation for confidence of confidence SAT,  $F(1, 39) = 29.79, p < .001, \eta_p^2 = .43$ , and of decision SAT,  $F(1, 39) = 14.37, p < .001, \eta_p^2 = .27$ , but not of the interaction between both,  $F(1, 39) = 1.04, p = .315, \eta_p^2 = .03$ . More specifically, participants tended to start evidence accumulation for confidence judgments at a higher level of evidence when asked to give fast confidence judgments ( $M = 0.64, SD = 0.10$ ) than when asked to think more carefully about them ( $M = 0.61, SD = 0.10$ ), but the starting point was also influenced by the decision SAT instructions (Figure 9C).

Notice that, different from Experiment 1, both confidence boundary urgencies were allowed to vary independently and thus are analyzed separately. Analysis of the upper confidence boundary urgency

revealed a significant effect of the confidence SAT,  $F(1, 39) = 38.98, p < .001, \eta_p^2 = .50$ . We did not observe an effect of the choice SAT,  $F(1, 39) = 3.66, p = .063, \eta_p^2 = .09$ , nor the interaction between both,  $F(1, 39) = 0.18, p = .674, \eta_p^2 = .01$ . As shown in Figure 9D, the slope was steeper when participants were instructed to make fast confidence judgments ( $M = 2.61, 1.20$ ) than when given the instruction to think carefully about their confidence judgments ( $M = 1.79, SD = 1.00$ ). Finally, for the lower confidence boundary, we found that urgency was not affected by confidence SAT,  $F(1, 39) = 1.17, p = .285, \eta_p^2 = .03, \eta_p^2 = .03$ , choice SAT,  $F(1, 39) = 0.20, p = .658, \eta_p^2 = .005$ , nor was there an interaction,  $F(1, 39) = 0.66, p = .422, \eta_p^2 = .02$  (Figure 9E).

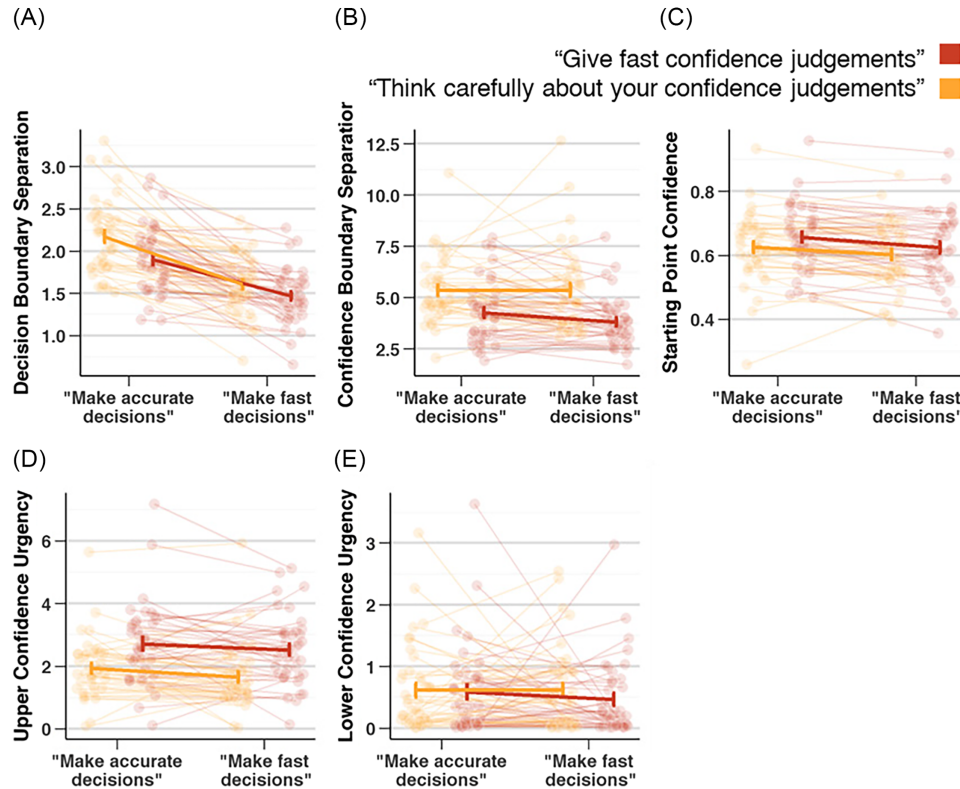
As a final sanity check, we again confirmed that the estimated drift rate scaled with motion coherence,  $F(1.14, 44.30) = 127.25, p < .001, \eta_p^2 = .77$ . All parameter estimates can be found in Table 3.

