**Peer Review and Communication History**

**MS Title**: When Response Selection Becomes Gambling: Post-error Slowing and Speeding in

Self-paced Colour Discrimination Tasks

**Author Names**: Charlotte Eben, Luc Vermeylen, Zhang Chen, Wim Notebaert, Ivan Ivanchei, Frederick Verbruggen

**Submitted:** Apr 22, 2022

**Editor First Decision**: Revise & Resubmit

Oct 30, 2022

Dear Charlotte Eben,

I have now received two reviews of your manuscript, “When response selection becomes gambling: post-error slowing and speeding in self-paced colour discrimination tasks”, and I apologise for the long time between submission and decision; I had a lot of trouble finding a second reviewer (and so a big thanks to the reviewers for taking the time to provide their expert opinions on the manuscript). The reviewers had mostly positive reactions to your manuscript, though they raised a lot of points that I would like to see addressed in a revision, which mostly involve a lot of re-writing of the main text of the manuscript. I also read the manuscript both before and after receiving the reviews, in order to form my own independent opinion and then consider the manuscript with the additional context of the reviews. My perspective was mostly in-line with that of the reviewers, and while I do not wish to repeat their specific points here, I think the following two broad aspects need to be addressed in the revision:

(1) The analysis methodology is quite inconsistent between experiments (e.g., frequentist vs Bayesian, plus other points made by Reviewer 2), and doesn’t quite conform to best practices, which can make the manuscript a bit tough to draw clear conclusions from. First, I would strongly recommend taking a “strength of evidence” approach (as suggested by most advocates of Bayesian model selection), rather than merely stating whether the evidence is “conclusive” or not (which I assume is determined by the BF > 10 or 1/10 threshold, but this also isn’t clear). While optional stopping is fine if the BF is interpreted according to the strength of relative evidence, the situation gets more difficult if the BF is used as some kind of pseudo-NHST (as Reviewer 1 alludes to regading error rates), and so I think the words “significant” and “conclusive” should be replaced with strength of evidence quantification when referring to BFs. Moreover, as Reviewer 2 points out, the analyses for Experiment 2 are extremely unclear (i.e., to me, the inferences seem to be based on the frequentist ANOVAs in this case, and I assume the BFs are inclusion Bayes factors (though using what method?) rather than BF10s), and I think need to be completely re-written.

(2) As pointed out by both reviewers, I think there’s a lot of extra motivation, cleaning up of definitions, and perhaps better choices of key words that could have been used throughout. Like Reviewer 1, I wasn’t sure whether “Controllability” was really a good description of the manipulation being used, and like Reviewer 2, I thought that the motivation for the study could have been stronger (and things like the discrepenacies in results between Experiment 2 and 3 could have been given a bit more of a discussion). The reviewers were quite comprehensive regarding these points, and I would like to see them addressed in the revision.

Therefore, I strongly encourage you to submit a revised version of the manuscript for further consideration at Collabra: Psychology. The revision should include a document with a point-by-point response to the reviewers’ comments, outlining each change made in your manuscript or providing a suitable rebuttal.

Overall, I believe that your manuscript is of excellent quality, and I hope that you will choose to revise it for further consideration at Collabra: Psychology. I look forward to receiving your revision.

Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This is the last opportunity for major editing, therefore please fully check your file prior to re-submission.

If you have any questions or difficulties during this process, please contact us at the editorial office editorialoffice@collabra.org. If you have any further questions about the reviews or revisions, then please feel free to contact me at nathan.j.evans@uon.edu.au.

We hope you can submit your revision within the next six weeks. If you cannot make this deadline, please let us know as early as possible.

Sincerely,

Nathan Evans

# Reviewer 1

##### Rating scale questions

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  | ✔ |  |  |

##### Open response questions

### Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

Summary: Experiments testing factors that determine the presence of PES are reported. E1a:self pacing – PES was seen in start-RT and in errors but not in RT. E1b: self-pacing + reward/punishment (did not change the pattern significantly). E2: two conditions varying between alternating blocks easy and hard (actually-impossible). No PES in easy, PES in hard-beginning and reversed PES in 2nd half. E3:
Evaluation: I liked the fact that all the studies were pre-registered, with a-priori sample size consideration and included BF analyses. The study is also informed by considerable literature. I think that the paper could benefit from dealing with the few issues which are listed below as well as from tightening and shortening. Shortening could be achieved by avoiding repetitions. For example, the general approach in the inferential statistics is the same, why repeat it in in and every experiment? Just write, “as in Exp 1” for example. The same holds for several other aspects (especially Methods and Results, of course). Additionally, I found the General Discussion too long.
More major:
Repeated increases in sample size. I am not sure this procedure of testing BF, and then increase sample size repeatedly does not violate the assumptions. I suspect it does. Specifically, with BF, there is (presumably) no bias favoring H1 against H0 as in NHT. However, there seems to be a bias towards conclusive H1/H0 outcomes and against inconclusive ones. (for example, in Bayes Factors Design Analysis one needs to a-priori specify whether significance testing be done repeatedly or only once all data were collected). Given the high threshold (BF>10) which the authors have adopted, I do not see this point as being detrimental, but for the aesthetics of the paper- the point should be dealt with.
Terms: “Controllability” seems like a sub-optimal label. Specifically, self-pacing has a control aspect to it (control of ITI), whereas in Exp 2 and 3, the pseudo-difficulty manipulation implies that there is no control over evidence accumulation (which also means no control of correctness via threshold setting). Changing the terms would have great implications on how the paper is eventually being written up, I assume.
E1: One of the (few) major conclusions from E1 is “Overall, this pattern of results indicates that making the task self-paced does not influence post-error slowing (too much).” This conclusion made me think that self-pacing was manipulated – but apparently it was not – i.e. all participants self-paced. Actually, the results of this experiment suggest (to me) that participants strategically employed self-pacing to deal with PES-related issues as evident from the start-RT results. This strategy was partly successful, since some (albeit non-significant) slowing was seen in RT and error-effects were not completely abolished.
P 17: “We also expected that as the experiment progressed, participants would realize that they were not in control in the ’hard’ blocks, and this would result in speeding (instead of slowing) after errors in these ’hard’ blocks (but still slowing in the ’easy’ blocks).” This prediction makes no sense (to me). If one realizes that this is pure gamble, one may speed-up overall, e.g., by pressing the same response key without making any choice. Why should that affect PES? Interestingly, I wrote this comment before seeing the results – which show that in the 1st part, (choice RT) Hard>easy, in the second part it was easy>hard (which sort of does not make sense, unless considering that participants shifted towards guessing).

