**Review of “*Is CPP an ERP marker of evidence accumulation in perceptual decision making? A multiverse study. [Stage 1 Registered Report]”***

Liu et al. plan to investigate the generalisability of the centro-parietal positivity (CPP) as a neural marker of evidence accumulation beyond simple perceptual decisions by conducting a reanalysis of publicly available datasets involving a variety of decision-making tasks. The authors plan to jointly model behavioural and neural data using the drift-diffusion model and evaluate the association between the CPP and a model-derived estimate of evidence accumulation, the DDM drift-rate. Moreover, the authors plan to take a “multiverse” approach to determine the robustness of their findings to different measurement definitions of the CPP and different methods of data pooling. Overall, this is an interesting proposal, and the results would be of value to CPP the literature. However, I have several comments/concerns:

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.
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.
3. 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:
	1. 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.
	2. 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.
	3. 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.
	4. 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 how the authors propose to perform model selection using LOO-CV based on their description.
	5. 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?
	6. 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.