## Discussion

The human ability to estimate and report the level of confidence in their decisions has been the central topic of many recent investigations (Rahnev et al., 2022). Despite a large number of studies examining how confidence is computed, the question of how people

**Figure 9**

*Influence of Choice and Confidence SAT on Decision Boundaries and Confidence Boundaries in Experiment 2*



*Note.* Instructing participants to make fast versus accurate choices influenced estimated decision boundaries (A) and the starting point of postdecision evidence accumulation (C), but not confidence boundary separation or upper/lower confidence urgency (B, D, E). Instructing participants to provide fast versus careful confidence ratings influenced the decision boundary separation, starting point of postdecision evidence accumulation, confidence boundary separation, and upper confidence boundary urgency (A)–(D), but not lower confidence boundary urgency (E). Same conventions as in Figure 3. SAT = speed–accuracy trade-off. See the online article for the color version of this figure.

decide *when* to provide a confidence rating has been unresolved. This is remarkable because the timing of confidence judgments can be highly diagnostic about the computations underlying decision confidence (Moran et al., 2015). In the current work, we quantitatively and qualitatively compare the stopping rules of five prominent postdecisional evidence accumulation models in their ability to fit empirical data. We considered the 2DSD and optional stopping models proposed by Pleskac and Busemeyer (2010), the CCB model by Moran et al. (2015), the race model proposed by Van Zandt and Maldonado-Molina (2004), and finally our novel FCB model in which postdecision accumulation terminates once it reaches one of two opposing slowly collapsing confidence boundaries. We manipulated the stopping rule for confidence judgments by providing participants with different instructions regarding the trade-off between speed and accuracy, both for decisions and for confidence judgments. In two experiments, we found that participants made faster and less accurate decisions when instructed to favor speed over accuracy, and that they made faster confidence judgments when instructed to favor speed over careful deliberation of confidence. Although the effects on average confidence were subtle or even absent (similar to how SAT instructions in perceptual tasks often have strong effects on RTs but small or even

nonsignificant effects on accuracy; e.g., Desender et al., 2022), in both experiments, the relation between confidence and accuracy (cf. confidence resolution) was stronger when participants were more cautious in their confidence ratings. Most importantly, across both experiments, we found through qualitative and quantitative model comparison that the newly proposed FCB model fitted the data better than the alternative models. Inspection of the estimated parameters of the winning FCB model showed that, as expected, SAT instructions about the decision influenced decision boundaries, and SAT instructions about confidence influenced confidence boundaries. Our findings have important consequences for the field of decision confidence, as they shed light on the stopping rule for confidence and thereby unravel the importance of considering the dynamics of confidence RTs.

### Modeling the Stopping Rule of Evidence Accumulation for Confidence

Previous work investigating the dynamics of decision confidence has mostly focused on explaining variations in confidence reports, with less focus on the speed with which those reports are given. When modeling confidence, the most common approach is to simply

