Dear Recommender and Reviewers:

We appreciate your letter and the reviewers' feedback regarding our manuscript entitled "Is CPP an ERP marker of evidence accumulation in perceptual decision-making? A multiverse study. " (Manuscript ID: #892). The comments provided are valuable and instrumental for revising and improving our paper. We went through the comments thoroughly and have made revisions which we hope address all the issues.

To increase the readability, we have taken several measures:

(1) We **bolded** the reviewers' comments and showcased the main **manuscript changes** in **blue** for convenience in this response letter.
(2) We provided two versions of manuscripts: one has tracked changes, where we highlighted the revision in blue and showed removed content in red with a strikethrough, allowing you to trace the alterations made; one cleaner version (the online version: https://osf.io/p6aum), where we only hightlighted the modification content in blue, to present a clear view of the final text.

Please note that page of manuscript referred in this letter is linked to the untracked manuscript.

**The reviewers highlight a number of areas that would benefit from revision, including the strength of the study rationale (including links to relevant background literature), precision of the hypotheses (and hypothesis 1 especially), clarity and precision of the overall inferential chain, the level of described detail concerning the multiverse analysis, and the control of potential bias due to prior data observation and analysis.**

**Response:** We sincerely thank the recommender for the summary of comments from the reviewers. In this revision, we will introduce our modification from these aspects generally:

1. Strength of the study rationale (including links to relevant background literature):

We have augmented the introduction section to provide a clearer description of the significance underpinning this study, emphasizing two key aspects: the generalizability and robustness of the relationship between CPP and evidence accumulation. Specifically, we systematically categorized the datasets into three types of decision-making tasks and clarified the decision nodes of multiverse analysis. Please see **Response 3.2** for details.

2. Precision of the hypotheses

We have revised our hypotheses to improve their clarity and precision. Particularly, we have developed three distinct hypotheses according to the hierarchy of perception. Please see **Response 2.1** and **Table 1** in the manuscript.

3. Clarity and precision of the overall inferential chain

We have formulated three hypotheses corresponding to levels of perception(Vetter et al., 2024). Guided by this framekwork, we classified the datasets into three levels: simple, mid, and complex perceptual decision-making and only pool effect sizes within each level. Also, we will test our hypothesis and make our inference at each level. For additional details, please consult **Response 1.2**, **Response 2.1**, and **Table 1** in the manuscript.

4. The level of described detail concerning the multiverse analysis

We have articulated the reasoning for the selection of two nodes in the multiverse analysis. After considering the equivalence of all available options, we decided to exclude the condition-wise pooling method. please consult **Response 1.1** and **2.13** in the manuscript.

5. The control of potential bias due to prior data observation and analysis

We have made a justification for the findings from prior publications involving our datasets, demonstrating minimal overlap with our research objectives. Additionally, to mitigate potential bias, we are actively seeking more datasets and will access these new datasets only after Stage 1 review is finished. For further details, please see **response 2.9** and **3.1** for details.

*Reviewer #1:*

**1.1 The point of a multiverse analysis is to examine and compare alternative analytic choices, and these alternatives ought to be sensible or defensible. In the context of this specific topic – the relationship between CPP and evidence accumulation – I think there is little reason to consider CPP quantification approaches (i.e., the other 8 pipelines) other than the combination of CPP build-up rate and trialwise pooling. Unless the authors are able to justify the other inclusions, I think that the proposed multiverse analysis (or at least some of the pipeline choices, such as CPP amplitude and bin-wise pooling) is not well motivated.**

> **Response 1.1**: We sincerely appreciate your insightful suggestion. We have carefully re-evaluated whether the nodes in our multiverse analysis represent principled equivalence decision nodes (Del Giudice & Gangestad., 2021). Based on our evaluation, we decided to keep the bin-wise pooling and three CPP metrics. We added justification for all paths of these two decision nodes.
>
> Specifically, for the node of CPP metrics, we kept all three options because they all appeared in previously published studies, and comparing whether they are homogeneous will be of value to the field. For instance, build-up rate was used by Kelly and colleagues (Kelly et al., 2021; Kelly & O'Connell, 2013), amplitude was used by Murphy and colleagues (Murphy et al., 2015; O'Connell et al., 2012), and peak amplitude was used by Twomey and colleagues (Twomey et al., 2015) (see ***Manuscript changes 1,*** Page 20, Line 348-370).
>
> For the node pooling method, we kept trial-wise and bin-wise pooling and removed condition-wise pooling. We included bin-wise pooling because it has been used in previous studies (de Gee et al., 2020; Kelly & O'Connell, 2013; O'Connell et al., 2012) and we are also interested in comparing this pooling method with the trial-wise pooling method, a typical method for hierarchical modeling. Note that we removed condition-wise pooling after sensitivity analysis because this method has a relatively low chance to detect the effect in our simulation (see our **Response 2.3** below, and ***Manuscript changes 2,*** Page 46, Line 778-783).

**Manuscript changes 1 (Page 20, Line 348-370)**:

The first decision node is the CPP metric, where we included three options used in previous studies: build-up rate (Kelly et al., 2021; Kelly & O'Connell, 2013), amplitude (Murphy et al., 2015; O'Connell et al., 2012), and peak amplitude (Twomey et al., 2015). The build-up rate measures the rate of CPP rise, which is analogous to the rate of evidence accumulation……The amplitude originates from a study by O'Connell et al. (2012)…... At the same time, the peak amplitude is a widely used method in ERP studies, treating CPP as equivalent to P300 (Twomey et al., 2015)…

The second decision node pertains to the pooling method in statistical analysis. This decision node has two possible options:trial-wise and bin-wise pooling. The first method analyzed CPP data on a trial-by-trial basis, maintaining the granularity of trial-level CPP measurements. The second method involved binning trials by CPP measurement values and averaging them within each bin, allowing for analysis of drift rate relationships across specific CPP value ranges. We included this approach because it has been used in previous studies (de Gee et al., 2020; Kelly & O'Connell, 2013; O'Connell et al., 2012). Another possible option is condition-wise pooling, however, we excluded this approach because of its low statistical power in our simulation (see supplementary materials, the section on Power analysis).

**Manuscript changes 2 (Page 46, Line 778-783):**

Additionally, we explored an alternative approach, called the condition-wise pooling method, in which CPP values were averaged across subjects within different experimental conditions. However, this method yielded exceedingly low statistical power, even for a large effect size of 0.5, with a detection rate of only 5% (i.e., the effect of CPP on the drift rate was detected in just 1 out of 20 iterations). Given this inadequacy, we deemed this approach unsuitable and excluded it from further consideration.

