

Journal of Experimental Psychology: General

Adaptive Regulation of Confidence Following Metacognitive Conflict: Evidence From Behavior and Computational Modeling

Jintao Xing, Stef Herregods, Luc Vermeylen, and Kobe Desender

Online First Publication, September 14, 2026. <https://dx.doi.org/10.1037/xge0001997>

CITATION

Xing, J., Herregods, S., Vermeylen, L., & Desender, K. (2026). Adaptive regulation of confidence following metacognitive conflict: Evidence from behavior and computational modeling. *Journal of Experimental Psychology: General*. Advance online publication. <https://dx.doi.org/10.1037/xge0001997>

Adaptive Regulation of Confidence Following Metacognitive Conflict: Evidence From Behavior and Computational Modeling

Jintao Xing, Stef Herregods, Luc Vermeulen, and Kobe Desender
Brain and Cognition, Katholieke Universiteit Leuven

It is well established that external feedback is effective not only for enhancing cognitive performance but also for regulating metacognitive monitoring. However, external feedback could be detrimental when it conflicts with one's internal feedback signal (i.e., the level of confidence we have in our decisions). Currently, however, it is unknown how discrepancies between internal and external feedback ("metacognitive conflict") influence cognitive and metacognitive performance. Here, we examined this question in a series of perceptual decision-making experiments in which participants reported both their level of confidence and received performance feedback. Metacognitive conflict occurs if participants have high confidence in their decision, but external feedback indicates that the decision was erroneous, or vice versa. Results revealed that metacognitive conflict triggers next-trial behavioral adjustments: Participants exhibit significantly lower confidence following metacognitive conflict. Furthermore, this behavioral adjustment was smaller in blocks with a high versus low proportion of metacognitive conflict trials, and confidence was generally reduced in blocks with a high proportion of metacognitive conflict. To better characterize the computational mechanisms underlying this effect, we deployed an evidence accumulation model with additional postdecision accumulation to quantify confidence. Model fits revealed that participants become more conservative in reporting confidence after experiencing metacognitive conflict, as evidenced by biased postdecision starting points. Jointly, our findings provide much-needed insight into the metacognitive consequences of metacognitive conflict.

Public Significance Statement

People often rely on both their own judgment and feedback from others. This study reveals that when these two sources conflict, individuals become more cautious in their next decision. By modeling the underlying decision processes, we show how metacognitive conflict biases confidence reporting. These findings provide new insights into how humans monitor and regulate their own decision making, with implications for learning, performance feedback, and adaptive behavior.

Keywords: metacognition, conflict processing, confidence, drift diffusion model, feedback

Supplemental materials: <https://doi.org/10.1037/xge0001997.supp>

The ability to recognize and introspect one's ongoing cognitive processes is referred to as "metacognition" (Flavell, 1979; Fleming & Frith, 2014; Yeung & Summerfield, 2012). This metacognitive access is highly adaptive, because it allows us to control these processes in case of error. For example, if students realize that they have misunderstood a formula while solving a math problem, they

might revisit their notes or seek clarification to ensure better performance in the future. This metacognitive monitoring process has been highlighted as a key driver of many adaptive behaviors including adjustment of decision-making strategies (Fleming, 2024; Frömer et al., 2021). The evaluation of one's own decision, often referred to as "decision confidence", is considered a metacognitive

Rachel Wu served as action editor.

Jintao Xing  <https://orcid.org/0000-0002-3446-7878>

Stef Herregods  <https://orcid.org/0000-0003-4105-3857>

Luc Vermeulen  <https://orcid.org/0000-0001-7370-7152>

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

A preprint of this article was posted on PsyArXiv prior to submission; see https://osf.io/preprints/psyarxiv/zc5bx_v1. All authors declared no competing interests. Jintao Xing is supported by the China Scholarship Council (Grant 202308510055). Luc Vermeulen is supported by an Fonds Wetenschappelijk Onderzoek postdoctoral fellowship by the Research Foundation Flanders, Belgium (Grant 1242924N). Stef Herregods is

supported by a project grant by the Research Foundation Flanders, Belgium (Grant G0B0521N).

Jintao Xing played a lead role in conceptualization, data curation, formal analysis, methodology, and writing—original draft and an equal role in funding acquisition. Stef Herregods played a supporting role in methodology and writing—review and editing. Luc Vermeulen played a supporting role in conceptualization, funding acquisition, and methodology. Kobe Desender played a lead role in supervision and writing—review and editing and an equal role in conceptualization, funding acquisition, and methodology.

Correspondence concerning this article should be addressed to Jintao Xing, Brain and Cognition, Katholieke Universiteit Leuven, Tiensestraat 102, 3000 Leuven, Belgium. Email: jintao.xing@kuleuven.be

assessment of the quality and sufficiency of evidence underlying a decision (Fleming & Daw, 2017; Fleming et al., 2012). Confidence enables individuals to recognize uncertainty in their decisions and, for example, resolve this uncertainty by seeking additional evidence or incorporating external feedback, thereby minimizing the likelihood of repeated errors (Ben Yehuda et al., 2024; Desender et al., 2018; Niccolo & Nicholas, 2020).

Similar to how confidence is described as a signal driving behavioral adjustment, a large body of literature has described how external feedback is one of the most influential factors shaping human behavior (Balzer et al., 1989; Fishbach et al., 2010; Raaijmakers et al., 2017). By providing information about whether a response was correct or erroneous, feedback helps individuals to reevaluate and adapt their behavior (Hattie & Timperley, 2007; Petrov et al., 2005). In the field of decision making, substantial evidence suggested that performance feedback improves behavioral performance by driving strategic selection or facilitating learning (Kahnt et al., 2011; Lee et al., 2023). A common behavioral consequence of receiving negative feedback is posterror slowing (PES), which indicates that negative feedback in decision-making tasks (i.e., following an incorrect decision) typically leads to increased response time (Laming, 1979; Rabbitt & Rodgers, 1977). A long-standing explanation for PES is that individuals increase their cautiousness to maintain relatively high accuracy following an error (Botvinick et al., 2001; Danielmeier & Ullsperger, 2011; Rabbitt & Rodgers, 1977). This account is further supported by evidence from computational modeling. The drift diffusion model explains the decision process as the noisy accumulation of evidence over time until one of two decision thresholds is reached (Ratcliff et al., 2016; Ratcliff & McKoon, 2008; Wiecki et al., 2013). Within this framework, empirical studies have shown that people indeed increase their decision thresholds after making an error, reflecting a more cautious decision strategy to prevent further mistakes (Dutilh et al., 2012; White et al., 2010).

From the above, it appears as if decision confidence and external feedback play highly similar and sometimes complementary roles in adaptive behavior. For example, consistent with PES, individuals exhibit response slowing following low-confidence decisions, even in the absence of external feedback (Desender et al., 2019). Likewise, neurobiological studies showed that neural regions sensitive to variations in confidence were similar to regions previously identified to be sensitive to external feedback (Garrison et al., 2013; Guggenmos et al., 2016). Importantly, however, it is well known that the sense of confidence is not always a reliable estimate of decision accuracy, as individuals frequently misestimate the accuracy of their decisions (Shekhar & Rahnev, 2021). This raises the intriguing question what happens when there is a (metacognitive) conflict between internal feedback (i.e., decision confidence) and external feedback. Specifically, metacognitive conflict occurs when participants have high confidence in a decision but receive external feedback indicating that the decision is erroneous. This can be defined as a negative prediction error (NPE) metacognitive conflict, given that performance is worse than expected. Conversely, metacognitive conflict also occurs when participants have low confidence in a decision but receive positive feedback indicating that the decision is correct. This can be defined as a positive prediction error (PPE) metacognitive conflict, given that performance is better than expected. Remarkably, although a large body of research has examined how internal and external feedback regulate behavior,

little attention has been paid to how metacognitive conflict between these two factors jointly shapes cognitive and metacognitive processes. In the present study, we aim to address this gap by combining both internal and external feedback within the context of a perceptual decision-making task.

To our knowledge, the current work is the first to investigate the cognitive and metacognitive consequences of metacognitive conflict. However, there are plenty of studies that have examined the influence of conflict at the cognitive level (Dreisbach & Fischer, 2015; Mayr, 2004). Cognitive conflict can occur in selective attention tasks (e.g., Stroop task, Simon task, or Flanker task) by introducing task-irrelevant information that interferes with the processing of task-relevant information (Braem et al., 2019). Performance in conflict tasks is marked by slower and less accurate responses to conflict trials than to nonconflict trials (i.e., the conflict effect). Relevant to the present study, this line of research has identified two key findings thought to express adaptive behavior: sequential (i.e., trial-level) and sustained (i.e., block-level) adaptation effects in response to conflict. At the trial level, the conflict effect is reduced following a conflict trial compared to a nonconflict trial (i.e., the conflict adaptation effect; Egner, 2007; Gratton et al., 1992). At the block level, the conflict effect is reduced in blocks where conflict is very prominent. Specifically, by manipulating the frequency of conflict and nonconflict trials within a block, researchers have demonstrated that the conflict effect is smaller in the mostly-conflict block than in the mostly-no-conflict block (i.e., the proportion congruency [PC] effect; Bugg & Crump, 2012; Crump et al., 2006; Logan & Zbrodoff, 1979). Note that conflict effects are different in nature compared to PES, given that PES is typically triggered by an external event (i.e., explicit feedback indicating that an error was committed), whereas conflict adaptation is triggered by an internal event (i.e., internal detection of conflicting responses triggering a change in attentional processing). Inspired by these trial-level and block-level adaptations to cognitive conflict, we investigated whether analogous trial-level and block-level adjustments occur in response to metacognitive conflict.

In summary, the main goal of the present study is to investigate the cognitive and metacognitive consequences of metacognitive conflict both at the trial level (Experiments 1A and 1B) and at the block level (Experiment 2). To this end, we conducted a series of three experiments in which participants performed a perceptual decision-making task, reported their confidence, and received feedback on choice accuracy. To unravel the computational mechanisms underlying metacognitive conflict-induced regulatory behavior, we fitted our empirical data using the flexible confidence boundary model, which extends the drift diffusion model by incorporating postdecision evidence accumulation to inform confidence (Desender et al., 2021; Herregods et al., 2025).