have a free parameter that controls the duration of postdecision evidence accumulation (Hellmann et al., 2021; Pleskac & Busemeyer, 2010; Yu et al., 2015). The current work is the first, to our knowledge, to test whether this approach can reliably fit confidence RTs, which typically show the same right-skewed distribution that is also characteristic of choice RTs, and compare this stopping rule to several competing alternatives. The most prominent alternative to this time-based stopping rule is confidence boundaries (i.e., an evidence-based stopping rule), which provide a plausible mechanism for the stopping rule of postdecision evidence accumulation. Indeed, some previous studies have already investigated the usefulness of confidence boundaries and found that these fit confidence RTs well (Moran et al., 2015; Van Zandt & Maldonado-Molina, 2004). Moran and colleagues proposed a single confidence boundary that collapses slowly over time, with the level of confidence being determined by the height of the boundary at the time of crossing. Our newly proposed FCB model differs from the CCB model by Moran and colleagues in the sense that the CCB model does not consider a lower confidence boundary; the model provides a confidence rating of .5 if the collapsing confidence boundary has not been reached before it collapses to .5. Contrastingly, the FCB model features both an upper and a lower confidence boundary, which can be mapped onto high versus low confidence (Experiment 1) or accounts for graded levels of confidence by further dividing the area in between the two confidence boundaries (Experiment 2). Van Zandt and Maldonado-Molina (2004) proposed that confidence can be computed based on the difference in postdecisional evidence between two independent (racing) accumulators. According to this model, a confidence judgment is given once one of the accumulators crosses a confidence boundary. Finally, Pleskac and Busemeyer (2010) proposed that confidence judgments result from postdecision evidence accumulation while crossing “optional stopping” thresholds of evidence. Crossing a threshold implicates a probability of giving the confidence judgment corresponding to the level of evidence accumulated. Results from our model comparison found unequivocal support, across two experiments, that the stopping rule for confidence judgments is best explained by an accumulation-to-bound mechanism. In both experiments, the 2DSD model provided a considerably worse fit to the confidence RT data because it was unable to capture intricate associations between confidence and confidence RTs. Additionally, from the different models considered here that did implement an accumulation-to-bound mechanism as the stopping rule for confidence, the newly proposed FCB model provided the best fit to the data. Again, the reason why this model was favored over the others was that it was able to flexibly allow for both positive, absent, negative, and even U-shaped associations between confidence and confidence RT (see Figure 3A in Herregods et al., 2024).

Interestingly, there has been a large body of work that modeled the stopping rule of evidence accumulation in tasks with simultaneous decision-making and confidence judgment tasks (Ratcliff & Starns, 2009, 2013). Given the simultaneous reporting of choice and confidence, a single boundary typically suffices to explain choice and confidence latencies, and this line of work proposes multiple accumulators to jointly explain choices and confidence. Such models can effectively account for the timing of choices and confidence; however, they do not consider temporal dissociations between choice and confidence. Our work differs substantially from this approach, as our main aim is to explain behavior in tasks where choices and confidence

judgments are given sequentially, thus allowing for the accumulation of postdecision evidence.

Note that in its current implementation, the FCB model predicts a rather close association between confidence and confidence RTs. Although the model captured the behavior well, this assumption could be relaxed by having additional metacognitive noise (Shekhar & Rahnev, 2021), allowing for a more variable mapping between accumulated evidence and confidence. Furthermore, for convenience, we modeled the starting point and the drift rate as being fixed across a series of trials. In the literature, adding variability to both of these parameters has been shown to induce fast and slow errors, respectively (Ratcliff & McKoon, 2008). Given that this implementation was shared across the different models, it seems unlikely that this modeling choice had an effect on our conclusions.

Finally, in Experiment 2, participants were asked to give a confidence judgment that reflects how probable they think they are correct, ranging from “sure correct” (i.e., 100%) to “sure error” (i.e., 0%). This situation differs from a considerable number of studies in which participants are only asked to judge the certainty of their response, ranging from “very sure” (100%) to “not sure” (50%). Note, however, that in the latter case of measuring confidence, it is unclear what a participant should do if they detect themselves making an error. Thus, in order to avoid the risk that different participants report their self-corrected errors in a different manner, it seems advisable in speeded decision-making tasks to include the option for participants to report (degrees of) self-detected errors.