Minor:
Figures: put all RT (start and RT) results on the same Y-scale to allow readers compare between slopes (across conditions/experiments).
Ex2: “we always presented the exact same stimulus (exactly between dark and light) …” The English does not make sense (to me). Perhaps you mean that these were the exact same dark/light stimuli? (which makes no sense to me) – or perhaps you meant that it was the exact same stimulus that was randomly associated with dark/light responses as correct responses? (this makes sense to me).

# Reviewer 2

##### Rating scale questions

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  |  | ✔ |

##### Open response questions

### Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

When response selection becomes gambling: post-error slowing and speeding in self-paced colour discrimination tasks
This empirical study runs a series of experiments that demonstrates a set of interesting sequential effects:
• Decreased controllability has a global influence on sequential effects where these effects disappear in the easy condition when hard blocks are intermixed.
• After eliciting post error slowing in an easy condition decreasing controllability (i.e., the hard condition) leads to post error speeding in perceptual decision-making tasks.
• Both slowing and speeding effects dissipate as errors accumulate.
I think this work should be published. However I found the justification for the study was not clear as well as some of the statistical decisions. I think at the end of the intro and start of the discussion you should more clearly map out why this study was run in the first place. Moreover, the explanations of these effects are not satisfying. Perhaps it is beyond this study to provide a single coherent account of the interesting effects found here, but what is currently provided has some logical flaws. I hope to see an improved version of this manuscript with the points above and below addressed and I hope the comments are useful to the author.
Gabriel Tillman
gabrieltillman.au@gmail.com
Intro
• What is meant by sub-optimal?
• “the amount of information that is required to make a decision, are increased to avoid subsequent errors”. Is this for the next trial only, for a window after the error, …?
• Is referring to the information required to make a decision as “task settings” correct given that these are the internal threshold settings of the participant?
• Williams et al. (2016): was there also an effect of payment vs no payment on cautiousness of participants? I can’t recall but does this influence your statement in the introduction?
• I was a little confused about the distinction between response speed errors and evidence quality errors. Are response speed errors due to participants having decision thresholds set too low? And are evidence quality errors due to the stimuli being the source of poor evidence or the participants internal evidence accumulation rate being poor? Difficult to distinguish the evidence quality account unless there was some sort of stimuli manipulation.
• Can objective controllability be more clearly defined in the present study section as well as the distinction of this studies work to Dyson?
Exp 1A
• Maybe best to cite Rouder’s paper on optional stopping to justify the BF calculated for 50 participants then a subsequent 50.
• Were the dark and light response mappings for the keyboard counter balanced? Perhaps not an issue unless you are interested in dark vs light response performance.
• “Numerically, we also found a post-error slowing effect in the choice RT (18 ms) but the evidence for this effect was inconclusive (BF10 = 0.619).” The Bayes factor here suggest no post error slowing (assuming 1 is the alternative and 0 is the null in the subscript), albeit inconclusive evidence. In any case it would be best to clear up the sign of the BF early. >1 evidence in favour of slowing and <1 evidence in favour of the null.
• You mention you will report p values but these are never reported. Given the optional stopping its probably best they are not reported, however reference to p-values should be removed as the reader will be looking for them.
Exp 1B
• In Table 1, I personally think it would be best to round to 2 decimal places for aesthetics. This could be done for statistics throughout the manuscript.
• What was the justification for stopping data collection after only 1 effect was decisive? For instance, if the BF is moving for both effects are moving towards evidence for the alternative would it not be best to collect data until conclusive evidence is obtained for all effects?
• “our main processing-“ this dash could just be a comma
• A bayes factor would be more appropriate for determining if there was no difference between size of slowing between experiments given you want to provide evidence for the null hypothesis.
• Given exp 1A and B have 6 responses, does Hick’s law have any influence on post error slowing?
Exp 2
• Was the seed fixed in the previous analyses?
• With the staircase procedure, if many of the errors were made after the change in stimulus difficulty, wouldn’t this mean that there is a confound correlated quite strongly with error occurrence and first half of experiment? Or was the definition of first half of the experiment AFTER the staircase procedure?
• “To determine statistical significance, we used an alpha of .05.” How does this relate to the Bayesian ANOVA?
• For main effects, were inclusion Bayes factors or p values being used as inferential evidence of the effects? The word significant is used in this section but it is not clear what statistics you are using for your inferential decisions given you ran a Bayesian ANOVA.
Exp 3
• Are the slopes for the error count analysis reported anywhere?
• In any of the experiments did you explore differences in response? For instance light and dark response mapping wasn’t counterbalanced so it would be good to check statistically if there were differences between the response type.
General Discussion
• You state “further tested the idea that objective controllability might play a critical role.” I think this is a useful research direction. But can it be clarified what you did that was above and beyond what is already known. For instance based on Dyson and the Damaso paper you could already conclude that participants who feel in control slow down after sub-optimal outcomes, whereas those who do not feel in control over the outcome do not slow down or speed up. Perhaps the Dyson study was in the context of gambling tasks and your study is in the context of perceptual decision making. In any case, can the justification of running these controllability studies be fleshed out more.
• In addition to justifying the focus of controllability more clearly, I think the theoretical treatment could be expanded on in the discussion. You note that controllability as a construct explains the effects of the controllability as a manipulation, however this seems circular and unsatisfying. What mechanism would need to change between easy and hard to account for the effect of controllability?
• Another effect that is left unexplained is the dissipation of the effects. To explain you note that maybe speeding occurs first because setting cautious thresholds appears to do nothing to the accuracy of their performance in the hard condition. Then over time, speeding appears to disappear because participants give up. However, why does giving up not appear as speeding through trials to end the experiment early?
• Perhaps it is beyond this study to explain all of the interesting effects of this work, but perhaps that should be more clearly stated.