Experiment 1A

Method

Participants

Sixty-seven participants (six men, 61 women; age: $M = 18.2$, $SD = 0.84$, range = 18–21) took part in return for course credit. Gender was self-reported by selecting from a list of options (man, woman, X). All participants reported normal or corrected-to-normal vision. All but seven participants were right-handed. One participant

was excluded because her performance was not different from chance level (as assessed by a binomial test). No participants were excluded for excessively (>90%) using a single button for confidence judgments. Participants provided written informed consent before participation. All procedures were approved by the KU Leuven Ethics Committee.

Stimuli and Apparatus

Stimuli were presented in white on a black background on a 24-in. LCD monitor, using Psychopy (Peirce et al., 2019) for Python. A total of 120 random moving white dots (diameter 4 pixels) were drawn in a 300-pixel circle centered on the fixation point. Each dot had a lifespan of five frames and moved at 1 pixel every frame. Critically, dot onset frames were randomized to ensure that lifespans were staggered, preventing a periodic visual refresh of the entire stimulus.

Procedure

Each experimental trial started with the display of a white fixation cross on a black background for 750 ms, followed by a random dot motion stimulus that lasted until a response was made, with a maximum of 3 s (see Figure 1). 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 did not respond within 3 s, the trial was aborted and a warning message appeared to remind the participant to respond more quickly on the next trial. Trial difficulty was controlled by manipulating the percentage of coherently moving dots, which was either 10%, 20%, or 40% during the main experiment. After making their choice, participants rated their confidence level on a 6-point scale using the keys 1, 2, 3, 8, 9, and 0. The scale was labeled as *sure wrong*, *probably wrong*, *guess wrong*, *guess correct*, *probably correct*, and *sure correct* (reversed order for half of the participants). If participants rate their confidence within 5 s, the feedback about accuracy (“correct” or “wrong”) was presented for 1,500 ms; otherwise, a warning message “Too slow, please respond faster” was shown on the screen. The experiment started with a practice session that consisted of three practice blocks (24 trials each) to get familiar with the task. The coherence of the first practice block was fixed at 80%, and the coherence of the subsequent two practice blocks was consistent with the formal session. After each practice block, participants received feedback about their average accuracy, average reaction time (RT), and average confidence RT of the preceding block. The formal session, consisting of 12 main blocks (60 trials each), began if accuracy in the practice blocks exceeded 60%; otherwise, the practice block was repeated.

Statistical Analysis

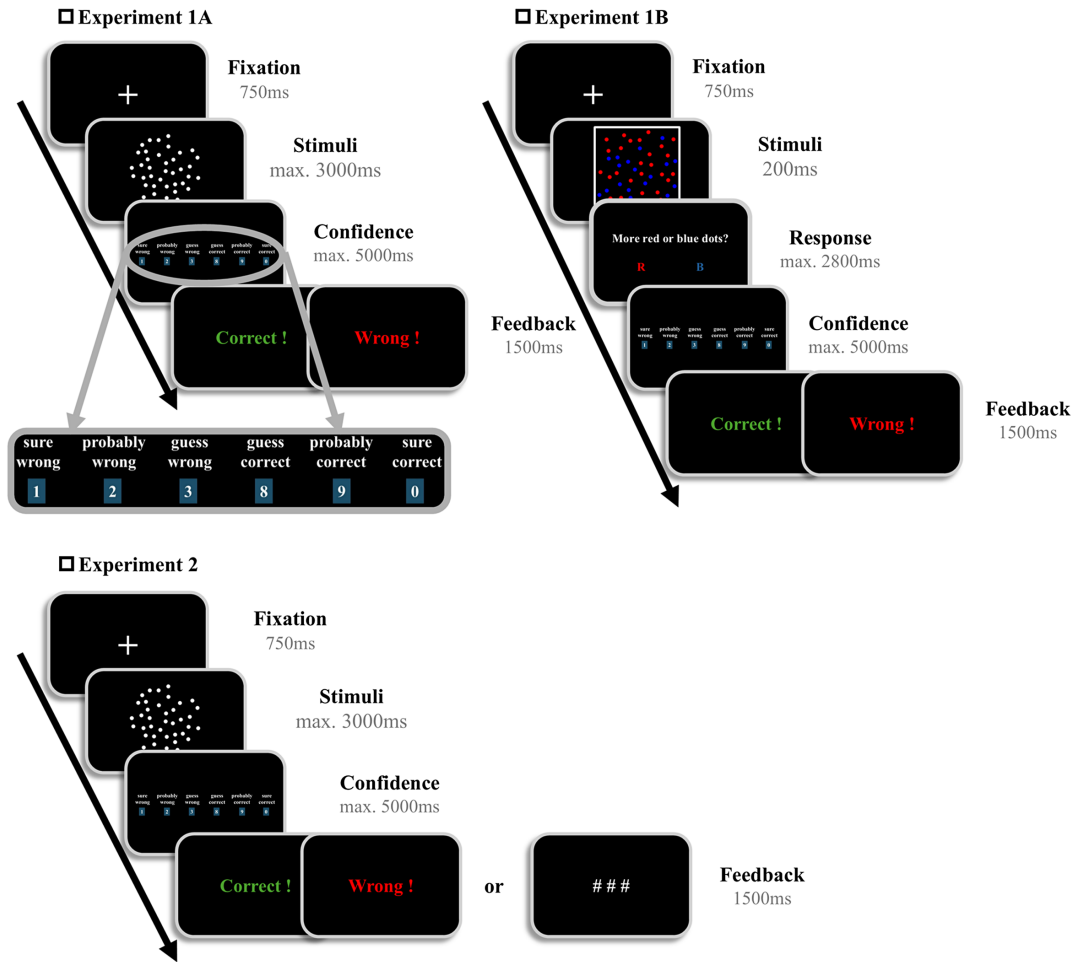
Data processing and model fitting were carried out in R (R Core Team, 2013), using the tidyverse (Wickham et al., 2019) and lme4 (Bates et al., 2015) packages. The first trial of each block and “too fast” trials (RT < 0.2 s) was excluded from the data set (3.16%). The “too slow” trials (>3 s) were aborted and not included in the original data set. RT and confidence-level data (only correct trials) were analyzed using linear mixed models after RT was log-transformed

and mean-centered. Accuracy data were analyzed using generalized linear mixed model with a logit link function. All models included at least a random intercept per participant, and three factors and their interactions as fixed effects, unless otherwise specified. Starting from this baseline model, we extended the random effects structure by adding random slopes for the fixed-effect predictors in order to improve model fit. Random slopes were added one at a time, following a consistent order across models, until the addition of random slopes led to a nonsignificant decrease in model fit or until the model was too complex to converge. The final models we used for all analyses (Experiments 1A, 1B, and 2) can be found in the Supplemental Materials. In order to assess effects of each factor of interest on our measures for the models, we calculated the *p* value based on the chi-square estimate using the Wald test from the car package (Fox & Weisberg, 2018). In addition to RT, accuracy, and confidence ratings, we also ran mixed model analyses on confidence RT, which are reported in the Supplemental Materials.

Drift Diffusion Model With Postdecision Confidence

To acquire further insight into the computational mechanisms underlying metacognitive conflict-induced strategy changes, we fitted our data using the flexible confidence boundary model (Herregods et al., 2025). This model simulates a random walk process that starts at z^*a (z was an unbiased starting point of .5) and terminates once the accumulated evidence reaches one of two opposing decision boundaries (0 or a). The strength of evidence accumulated is controlled by a drift rate v , nondecision components are captured by ter . After the accumulated evidence reached 0 or a (and a choice was made), evidence continued to accumulate until one of two opposing confidence boundaries (with the distance between both confidence boundaries determined by parameter a_2) was reached. These confidence boundaries slowly collapse over time, with the strength of the collapse being under the control of separate parameters for the upper (u_{upper}) and lower (u_{lower}) confidence boundaries. This flexibility of having independent collapse rates for both confidence boundaries allows the model to flexibly compute confidence depending on whether the confidence scale runs from “guess correct” to “sure correct” or from “guess wrong” to “sure wrong” or depending on how accurate reporting of confidence is rewarded (for a more detailed treatment, see Herregods et al., 2024). By dividing the space in between the two confidence boundaries into six categories, the model produced a level of confidence between 1 and 6 depending on the state of the accumulated evidence. The onset of the postdecision evidence accumulation can be biased, and this is captured by a free parameter controlling potential biases toward one of the confidence boundaries (postdecision starting point bias, z_2). The strength of postdecision accumulation was controlled by v_2 . By separating the decision drift rate from the postdecision drift rate, the model could account for individual differences in metacognitive accuracy (Desender et al., 2022). Finally, ter_2 captures non-accumulation-related processes in confidence.

To better understand the computational mechanisms of metacognitive conflict, per participant we fitted the data separately for three types of trials, categorized based on the preceding trial: (a) trials following a no-metacognitive conflict trial (i.e., feedback and confidence were congruent), (b) trials where the previous trial involved metacognitive conflict resulting from a PPE (confidence

Figure 1*Schematic of the Perceptual Decision-Making Task With Confidence Reports and Feedback*

Note. Each trial started with a fixation cross for 750 ms, followed by a cloud of moving dots (Experiment 1A, Experiment 2) or a box containing a mix of red and blue dots (Experiment 1B). Moving dots were presented until participants entered their response (up to 3,000 ms). The box containing red and blue dots was presented for 200 ms, after which participants had an additional 2,800 ms to respond. After their perceptual decisions, observers were cued to make a confidence rating on a 6-point scale (within 5,000 ms). In Experiments 1A and 1B, evaluative feedback (either negative or positive) was provided after the confidence rating on each trial. In Experiment 2, either evaluative feedback or neutral feedback was presented, which allowed to manipulate the proportion of metacognitive conflict. max. = maximum. See the online article for the color version of this figure.

rating of 1, 2, or 3, but the response was in fact correct), and (c) trials where the previous trial involved metacognitive conflict resulting from a NPE (confidence rating of 4, 5, or 6, but the response was in fact incorrect). Data were modeled separately for NPE and PPE metacognitive conflict to assure that findings are specific to metacognitive conflict (i.e., occurring for both NPE and PPE) and do not reflect, for example, PES in disguise (i.e., only occurring for NPE). In order to obtain stable parameter estimates, we only fitted the model if there were a minimum of at least 30 trials for each trial type, which was the case for $N = 43$ participants. In total, this resulted on average in 482.2 trials ($SD = 54.7$, range = 369–587) for the no-metacognitive conflict condition, 80.7 trials ($SD = 40.3$, range = 34–206) for the PPE metacognitive conflict condition, and 133.5 trials ($SD = 44.4$, range = 53–237) for the NPE condition.