### The Stopping Rule for Confidence Is Under Strategic Control

The FCB Model, which best fitted the data in both experiments, assumes that confidence reports are quantified as soon as a process of postdecision accumulation reaches one of two opposing confidence boundaries. Importantly, these confidence boundaries can slowly collapse as time passes. Therefore, this model predicts that confidence depends both on the height and the collapse rate of these confidence boundaries. SATs in choices can be implemented by either changing the overall height or by changing the collapse rate of the decision boundary, whereas instructions regarding the trade-off between speed and decision accuracy modulate the height of the decision boundary. Providing participants with a response deadline (e.g., respond within 1 s) modulates the rate of collapse of the decision boundary (Katsimpokis et al., 2020; Murphy et al., 2016). Interestingly, the same two mechanisms seem to apply in SATs for confidence judgments. In the current work, we have shown that instructions to modulate the confidence SAT influence the height of the confidence boundaries, while the rate of collapse was mostly unaffected. In a direct follow-up of the current work, we have shown that providing participants with a deadline for their confidence response selectively modulates the rate of collapse of the confidence boundaries while leaving their height intact (Grogan et al., 2025). Collectively, these findings demonstrate that the same principles that govern the setting of boundary height and collapse rate also apply in the context of confidence boundaries. The results concerning the urgency of the confidence boundaries of the current work require a bit more elaboration. In Experiment 1, urgency for the confidence boundaries did not recover well, suggesting that parameter estimates should not be interpreted. Urgency parameters showed good recovery for Experiment 2. This was likely because using a six-option



confidence scale provides much more informative data that can be used to constrain the collapse rate of the confidence boundaries. In Experiment 2, we found that in response to instructions requiring fast confidence responses, participants increased the level of urgency for the upper confidence boundary, but not the lower confidence boundary. Having confidence boundaries that can collapse at independent rates is a key feature of the FCB model, which allows it to flexibly compute confidence depending on environmental constraints. Simulations show that, depending on the specific reward scheme, it is optimal to independently collapse the upper and lower confidence boundaries (see Supplemental Figure S12).

### Characteristics of Postdecision Processing

If confidence can be understood as an accumulation-to-bound signal, it follows that the reported level of confidence should depend on the height of the confidence boundary. Similar to how decreasing the decision boundary induces faster RTs and less accurate responses, it follows that, given a positive drift rate, decreasing the confidence boundaries should induce faster confidence RTs and lower confidence. Contrary to this, in the estimated parameters of the FCB model, we did not observe a clear influence of confidence SAT on average confidence despite a clear difference in the height of the confidence boundary. As can be seen in Table 3, the FCB model explained these data by assuming that decreasing the confidence boundaries was associated with increased postdecision drift rates. Future work might examine whether this prediction holds in postdecision centroparietal electroencephalography signals, which are thought to reflect the postdecision accumulation-to-bound signal (Desender et al., 2021). Although we did not find an effect on average confidence, there was a clear effect on confidence resolution: The relation between confidence and accuracy was much stronger when participants increased the confidence boundary. This finding could be anticipated because increasing the confidence boundaries effectively requires collecting more postdecision evidence before reporting confidence, that is, making a more informed confidence judgment. This finding adds to a number of reports showing that measures of metacognitive accuracy critically depend on the timing of confidence reports (Rosenbaum et al., 2022; Yu et al., 2015).

As shown in Supplemental Figure S9, which visualizes the fitted parameter values of the FCB<sub>full</sub> model, when participants report that they are certain that they made an error, this is on average the case when participants collect a substantial degree of postdecision evidence that conflicts with their initial choice; that is, the confidence boundary associated with “certainly wrong” is further away from the decision boundary associated with the current choice than with the unchosen decision boundary. This could point toward a difference in reference frame based on which pre- and postdecision evidence is evaluated. Although postdecision accumulation is often treated as a simple continuation of predecision accumulation, some authors have proposed that postdecision accumulation might induce a change in reference frame, for example, selectively accumulating evidence in favor of having committed an error (Desender et al., 2021; Murphy et al., 2015). More specifically, according to the FCB<sub>full</sub> model, a participant starts accumulating evidence at  $z \times a$  and makes a decision once a decision boundary at  $a$  or  $0$  is reached. If postdecision evidence continues to accumulate until hitting a confidence boundary at exactly  $z \times a$ , this would refer to an “uncertain” response, but in a “novel”

reference frame, this could actually reflect strong evidence of having made an error.