**1.2 I also think the analysis plan for Hypothesis 1 needs improvement. Here, the link between CPP buildup and drift rate is tested separately for each dataset – if the authors want to comment on whether evidence accumulation models are applicable to a wide variety of more complex perceptual tasks, the analysis should include a statistical comparison of equivalent measures from the different datasets. As alternative analysis plans, I have two suggestions: (1) a "simpler" meta-analytical approach, i.e., comparison of measures of effect sizes (or equivalent) for each dataset/experiment. (2) A multiverse analysis in which the different datasets/tasks (and possibly alternative models) are "universes" – the question being asked here is whether the various paradigms produce an equivalent CPP that is an indicator of evidence accumulation.**

**Response 1.2**: We sincerely appreciate your insightful suggestions. As our study focuses on examining the generalizability of CPP as an ERP marker of evidence accumulation at different levels of perceptual decision-making, now we implemented a two-stage framework to systematically compare results from different datasets. Firstly, we categorize datasets into three levels based on the complexity of tasks: simple, intermediate, and complex guided by the hierarchy of perceptual processing (Vetter et al., 2024). Please see **Manuscript changes 1**, Page 5, Line 125-134.

Secondly, we treat datasets at the same level as equivalent and will synthesize their findings using meta-analytical approaches. Subsequently, we test hypotheses for each level separately. Please see below for the revision. (see **Manuscript changes 2**, Page 19, Line 332-341).

**Manuscript changes 1 (Page 5, Line 125-134):**

Specifically, we leveraged four publicly available datasets (see Methods for details), which encompassed tasks such as random-dot motion discrimination, face matching, fear-happy judgment, and memory-based decision-making. Based on the hierarchy of perceptual processing (Vetter et al., 2024), these tasks can be categorized into three levels, simple, intermediate, and complex perceptual processing. The inclusion of datasets involving low-level perceptual decision-making allowed us to assess the replicability of the relationship between CPP and evidence accumulation across diverse perceptual decision-making paradigms. Conversely, datasets involving other two decision-making tasks enabled us to evaluate the generalizability of this relationship to higher-level perceptual processes. This approach ensures a comprehensive examination of the robustness and scope of the observed correlations.

**Manuscript changes 2 (Page 19, Line 332-341):**

We will conduct a Robust Bayesian Meta-Analysis (Bartoš et al., 2025) to pool the effect of CPP build-up rate on drift rate across datasets at the same perceptual level, implemented via the R package RoBMA. Effect sizes and standard errors will be collated from the posterior distribution of the effect of CPP build-up rate on drift rate in each dataset. The framework integrates fixed-effect, random-effects, and publication bias-adjusted models (e.g., selection models) through Bayesian model averaging. Weakly informative priors are used: a normal distribution (μ: mean = 0, SD = 1) for the pooled effect, a truncated normal (τ: mean = 0, SD = 1, lower = 0) or inverse Gamma (shape = 1, scale = 0.15) for heterogeneity, and Beta priors for publication bias parameters. Posterior distributions will be estimated via MCMC sampling, with pooled effect mean and 95% credible intervals reported.

**1.3 The third author's first/last name order is inconsistent.**

**Response 1.3**: We sincerely appreciate your attention to this detail. The third author, Hu Chuan-Peng, preferred to present his name in accordance with the Chinese naming order, where the family name precedes the given name. The other authors agree with this naming order.

**1.4 Pg 13 "(6) Calculate CPP": Description could be more detailed – I assume "fitting a linear trend" is a least-squares linear regression, done for each participant and trial?**

**Response 1.4**: We sincerely appreciate your careful review and valuable feedback. We employed the least-squares method to fit a linear trend for each participant and each trial, implemented using the Python function *numpy.polyfit()*. We added this detail now.

**Manuscript changes (Page 20, Line 353-358):**

The build-up rate measures the rate of CPP rise, which is analogous to the rate of evidence accumulation. It was determined by applying a linear fit using the least-squares method, implemented with the Python function *numpy.polyfit()*. This calculation was performed within a 100 ms window from -180 ms to -80 ms before the response, after smoothing the EEG data with a 51-point moving average filter.

**1.5 Pg 19, Fig 4A: I am assuming that this part described in the caption was not completed: "Each column represents different measurements, and each row corresponds to different pooling methods" – is the intended plot something like Fig S1?**

**Response 1.5**: Thank you for your comments. We have revised the caption by removing the redundant content and ensuring it accurately reflects the current figure. The updated caption can now be found alongside the figure in the revised manuscript.

**Manuscript changes (Page 24, Line 412):**

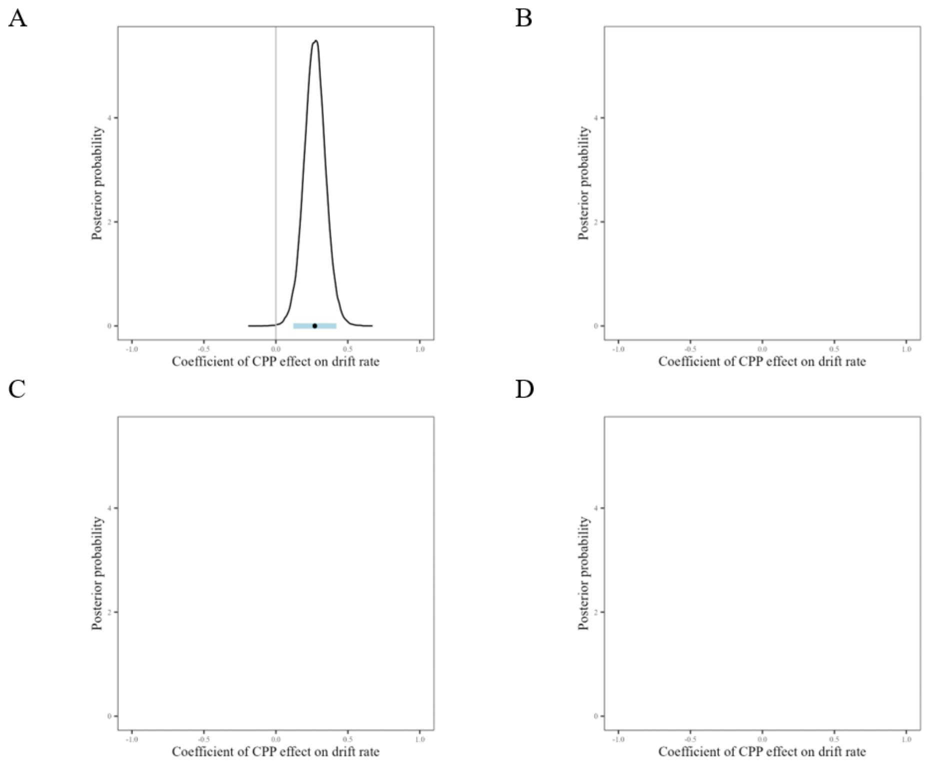

**Figure 4.** The posterior distribution of the coefficients for the CPP effect on drift rate across all four datasets. **A.** The posterior probability distribution for the coefficient of the CPP effect on drift rate in Dataset 1. The x-axis represents the coefficient of the CPP effect on drift rate, while the y-axis represents the posterior probability. The black point indicates the mean drift rate, and the blue bar represents the 95% highest density interval (HDI) of the drift rate. The vertical line denotes zero. If the 95% HDI of the coefficient of CPP measurements on the drift rate does not include zero, it indicates a stable effect. **B. C. D.** Similar analyses for datasets 2, 3, and 4, though these datasets have not yet been analyzed.