Model parameters were estimated separately for each participant and within each metacognitive conflict condition (NPE metacognitive conflict, PPE metacognitive conflict, and no-metacognitive conflict) by jointly minimizing a composite cost function based on quantiles of the RT and confidence RT distributions (Ratcliff & Tuerlinckx, 2002). RT quantiles were computed separately for correct and error trials in both the observed and simulated data. Critically, the observed interquantile proportions were weighted by the relative frequencies of correct and error responses. Specifically, for each response category (correct or error), the interquantile proportions $[.1, .2, .2, .2, .2, .1]$ were multiplied by the proportion of trials belonging to that category (e.g., if 70% of trials were correct, the correct interquantile proportions became $0.7 \times [.1, .2, .2, .2, .2, .1]$). Confidence RT quantiles were computed separately for correct

and error trials and for each of the six confidence choices in the same weighting procedure. The resulting parts of the cost function are shown in the following equations:

$$MSE_{RT} = \frac{\sum (oRT_{i,q} - sRT_{i,q})^2}{sRT_{i,q}}. \quad (1)$$

$$MSE_{RTconf1} = \frac{\sum (oRTconf_{i,q} - sRTconf_{i,q})^2}{sRTconf_{i,q}}. \quad (2)$$

$$MSE_{RTconf2} = \frac{\sum (oRTconf_{i,cj} - sRTconf_{i,cj})^2}{sRTconf_{i,cj}}. \quad (3)$$

Here, oRT and sRT refer to the observed and simulated RT proportions, while oRTconf and sRTconf refer to the observed and simulated confidence RT proportions. These are calculated across multiple quantiles (q : .1, .3, .5, .7, and .9) for correct and error trials (i), also for all levels of confidence (cj) separately. We minimized the sum of the error functions given by Equations 1–3 using differential evolution optimization, as implemented in the DEoptim package (Mullen et al., 2011), to estimate all parameters. The number of iterations was set to 500, and the number of simulated trials was set to 1,000. To assess model fits, we simulated choices, RTs, confidence choices, and confidence RTs from the estimated parameters. To visualize model fits, we simulated decisions, RTs, and confidence from the estimated parameters. Full details about the model and fitting approach can be found in Herregods et al. (2025).

Transparency and Openness

We report all manipulations, all data exclusions, and all measures in the study. All data and analysis code for all three experiments are available on the Open Science Framework (Xing & Desender, 2026).

Results

Behavioral Results

Confidence. We submitted confidence reports on the current trial to a full-factorial linear mixed effects model with current coherence, previous confidence, and previous accuracy as predictors. The main effect of coherence reached significance (see Supplemental Figure S3), $\chi^2(2) = 2142.30$, $p < .001$, reflecting higher confidence with increasing motion coherence (10% coherence: $M = 4.38$; 20% coherence: $M = 4.62$; 40% coherence: $M = 4.98$). The results also showed a significant main effect of previous confidence, $\chi^2(1) = 146.29$, $p < .001$, with higher confidence on the current trial after participants reported higher confidence on the previous trial. Also, the main effect of previous feedback was significant, $\chi^2(1) = 36.10$, $p < .001$, with higher confidence after participants received correct compared to wrong feedback. Most importantly, the two-way interaction between previous feedback and previous confidence was significant, $\chi^2(1) = 85.42$, $p < .001$, reflecting a metacognitive conflict effect (see Figure 2A). Follow-up analyses showed that following NPE metacognitive conflict (i.e., confidence ≥ 4 followed by incorrect feedback), confidence was reduced compared to when the previous decision was correct, $\chi^2(1) = 151.26$, $p < .001$. Although confidence was also numerically lower following a PPE metacognitive conflict (i.e., confidence ≤ 3 followed by correct feedback) compared to when the previous decision was incorrect, this effect was not significant,

$\chi^2(1) < 0.01$, $p = .976$. Whereas the current analyses only report the influence of previous-trial metacognitive conflict, see the Supplemental Materials for an analysis of this effect up to seven trials back. There was also a two-way interaction between previous confidence and coherence, $\chi^2(2) = 15.02$, $p < .001$. Finally, although there was a significant three-way interaction between previous feedback, previous confidence level, and coherence ($p = .017$), the metacognitive conflict effect described earlier remained significant across all three coherence levels, $ps < .001$.

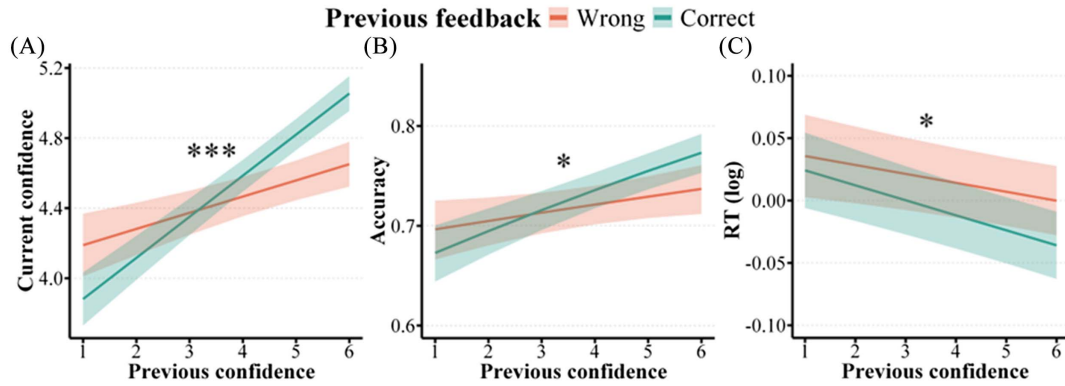
Accuracy. The main effect of coherence was significant (see Supplemental Figure S3), $\chi^2(2) = 1983.53$, $p < .001$, reflecting greater accuracy with increasing motion coherence (10% coherence: $M = 0.61$; 20%: $M = 0.71$; 40%: $M = 0.84$). The main effect of previous feedback also reached significance, $\chi^2(1) = 8.52$, $p = .004$, with greater accuracy after participants received correct compared to wrong feedback. There was also a main effect of previous confidence level, $\chi^2(1) = 26.19$, $p < .001$, with greater accuracy after higher confidence report. Most importantly, the two-way interaction between previous feedback and previous confidence level was significant, $\chi^2(1) = 6.35$, $p = .012$, indicating an effect of metacognitive conflict (see Figure 2B). To dissect this interaction, we estimated separate models for trials where participants thought they made a mistake (previous confidence: 1, 2, and 3) and trials where participants thought they were correct (previous confidence: 4, 5, and 6). This revealed that following NPE metacognitive conflict (i.e., confidence ≥ 4 followed by incorrect feedback), accuracy was reduced compared to when the previous decision was correct, $\chi^2(1) = 23.27$, $p < .001$. Although accuracy was also numerically lower following a PPE metacognitive conflict (i.e., confidence ≤ 3 followed by correct feedback) compared to when the previous decision was incorrect, this effect was not significant, $\chi^2(1) = 1.42$, $p = .234$. The results also showed a significant two-way interaction between coherence and previous confidence level, $\chi^2(2) = 20.06$, $p < .001$. Additionally, the three-way interaction effect between previous feedback, previous confidence level, and coherence was also significant, $\chi^2(2) = 13.34$, $p = .001$. This was explained by the two-way interaction between previous confidence level and previous feedback only reaching significance in the low and high difficulty conditions ($ps < .028$), not in the medium difficulty conditions ($p = .649$).

Choice RT. The main effect of coherence reached significance (see Supplemental Figure S3), $\chi^2(2) = 1339.43$, $p < .001$, reflecting shorter choice RT with increasing motion coherence (10% coherence: $M = 1.11$ s; 20% coherence: $M = 1.03$ s; 40% coherence: $M = 0.93$ s). There was also a significant main effect of previous confidence level, $\chi^2(1) = 32.80$, $p < .001$, with shorter choice RT after participants reported higher confidence level on the previous trial. Also, the main effect of previous feedback was significant, $\chi^2(1) = 74.81$, $p < .001$, with shorter choice RT after participants received correct compared to wrong feedback. Critically, the two-way interaction between previous feedback and previous confidence level was significant, $\chi^2(1) = 4.61$, $p = .032$, showing a metacognitive conflict effect (see Figure 2C). Follow-up analysis revealed that following NPE metacognitive conflict (i.e., confidence ≥ 4 followed by incorrect feedback), choice RT was longer compared to when the previous decision was correct, $\chi^2(1) = 89.98$, $p < .001$. However, the choice RT was shorter following a PPE metacognitive conflict (i.e., confidence ≤ 3 followed by correct feedback) compared to when the previous decision was incorrect, $\chi^2(1) = 4.51$, $p = .034$. The two-way

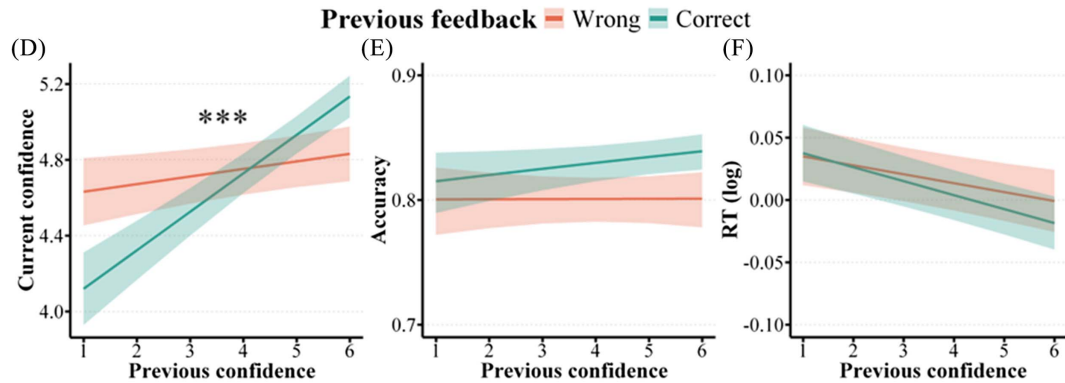
Figure 2

Metacognitive Conflict Between Internal and External Feedback Influences Subsequent Confidence and Cognitive Performance

□ **Experiment 1A**



□ **Experiment 1B**



Note. In Experiment 1A, the two-way interaction effect between previous confidence level and previous feedback was significant for all three dependent variables (accuracy, confidence, and RT), with lower confidence and accuracy after metacognitive conflict (A–C). In Experiment 1B, the two-way interaction effect between the previous confidence level and previous feedback (i.e., showing metacognitive conflict) was significant only for confidence (D–F). The error bars represent the 95% confidence intervals. RT = reaction time. See the online article for the color version of this figure.