**Author Response**
Jan 11, 2023

Dear Dr. Evans,

Thank you very much for giving us the opportunity to revise our manuscript. We would also like to thank you, and the reviewers for the constructive reviews provided. We appreciate the amount and detail of the comments and have incorporated them in the revised manuscript. You will find detailed responses to every comment below. Novel sections are highlighted in yellow (here and in the manuscript).

Comments:

1. *The analysis methodology is quite inconsistent between experiments (e.g., frequentist vs Bayesian, plus other points made by Reviewer 2), and doesn’t quite conform to best practices, which can make the manuscript a bit tough to draw clear conclusions from. First, I would strongly recommend taking a “strength of evidence” approach (as suggested by most advocates of Bayesian model selection), rather than merely stating whether the evidence is “conclusive” or not (which I assume is determined by the BF > 10 or 1/10 threshold, but this also isn’t clear). While optional stopping is fine if the BF is interpreted according to the strength of relative evidence, the situation gets more difficult if the BF is used as some kind of pseudo-NHST (as Reviewer 1 alludes to regarding error rates), and so I think the words “significant” and “conclusive” should be replaced with strength of evidence quantification when referring to BFs.*

**Reply:** We have adjusted the analyses and omitted any reference to frequentist analyses (for completeness, we kept the p-values in our tables).

1. *Moreover, as Reviewer 2 points out, the analyses for Experiment 2 are extremely unclear (i.e., to me, the inferences seem to be based on the frequentist ANOVAs in this case, and I assume the BFs are inclusion Bayes factors (though using what method?) rather than BF10s), and I think need to be completely re-written*.

**Reply:** We have re-written the analyses of Experiment 2 in line with your suggestion about the strength of evidence approach (p.21/22).

“We conducted a Bayesian repeated measures ANOVA with the factors condition
(hard vs. easy), previous outcome (error vs. correct) and part of the experiment (first half vs. second half). For the Bayes factors, we used the default prior widths as used in R and JASP. For Bayesian ANOVAs the prior width is 0.5. We conducted the Bayesian ANOVA with JASP, using the Bayesian repeated measures ANOVA calculating the BF with effects compared to a null model. Specifically, we report the inclusion Bayes factors across matched models. The preregistered follow-up t-tests can be found in the online supplementary material.

The descriptive statistics can be found in Figure 2. First, the evidence for an effect of previous outcome was anecdotal (BF10 = 0.48), showing no overall difference between trials following errors (M = 312 ms; SD = 185 ms) and trials following correct trials (M = 302 ms; SD = 170 ms). Second, we found moderate evidence (BF10 = 3.48) for an effect of condition, indicating faster start latencies in the easy condition (M = 301 ms; SD = 169 ms) compared with the hard condition (M = 315 ms; SD = 185 ms). Third, we found extreme evidence (BF10 > 100) for an effect of part of the experiment, showing slower start latencies in the first half of the experiment (M = 330 ms; SD = 193 ms) compared to the second half (M = 286 ms; SD = 155 ms). Fourth, the interaction between condition and part of the experiment yielded anecdotal (BF10 = 0.38) evidence. Finally, the interaction between previous outcome and part of the experiment (BF10 = 0.13), the interaction between previous outcome and condition (BF10 = 0.23) and the three-way interaction (BF10 = 0.18) yielded anecdotal evidence for the null hypothesis. For detailed inferential statistics, see Table 2.