**1.6 Pg 26-end: References for supplementary methods were not provided.**

**Response 1.6**: We sincerely appreciate your attention to this detail. We have now included the appropriate references for the supplementary methods in the revised manuscript.

*Reviewer #2:*

Key issues as recommended by the peer community site

**2.1. 'Does the research question make sense in light of the theory or applications? Is it clearly defined? Where the proposal includes hypotheses, are the hypotheses capable of answering the research question?'**

**a. The hypothesis noted in the final paragraph of the introduction is well defined and testable statement, that aligns with the theoretical framework. However, the summary in Table 1 loses this precision. I recommend that the authors amend the question and hypothesis within Table 1 to align with the precision provided within the introduction. For example, question - Is CPP a consistent ERP marker for evidence accumulation at the trial level across multiple perceptual decision-making tasks?; hypothesis - If CPP is a generalisable ERP marker of evidence accumulation, then CPP build-up rate will show a statistically significant positive correlation with the drift rate across multiple perceptual tasks.**

> **Response 2.1:** We sincerely appreciate your insightful suggestion. We have carefully revised both the research question and hypothesis to ensure they align more precisely with the detailed rationale provided in the introduction.
>
> Furthermore, as we have categorized datasets into three perceptual levels based on their stimuli, we have applied these modifications to each category separately to ensure consistency and clarity. Please refer to Table 1 for details. **(Page 14, Line 236)**.

**2.2. 'Is the protocol sufficiently detailed to enable replication by an expert in the field, and to close off sources of undisclosed procedural or analytic flexibility?'**

**a. It would be more transparent if the authors stated the decisions taken for the following decision points in the workflow: unaccepted task performance (if participants were not removed based on task performance, it would be clear to state this. If they were, please report the threshold used); whether variables were normalized and/or centered; were there adjustments for multiple testing; were bad channels in the EEG datasets identified and, if so, how were they handled; were bad data segments in the EEG datasets identified and, if so, how were they handled.**

> **Response 2.2**: We sincerely thank you for your thorough review of our manuscript. We have added the following details to the revised manuscript.

**Manuscript changes 1 (Page 15, Line 244-247)**:

**Data preprocessing**

We excluded subjects from the four datasets who lacked either behavioral or EEG data, as the subsequent joint modeling requires both types of data to be analyzed together. Specifically, Subject 2 was excluded due to missing behavioral data in Dataset 1. For Dataset 2, XXX. For Dataset 3, XXX. For Dataset 4, XXX.

**Manuscript changes 2 (Page 15, Line 252-281)**:

***EEG preprocessing*** The raw EEG signals were processed using the MNE-python software (Gramfort, 2013). The preprocessing protocol for all datasets included the following steps (Pernet et al., 2020):

(1) …

(2) …

(3) …

(4) Checking for and removing bad channels. Bad channels were identified based on criteria such as high impedance, excessive noise, or abnormal signal amplitude by visual inspection. Identified channels were interpolated using data from surrounding channels to ensure data quality. In Dataset 1, no bad channels were identified. For Dataset 2, XXX. For Dataset 3, XXX. For Dataset 4, XXX.

(5) …

(6) Epoch the EEG data. EEG data was epoched based on each dataset's experimental procedure. … To retain as many trials as possible, we did not reject any trials after artifact correction.

(7) …

**Manuscript changes 3 (Page 17, Line 296-297)**:

***Model specification of joint models*** We employed a hierarchical modeling approach … It is important to note that, when incorporating the CPP build-up rate as a covariate, we normalized it to prevent the drift rate from exceeding its valid range during sampling:

$$\mathrm{v}_j = \beta_{0,\mathrm{j}} + \beta_{1,\mathrm{j}}\mathrm{X}_{condition} + \beta_{2,\mathrm{j}}\mathrm{X}_{CPP} + \beta_{3,\mathrm{j}}\mathrm{X}_{condition}\mathrm{X}_{CPP}$$

....

**2.3. 'Is there an exact mapping between the theory, hypotheses, sampling plan (e.g. power analysis, where applicable), preregistered statistical tests, and possible interpretations given different outcomes?'**

**a. The recommended amendment to the hypothesis at point 1 above would improve the direct mapping of the theoretical background to the hypothesis. It is noted that the authors use previously collected datasets, therefore an á prior power analysis is not applicable. However, the authors could report a sensitivity analysis to determine the smallest effect size that the existing sample sizes could reliably detect with a desired level of power (e.g., 80%), or commit to calculating the observed power based on the effect size obtained after conducting the analyses. The statistical tests are specified in advance and align with the hypothesis.**

> **Response 2.3**: Thank you for your insightful suggestion. We conducted a sensitivity analysis to ensure that the sample sizes provide sufficient statistical power (80%) to reliably detect the effect of the specified size. We used the pilot data (Dataset 1), which has the smallest sample size (N = 16 and 288 trials) in all four datasets (N = 23 and 269 / 311 trials, N = 80 and 288 trials, N = 23 and 252 trials for Dataset 2, 3 and 4, respectively).
>
> Our sensitivity analysis showed that, with 16 subjects and 288 trials, we have an 80% chance of detecting a regression coefficient of 0.2. We anticipate that with a greater number of subjects and trials, our models are sensitive enough to detect a relatively small effect (0.2) **(Page 42, Line 700-763)**.

**Manuscript changes (Page 42, Line 700-763)**:

We conducted a sensitivity analysis to assess whether our model, sample size, and trial numbers have enough chance (> 80%) to detect a relatively small effect. Given that we use the Bayesian hierarchical model to estimate the relationship between CPP and drift rate, we employed anparameter recovery approach for the sensitivity analysis. More specifically, we used information from pilot data (Dataset 1) as the benchmark because it has the smallest sample size among all four datasets. That said, in the simulation, we set the number of participants at 16 and the number of trials per participant fixed at 288. In parameter recovery, the model specification was the same as we used to fit the data (detailed model specifications can be found in the Method section). The only parameter we varied during the simulation was the key effect, the regression coefficient of CPP as a predictor of drift rate in HDDM, with a range of [0.1, 0.5] and a step of 0.1.

The parameter recovery was conducted by the following steps.

First, data generation. We used the function *hddm.generate.gen_rand_data()* to generate simulated data (reaction times and choice), with the Wiener first passage time function:

$$\mathrm{rt_{i,j}}, choice_{i,j} \sim wfpt(v_j, a_j, z_j, t_{0\,j})$$

The *wfp*t function has four parameters: $v_j, a_j, z_j, t_{0\,j}$. To optimize computational efficiency during data simulation, we fixed values across participants of parameters that are irrelevant to our goal. The exact values for these condition-irrelevant parameters, $a_j, z_j$ were selected based on (Wiecki et al., 2013) to ensure they fell within a plausible range. For parameters directly relevant to our hypotheses, $v_j, t_{0\,j}$, were calibrated using empirical findings from the pilot dataset.