* $p < .05$. *** $p < .001$.

interaction between previous confidence and coherence also reached significance, $\chi^2(2) = 16.41$, $p < .001$.

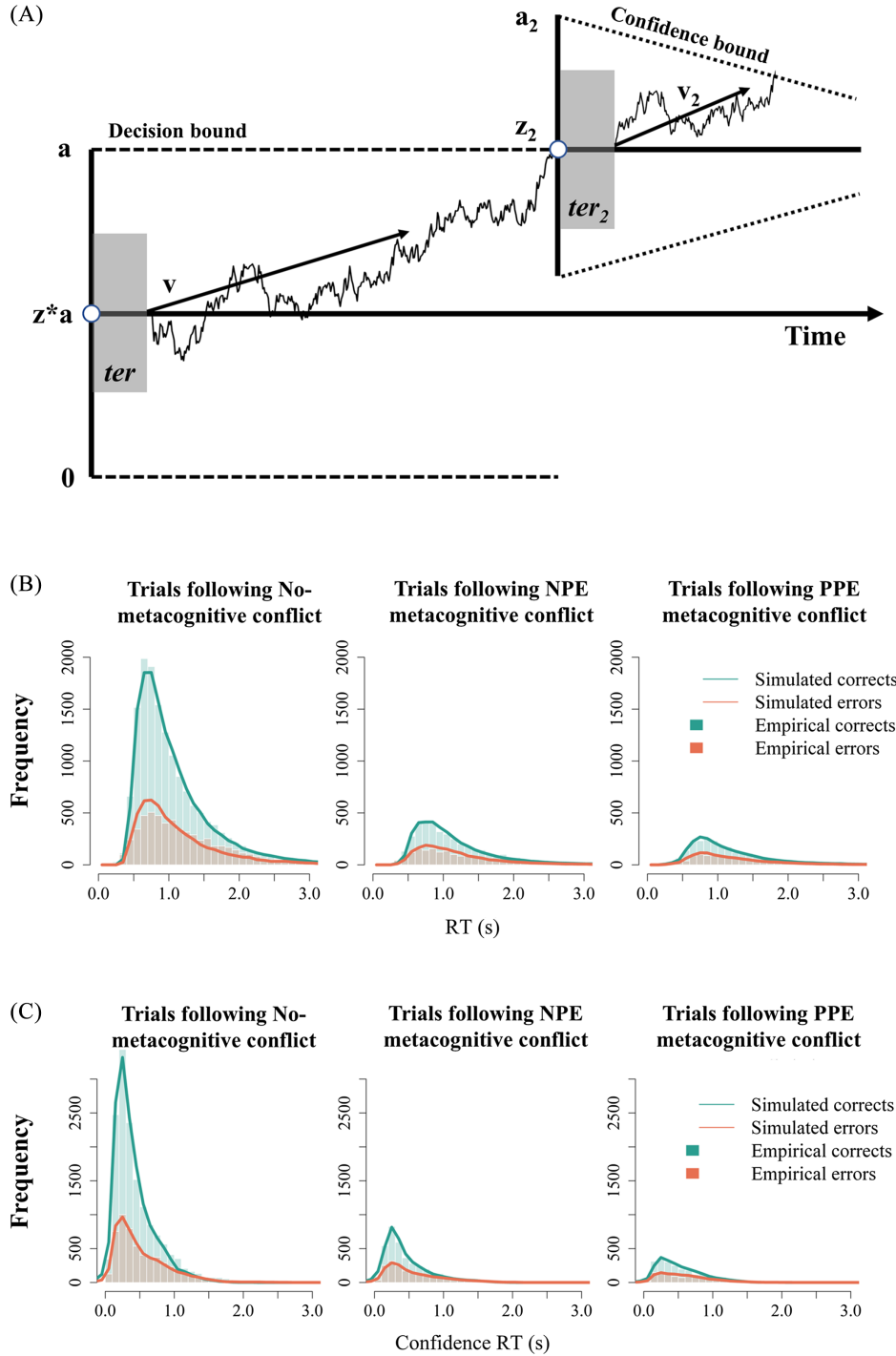
Drift Diffusion Model With Postdecision Confidence

Model Fit. Behavioral results showed that participants adapt cognitive and metacognitive behavior after experiencing a metacognitive conflict. Most prominently, confidence was reduced after a metacognitive conflict trial. Importantly, these behavioral findings do not reveal the mechanisms underlying this effect. Therefore, we fitted our data using a drift diffusion model with postdecision evidence accumulation, which informs confidence (i.e., as implemented in the flexible confidence boundary model by Herregods et al., 2025). Following the accumulation-to-bound process for decisions, evidence accumulation persists and terminates upon reaching one of two confidence boundaries (see Figure 3A). Confidence is determined by the

level of accumulated evidence at the time the confidence boundary is reached. As can be seen in Figure 3B, this model provided a good fit in explaining decisions, RTs, and confidence. Mean square error values for each condition are also reported in the Supplemental Materials.

Model Parameters. Having confirmed that the drift diffusion model with postdecision confidence by Herregods et al. (2025) provided a good fit for our data, we proceeded to investigate which estimated parameters varied as a function of metacognitive conflict. To this end, we performed repeated-measures analysis of variance to explore whether the metacognitive conflict on the previous trial led to changes in current trial parameters. Furthermore, we applied Bonferroni correction for multiple comparisons. To account for seven core model parameters (a , a_2 , ter , ter_2 , z_2 , v , v_2), the significance threshold α was adjusted to $.05/7 \approx .007$. First, we found a significant effect of previous conflict type on the estimated starting point for postdecision accumulation, $F(2, 84) = 15.59$, $p < .001$ (see

Figure 3
Drift Diffusion Model With Postdecision Confidence



Note. Panel A shows the drift diffusion model with postdecision confidence (Herregods et al., 2025). Evidence accumulates departing from starting point z^*a until it reaches one of two opposing boundaries (0 or a). The strength of the accumulation process is captured by the drift rate (v). After the accumulated evidence reaches a decision boundary, evidence continues to accumulate until one of the two confidence boundaries is reached (the distance between both is determined by parameter a_2), which slowly collapse over time (controlled by parameters u_{upper} and u_{lower}). The postdecision accumulation process can have its own biased starting point z_2 . The strength of postdecision accumulation is controlled by v_2 . Non-accumulation-related

(figure continues)

Figure 4A). Post hoc tests confirmed that when the previous trial contained PPE or NPE metacognitive conflict, the current post-decision starting point was lower relative to when there was no metacognitive conflict on the previous trial, $t_s \geq 4.53$, $p_s \leq .001$. There was no significant difference between the two metacognitive conflict types, $t = 1.37$, $p = .177$. This finding effectively shows that whenever participants experienced metacognitive conflict (regardless of whether this was a PPE or NPE metacognitive conflict), on the next trial, they implemented a biased starting point for post-decision accumulation, starting closer to (and thus making it more likely to terminate at) the confidence boundary associated with lower confidence. Second, we found significant effects of previous metacognitive conflict type on the estimated current trial nondecision time (ter), $F(2, 84) = 6.11$, $p = .007$. This finding indicates that participants spend more time on processes unrelated to the decision (e.g., stimulus encoding, pressing the choice button) for the perceptual decision following metacognitive conflict. Post hoc tests showed that nondecision time was longer following both PPE and NPE conflicts relative to no-metacognitive conflict trials ($t_s \geq 2.89$, $p_s \leq .012$). Furthermore, there was no significant difference between the two types of metacognitive conflict (PPE vs. NPE metacognitive conflict), $t = 1.09$, $p = .284$. None of the other parameters met the Bonferroni-corrected threshold, $p_s > .018$.

Experiment 1B

The aim of Experiment 1B was to replicate the findings of Experiment 1A in a preregistered approach (Xing & Desender, 2024).

Method

Participants

Fifty-one participants (three men, 48 women; age: $M = 18.6$, $SD = 2.62$, range = 17–35) took part in return for course credit. All participants reported normal or corrected-to-normal vision. All but five participants were right-handed. Two participants were excluded because they gave the same confidence rating in more than 90% of the trials. Additionally, although not preregistered, two participants were excluded because they made more than 50% metacognitive errors. Participants provided written informed consent before participation. All procedures were approved by the KU Leuven Ethics Committee.

Stimuli and Apparatus

Stimuli were presented in white on a black background on a 24-in. LCD monitor using Psychopy (Peirce et al., 2019) for Python. Eighty colored dots were drawn in a white square, with random positions generated on each trial. Trial difficulty was controlled by manipulating the difference in number of trials between both boxes,

which were either 44 versus 36, 47 versus 33, or 53 versus 27 during the main experiment.

Statistical Analysis

Data preprocessing followed the same procedures as in Experiment 1A, with 2.11% of the trials excluded from the data set.

Procedure

The procedure was identical to Experiment 1A, except that the task was changed to a color discrimination task (i.e., instead of the dot motion task). On each trial, a box filled with 80 colored dots was presented, and participants indicated whether there were more blue or red dots. The dot color stimulus was displayed for 200 ms, after which participants could make their choice. If participants did not respond within 2,800 ms, the trial was aborted and a warning message appeared instructing participants to respond more quickly on the next trial.