In the choice RT the pattern was a bit different. First, we found moderate evidence (BF10 = 3.35) for an effect of previous outcome, suggesting faster response times for trials following error trials (M = 701 ms; SD = 304 ms) compared to trials following correct trials (M = 722 ms; SD = 292 ms). For effect of condition, we found moderate (BF10 = 0.28) evidence in favor of the null hypothesis, suggesting no difference between easy (M = 717 ms; SD = 262 ms) and hard trials (M = 705 ms; SD = 331 ms). Third, evidence for an effect of part of the experiment was extreme (BF10 > 100), showing the same pattern as in the start RT: participants were slower in the first part of the experiment (M = 760 ms; SD = 285 ms) compared with the second part of the experiment (M = 662 ms; SD = 286 ms). Fourth, the interaction between condition and part of the experiment yielded anecdotal (BF10 = 1.42) evidence. Finally, all other interactions yielded moderate evidence in favor of the null hypothesis. For detailed inferential statistics, see Table 2.”

1. *As pointed out by both reviewers, I think there’s a lot of extra motivation, cleaning up of definitions, and perhaps better choices of key words that could have been used throughout. Like Reviewer 1, I wasn’t sure whether “Controllability” was really a good description of the manipulation being used, and like Reviewer 2, I thought that the motivation for the study could have been stronger (and things like the discrepancies in results between Experiment 2 and 3 could have been given a bit more of a discussion). The reviewers were quite comprehensive regarding these points, and I would like to see them addressed in the revision.*

**Reply:** We have responded to each point of the reviewers in detail for these general points.

We also added a sentence on the results of Experiment 2 in comparison to the results in Experiment 3 (p. 30):

“Our between-group comparison revealed that participants who had control over the
outcome (i.e., the ’easy’ group) slowed down after committing an error. However, when participants did not have control over the outcome (i.e., the ’hard but doable’ and ’impossible’ groups), they sped up after errors.

This difference between conditions further supports the conclusion of Experiment 2 that introducing trials on which participants did not have control over the outcome also influenced performance in the condition in which they *did* have control over the outcome. After all, when controllability over the outcome was manipulated within participants (Experiment 2), we observed post-error speeding in both conditions (i.e., there was no effect of condition); but when controllability over the outcome was manipulated between participants (Experiment 3), post-error speeding/slowing did depend on the condition. In other words, the overall task context also seems to play a role.“

**Reviewer 1:**

1. *Evaluation: I liked the fact that all the studies were pre-registered, with a-priori sample size consideration and included BF analyses. The study is also informed by considerable literature.*

**Reply:** Thank you very much.

1. *I think that the paper could benefit from dealing with the few issues which are listed below as well as from tightening and shortening. Shortening could be achieved by avoiding repetitions. For example, the general approach in the inferential statistics is the same, why repeat it in in and every experiment? Just write, “as in Exp 1” for example. The same holds for several other aspects (especially Methods and Results, of course). Additionally, I found the General Discussion too long.*

**Reply:** We have made sure that we do not repeat things that have already been said. But given that the experiments differ quite a bit from each other, we needed to explain several things in the method sections of each experiment to ensure that our experiments are replicable.

1. *Repeated increases in sample size. I am not sure this procedure of testing BF, and then increase sample size repeatedly does not violate the assumptions. I suspect it does. Specifically, with BF, there is (presumably) no bias favoring H1 against H0 as in NHT. However, there seems to be a bias towards conclusive H1/H0 outcomes and against inconclusive ones. (for example, in Bayes Factors Design Analysis one needs to a-priori specify whether significance testing be done repeatedly or only once all data were collected). Given the high threshold (BF>10) which the authors have adopted, I do not see this point as being detrimental, but for the aesthetics of the paper- the point should be dealt with.*

**Reply:** Thank you for this comment. We have expanded our explanation about the testing procedure and the robustness against repeated testing (p.9 and p. 12):

“Importantly, the Bayesian sequential testing procedure allows us to interpret Bayes Factors despite optional testing (Rouder, 2014; Schönbrodt & Wagenmakers, 2018; Schönbrodt et al., 2017; but see also de Heide and Grünwald, 2021).”

“We report Bayes factors for statistical inferences. For the Bayesian analyses, we report the Bayes Factor BF10, which quantifies the evidence for the alternative hypothesis against the null hypothesis. A Bayes Factor > 1 is in favor for the alternative hypothesis, whereas a Bayes Factor < 1 is in favor of the null hypothesis. A Bayes Factor around 1 yields inconclusive evidence. The size of the Bayes Factor determines the strength of the evidence: anecdotal (1/3 − 1; 1-3), moderate (1/3 − 1/10; 3-10), strong (1/10 − 1/30; 10-30), very strong (1/30 − 1/100; 30-100) or extreme (< 1/100; > 100). We used the default prior width as used in R. For Bayesian t-tests this prior width is 0.707, corresponding to a medium effect.”

*4. Terms: “Controllability” seems like a sub-optimal label. Specifically, self-pacing has a control aspect to it (control of ITI), whereas in Exp 2 and 3, the pseudo-difficulty manipulation implies that there is no control over evidence accumulation (which also means no control of correctness via threshold setting). Changing the terms would have great implications on how the paper is eventually being written up, I assume.*

**Reply:** This is a good point; people can indeed have control over several elements in our task. But here we were interested in the ‘control over the outcome’, as this is a crucial difference between games of chance and the choice tasks that are normally used in the cognitive field. We have clarified this in the manuscript, and now consistently refer to ‘control over the outcome’.