The decision threshold $a_j$ and starting point bias $z_j$ were fixed at 1 and 0.5, respectively, across all subjects. The drift rate $v_j$ and non-decision time $t_{0\,j}$ were modeled as linear combinations of relevant factors to model the effects of experimental conditions and centro-parietal positivity (CPP). For the drift rate, we included the experimental conditions, CPP values, and their interaction:

$$\mathrm{v}_j = \beta_{0,\mathrm{j}} + \beta_{1,\mathrm{j}}\mathrm{X}_{coherence} + \beta_{2,\mathrm{j}}\mathrm{X}_{CPP} + \beta_{3,\mathrm{j}}\mathrm{X}_{coherence}\mathrm{X}_{CPP}$$

Here, $\beta_{0,\mathrm{j}}$ represents the baseline drift rate, which was fixed at 3. The coefficients $\beta_{1,\mathrm{j}}, \beta_{2,\mathrm{j}}, \beta_{3,\mathrm{j}}$ across subjects were drawn from the following distributions:

$$\beta_{1,\mathrm{j}} \sim N(1,0.4),$$

$$\beta_{2,\mathrm{j}} \sim N(effect\ size, 0.1),$$

$$\beta_{3,\mathrm{j}} \sim N(0,0.1).$$

The CPP values, $\mathrm{X}_{CPP}$, were drawn from a standard normal distribution (mean = 0, SD = 1), which is consistent with the standard the CPP indices used in our model. The $\mathrm{X}_{coherence}$ here represents motion coherence, with two levels 0 and 1, corresponding to high and low coherence in the experimental design of Dataset 1. Each condition has 144 trials.

Similarly, the non-decision time $t_{0\,j}$ was modeled to account for the influence of spatial prioritization:

$$t_{0\,j} = \gamma_{0,\mathrm{j}} + \gamma_{1,\mathrm{j}}\mathrm{X}_{prioritization}$$

Where $\gamma_{0,\mathrm{j}}$ represents the baseline non-decision time and was fixed at 0.3 across all subjects, which captures the effect of spatial prioritization across subjects were drawn from：

$$\gamma_{1,\mathrm{j}} \sim N(-0.02,0.01)$$

The $\mathrm{X}_{prioritization}$, which represents spatial cue in the experimental design, was coded as 0 or 1, representing valid cue or invalid neutral cue conditions. As in Dataset 1, these two conditions have an equal number of trials.

With the above parameter settings, we generated simulated data that has 16 participants and 288 trials per participant. The number of trials for a combination of experimental conditions also aligns with the experimental design in Dataset 1.

Second, parameter estimation based on the simulated data. We fitted simulated data using *hddm.HDDMRegressor( m = hddm.HDDMRegressor(data = df, models = [{'model': 'v ~ 1 + coherence + cpp + coherence: cpp', 'link_func':lambda x:x}{'model': 't ~ 1 + prioritization', 'link_func':lambda x:x }], include = ['v', 'a', 't', 'z'], group_only_regressors=False, keep_regressor_trace=True))*. Posterior distributions were sampled 6,000 times, with the first 3,000 discarded as burn-in, across four independent chains for robustness.

Third, inference. If the 95% highest density interval (HDI) of the CPP regression coefficient excluded zero, we inferred that the model detected the effect, otherwise, we inferred the model did not detect the effect.

We repeated the above three steps for 30 times. Statistical power[1] = (the number of simulations that detected the effect)/30.

Results showed statistical power of 37%, 83%, 100%, 100%, and 100% for effect sizes of 0.1 to 0.5, respectively (see Figure S3). These findings indicate that, except for the smallest effect size (0.1), the model reliably detects the CPP-drift rate relationship with high power.

[1] We aware that statistical power is a term primarily from Frequentist statistics. Here we used the similar logical for sensitivity analysis, thus, we used the term here for simplicity.

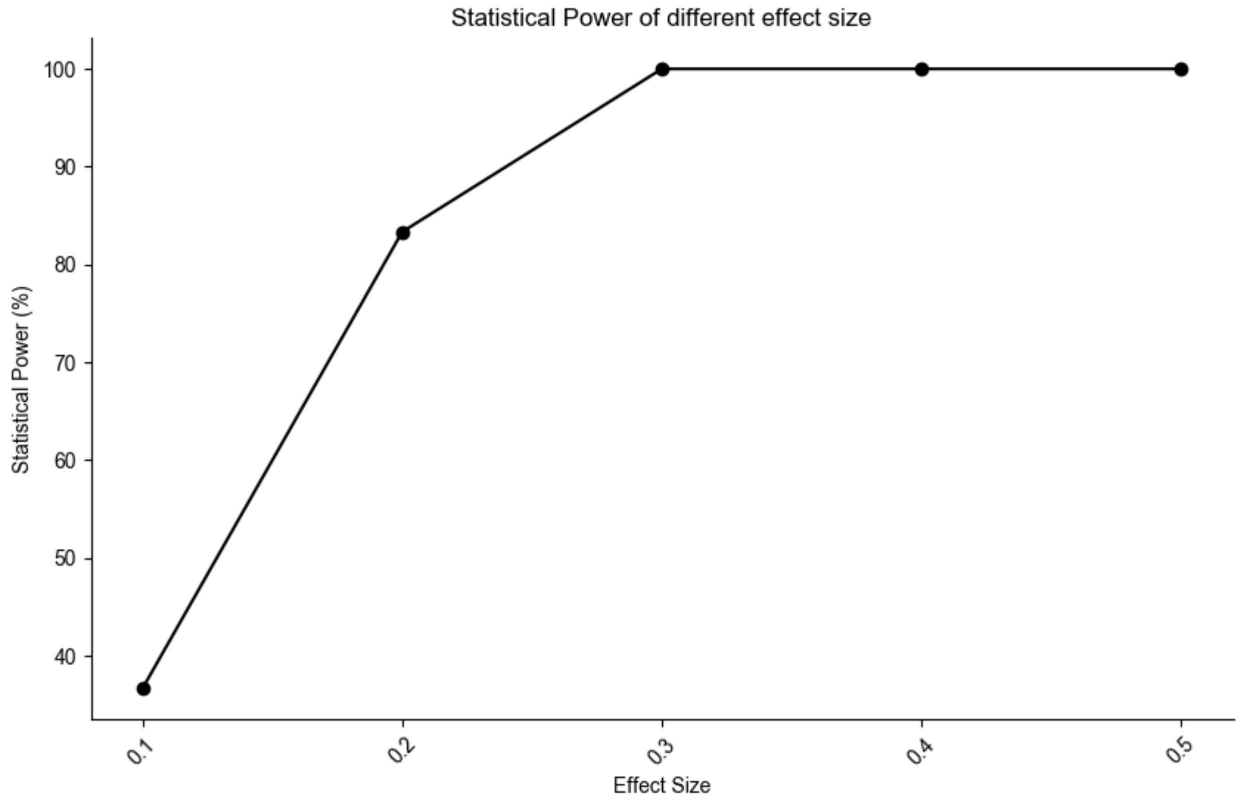

**Figure S3. Statistical Power of different effect sizes.** Each point on the lines corresponds to the statistical power at a specific effect size. The effect sizes on the x-axis are fixed at 0.1, 0.2, 0.3, 0.4, 0.5. The statistical powers are 37%,83%, 100%,100%,100%.