Inspection of the estimated FCB model parameters in Table 3 reveals an interesting difference in magnitude between the nondecision component associated with the decision,  $Ter$ , and that associated with the confidence report,  $Ter_2$ . In line with the literature, values of  $Ter$  are in the range of .4 s–.5 s on average, suggesting that this is the time participants spend on processes unrelated to the actual decision (e.g., stimulus processing, motor components). These estimates are by definition positive. Contrary to this, values of  $Ter_2$  are very low for Experiment 1 and even negative for Experiment 2. Although negative values of  $Ter_2$  might seem counterintuitive at first, they suggest that “postdecision” processing already initiates prior to the execution of the decision motor response (e.g., Verdonck et al., 2021). In line with this observation, there is some work that has suggested that prechoice or perichoice neural signals contribute to the computation of decision confidence (Feuerriegel et al., 2022; Gherman & Philastides, 2015; Murphy et al., 2015).

### Conclusion

We demonstrated that the stopping rule for confidence judgments is well described by an accumulation-to-bound process terminating postdecision processing, similar to how choice formation occurs. Similar to the decision boundaries, these confidence boundaries are under strategic control and can be increased or decreased by instructing participants to make very careful or very fast confidence judgments, respectively. This implementation of the stopping rule fits the data better compared to the popular 2DSD model (in which postdecision processing terminates after a fixed amount of time has passed) and three other competing models. Taken together, the current work unravels the stopping rule for postdecision processing that informs the computation of confidence.

### References

- Balci, F., Simen, P., Niyogi, R., Saxe, A., Hughes, J. A., Holmes, P., & Cohen, J. D. (2011). Acquisition of decision making criteria: Reward rate ultimately beats accuracy. *Attention, Perception & Psychophysics*, 73(2), 640–657. <https://doi.org/10.3758/s13414-010-0049-7>
- Balsdon, T., Wyart, V., & Mamassian, P. (2020). Confidence controls perceptual evidence accumulation. *Nature Communications*, 11(1), Article 1753. <https://doi.org/10.1038/s41467-020-15561-w>
- Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- 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>
- Bogacz, R., Hu, P. T., Holmes, P. J., & Cohen, J. D. (2010). Do humans produce the speed–accuracy trade-off that maximizes reward rate? *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 63(5), 863–891. <https://doi.org/10.1080/17470210903091643>
- Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U., & Nieuwenhuis, S. (2010). The neural basis of the speed–accuracy tradeoff. *Trends in Neurosciences*, 33(1), 10–16. <https://doi.org/10.1016/j.tins.2009.09.002>
- Boldt, A., & Yeung, N. (2015). Shared neural markers of decision confidence and error detection. *The Journal of Neuroscience*, 35(8), 3478–3484. <https://doi.org/10.1523/JNEUROSCI.0797-14.2015>