*5. E1: One of the (few) major conclusions from E1 is “Overall, this pattern of results indicates that making the task self-paced does not influence post-error slowing (too much).” This conclusion made me think that self-pacing was manipulated – but apparently it was not – i.e. all participants self-paced. Actually, the results of this experiment suggest (to me) that participants strategically employed self-pacing to deal with PES-related issues as evident from the start-RT results. This strategy was partly successful, since some (albeit non-significant) slowing was seen in RT and error-effects were not completely abolished.*

**Reply:** In Experiment 1, we just wanted to determine if the difference between choice and gambling tasks was not merely due to the self-paced nature of gambling tasks (p.13).

“Overall, this pattern of results indicates that making the task self-paced does not make participants speed up (rather than slow down) after an error.”

After ruling out this alternative hypothesis, we could test our main hypothesis that controllability over the outcome could explain the discrepancy between the findings in the gambling and cognitive (control) fields.

*6. P 17: “We also expected that as the experiment progressed, participants would realize that they were not in control in the ’hard’ blocks, and this would result in speeding (instead of slowing) after errors in these ’hard’ blocks (but still slowing in the ’easy’ blocks).” This prediction makes no sense (to me). If one realizes that this is pure gamble, one may speed-up overall, e.g., by pressing the same response key without making any choice. Why should that affect PES? Interestingly, I wrote this comment before seeing the results – which show that in the 1st part, (choice RT) Hard>easy, in the second part it was easy>hard (which sort of does not make sense, unless considering that participants shifted towards guessing).*

**Reply:** We based this assumption on previous findings and added this clarification (p.17):

“We also expected that as the experiment progressed, participants would realize that they were not in control over the outcome in the ’hard’ blocks. Based on previous literature (Damaso et al., 2020), we predicted that not having control over the outcome in the hard blocks would result in post-error *speeding* (in contrast to the ’easy’ blocks, in which we still predicted post-error *slowing*).”

For a discussion of a general speeding effect, please see also our response to the last comment of **reviewer** 2:

Participants might have initially cared about the outcome, which is reflected by speeding after errors compared with correct trials. A general speeding towards the end combined with the dissipation of the post-error effect indicates that eventually participants simply gave up and clicked through the experiment (p. 35/36).

“Surprisingly, Experiment 3 showed that the effects in all three groups dissipated over time. Although the decreasing post-error slowing in the easy condition can be explained in terms of decreased orienting to accumulating errors, the pattern in the two difficult conditions is more surprising. One tentative explanation is that post-error speeding dissipates over time due to ‘learned helplessness’ (following the initial frustration). For instance, Mikulincer (1988) showed that participants who were repeatedly exposed to unsolvable problems initially tried harder but eventually gave up. We observed a similar pattern in an earlier gambling study (Eben et al., 2020): in a not very engaging gambling task, we found general speeding on gambling trials (compared with non-gambling trials); furthermore, participants initially showed more speeding after (gambled) losses compared to (gambled) wins, but this effect dissipated over time. It seems that we observe a similar pattern here. As the task in Experiment 3 was not very engaging, participants might have stopped caring about the outcomes of their actions as the experiment progressed, resulting in reduced difference between post-error and post-correct trials. Another indication that participants eventually gave up on doing the task is a general speeding towards the end of the experiment (see Figures 2 and 3 in the supplementary materials). Whether this motivational account (or more specifically, reduced motivation as the experiment progresses) could also explains the reduced post-error slowing in the easy condition needs further testing though.”

*7. Figures: put all RT (start and RT) results on the same Y-scale to allow readers compare between slopes (across conditions/experiments).*

**Reply:** We have used the same range (i.e., 500 ms) of the y-axis for all figures to make them comparable.

*8. Ex2: “we always presented the exact same stimulus (exactly between dark and light) …” The English does not make sense (to me). Perhaps you mean that these were the exact same dark/light stimuli? (which makes no sense to me) – or perhaps you meant that it was the exact same stimulus that was randomly associated with dark/light responses as correct responses? (this makes sense to me).*

**Reply:** It was one stimulus (RGB 127 127 127) for the entire hard blocks, but we told participants that there were two stimuli (which were hard to distinguish). We deleted the ‘exactly between dark and light’ addition to avoid confusion.

**Reviewer 2:**

*1. I think this work should be published. However I found the justification for the study was not clear as well as some of the statistical decisions. I think at the end of the intro and start of the discussion you should more clearly map out why this study was run in the first place.*

*Moreover, the explanations of these effects are not satisfying. Perhaps it is beyond this study to provide a single coherent account of the interesting effects found here, but what is currently provided has some logical flaws. I hope to see an improved version of this manuscript with the points above and below addressed and I hope the comments are useful to the author.*

**Reply:** We worked on your single points and tried to make the entire manuscript clearer.

*2. What is meant by sub-optimal?*

**Reply:** We added an explanation to make clear what we mean by sub-optimal (p.3).

“(…) sub-optimal outcomes (e.g., losses, outcomes that are worse than the best possible outcome, or outcomes that come with serious costs) (…).”

*3. “the amount of information that is required to make a decision, are increased to avoid subsequent errors”. Is this for the next trial only, for a window after the error, …?*

**Reply:** Yes, we added that this is meant to be for several trials after the error (p.3).

“If the action outcome is sub-optimal, which is the case for errors, participants adjust the ‘task set’ or ‘task parameters’ (Logan & Gordon, 2001), such as the amount of information that is required to make a decision, to avoid subsequent errors. Such adjustments might persist for multiple trials (e.g., Forster & Cho, 2014)”

 *4. Is referring to the information required to make a decision as “task settings” correct given that these are the internal threshold settings of the participant?*

**Reply**: A popular concept in the control literature is the ‘task set’. Following Logan and Gordon (2001), we define a task set as a collection of ‘parameters’ that program the cognitive system to perform a certain task. Task parameters can include, e.g., which stimulus feature to attend to, but also how much information is required to make a decision. Furthermore, in line with Logan and Gordon (2001), we assume that a main function of the cognitive control system is to adjust these parameters when required.