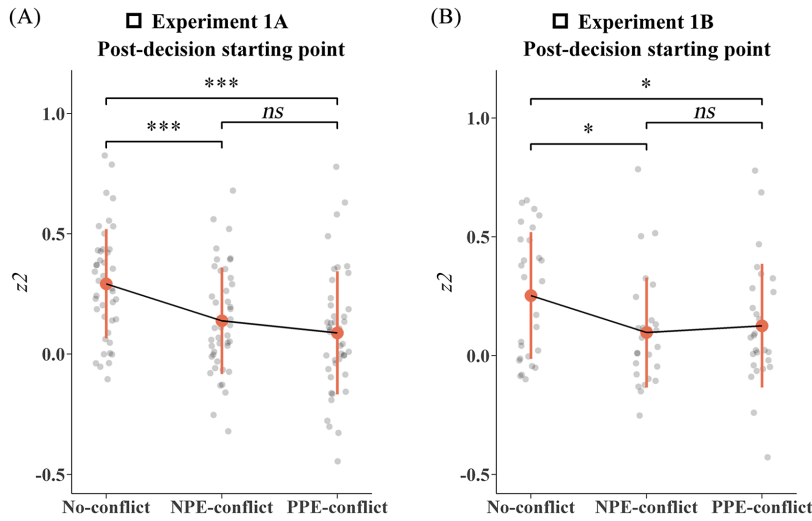
Results

Behavioral Results

Confidence Level. We replicated the metacognitive conflict effect, as the two-way interaction between previous feedback and previous confidence level was significant, $\chi^2(1) = 68.80$, $p < .001$, reflecting a strong metacognitive conflict effect (see Figure 2D). Follow-up analysis showed that after perceived correct trials (previous confidence rating: 4, 5, or 6), there was a significant main effect of previous feedback, with higher confidence after participants received correct versus wrong feedback, $\chi^2(1) = 52.93$, $p < .001$, whereas this effect was reversed for the perceived error trials (previous confidence rating: 1, 2, or 3), $\chi^2(1) = 3.89$, $p = .048$. Additionally, there was a significant main effect of difficulty (see Supplemental Figure S4), $\chi^2(2) = 1281.34$, $p < .001$, reflecting lower confidence with higher difficulty levels (high difficulty: $M = 4.64$; medium difficulty: $M = 4.84$; low difficulty: $M = 5.13$). The main effect of the previous confidence level was also significant, $\chi^2(1) = 24.63$, $p < .001$, with higher confidence on the current trial after participants reported higher confidence on the previous trial. There was also a significant two-way interaction between previous feedback and trial difficulty, $\chi^2(1) = 9.00$, $p = .011$. The two-way interaction between previous confidence level and trial difficulty was also significant, $\chi^2(2) = 34.03$, $p < .001$.

Accuracy. We did not replicate the metacognitive conflict effect for accuracy; the two-way interaction effect between previous feedback and previous confidence level was not significant, $p = .394$. There was a significant main effect of difficulty (see Supplemental Figure S4), $\chi^2(2) = 1,598.44$, $p < .001$, reflecting lower accuracy

Figure 3 Note (continued). components are captured by ter and ter_2 . Panel B shows the model fits of RT in Experiment 1A. Simulated and empirical RTs separately for correct and error trials following no/NPE/PPE metacognitive conflict. Correct/error responses correspond to the accuracy of the current trial. Panel C shows the model fits of confidence RT fits in Experiment 1A. Simulated and empirical confidence RTs separately for correct and error trials following no/NPE/PPE metacognitive conflict. Inspection of the model fit reveals that our model accurately captures the distributions in RTs and confidence RTs seen in the data across all three metacognitive conflict conditions. PPE = positive prediction error; NPE = negative prediction error; RT = reaction time; ter = nondecision time. See the online article for the color version of this figure.

Figure 4*Adaptation of Postdecision Starting Point Following Metacognitive Conflict*

Note. The figure shows the computational modeling results for Experiments 1A and 1B. Participants adapted postdecision parameters following metacognitive conflict. The postdecision starting point was higher following metacognitive conflict trials compared to no-metacognitive conflict trials. All results adhere to a strict Bonferroni-corrected threshold. The error bars represent the standard deviation. PPE = positive prediction error; NPE = negative prediction error; *ns* = not significant. See the online article for the color version of this figure.

* $p < .05$. *** $p < .001$.

with higher difficulty level (high difficulty: $M = 0.69$; medium difficulty: $M = 0.80$; low difficulty: $M = 0.91$). The main effect of previous feedback was also significant, $\chi^2(1) = 21.18$, $p < .001$, with greater accuracy after participants received correct feedback than wrong feedback. Also, the two-way interaction effect between previous feedback and difficulty reached significance, $\chi^2(2) = 15.18$, $p < .001$.

Choice RT. We did not replicate the metacognitive conflict effect for RTs; the two-way interaction effect between previous feedback and previous confidence level was not significant, $p = .098$. There was a significant main effect of difficulty (see Supplemental Figure S4), $\chi^2(2) = 488.96$, $p < .001$, reflecting longer choice RT when trial difficulty was higher (high difficulty: $M = 0.86$ s; medium difficulty: $M = 0.82$ s; low difficulty: $M = 0.76$ s). The main effect of the previous confidence level also reached significance, $\chi^2(1) = 38.34$, $p < .001$, with shorter RT after participants reported higher confidence on the previous trial. The two-way interaction effect between previous feedback and difficulty was significant, $\chi^2(2) = 9.95$, $p = .007$.

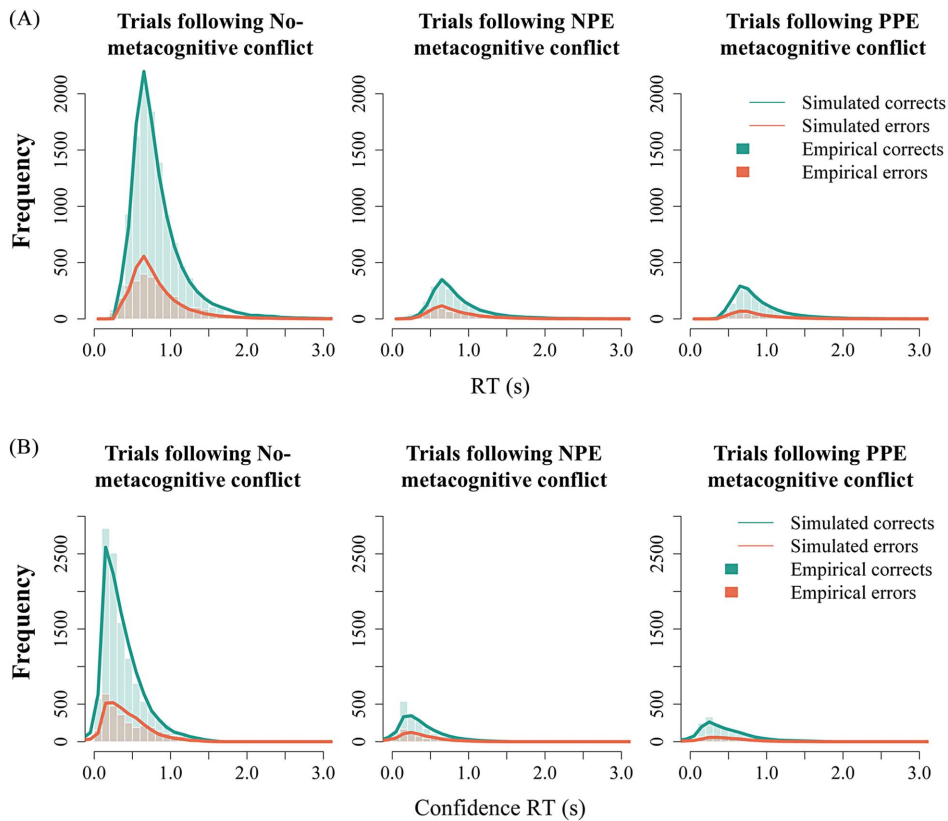
Drift Diffusion Model With Postdecision Confidence

As planned, we again fitted the drift diffusion model with postdecision accumulation for confidence to the data of Experiment 1B. However, despite using the same implementation as for Experiment 1A, the model fits failed to properly converge for these data (both at 500 iterations, as in 1A, as well as when increasing the total iterations to 2,000). One reason for this could be that the number of trials was somewhat lower compared to Experiment 1A. In total, on average,

we retained 537.3 trials ($SD = 58.6$, range = 425–608) for the no-metacognitive conflict condition, 73.1 trials ($SD = 45.1$, range = 32–179) for the PPE conflict condition, and 93.9 trials ($SD = 40.2$, range = 42–197) for the NPE conflict condition. To ensure proper model convergence, we decided to fix the urgency parameters of the confidence boundaries, u_{upper} and u_{lower} , to the mean value obtained from fitting Experiment 1A. Having two fewer free parameters effectively solved the convergence issues encountered when fitting the full model. As shown in Figure 5, this simplified model provided a good fit for decisions, RTs, and confidence. We then performed repeated-measures analysis of variance and applied Bonferroni correction for multiple comparisons, similar to Experiment 1A, to examine whether metacognitive conflict on the previous trial influenced current trial parameters. In general, the results of Experiment 1B replicated the findings of Experiment 1A, demonstrating that metacognitive conflict on the previous trial biased the starting point of postdecision evidence accumulation toward the lower confidence boundary, reflecting a more cautious confidence reporting strategy.

Specifically, consistent with Experiment 1A, we found a significant effect of previous conflict type on the estimated starting point for postdecision accumulation, $F(2, 52) = 5.82$, $p = .005$ (see Figure 4B). Post hoc tests showed that both when the previous trial contained a PPE or a NPE metacognitive conflict, the current postdecision starting point was lower relative to when there was no metacognitive conflict on the previous trial, $t_s \geq 2.70$, $p_s \leq .029$. Additionally, no significant difference was observed between PPE and NPE metacognitive conflict trials, $t = 0.68$, $p = .501$. There were no significant effects for any of the other parameters, $p_s \geq .018$.

Figure 5
Flexible Confidence Boundary Fits for Experiment 1B



Note. Panel A shows the model fits of RT in Experiment 1B. Simulated and empirical RTs separately for correct and error trials following no/NPE/PPE metacognitive conflict. Correct/error responses correspond to the accuracy of the current trial. Panel B shows the model fits of confidence RT in Experiment 1B. Simulated and empirical confidence RTs separately for correct and error trials following no/NPE/PPE metacognitive conflict. Inspection of the model fit reveals that our model accurately captures the distributions in RTs and confidence RTs seen in the data across all three metacognitive conflict conditions. PPE = positive prediction error; NPE = negative prediction error; RT = reaction time. See the online article for the color version of this figure.

Interim Discussion

In two experiments, we demonstrated that participants exhibit trial-by-trial adaptations in response to metacognitive conflict. In Experiments 1A and 1B, we observed a strong influence of metacognitive conflict on subsequent confidence reports: Participants reported lower confidence following metacognitive conflict trials. While Experiment 1A revealed subtle effects of metacognitive conflict on decision RT and accuracy, these patterns were not replicated in Experiment 1B, suggesting that metacognitive conflict predominantly affects the subsequent computation of confidence rather than decision accuracy. Results from computational modeling revealed that this was mostly explained by a downward shift in the postdecision starting point following metacognitive conflict, making participants more conservative in their confidence judgments. Having established that participants strategically adapt to metacognitive conflict at the trial-by-trial level, we next investigated whether a similar effect emerges at the block level (Ridderinkhof, 2002). To address this, we created high metacognitive conflict (HMC) blocks and low

metacognitive conflict (LMC) blocks and tested their influence on both cognitive performance and metacognitive monitoring.