- Calder-Travis, J., Charles, L., Bogacz, R., & Yeung, N. (2024). Bayesian confidence in optimal decisions. *Psychological Review*, 131(5), 1114–1160. <https://doi.org/10.1037/rev0000472>
- Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological Science*, 29(5), 761–778. <https://doi.org/10.1177/0956797617744771>
- Desender, K., Murphy, P., Boldt, A., Verguts, T., & Yeung, N. (2019). A postdecisional neural marker of confidence predicts information-seeking in decision-making. *The Journal of Neuroscience*, 39(17), 3309–3319. <https://doi.org/10.1523/JNEUROSCI.2620-18.2019>
- Desender, K., Ridderinkhof, K. R., & Murphy, P. R. (2021). Understanding neural signals of post-decisional performance monitoring: An integrative review. *eLife*, 10, Article e67556. <https://doi.org/10.7554/eLife.67556>
- Desender, K., Vermeulen, L., & Verguts, T. (2022). Dynamic influences on static measures of metacognition. *Nature Communications*, 13(1), Article 4208. <https://doi.org/10.1038/s41467-022-31727-0>
- Donner, T. H., Siegel, M., Fries, P., & Engel, A. K. (2009). Buildup of choice-predictive activity in human motor cortex during perceptual decision making. *Current Biology*, 19(18), 1581–1585. <https://doi.org/10.1016/j.cub.2009.07.066>
- Feuerriegel, D., Murphy, M., Konski, A., Mepani, V., Sun, J., Hester, R., & Bode, S. (2022). Electrophysiological correlates of confidence differ across correct and erroneous perceptual decisions. *NeuroImage*, 259, Article 119447. <https://doi.org/10.1016/j.neuroimage.2022.119447>
- Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, 329(5998), 1541–1543. <https://doi.org/10.1126/science.1191883>
- Folke, T., Jacobsen, C., Fleming, S. M., & De Martino, B. (2016). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, 1, Article 0002. <https://doi.org/10.1038/s41562-016-0002>
- Fox, J., & Weinberg, S. (2019). *An R companion to applied regression* (3rd ed.). SAGE Publications.
- Gherman, S., & Philastides, M. G. (2015). Neural representations of confidence emerge from the process of decision formation during perceptual choices. *NeuroImage*, 106, 134–143. <https://doi.org/10.1016/j.neuroimage.2014.11.036>
- Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, 30(1), 535–574. <https://doi.org/10.1146/annurev.neuro.29.051605.113038>
- Grogan, J. P., Vermeulen, L., Mannion, S. L., McCabe, C., Monakhovych, D., Desender, K., & O'Connell, R. G. (2025). *Neurally-informed modelling unravels a single evidence accumulation process for choices and subsequent confidence reports*. bioRxiv. <https://doi.org/10.1101/2025.06.05.658071>
- Hellmann, S., Zehetleitner, M., & Rausch, M. (2021). *Simultaneous modeling of choice, confidence and response time in visual perception*. bioRxiv.
- Hellmann, S., Zehetleitner, M., & Rausch, M. (2023). Simultaneous modeling of choice, confidence, and response time in visual perception. *Psychological Review*, 130(6), 1521–1543. <https://doi.org/10.1037/rev0000411>
- Herregods, S., & Desender, K. (2021). *Confidence bounds in drift diffusion models*. <https://doi.org/10.17605/OSF.IO/Z2UCM>
- Herregods, S., Vermeulen, L., & Desender, K. (2024). Flexible relations between confidence and confidence RTs in post-decisional models of confidence: A reply to Chen and Rahnev. *Journal of Vision*, 24(12), Article 9. <https://doi.org/10.1167/jov.24.12.9>
- Hester, R., & Garavan, H. (2005). Neural correlates of error detection with and without awareness. *Memory and Cognition*, 33, 221–223. [https://www.researchgate.net/publication/284497182\\_Neural\\_correlates\\_of\\_error\\_detection\\_with\\_and\\_without\\_awareness](https://www.researchgate.net/publication/284497182_Neural_correlates_of_error_detection_with_and_without_awareness)
- Hollands, J. G., & Dyre, B. P. (2000). Bias in proportion judgments: The cyclical power model. *Psychological Review*, 107(3), 500–524. <https://doi.org/10.1037/0033-295X.107.3.500>
- Kassambara, A. (2023). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests* [Computer software]. R package Version 0.7.2. <https://CRAN.R-project.org/package=rstatix>
- Katsimpokis, D., Hawkins, G. E., & van Maanen, L. (2020). Not all speed–accuracy trade-off manipulations have the same psychological effect. *Computational Brain & Behavior*, 3(3), 252–268. <https://doi.org/10.1007/s42113-020-00074-y>
- Kvam, P. D., Marley, A. A. J., & Heathcote, A. (2023). A unified theory of discrete and continuous responding. *Psychological Review*, 130(2), 368–400. <https://doi.org/10.1037/rev0000378>
- Maniscalco, B., Charles, L., & Peters, M. A. K. (2025). Optimal meta-cognitive decision strategies in signal detection theory. *Psychonomic Bulletin & Review*, 32(3), 1041–1069. <https://doi.org/10.3758/s13423-024-02510-7>
- 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>
- Mullen, K., Ardia, D., Gil, D., Windover, D., & Cline, J. (2011). 'DEoptim': An R package for global optimization by differential evolution. *Journal of Statistical Software*, 40(6), 1–26. <https://doi.org/10.18637/jss.v040.i06>
- Murphy, P. R., Boonstra, E., & Nieuwenhuis, S. (2016). Global gain modulation generates time-dependent urgency during perceptual choice in humans. *Nature Communications*, 7, Article 13526. <https://doi.org/10.1038/ncomms13526>
- Murphy, P. R., Robertson, I. H., Harty, S., & O'Connell, R. G. (2015). Neural evidence accumulation persists after choice to inform metacognitive judgments. *eLife*, 4, Article e11946. <https://doi.org/10.7554/eLife.11946>
- O'Connell, R. G., Dockree, P. M., & Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature Neuroscience*, 15(12), 1729–1735. <https://doi.org/10.1038/nn.3248>
- Palminteri, S., Wyart, V., & Koehlin, 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., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1), 195–203. <https://doi.org/10.3758/s13428-018-01193-y>
- 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), Article 3261. <https://doi.org/10.1038/s41467-021-23540-y>
- Pereira, M., Perrin, D., & Faivre, N. (2022). A leaky evidence accumulation process for perceptual experience. *Trends in Cognitive Sciences*, 26(6), 451–461. <https://doi.org/10.1016/j.tics.2022.03.003>
- Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, 117(3), 864–901. <https://doi.org/10.1037/a0019737>
- R Core Team. (2021). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org>
- Rafiei, F., & Rahnev, D. (2021). Qualitative speed–accuracy tradeoff effects that cannot be explained by the diffusion model under the selective influence assumption. *Scientific Reports*, 11(1), Article 45. <https://doi.org/10.1038/s41598-020-79765-2>
- Rahnev, D., Balsdon, T., Charles, L., de Gardelle, V., Denison, R., Desender, K., Faivre, N., Filevich, E., Fleming, S. M., Jehee, J., Lau, H., Lee, A. L. F., Locke, S. M., Mamassian, P., Odegaard, B., Peters, M., Reyes, G., Rouault, M., Sackur, J., ... Zylberberg, A. (2022). Consensus goals in the field of visual metacognition. *Perspectives on*