We now refer to Logan and Gordon (2001) in the introduction. But when we refer to increasing the amount of information required to make a decision, we have replaced task settings with ‘response threshold’ (p.3), as this is more specific.

 *5. Williams et al. (2016): was there also an effect of payment vs no payment on cautiousness of participants? I can’t recall but does this influence your statement in the introduction?*

**Reply:** Yes, they only found it in unpaid participants. We added the word ‘unpaid’ to make that clear. Please note that in the previous version of the manuscript, we already discussed the effect of payment when introducing our study and more specifically, why we ran Experiment 1A and 1B (p.8):

“The effect of time (or pace) and reward/punishment might even interact in unexpected ways. For example, some studies found that in fast-paced tasks participants slowed down more after errors when they were rewarded for correct responses and punished for incorrect responses (Riesel et al., 2012; Stürmer, 2011). By contrast, in their slow-paced task, Williams et al. (2016) observed the most pronounced post-error effects in the unrewarded condition. Importantly though, Williams et al. (2016) observed speeding after errors rather than slowing. Given this pattern of results, we first attempted to replicate post-error slowing in a self-paced task (Experiment 1A), in which reward and punishment were delivered (Experiment 1B). “

  *6. I was a little confused about the distinction between response speed errors and evidence quality errors. Are response speed errors due to participants having decision thresholds set too low? And are evidence quality errors due to the stimuli being the source of poor evidence or the participants internal evidence accumulation rate being poor? Difficult to distinguish the evidence quality account unless there was some sort of stimuli manipulation.*

**Reply**: We added a clarification as suggested by you (p.6):

*“*In particular, it seems that control over the outcome might determine to what
extent participants slow down, or by contrast, speed up when something goes wrong. For example, Damaso et al. (2020) investigated changes in response speed after different types of errors. They distinguished between ’response speed errors’ and ’evidence quality errors’. Similar to Yeung and Summerfield (2012), they defined ’response speed errors’ as errors in which participants simply responded too quickly (i.e., they did not sample enough information to make a correct choice; in other words, their decision threshold was too low). On the other hand, ’evidence quality errors’ were defined as (very) slow errors, which were committed due to the very poor quality of the stimulus (i.e., there was insufficient evidence for either response; for a similar distinction see also Beatty et al., 2018).”

 *7. Can objective controllability be more clearly defined in the present study section as well as the distinction of this studies work to Dyson?*

**Reply:** This is a good point; People can have indeed control over several elements in our task. But here we were interested in the ‘control over the outcome’ and thus clarified this in the manuscript: We have adjusted the label as ‘control over the outcome’ throughout the entire manuscript to make it clearer that we are talking about control over the outcome in contrast to games of chance.

With regard to the distinction to Dyson et al 2018, we now write on p.7/8:

“Thus, to further investigate the role of overall task context, in the present study, we investigated whether the objective controllability over the outcome indeed determines whether participants speed up or slow down after sub-optimal outcomes. In contrast to Dyson et al. (2018), we did not test this in a gambling-related context, but in a task that typically allows (at least some) control over the outcome. But before we could test our ‘control over outcome’ hypothesis, we first needed to rule out other task features that could influence speeding vs. slowing. First of all, gambling tasks are often self-paced, which means participants press a key to start the next trial. However as mentioned above, the post-error slowing literature suggests that the pace of a task can influence the extent to which people slow down (and even speed up) after errors, regardless of controllability over the outcome. Second, in gambling tasks participants can usually win or lose points (or money). Thus, negative outcomes in gambling tasks might be more motivationally salient compared to errors in most choice reaction time experiments, in which no reward or punishment is delivered after every outcome.”

*8. Maybe best to cite Rouder’s paper on optional stopping to justify the BF calculated for 50 participants then a subsequent 50.*

**Reply**: We added this in our sample size rationale (p.9):

“Importantly, the Bayesian sequential testing procedure allows us to interpret Bayes Factors despite optional testing (Rouder, 2014; Schönbrodt & Wagenmakers, 2018; Schönbrodt et al., 2017; but see also de Heide and Grünwald, 2021).”

 *9. Were the dark and light response mappings for the keyboard counter balanced? Perhaps not an issue unless you are interested in dark vs light response performance.*

**Reply**: No, they were not counter-balanced. We added a sentence to make that clear (p.10):

*“*Note that we did not counterbalance the stimulus response mapping.”

 *10. “Numerically, we also found a post-error slowing effect in the choice RT (18 ms) but the evidence for this effect was inconclusive (BF10 = 0.619).” The Bayes factor here suggest no post error slowing (assuming 1 is the alternative and 0 is the null in the subscript), albeit inconclusive evidence. In any case it would be best to clear up the sign of the BF early. >1 evidence in favour of slowing and <1 evidence in favour of the null.*

**Reply**: We added an explanation about Bayesian testing and also changed that description (p.12):

“A Bayes Factor > 1 is in favor for the alternative hypothesis, whereas a Bayes Factor < 1 is in favor of the null hypothesis. A Bayes Factor around 1 yields inconclusive evidence. The size of the Bayes Factor determines the strength of the evidence: anecdotal (1/3 − 1; 1-3), moderate (1/3 − 1/10; 3-10), strong (1/10 − 1/30; 10-30), very strong (1/30 − 1/100; 30-100) or extreme (< 1/100; > 100. We used the default prior width as used in R. For Bayesian t-tests this prior width is 0.707, corresponding to a medium effect.”