Experiment 2

Method

Participants

Fifty-four participants (14 men, 40 women; age: $M = 18.3$, $SD = 3.08$, range = 18–29) took part in return for course credit. All participants reported normal or corrected-to-normal vision. All but nine participants were right-handed. Six participants were excluded because their performance was not different from the chance level (as assessed by a binomial test). One more participant was excluded because she gave the same confidence rating in more than 90% of the trials. Participants provided written informed consent before participation. All procedures were approved by the KU Leuven Ethics Committee.

Procedure

Stimuli and procedure were similar to Experiment 1A (i.e., a dot motion task), with the following exception: Instead of providing evaluative feedback (correct or wrong) on every trial, we also presented neutral feedback (three hashtags, “###”) on some trials (see Figure 2). Because of these neutral feedback trials, it is possible to manipulate the level of metacognitive conflict by adjusting the ratio of neutral to evaluative feedback. To examine whether the metacognitive effect varies across different levels of conflict, we implemented two conditions: HMC and LMC. The formal phase of the experiment consisted of six HMC blocks and six LMC blocks. The block type alternated every three blocks and was counter-balanced across participants. Given that the average proportion of metacognitive conflict trials in Experiment 1A was 30%, we accordingly preset the ratio of evaluative feedback in the HMC and LMC conditions of Experiment 2, so that the overall proportion of evaluative feedback would be balanced across the two conditions (see Table 1). Specifically, in the HMC blocks, participants received evaluative feedback on 90% of metacognitive conflict trials and 10% of no-metacognitive conflict trials. In the remaining trials within these blocks, participants received neutral feedback. In the LMC blocks, on 30% of metacognitive conflict trials and on 40% of no-metacognitive conflict trials, participants received evaluative feedback, whereas on the remainder, they received neutral feedback. In this design, the actual proportion of evaluative feedback in HMC and LMC blocks was effectively balanced at 37.7% and 33.7%, respectively. To rule out the possibility that participants interpreted neutral feedback as correct, in the Supplemental Materials, we report an additional analysis showing that the effect of previous-trial neutral feedback on current-trial confidence differs from the effect of evaluative feedback (correct or wrong).

Statistical Analysis

Data processing was like Experiment 1A, with 3.71% of the trials removed from the data set. To examine the block effects induced by metacognitive conflict, the data analysis in Experiment 2 was conducted separately for trials where the preceding feedback was either neutral or evaluative. Furthermore, a block number variable was included to account for the sequence of the three blocks within each condition (HMC or LMC) and account for the fact that the experiment alternated between both types of blocks. For the trials following neutral feedback, three factors—coherence, block type, and block number—and their interactions were included as fixed effects. For the trials following evaluative feedback, three factors—previous feedback, previous confidence level, and block type—and their interactions were included as fixed effects (coherence was excluded from the model to ensure sufficient statistical power for the

current analysis). Model fitting was consistent with those in Experiment 1, unless otherwise specified.

Results

Trials Following Neutral Feedback

To investigate metacognitive conflict block effects, we analyzed trials preceded by neutral feedback to assess whether performance systematically differed across experimental blocks (see Figure 6A–6C).

Confidence. The main effect of block type was not significant, $p = .082$. The main effect of coherence was significant (see Supplemental Figure S5), $\chi^2(2) = 1,048.51$, $p < .001$, reflecting higher confidence level with increasing motion coherence (10% coherence: $M = 4.19$; 20% coherence: $M = 4.39$; 40% coherence: $M = 4.78$). Most importantly, the two-way interaction between block type and block number was significant, $\chi^2(2) = 13.816$, $p = .001$. Follow-up analysis showed that there was a significant main effect of block type in the third block, with higher confidence for LMC blocks than HMC blocks (see Figure 6A), $\chi^2(1) = 13.760$, $p < .001$, whereas this main effect of block type was not significant in the first and second blocks, $ps \geq .143$.

Accuracy. The main effect of block type was not significant, $p = .712$. There was a significant main effect of coherence (see Supplemental Figure S5), $\chi^2(2) = 958.30$, $p < .001$, reflecting greater accuracy with increasing motion coherence (10% coherence: $M = 0.62$; 20% coherence: $M = 0.73$; 40% coherence: $M = 0.86$). No other effects reached statistical significance, $ps \geq .068$.

Choice RT. The main effect of block type was not significant, $p = .903$. The main effect of coherence reached significance (see Supplemental Figure S5), $\chi^2(2) = 657.06$, $p < .001$, reflecting shorter RT with increasing motion coherence (10% coherence: $M = 1.21$ s; 20% coherence: $M = 1.12$ s; 40% coherence: $M = 0.98$ s). The main effect of block number was significant, $\chi^2(2) = 46.44$, $p < .001$, with shorter RT for later blocks. Also, the two-way interaction between block number and block type was significant, $\chi^2(2) = 13.823$, $p = .001$. However, follow-up analysis revealed that the simple main effect of block type was not significant within any of the three blocks, $ps \geq .180$.

Trials Following Evaluative Feedback

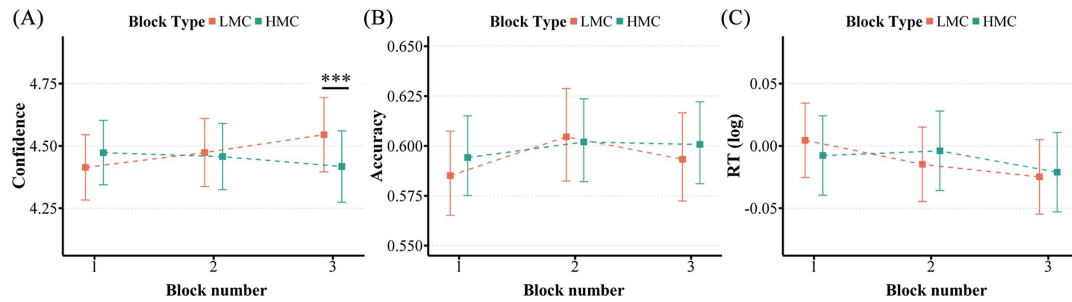
Next, for completeness, we tested whether the trial-by-trial effects observed in Experiments 1A and 1B could be replicated in the subset of trials where participants did receive evaluative feedback on the previous trial. Furthermore, we tested whether these effects were modulated by the overall proportion of metacognitive conflict (i.e., HMC vs. LMC conditions).

Table 1
The Predicted Ratio of Evaluative Feedback

Trial type	HMC condition	LMC condition
Metacognitive conflict trials (30%)	90% $\times$ 30% = 27%	30% $\times$ 30% = 9%
No-metacognitive conflict trials (70%)	10% $\times$ 70% = 7%	40% $\times$ 70% = 28%
Total	34%	37%

Note. Values in bold indicate the proportion of trials receiving evaluative feedback. HMC = high metacognitive conflict; LMC = low metacognitive conflict.

Figure 6
Mean Confidence, Accuracy, and RT Across LMC and HMC Blocks



Note. Mean confidence (A), accuracy (B), and RT (C) were shown for LMC and HMC across three blocks. No significant differences were observed between the two block types across these measures. However, there was a significant main effect of block type in the third block, with higher confidence for LMC blocks than HMC blocks. The error bars represent the 95% confidence intervals. HMC = high metacognitive conflict; LMC = low metacognitive conflict; RT = reaction time. See the online article for the color version of this figure.

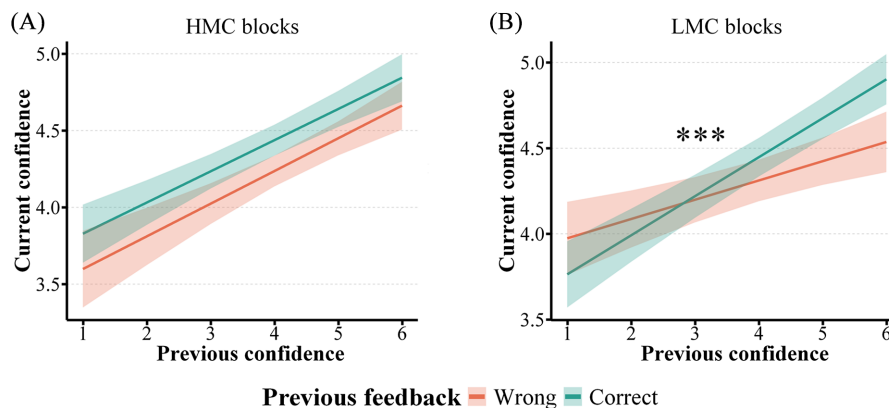
*** $p < .001$.

Confidence. The main effect of previous confidence level reached significance, $\chi^2(1) = 81.78, p < .001$, with higher confidence on the current trial after participants reported higher confidence on the previous trial. Also, the main effect of previous feedback was significant, $\chi^2(1) = 48.56, p < .001$, with higher confidence after participants received correct compared to wrong feedback. Crucially, there was a two-way interaction between previous feedback and previous confidence level, $\chi^2(2) = 8.93, p = .003$, reflecting a metacognitive conflict effect. Follow-up analyses showed that for the perceived correct trials (previous confidence rating: 4, 5, or 6), participants reported higher confidence after receiving correct versus wrong feedback, $\chi^2(1) = 43.50, p < .001$, whereas this effect was nonsignificantly reversed for perceived error trials (previous

confidence rating: 1, 2, or 3), $\chi^2(1) = 0.07, p = .785$. The three-way interaction between previous confidence level, previous feedback, and block type was also significant, $\chi^2(1) = 8.63, p = .003$, which reflected that the two-way interaction between previous confidence level and previous feedback only reached significance in the LMC blocks, $\chi^2(1) = 16.04, p < .001$, not in the HMC blocks, $\chi^2(1) = 0.08, p = .775$ (see Figure 7). These results show that the trial-by-trial adaptation of confidence following metacognitive conflict effectively depends on the frequency of metacognitive conflict in a block.