- Psychological Science*, 17(6), 1746–1765. <https://doi.org/10.1177/17456916221075615>
- Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 20(4), 873–922. <https://doi.org/10.1162/neco.2008.12-06-420>
- Ratcliff, R., & Starns, J. J. (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>
- 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>
- Rosenbaum, D., Glickman, M., Fleming, S. M., & Usher, M. (2022). The cognition/metacognition trade-off. *Psychological Science*, 33(4), 613–628. <https://doi.org/10.1177/09567976211043428>
- Shekhar, M., & Rahnev, D. (2021). Sources of metacognitive inefficiency. *Trends in Cognitive Sciences*, 25(1), 12–23. <https://doi.org/10.1016/j.tics.2020.10.007>
- Solway, A., & Botvinick, M. M. (2015). Evidence integration in model-based tree search. *Proceedings of the National Academy of Sciences of the United States of America*, 112(37), 11708–11713. <https://doi.org/10.1073/pnas.1505483112>
- Tuerlinckx, F., Maris, E., Ratcliff, R., & De Boeck, P. (2001). A comparison of four methods for simulating the diffusion process. *Behavior Research Methods, Instruments, & Computers*, 33(4), 443–456. <https://doi.org/10.3758/BF03195402>
- 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 Zandt, T., & Maldonado-Molina, M. M. (2004). Response reversals in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(6), 1147–1166. <https://doi.org/10.1037/0278-7393.30.6.1147>
- Verdonck, S., Loossens, T., & Philiastides, M. G. (2021). the leaky integrating threshold and its impact on evidence accumulation models of choice response time (RT). *Psychological Review*, 128(2), 203–221. <https://doi.org/10.1037/rev0000258>
- Vickers, D. (1979). *Decision processes in visual perception*. Academic Press.
- Yu, S., Pleskac, T. J., & Zeigenfuse, M. D. (2015). Dynamics of post-decisional processing of confidence. *Journal of Experimental Psychology: General*, 144(2), 489–510. <https://doi.org/10.1037/xge0000062>
- Zylberberg, A., Fetsch, C. R., & Shadlen, M. N. (2016). The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. *eLife*, 5, Article e17688. <https://doi.org/10.7554/eLife.17688>

Received October 24, 2024

Revision received September 24, 2025

Accepted October 4, 2025 ■