“Numerically, we also found a post-error slowing effect in the choice RT (18 ms) but the evidence for this effect was (inconclusive) in favor of the null hypothesis (BF10 = 0.619).”

 *11. You mention you will report p values but these are never reported. Given the optional stopping its probably best they are not reported, however reference to p-values should be removed as the reader will be looking for them.*

**Reply:** We omitted all reference to the p-values in the main text, but we kept them in the tables for completeness. *12. In Table 1, I personally think it would be best to round to 2 decimal places for aesthetics. This could be done for statistics throughout the manuscript.*

**Reply:** We have now rounded the statistics to 2 decimals.

 *13. What was the justification for stopping data collection after only 1 effect was decisive? For instance, if the BF is moving for both effects are moving towards evidence for the alternative would it not be best to collect data until conclusive evidence is obtained for all effects?*

**Reply**: We were only interested in whether we would observe speeding or slowing when simply adding reward. As both effects pointed in the same direction in the previous experiment and as our cut-off for the Bayes Factor was quite conservative, we decided that strong evidence for one of the dependent variables would be sufficient. Eventually, both effects reached the crucial BF (p.14).

*“*We only wanted to know whether we can observe slowing or speeding after errors when simply adding reward. As both effects pointed in the same direction in the previous experiment and as our cut-off for the Bayes Factor was quite conservative, we decided that strong evidence for one of the dependent variables (i.e., start RT or choice RT) would be sufficient. Eventually, both effects reached the crucial BF (for details on the Bayesian t-tests see online supplementary material).”

 *14. “our main processing-“ this dash could just be a comma*

**Reply**: Yes, thanks. We changed that.

 *15. A bayes factor would be more appropriate for determining if there was no difference between size of slowing between experiments given you want to provide evidence for the null hypothesis.*

**Reply**: Indeed.As we did not specify in our preregistration how we would test between-experiment differences, we moved the initial chi² test to the supplementary material and report a between-experiment Bayesian ANOVA instead (p.16):

“However, Bayesian ANOVAs with experiment as between-subject factor indicated that there was no difference in the amount of slowing between the experiment without reward/punishment (Experiment 1A; for further information see OSF) and the experiment with reward/punishment (Experiment 1B). Specifically, the ANOVA yielded moderate evidence for the null hypothesis (BF10 = .0169) in the start RT and anecdotal evidence (BF10 =.308) in the choice RT for the interaction between experiment and previous outcome. Thus, the delivery of reward and punishment could not explain the discrepancies between studies, or more generally, why people slow down or speed up after errors or sub-optimal outcomes. For this reason, we also omitted the reward/punishment manipulation from the remaining experiments.”

 *16. Given exp 1A and B have 6 responses, does Hick’s law have any influence on post error slowing?*

**Reply:** Generally, yes. Hick’s law influences error awareness but the number of choice options does not seem to influence slowing much (at least, for ‘aware’ errors; Yeung et al., 2004; Maier et al., 2010). Here we presented feedback, thus we assume that Hick’s law does not play a role here. In any case, we still observed post-error slowing in the experiments with 6 response options (which is for our purposes, the most important finding).

 *17. Was the seed fixed in the previous analyses?*

**Reply:** The analyses of Experiment 1 were Bayesian t-tests; for these tests, the BF results are relative stable, so we did not fix the seed. In Experiment 2, we also ran Bayesian ANOVAs. Bayesian ANOVAs require running simulations; this can result in different Bayes factors when the analyses are rerun. To avoid this, we fixed the seed for the ANOVAs.

*18. With the staircase procedure, if many of the errors were made after the change in stimulus difficulty, wouldn’t this mean that there is a confound correlated quite strongly with error occurrence and first half of experiment? Or was the definition of first half of the experiment AFTER the staircase procedure?*

**Reply:** The experimental part was only determined in the experimental blocks, so after the staircase procedure training blocks. We added a sentence on this (p./18):

“Note that the proper experiment only started after the individual difficulty level was determined for every participant. Thus, our staircase procedure blocks were not included in the analyses.”

 *19. “To determine statistical significance, we used an alpha of .05.” How does this relate to the Bayesian ANOVA?*

**Reply**: Originally, we reported frequentist results alongside the BF for completeness. In order to address your comment and the comment of the editor, we omitted all frequentist results from the main text (except in our tables and the above-mentioned sentence).