**2.4. ‘For proposals that test hypotheses, have the authors explained precisely which outcomes will confirm or disconfirm their predictions?’**

*a. Yes.*

*NA*

**2.5. ‘Is the sample size sufficient to provide informative results?’**

**a. As explained under point 3, this remains unclear until the authors either report a sensitivity analysis or commit to calculating the observed power.**

**Response 2.5**: Please refer to **response 2.3** for clarification on this matter.

**2.6. 'Where the proposal involves statistical hypothesis testing, does the sampling plan for each hypothesis propose a realistic and well justified estimate of the effect size?'**

**a. The authors analyse preexisting datasets. While they do not report the sampling approaches, they refer the readers to the original studies for further details.**

*NA*

**2.7. 'Have the authors avoided the common pitfall of relying on conventional null hypothesis significance testing to conclude evidence of absence from null results? Where the authors intend to interpret a negative result as evidence that an effect is absent, have authors proposed an inferential method that is capable of drawing such a conclusion, such as Bayesian hypothesis testing or frequentist equivalence testing?'**

**a. They interpret the 95% highest density interval of the posterior distribution for the effect of CPP build-up rate on drift rate, to allow probabilistic statements about parameter estimates rather than relying on p-values. The authors specify a criterion for concluding a positive effect: if the lower bound of the 95% HDI is above zero, they interpret this as evidence of a positive correlation between CPP and drift rate, implying that they would not necessarily conclude the absence of an effect but instead interpret this as insufficient evidence to support a positive correlation.**

*NA*

**2.8. 'Have the authors minimised all discussion of post hoc exploratory analyses, apart from those that must be explained to justify specific design features? Maintaining this clear distinction at Stage 1 can prevent exploratory analyses at Stage 2 being inadvertently presented as pre-planned.'**

**a. The authors have detailed a clear, pre-specified approach, with justification for the structured analysis plan. Authors report a predefined criterion for evaluating CPP build up effect on drift rate.**

*NA*

**2.9. 'Have the authors clearly distinguished work that has already been done (e.g. preliminary studies and data analyses) from work yet to be done?'**

**a. It is not immediately clear which analyses were completed by prior publications using the datasets. Related to this, a clear justification for the datasets selected from those available for the present study is required.**

> **Response 2.9:** Thank you for your valuable comment. We have added a clarification in the Methods section, specifying which analyses were conducted in prior studies and which are newly performed in the present work. (see **Supplementary materials at Page 32, Line 549-576**).

**Manuscript changes (Supplementary materials at Page 32, Line 549-576)**:

### Prior Analyses of Datasets

To clearly distinguish our study's novel contributions from previous analyses, we have summarized the findings of prior work related to our datasets, highlighting their scope and limitations relative to our objectives.

#### Dataset 1 (Georgie et al., 2018)

The original study examined behavioral and EEG data but did not investigate the relationship between CPP and drift rate, nor did it report estimates for CPP or drift rate. A subsequent analysis by Ghaderi-Kangavari and colleagues (Ghaderi-Kangavari et al., 2023) explored this relationship using a conventional two-step approach, laying the groundwork for our study. However, it did not systematically evaluate alternative CPP measurements or pooling methods, both of which we address in our work.

#### Dataset 2 (Van Vugt et al., 2019)

This study investigated the relationship between CPP and drift rate but restricted its analysis to CPP slope (one of three possible measurements) and employed a bin-wise pooling method (one of two available methods). Specifically, the dataset was divided into two bins—high and low CPP slope trials—and drift rates for each bin per participant were estimated using the DMA toolbox (Vandekerckhove & Tuerlinckx, 2008), followed by a *t*-test comparing them. While the findings offered preliminary insights, the study did not evaluate the robustness of the relationship between CPP and drift rate.

#### Dataset 3 (Newman et al., 2017)

The original study reported CPP amplitudes (one of three possible measurements) and behavioral indices such as response time and accuracy, but it did not explicitly examine the relationship between CPP and drift rate. Consequently, there is no overlap between the original findings and our research objectives, making this dataset particularly valuable for extending prior work.

Dataset 4 (Sun et al., 2023)

The original study employed multiple neuroimaging techniques, including EEG, fMRI, single-neuron recordings, and eye-tracking, but did not analyze the CPP or drift rate. As such, there is no overlap between the original study and our research objectives here.

**2.10. 'Have the authors prespecified positive controls, manipulation checks or other data quality checks? If not, have they justified why such tests are either infeasible or unnecessary? Is the design sufficiently well controlled in all other respects?'**

**a. This is not reported in the present stage 1 manuscript.**

**Response 2.10:** Thank you for your valuable comment. Rather than specifying positive controls or manipulation checks, we ensured data quality through a two-step process in our study. First, we exclusively selected data from experimental tasks engaging different levels of perceptual decision-making. Second, we implemented standard data quality control procedures during ERP preprocessing and for identifying behavioral outliers, ensuring the integrity of both ERP and behavioral data. These details are comprehensively described in the Methods section. (see **Page 9, Line 167-222**).

**2.11. 'When proposing positive controls or other data quality checks that rely on inferential testing, have the authors included a statistical sampling plan that is sufficient in terms of statistical power or evidential strength?'**

**a. This is covered in my response to 3 and 10.**

*NA*

**2.12. 'Does the proposed research fall within established ethical norms for its field? Regardless of whether the study has received ethical approval, have the authors adequately considered any ethical risks of the research?'**

**a. Yes, the proposed research falls within established ethical norms for the field.**

*NA*

**2.13. It is encouraging to see that the authors wish to report uncertainty and assess the robustness of results to variations in data analysis decisions. Multiverse analyses should be**

**systematic and decisions transparent. Therefore, the authors should (1) specify which element of the workflow is subjected to a multiverse analysis (i.e. two decision nodes in the analytical procedure are forked, whereas a multiverse analysis in general could refer to forking behavioral and EEG data preprocessing decisions also); (2) for the decision nodes that are forked, there should be transparency in the options that were considered at each decision node, including those that were not included, and the decision-making procedure to include those that are included. This will help readers to identify potential bias in the reported multiverse of results. (3) The authors should state whether the options included are equivalent (e.g. a principled multiverse, Del Guidice & Gangestad, 2021) and, if so, on which criteria are they deemed equivalent (e.g., comparable validity, examine the same effect, or estimate the effect with comparable precision).**

**Response 2.13**: We sincerely appreciate your constructive suggestions.

(1) We identified two decision nodes for the multiverse analysis: (a) measurement variability in CPP quantification and (b) pooling methods for handling EEG noise. These nodes were selected because they represent key sources of variability in the CPP quantification process. Importantly, we constrained our multiverse analysis to decisions directly related to CPP quantification, rather than extending to broader behavioral or EEG preprocessing steps, to specifically examine the robustness of the relationship between CPP and evidence accumulation.

(2) For each decision node, we have now provided a detailed explanation of the options considered, including those ultimately excluded. Please see our **Response 1.1** above for details.