Accuracy. The main effect of the previous confidence level was significant, $\chi^2(1) = 9.30, p = .002$, with greater accuracy after higher confidence report. The main effect of previous feedback also reached significance, $\chi^2(1) = 22.63, p < .001$, with greater accuracy

Figure 7
The Two-Way Interaction Effects Between Previous Confidence and Previous Feedback on Confidence Reports



Note. The figure shows the two-way interaction effects between previous confidence and previous feedback on confidence reports for HMC blocks (A) and LMC blocks (B). The two-way interaction only reached significance in the LMC blocks. The error bars represent the 95% confidence intervals. HMC = high metacognitive conflict; LMC = low metacognitive conflict. See the online article for the color version of this figure.

*** $p < .001$.

after participants received correct compared to wrong feedback. The two-way interaction between previous feedback and previous confidence level was not significant, $p = .886$.

Choice RT. The main effect of previous feedback was significant, $\chi^2(1) = 23.54$, $p < .001$, with shorter choice RT after participants received correct compared to wrong feedback. The main effect of the previous confidence level was also significant, $\chi^2(1) = 3.95$, $p = .047$, with shorter choice RT after higher confidence report. Also, the two-way interaction between previous feedback and previous confidence level was significant, $\chi^2(1) = 4.18$, $p = .041$, reflecting a metacognitive conflict effect. The follow-up showed that the main effect of previous feedback reached significance for the perceived correct group, with shorter choice RT after participants received correct feedback than wrong feedback, $\chi^2(1) = 39.74$, $p < .001$. This effect was nonsignificantly reversed in the perceived error group, $\chi^2(1) = 0.84$, $p = .359$.

Discussion

The present study reports a series of three experiments investigating the cognitive and metacognitive consequences of metacognitive conflict both at the behavioral and at the computational levels. Our study showed that metacognitive conflict (the discrepancy between confidence and external feedback) elicits systematic metacognitive adjustments. Following metacognitive conflict, participants became less confident, which was explained by decreased postdecision starting points leading to more conservative confidence judgments. Furthermore, this effect was smaller in blocks with higher proportion of metacognitive conflict trials than blocks with a lower proportion, and participants generally reported lower confidence in blocks with a high proportion of metacognitive conflict trials. These findings suggest that discrepancies between confidence and feedback primarily disrupt subsequent metacognitive monitoring.

Metacognitive Conflict as a Prediction Error Signal

Behavioral analyses showed that after metacognitive conflict (receiving feedback inconsistent with the confidence reports), participants exhibited lower confidence in their decisions. These post-metacognitive conflict effects show a similar pattern to impairments in cognitive performance (e.g., reduced accuracy, slower responses) following cognitive conflict (Botvinick et al., 2001; Braem et al., 2014; Van der Borgh et al., 2014), suggesting partially similar mechanisms for behavioral adaptation between cognitive and metacognitive levels. To further examine this similarity, we take inspiration from the theoretical framework of the predicted response outcome model. The predicted response outcome model was inspired by the key role of the medial prefrontal cortex in a wide range of adaptive behaviors, such as error monitoring, conflict resolution, and error likelihood estimation (Brown & Braver, 2005; Ito et al., 2003; Matsumoto et al., 2007; Rudebeck et al., 2008). The predicted response outcome model posits that conflict effects arise from the simultaneous prediction of competing response-outcome contingencies, rather than direct response competition (Alexander & Brown, 2011, 2014). Specifically, cognitive conflict is resolved by suppressing inappropriate action plans before response execution, enabling adaptive adjustments in subsequent trials (Alexander & Brown, 2011; Silvetti et al., 2014). Analogously, we propose that metacognitive conflict may reflect a higher order prediction error: PPE metacognitive

conflict corresponds to unexpected correct feedback, while NPE metacognitive conflict corresponds to unexpected incorrect feedback. This correspondence raises the intriguing question whether also at the neural level, metacognitive conflict and cognitive conflict share similar neural signatures in the medial prefrontal cortex and especially the anterior cingulate. Apart from these similarities, however, it should also be noted that whereas previous-trial cognitive conflict is known to reliably affect current trial accuracy and current-trial RTs, in our study, previous-trial metacognitive conflict reliably affected current-trial confidence but its influence on confidence RTs was less consistent across the three studies. Thus, this further calls for additional work clarifying the degree of overlap between cognitive and metacognitive conflict adaptation.

To acquire more insight into the underlying computational processes implementing changes triggered by metacognitive conflict, we relied on an evidence accumulation framework that accounts for both decisions and confidence reports (Herregods et al., 2025). The computational findings showed that participants adopted biased postdecision starting points following metacognitive conflict, which effectively made them more conservative in their reporting of confidence (i.e., requiring more postdecision evidence before giving confidence judgment). In addition, there was some modest evidence that metacognitive conflict induced prolonged nondecision times, although this effect did not survive Bonferroni correction.

Distinct Metacognitive Strategies in HMC and LMC

As shown in Experiments 1A and 1B, metacognitive conflict elicited robust trial-level adjustments. Similarly, we observed block-level adaptation effects in Experiment 2. The metacognitive conflict effect we found in Experiments 1A and 1B varied as a function of the proportion of metacognitive conflict trials. The effect was smaller in blocks with a higher proportion of metacognitive conflict trials (HMC blocks) than blocks with a lower proportion (LMC blocks). Furthermore, the analysis restricted to trials following neutral feedback revealed a global adaptation of confidence, showing that participants generally reported higher confidence in LMC blocks than HMC blocks (though this effect was only observed in the third block, suggesting it takes time for this effect to emerge).

These observations likely reflect distinct information processing mechanisms across the two conditions (LMC vs. HMC). In the LMC condition, where metacognitive conflict was infrequent (only 30% of evaluative feedback triggering metacognitive conflict), participants reported significantly lower confidence following metacognitive conflict trials compared to no-metacognitive conflict trials, mirroring the pattern observed in Experiment 1A. This trial-by-trial adjustment likely occurs because infrequent events (i.e., metacognitive conflict) are known to elicit a strong orienting response (Notebaert et al., 2009). Consequently, participants tended to adjust their confidence dynamically, lowering their reported confidence after experiencing an infrequent metacognitive conflict event. In the HMC condition, however, where conflict was more frequent (90% of evaluative feedback trials were metacognitive conflict), participants appeared to shift toward a conservative confidence judgment strategy at the block level, generally reporting low confidence, but no longer rapidly adapting their confidence reporting at the trial level.

In selective attention paradigms (e.g., Stroop task), researchers have likewise found that the cognitive conflict effect is reduced in blocks where conflict is very prominent, which is referred to as the

PC effect (Lowe & Mitterer, 1982; Verguts & Notebaert, 2008). PC effects can be attributed to a global strategic adaptation in which PC determines the strategy (e.g., attention setting) to process task-relevant and task-irrelevant information in cognitive conflict tasks (Abrahamse et al., 2013; Torres-Quesada et al., 2013). Specifically, as the ratio of congruent trials to incongruent trials increases, attention to the relevant dimension can be attenuated such that the influence of the irrelevant dimension increases. Similarly, in the HMC condition of the present study, participants likely adopted a globally conservative bias after realizing the discrepancy between their confidence and feedback. Our findings suggest that confidence is modulated not only by trial-by-trial metacognitive conflict but also by contextual information, where infrequent metacognitive conflict leads to trial-by-trial confidence adjustment, but frequent metacognitive conflict drives a block-level bias toward low confidence (Lebreton et al., 2019; Van Marcke & Desender, 2025).

Constraints on Generality

The present findings are based on a sample of healthy university students. As such, the generalizability of the results may be limited to young adults with relatively high levels of education and typical cognitive functioning. Given the limitation of recruitment methods, these findings may not extend to populations from different age groups, those with lower educational attainment, or those with cognitive impairments. Further research is needed to determine whether the observed effects generalize to more diverse populations.

Conclusion

Together, our results extend the concept of cognitive conflict to the metacognitive level by demonstrating that metacognitive conflict leads to next-trial changes in the underlying computations of confidence, leading to lower reported confidence. Computational modeling further revealed that participants adopted a more conservative confidence-reporting strategy following metacognitive conflict, characterized by biased postdecision starting points. Crucially, this adjustment is not purely restricted to the trial level but is modulated by the proportion of metacognitive conflict in the task. While infrequent metacognitive conflict (LMC) triggers a trial-by-trial biased confidence report, HMC environments drive a global, block-level adaptation.