*Reviewer #3:*

**3.1 My primary concern for this proposal is its suitability for a PCI RR given the use of existing data. The authors plan to reanalyse four publicly available data sets, and state that they have already analysed one of the datasets (Dataset 1) to demonstrate their analytical pipeline. My understanding of the levels of bias control recognised by PCI RR is that, having already analysed part of the data, this proposal is at Level 0, making it ineligible for consideration as a PCI RR? However, if the authors consider the already analysed dataset to be pilot data demonstrating feasibility of their proposed analysis pipeline, my understanding is that the results from Dataset 1 must be clearly distinguished from the other three datasets at latter stages of review.**

> **Response 3.1**: Thank you for your question. We classify our proposal as PCI-RR Level 3 because, while we have access to the datasets, we have exclusively examined the pilot data (Dataset 1) and deliberately avoided inspecting the others.
>
> Regarding Dataset 1, it serves as pilot data to demonstrate the feasibility of our analysis pipeline; however, our conclusion will predominantly rely on the unanalyzed data. That said, results from Dataset 1 will be treated separately from the remaining dataset.
>
> To further enhance our work, we explored additional potential datasets but did not download any of them to blind us from peeking into these datasets. Please see our response to your next point.

**3.2 The authors' key research question is "whether the relationship between CPP and evidence accumulation observed in simple perceptual tasks can be generalised to more complex perceptual decision-making tasks." This is a scientifically valid question that stems from previous research. However, this question has been addressed in previous studies (e.g., van Vugt et al. (2019); Pisauro et al. (2017). Nature Communications. https://doi.org/10.1038/ncomms15808). The proposal would be strengthened if the authors included discussion of previous research in the Introduction section and addressed how their proposal extends past work (e.g., through joint modelling and a multiverse analysis approach). Relatedly, the datasets the authors propose to analyse could be selected to more convincingly address the research question. Namely, Dataset 2 comes from a study by van Vugt et al. (2019) which has already addressed the same research question (albeit with a different methodology), while Dataset 3 involves a random dot motion task, and so does not address the question of generalisability beyond tasks involving simple perceptual features. If Dataset 1 is excluded on the basis of being pilot data/already analysed, then only Dataset 4 is of particular interest.**

**Response 3.2**: We sincerely thank you for these thoughtful comments. We have revised the standard for selecting datasets and distinguished our works and previous studies from which the original datasets were generated.

Firstly, we adopted the hierarchical framework for perception (Firestone & Scholl, 2016; Newen & Vetter, 2017, 2017; Pylyshyn, 1999; Teufel & Nanay, 2017; Vetter et al., 2024). That is, perceptual decision-making could be categorized into three distinct levels based on stimulus complexity and how they are processed in the brain: low-level, mid-level, and high-level. The low-level perceptual decisions are characterized by basic sensory features, mid-level by complex features and objects, and high-level by complex scenes and body representations. Applying these standards, the pilot data (Dataset 1) analyzed in our Stage 1 manuscript is classified as mid-level. This is because the stimuli—cars and faces—integrate simple physical features (e.g., edges, shapes, orientations) into more complex, yet not fully abstract, perceptual forms. Similarly, Dataset 2 is also mid-level, given its use of facial stimuli. Dataset 3 is deemed low-level, as its task requires judging the motion direction of random dots, a fundamental sensory feature. In contrast, Dataset 4 is classified as high-level, as emotions reflect complex body representations integrating perceptual and affective processing. Moreover, to strengthen our study, we will search more datasets from platforms such as OpenNeuro and include datasets that meet our criteria after Stage 1. These additional datasets, if reusable, will further improve the robustness and generalizability of our findings in Stage 2.-(see **Manuscript changes 1**, **Page 4, Line 92-107**).

Secondly, we distinguished our method from the original studies of the datasets used here. More specifically, we applied a unified approach to define CPP by pre-specified electrode locations and implemented joint modeling techniques to quantify the relationship between CPP and evidence accumulation. (see **Manuscript changes 2**, **Page 5, Line 108-120**).

| | CPP electrodes | CPP metric | quantify CPP and drift rate |
|---|---|---|---|
| Our work | Pre-defined CPz, CP1 and CP2 | CPP amplitude, CPP slope, CPP peak amplitude | Standardized regression coefficients of effect of CPP metrics on drift rate by joint model |
| van Vugt et al. (2019) | Pre-defined CPz, CP1 and CP2 | CPP slope | *t*-test of drift rate between trials with high CPP slopes and low CPP slopes |
| Pisauro et al. (2017) | Centro-parietal electrode cluster defined by EEG traces and model Dynamics | CPP amplitude | Correlation between CPP amplitude and model-simulated amplitudes |

**Manuscript changes 1 (Page 4, Line 92-107)**:

Nevertheless, real-life perceptual decisions often require more complex processing. For example, to accurately detect someone's emotions, it is necessary not only to recognize facial lines and orientations but also to integrate these features with facial expressions. This complex information processing demands coordination across multiple brain regions. In fact, complex tasks can be further categorized into mid-level and high-level perceptual processing (Vetter et al., 2024). While researches have explored the relationship between CPP and evidence accumulation beyond perceptual decision-making, such as value-based decision-making (Pisauro et al., 2017) and social decision-making (Arabadzhiyska et al., 2022), this replationship was less explored in complex perceptual tasks. Thus, the generalizability of the relationship between CPP and evidence accumulation remains unclear.

**Manuscript changes 2 (Page 5, Line 108-120)**:

Methodological heterogeneity—including inconsistent CPP metrics (e.g., variable electrode montages) and divergent computational frameworks— further complicated the issue. Unlike the pioneering studies by O’Connell et al. (2012), where CPP amplitude was used to explore evidence accumulation, follow-up studies adopted various methods. For example, while CPP is derived from the parietal lobe, electrodes selected for CPP varied across studies: some used CPz (Kelly &

O’Connell, 2013; Van Vugt et al., 2019), others used Pz (Murphy et al., 2015; Newman et al., 2017). Furthermore, methods for quantifying CPP differed —O’Connell et al. (2012) focused on amplitude, Kelly & O’Connell (2013) incorporated slope, and Murphy et al. (2015) combined slope, amplitude, and peak latency. These discrepancies in electrode selection and computational approaches underscore the challenge of comparing findings across studies, highlighting the need for a standardized framework to characterize the CPP-evidence accumulation relationship. The methodological heterogeneity calls for a systematic evaluation of the relationship between CPP and evidence accumulation across different perceptual decision-making tasks.

**The authors plan to use the Hierarchical Drift-Diffusion Model (HDDM) Python package to jointly model the CPP and behaviour and establish the relationship between the CPP and drift-rate. The proposed rationale for (dis)confirming the hypothesis that “CPP build-up rate is positively correlated with the drift rate” is to use the Bayesian 95% highest density interval (HDI) for the CPP-drift-rate coefficient, such that a 95% HDI > 0 will be taken as evidence of a positive correlation. Overall, this is a sound analysis plan with a logical and plausible hypothesis given previous research. However, greater clarity is needed around the modelling and multiverse procedure:**

**3.3 Dataset 1 (Georgie et al., 2018) includes only 288 trials in total (72 per condition) from eight participants, and Dataset 4 included only 252 trials from 23 participants. These seem like very small amounts of data for the proposed HDDM analysis pipeline. It will be important for the authors to demonstrate that their models successfully converge and that the parameter estimates are reliable.**

> **Response 3.3**: Thank you for your careful review. We have conducted a sensitivity analysis which revealed that the current number of participants and trials are sensitive to detect an effect size of 0.2, please refer to our **response 2.3** above. Our analysis of the pilot data revealed that the r-hat, an index for MCMC convergence in Bayesian analysis, of all parameters was more than 1.1, which indicates that the model has converged.

**3.4 The justification for each model specification is unclear. The authors should provide greater guidance as to why they have chosen to compare the models they have for each dataset.**

> **Response 3.4**: Thanks for this suggestion. Now we have re-checked all models specified and balanced the complexity and interpretability of each model. The parameters in the

competing models for each dataset now should have clear theoretical explanations, otherwise, they are fixed at participant level. The key effects tested by each model (as compared to simpler models) are also listed in Table S1 for clarity. Please see below our justifications for model specification (see supplementary materials **Page 33, Line 577-625**).

**Manuscript changes (Page 33, Line 577-625)**:

We specified competing models by balancing the complexity and interpretability of model parameters. The parameters in the competing models for each dataset should have clear theoretical explanations, otherwise, they are fixed at participant level (see Table S1 for full specifications across datasets).

**Table S1.** Model specifications for behavioral data of all 4 datasets.

| Dataset | Index | Model specification | Key effect tested by the model | DIC | LOO-CV |
|---|---|---|---|---|---|
| Georgie et al. (2018) | 1 | _hddm.HDDMRegressor (v, a, t, z)_ |  | -2981 | -3990 |
|  | 2 | _hddm.HDDMRegressor(**v ~ coherence**, a, t, z)_ | Coherence's effect on _v_ | -3391 | -4368 |
|  | 3 | _hddm.HDDMRegressor(v ~ coherence, **z~prioritization**, a, t)_ | Spatial prioritization's effect on _z_ | -3464 | -4465 |
|  | 4 | _hddm.HDDMRegressor(v ~ coherence, **t ~ prioritization, a, z)**_ | Spatial prioritization's effect on _t_ | -3460 | -4464 |
| Van Vugt et al. (2019) | 1 | _hddm.HDDMRegressor (v, a, t, z)_ |  |  |  |
| Experiment 1 | 2 | _hddm.HDDMRegressor(**v ~ the similarity of faces**, a, t, z)_ | The similarity of faces' effect on _v_ |  |  |
|  | 3 | _hddm.HDDMRegressor(v, **a ~ the similarity of faces**, t, z)_ | The similarity of faces' effect on _a_ |  |  |
|  | 4 | _hddm.HDDMRegressor(v, a, **t ~ the similarity of faces**, z)_ | The similarity of faces' effect on _t_ |  |  |
| Van Vugt et al. (2019) | 1 | _hddm.HDDMRegressor (v, a, t, z)_ |  |  |  |
| Experiment 2 | 2 | _hddm.HDDMRegressor(**v ~ the similarity of faces**, a, t, z)_ | The similarity of faces' effect on _v_ |  |  |
|  | 3 | _hddm.HDDMRegressor(v, **a ~ the similarity of faces**, t, z)_ | The similarity of faces' effect on _a_ |  |  |
|  | 4 | _hddm.HDDMRegressor(v, a, **t ~ the similarity of faces**, z)_ | The similarity of faces' effect on _t_ |  |  |
| Newman et al. (2017) | 1 | _hddm.HDDMRegressor (v, a, t, z)_ |  |  |  |
|  | 2 | _hddm.HDDMRegressor(v, a, **t ~ hemisphere**, z)_ | Visual hemifields' effect on _t_ |  |  |
|  | 3 | _hddm.HDDMRegressor(v, a, t**, z ~ hemisphere**)_ | Visual hemifields' effect on _z_ |  |  |
| Sun et al. (2023) | 1 | _hddm.HDDMRegressor (v, a, t, z)_ |  |  |  |
|  | 2 | _hddm.HDDMRegressor(**v ~ percentage of happy face**, a, t, z)_ | Percentage of happy face's effect on _v_ |  |  |
|  | 3 | _hddm.HDDMRegressor(v, a, t, **z ~ percentage of happy face**)_ | Percentage of happy face's effect on _z_ |  |  |

For Dataset 1, we defined 4 competing models for Dataset 1. The simplest model (model 1) assumed that there was no effect of experimental manipulations, so it serves as a baseline model, including four parameters of DDM. Then, in model 2, we tested the effect of coherence, one of two experimental manipulations, on drift rate by allowing the drift rate to vary at different coherence levels (Kelly & O'Connell, 2013; Philiastides et al., 2006, 2014), while keeping the other parameters as in model 1. Model 3 was built on model 2, in which we further tested the effect of the spatial cue on starting point *z* by allowing *z* to vary with different spatial cue conditions (Sagar et al., 2019) and kept the other parameters as in model 2. Similarly, in model 4, we tested whether spatial cue also affects non-decision time (Ghaderi-Kangavari et al., 2023). Model 4 was similar to model 3 but the spatial cue's effect was on non-decision time *t* (see **table S1** for the specification of these four models).

For Dataset 2, we defined 4 competing models for both tasks. The baseline model (model 1) includes four parameters of DDM. Since facial similarity arises from the interaction between two faces, it would be unreasonable to assume a bias toward only one face affecting the starting point (*z*). Consequently, we excluded this assumption from consideration. Thus, building on model 1, we constructed model 2 to test whether facial similarity affects drift rate *v* by enabling the drift rate *v* to vary with the levels of facial similarity while keeping the other parameters as in model 1. Similarity, we established model 3 to assess whether facial similarity impacts threshold *a* by allowing the threshold *a* vary with the effect of similarity of the face. Comparably, we formulated model 4 to investigate whether the similarity of the face influences non-decision-time *t* by letting the non-decision-time *t* vary with the effect of similarity of the face (see **table S1** for the specification of these four models).

For Dataset 3, we established 3 competing models for Dataset 3. The baseline model (model 1) incorporates four parameters of DDM. Previous researches indicate its influence arises from two primary mechanisms: attentional asymmetries (Corbetta & Shulman, 2011) or variations in the onset of evidence accumulation (Newman et al., 2017). Accordingly, we focused exclusively on these two possibilities. Building on model 1, we designed model 2 to determine whether the visual hemispheres where the stimulus appears impact non-decision time *t* by enabling this parameter to vary according to the different visual hemispheres while keeping the other parameters as in model 1. In model 3, similar to Model 2, we assess the effect of visual hemispheres on the starting point *z* by permitting this parameter to vary according to the different visual hemispheres (see **Table S1** for the specification of these three models).