References

- Abrahamse, E. L., Duthoo, W., Notebaert, W., & Risko, E. F. (2013). Attention modulation by proportion congruency: The asymmetrical list shifting effect. *Journal of Experimental Psychology: Learning Memory and Cognition*, 39(5), 1552–1562. <https://doi.org/10.1037/a0032426>
- Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, 14(10), 1338–1344. <https://doi.org/10.1038/nn.2921>
- Alexander, W. H., & Brown, J. W. (2014). A general role for medial prefrontal cortex in event prediction. *Frontiers in Computational Neuroscience*, 8, Article 69. <https://doi.org/10.3389/fncom.2014.00069>
- Balzer, W. K., Doherty, M. E., & O'Connor, R. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin*, 106(3), 410–433. <https://doi.org/10.1037/0033-2909.106.3.410>
- Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Ben Yehuda, M., Murphy, R. A., Le Pelley, M. E., Navarro, D. J., & Yeung, N. (2024). Confidence regulates feedback processing during human probabilistic learning. *Journal of Experimental Psychology: General*, 154(1), 80–95. <https://doi.org/10.1037/xge0001669>
- Botvinick, M. M., Carter, C. S., Braver, T. S., Barch, D. M., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108(3), 624–652. <https://doi.org/10.1037/0033-295X.108.3.624>
- Braem, S., Abrahamse, E. L., Duthoo, W., & Notebaert, W. (2014). What determines the specificity of conflict adaptation? A review, critical analysis, and proposed synthesis. *Frontiers in Psychology*, 5, Article 1134. <https://doi.org/10.3389/fpsyg.2014.01134>
- Braem, S., Bugg, J. M., Schmidt, J. R., Crump, M. J. C., Weissman, D. H., Notebaert, W., & Egner, T. (2019). Measuring adaptive control in conflict tasks. *Trends in Cognitive Sciences*, 23(9), 769–783. <https://doi.org/10.1016/j.tics.2019.07.002>
- Brown, J. W., & Braver, T. S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science*, 307(5712), 1118–1121. <https://doi.org/10.1126/science.1105783>
- Bugg, J. M., & Crump, M. J. C. (2012). In support of a distinction between voluntary and stimulus-driven control: A review of the literature on proportion congruent effects. *Frontiers in Psychology*, 3, Article 367. <https://doi.org/10.3389/fpsyg.2012.00367>
- Crump, M. J. C., Gong, Z., & Milliken, B. (2006). The context-specific proportion congruent Stroop effect: Location as a contextual cue. *Psychonomic Bulletin and Review*, 13(2), 316–321. <https://doi.org/10.3758/BF03193850>
- Danielmeier, C., & Ullsperger, M. (2011). Post-error adjustments. *Frontiers in Psychology*, 2, Article 233. <https://doi.org/10.3389/fpsyg.2011.00233>
- Desender, K., Boldt, A., Verguts, T., & Donner, T. H. (2019). Confidence predicts speed–accuracy tradeoff for subsequent decisions. *eLife*, 8, 1–25. <https://doi.org/10.7554/eLife.43499>
- 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., Donner, T. H., & Verguts, T. (2021). Dynamic expressions of confidence within an evidence accumulation framework. *Cognition*, 207, Article 104522. <https://doi.org/10.1016/j.cognition.2020.104522>
- 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>
- Dreisbach, G., & Fischer, R. (2015). Conflicts as aversive signals for control adaptation. *Current Directions in Psychological Science*, 24(4), 255–260. <https://doi.org/10.1177/0963721415569569>
- Dutilh, G., Vandekerckhove, J., Forstmann, B. U., Keuleers, E., Brysbaert, M., & Wagenmakers, E. J. (2012). Testing theories of post-error slowing. *Attention, Perception, and Psychophysics*, 74(2), 454–465. <https://doi.org/10.3758/s13414-011-0243-2>
- Egner, T. (2007). Congruency sequence effects and cognitive control. *Cognitive, Affective, & Behavioral Neuroscience*, 7(4), 380–390. <https://doi.org/10.3758/CABN.7.4.380>
- Fishbach, A., Eyal, T., & Finkelstein, S. R. (2010). How positive and negative feedback motivate goal pursuit. *Social and Personality Psychology Compass*, 4(8), 517–530. <https://doi.org/10.1111/j.1751-9004.2010.00285.x>
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906–911. <https://doi.org/10.1037/0003-066X.34.10.906>
- Fleming, S. M. (2024). Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, 75(1), 241–268. <https://doi.org/10.1146/annurev-psych-022423-032425>
- Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, 124(1), 91–114. <https://doi.org/10.1037/rev0000045>
- Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: Computation, biology and function. *Philosophical Transactions of the*

- Royal Society B: Biological Sciences, 367(1594), 1280–1286. <https://doi.org/10.1098/rstb.2012.0021>
- Fleming, S. M., & Frith, C. D. (2014). Metacognitive neuroscience: An introduction. In S. M. Fleming & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 1–6). Springer. https://doi.org/10.1007/978-3-642-45190-4_1
- Fox, J., & Weisberg, S. (2018). *An R companion to applied regression*. Sage Publications.
- Frömer, R., Nassar, M. R., Bruckner, R., Stürmer, B., Sommer, W., & Yeung, N. (2021). Response-based outcome predictions and confidence regulate feedback processing and learning. *eLife*, 10, Article e62825. <https://doi.org/10.7554/ELIFE.62825>
- Garrison, J., Erdeniz, B., & Done, J. (2013). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, 37(7), 1297–1310. <https://doi.org/10.1016/j.neubiorev.2013.03.023>
- Gratton, G., Coles, M. G. H., & Donchin, E. (1992). Optimizing the use of information: Strategic control of activation of responses. *Journal of Experimental Psychology: General*, 121(4), 480–506. <https://doi.org/10.1037/0096-3445.121.4.480>
- Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *eLife*, 5, Article e13388. <https://doi.org/10.7554/eLife.13388>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- 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://doi.org/10.1037/rev0000603>
- Herregods, S., Vermeylen, 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>
- Ito, S., Stuphorn, V., Brown, J. W., & Schall, J. D. (2003). Performance monitoring by the anterior cingulate cortex during saccade countermanding. *Science*, 302(5642), 120–122. <https://doi.org/10.1126/science.1087847>
- Kahnt, T., Grueschow, M., Speck, O., & Haynes, J. D. (2011). Perceptual learning and decision-making in human medial frontal cortex. *Neuron*, 70(3), 549–559. <https://doi.org/10.1016/j.neuron.2011.02.054>
- Laming, D. (1979). Choice reaction performance following an error. *Acta Psychologica*, 43(3), 199–224. [https://doi.org/10.1016/0001-6918\(79\)90026-X](https://doi.org/10.1016/0001-6918(79)90026-X)
- Lebreton, M., Baci, K., Palminteri, S., & Engelmann, J. B. (2019). Contextual influence on confidence judgments in human reinforcement learning. *PLOS Computational Biology*, 15(4), Article e1006973. <https://doi.org/10.1371/journal.pcbi.1006973>
- Lee, H. J., Lee, H., Lim, C. Y., Rhim, I., & Lee, S. H. (2023). Corrective feedback guides human perceptual decision-making by informing about the world state rather than rewarding its choice. *PLOS Biology*, 21(11), Article e3002373. <https://doi.org/10.1371/journal.pbio.3002373>
- Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. *Memory & Cognition*, 7(3), 166–174. <https://doi.org/10.3758/BF03197535>
- Lowe, D. G., & Mitterer, J. O. (1982). Selective and divided attention in a Stroop task. *Canadian Journal of Psychology*, 36(4), 684–700. <https://doi.org/10.1037/h0080661>
- Matsumoto, M., Matsumoto, K., Abe, H., & Tanaka, K. (2007). Medial prefrontal cell activity signaling prediction errors of action values. *Nature Neuroscience*, 10(5), 647–656. <https://doi.org/10.1038/nn1890>
- Mayr, U. (2004). Conflict, consciousness, and control. *Trends in Cognitive Sciences*, 8(4), 145–148. <https://doi.org/10.1016/j.tics.2004.02.006>
- Mullen, K. M., Ardia, D., Gil, D. L., 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>
- Niccolo, P., & Nicholas, Y. (2020). The role of decision confidence in advice-taking and trust formation. *Journal of Experimental Psychology: General*, 150(3), 507–526. <https://doi.org/10.1037/xge0000960.supp>
- Notebaert, W., Houtman, F., Van Opstal, F., Gevers, W., Fias, W., & Verguts, T. (2009). Post-error slowing: An orienting account. *Cognition*, 111(2), 275–279. <https://doi.org/10.1016/j.cognition.2009.02.002>
- 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>
- Petrov, A. A., Doshier, B. A., & Lu, Z. L. (2005). The dynamics of perceptual learning: An incremental reweighting model. *Psychological Review*, 112(4), 715–743. <https://doi.org/10.1037/0033-295X.112.4.715>
- Raaijmakers, S. F., Baars, M., Schaap, L., Paas, F., & van Gog, T. (2017). Effects of performance feedback valence on perceptions of invested mental effort. *Learning and Instruction*, 51, 36–46. <https://doi.org/10.1016/j.learninstruc.2016.12.002>
- Rabbitt, P., & Rodgers, B. (1977). What does a man do after he makes an error? An analysis of response programming. *Quarterly Journal of Experimental Psychology*, 29(4), 727–743. <https://doi.org/10.1080/14640747708400645>
- 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., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, 20(4), 260–281. <https://doi.org/10.1016/j.tics.2016.01.007>
- Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, 9(3), 438–481. <https://doi.org/10.3758/BF03196302>
- R Core Team. (2013). *R: A language and environment for statistical computing* (Version 4.4.1) [Computer software]. R Foundation for Statistical Computing. <https://www.r-project.org/>
- Ridderinkhof, K. R. (2002). Micro- and macro-adjustments of task set: Activation and suppression in conflict tasks. *Psychological Research*, 66(4), 313–323. <https://doi.org/10.1007/s00426-002-0104-7>
- Rudebeck, P. H., Behrens, T. E., Kennerley, S. W., Baxter, M. G., Buckley, M. J., Walton, M. E., & Rushworth, M. F. S. (2008). Frontal cortex subregions play distinct roles in choices between actions and stimuli. *Journal of Neuroscience*, 28(51), 13775–13885. <https://doi.org/10.1523/JNEUROSCI.3541-08.2008>
- 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>
- Silvetti, M., Alexander, W., Verguts, T., & Brown, J. W. (2014). From conflict management to reward-based decision making: Actors and critics in primate medial frontal cortex. *Neuroscience & Biobehavioral Reviews*, 46, 44–57. <https://doi.org/10.1016/j.neubiorev.2013.11.003>
- Torres-Quesada, M., Funes, M. J., & Lupiáñez, J. (2013). Dissociating proportion congruent and conflict adaptation effects in a Simon-Stroop procedure. *Acta Psychologica*, 142(2), 203–210. <https://doi.org/10.1016/j.actpsy.2012.11.015>
- Van der Borgh, L., Braem, S., & Notebaert, W. (2014). Disentangling posterior and postconflict reduction of interference. *Psychonomic Bulletin & Review*, 21(6), 1530–1536. <https://doi.org/10.3758/s13423-014-0628-z>
- Van Marcke, H., & Desender, K. (2025). Context-dependent role of confidence in information-seeking. *Cognition*, 263, Article 106219. <https://doi.org/10.1016/j.cognition.2025.106219>

- Verguts, T., & Notebaert, W. (2008). Hebbian learning of cognitive control: Dealing with specific and nonspecific adaptation. *Psychological Review*, 115(2), 518–525. <https://doi.org/10.1037/0033-295X.115.2.518>
- White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010). Using diffusion models to understand clinical disorders. *Journal of Mathematical Psychology*, 54(1), 39–52. <https://doi.org/10.1016/j.jmp.2010.01.004>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., ... Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), Article 1686. <https://doi.org/10.21105/joss.01686>
- Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in Neuroinformatics*, 7, Article 14. <https://doi.org/10.3389/fninf.2013.00014>
- Xing, J., & Desender, K. (2024). *The metacognitive conflict between internal feedback and external feedback*. Open Science Framework. <https://doi.org/10.17605/OSF.IO/B548M>
- Xing, J., & Desender, K. (2026). *The metacognitive conflict between internal feedback and external feedback*. Open Science Framework. <https://osf.io/52qne>
- Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594), 1310–1321. <https://doi.org/10.1098/rstb.2011.0416>

Received August 25, 2025

Revision received June 28, 2026

Accepted July 11, 2026 ■