For Dataset 4, we constructed 3 competing models for Dataset 4. The baseline model (model 1) incorporates four parameters of DDM. We hypothesize that the percentage of happy faces affects behavior through two key mechanisms: task difficulty to differentiate different emotions (it is happiness or fearness in this dataset) (Ashby et al., 1999), and response bias to specific emotions (Fazio, 2001). Therefore, in model 2, we tested the effect of the percentage of happy faces on the drift rate *v* by allowing *v* to vary with percentages while keeping other parameters as in model 1.

Likewise, in model 3, instead of the drift rate *v*, we formulated Model 3 to examine whether the same percentage influences starting point *z* by enabling starting point *z* to vary with percentages (see **Table S1** for the specification of these three models).

**3.5 Relatedly, why did the authors switch between hddm.HDDM() and hddm.HDDMRegressor() within each dataset? My understanding is that hddm.HDDMRegressor() can still be used to estimate Model 1 (i.e., a a ~ 1 v ~ 1, t ~ 1) and Model 2 (a ~ 1 v ~ 1, t ~ 1 and z ~ 1) for each dataset.**

**Response 3.5**: We now only use *hddm.HDDMRegressor()* in our data analysis. (see Page 34, Line 583).

**3.6 Further details are required about the leave-one-out cross-validation (LOO-CV) procedure. Perhaps this is just my lack of familiarity with this technique, but I am unclear on how the authors propose to perform model selection using LOO-CV based on their description.**

**Response 3.6**: Thank you for your suggestion. We have added further clarification on the leave-one-out cross-validation (LOO-CV) procedure. (see **Page 37, Line 633-643**).

**Manuscript changes (Page 37, Line 633-643)**:

LOO-CV is a model evaluation and selection technique based on cross-validation, where the key idea is to systematically omit one observation $y_i$ from the dataset and train the model on the remaining data $D_{-i}$. The trained model is then used to predict the omitted observation $y_i$, and the log predictive density log $p(y_i \mid D_{-i})$ for that observation is calculated. This procedure is iterated through all observations. Then, the Expected Log Predictive Density (ELPD), calculated by summing the log predictive densities for all observations, serves as a measure of the model's predictive performance. In our study, we used the LOO-CV algorithm implemented in the Python library Arviz (Kumar et al., 2019), which incorporates Pareto-smoothed importance sampling (PSIS) specifically developed for Bayesian methods (Vehtari et al., 2017). To ensure the accuracy of the LOO-CV algorithm, we excluded trials (amounting to 4% of all trials in Dataset 1) with a diagnostic value *k*-hat exceeding 0.7, as recommended by (Vehtari et al., 2017).

**3.7 For the bin-wise and condition-wise CPP pooling methods, it is unclear what will actually be entered into the HDDMRegressor() function as a covariate on each trial. Are the same few aggregate values based on condition- or bin-wise averaging to be used? If so,**

**doesn’t this defeat the purpose of providing a trial-wise covariate because the variability is now removed?**

**Response 3.7**: It is correct that the same few averaged value of bins or conditions will be used for bin-wise or condition-wise pooling method. However, these methods may not defeat trial-wise method because averaging not only removed variability but may also remove valuable information. Our sensitivity analysis revealed that the "statistical power" of bin-wise approach is comparable to trial-wise but the condition-wise is much worse. See our **Response 1.1**. above and supplementary materials, the section on Power analysis.

**3.8 In instances where the winning base-model for the behavioural data allows drift-rate to vary by task dependent variables, are the authors also planning to model the interaction between the CPP and the task dependent variables? The formula on Page 14 seems to indicate this, but the results for Dataset 1 presented on Page 18 do not include the interaction effect.**

> **Response 3.8**: Thank you for your careful review. Now we added the interaction in the supplementary materials. (see **Page 38, Line 670-688**).

**Manuscript changes (Page 38, Line 670-688):**

Additionally, none pipelines demonstrated a significant interaction effect between CPP and coherence on drift rate.

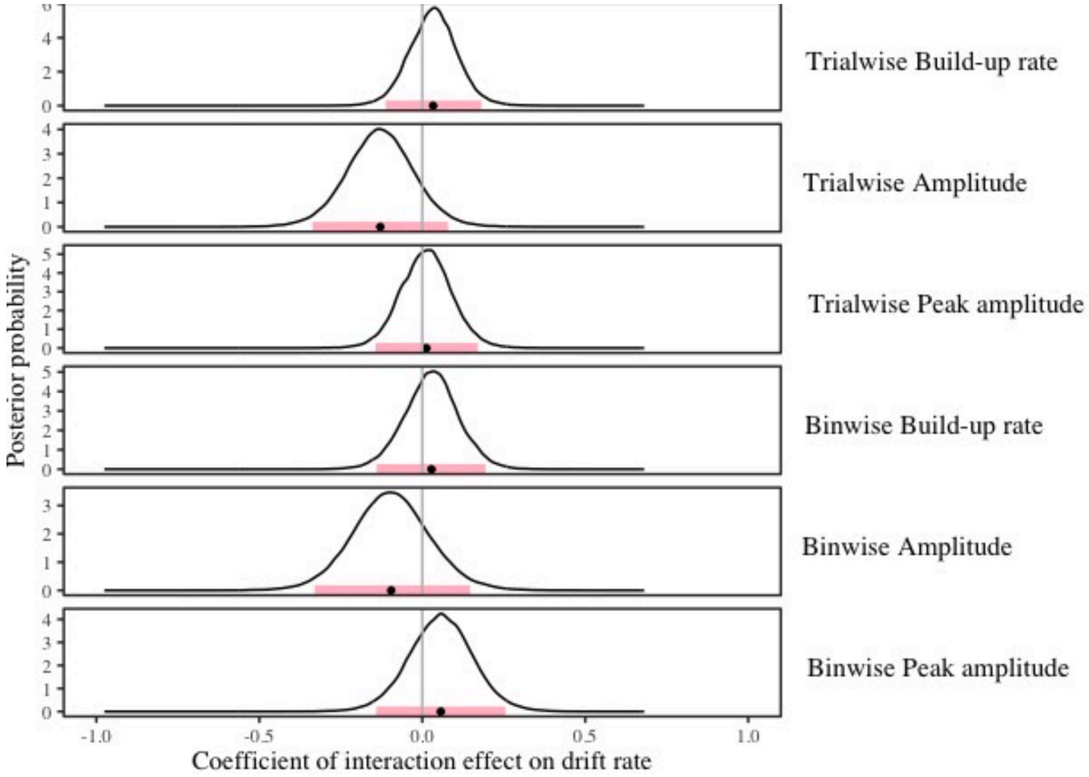

**Figure S2.** The posterior distribution of coefficient of the interaction effect between CPP and coherence on drift rate. The group-level posterior probability of the coefficient of the interaction effect between CPP and coherence on drift rate is depicted. The x-axis represents the coefficient of the interaction effect between CPP and coherence on drift rate, while the y-axis represents the posterior probability. Each column represents different measurements and pooling methods. The black point indicates the mean drift rate, and the pink bar represents the 95% highest density interval (HDI) of the drift rate. The vertical line denotes zero. If the 95% HDI of the coefficient of the interaction effect between CPP and coherence on the drift rate does not include zero, it indicates a stable effect.