 *20. For main effects, were inclusion Bayes factors or p values being used as inferential evidence of the effects? The word significant is used in this section but it is not clear what statistics you are using for your inferential decisions given you ran a Bayesian ANOVA.*

**Reply:** Please see our response to your previous comment.

 *21. Are the slopes for the error count analysis reported anywhere?*

**Reply:** They are reported in the revised version of the manuscript (p. 29):

In order to explore post-error slowing and speeding effects as errors accumulate, we measured the difference in start RT between post-error and post-correct trials by error count per group. In order for the model to converge, we used z-scores of the error counts. In the start RT in the ’easy’ group, we found an effect of error count, χ2(1) = 7.09 p = .007, showing that participants slowed down after errors in the beginning of the experiment but less so when errors accumulated (b = -13.12). Surprisingly, participants in the ’hard but doable’ group sped up after error trials in the beginning but this effect was also reduced towards the end of experiment (b = 16.73), χ2(1) = 10.86, p < .001. The ’impossible’ group showed a similar numerical pattern as the ’hard but doable’ group (b = 8.98), which was however not significant, χ2(1) = 2.70, p = .100. In the choice RT, the numerical pattern was the same as in the start RT but the effect of error count was only significant in the ’easy’ group (b = -10.22), χ2(1) = 5.47, p = .019, but not for the ’hard but doable’ group (b = 6.31), χ2(1) = .92, p = .337, or the ’impossible’ group (b = -2.45), χ2(1) = 0.19, p = .659. “

 *22. In any of the experiments did you explore differences in response? For instance light and dark response mapping wasn’t counterbalanced so it would be good to check statistically if there were differences between the response type.*

**Reply:** Yes, it is possible that there are differences between responses to dark and light stimuli (e.g., most people are right-handed, and we assume people to be faster with their dominant hand). We did not counterbalance the mapping though. Instead, we opted for a ‘natural’ mapping, as dark is usually presented on the left and light on the right. If we would have counterbalanced the mapping, we could have introduced compatibility effects (which we wanted to avoid).

*23. You state “further tested the idea that objective controllability might play a critical role.” I think this is a useful research direction. But can it be clarified what you did that was above and beyond what is already known. For instance based on Dyson and the Damaso paper you could already conclude that participants who feel in control slow down after sub-optimal outcomes, whereas those who do not feel in control over the outcome do not slow down or speed up. Perhaps the Dyson study was in the context of gambling tasks and your study is in the context of perceptual decision making. In any case, can the justification of running these controllability studies be fleshed out more.*

**Reply:** Please see the **response to one of your comment # 7.**

*24. In addition to justifying the focus of controllability more clearly, I think the theoretical treatment could be expanded on in the discussion. You note that controllability as a construct explains the effects of the controllability as a manipulation, however this seems circular and unsatisfying. What mechanism would need to change between easy and hard to account for the effect of controllability?*

**Reply:** We have added what mechanism might be at play here (p. 32/33):

“We found that participants from the easy group – who had indeed control over the outcome – slowed down after errors. These participants could improve their performance on the task by accumulating more evidence (i.e., increasing response thresholds). However, participants in the other two groups – who had an error rate of 50% and no control over the outcome as the outcome was predetermined – sped up after error trials. These participants might have realized that no matter how much evidence they accumulated, they would not be able to improve their performance on the task and therefore sped up.”

 *25. Another effect that is left unexplained is the dissipation of the effects. To explain you note that maybe speeding occurs first because setting cautious thresholds appears to do nothing to the accuracy of their performance in the hard condition. Then over time, speeding appears to disappear because participants give up. However, why does giving up not appear as speeding through trials to end the experiment early?*

**Reply:** Participants might have initially cared about the outcome, which is reflected by speeding after errors compared with correct trials. A general speeding towards the end combined with the dissipation of the post-error effect indicates that eventually participants simply gave up and clicked through the experiment (p. 35).

“Surprisingly, Experiment 3 showed that the effects in all three groups dissipated over time. Although the decreasing post-error slowing in the easy condition can be explained in terms of decreased orienting to accumulating errors, the pattern in the two difficult conditions is more surprising. One tentative explanation is that post-error speeding dissipates over time due to ‘learned helplessness’ (following the initial frustration). For instance, Mikulincer (1988) showed that participants who were repeatedly exposed to unsolvable problems initially tried harder but eventually gave up. We observed a similar pattern in an earlier gambling study (Eben et al., 2020): in a not very engaging gambling task, we found general speeding on gambling trials (compared with non-gambling trials); furthermore, participants initially showed more speeding after (gambled) losses compared to (gambled) wins, but this effect dissipated over time. It seems that we observe a similar pattern here. As the task in Experiment 3 was not very engaging, participants might have stopped caring about the outcomes of their actions as the experiment progressed, resulting in reduced difference between post-error and post-correct trials. Another indication that participants eventually gave up on doing the task is a general speeding towards the end of the experiment (see Figures 2 and 3 in the supplementary materials). Whether this motivational account (or more specifically, reduced motivation as the experiment progresses) could also explains the reduced post-error slowing in the easy condition needs further testing though.”

References (not mentioned in the main manuscript):

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological review*, *111*(4), 931.

Maier, M. E., Steinhauser, M., & Hübner, R. (2010). Effects of response-set size on error-related brain activity. *Experimental Brain Research*, *202*(3), 571-581.

**Editor Final Decision:** Accept

Feb 21, 2023

Dear Charlotte Eben,

I have now had a chance to read over your manuscript “When response selection becomes gambling: post-error slowing and speeding in self-paced colour discrimination tasks”, along with the letter describing the changes you made. I was also able to receive a review from one of the original reviewers. Both the reviewer and I were very happy with your revision and your responsiveness to all of the concerns, and so I am happy to say that your paper is now officially accepted for publication in Collabra: Psychology. Congratulations on this excellent work, I think it will make an important contribution to the literature and I look forward to seeing it published! I hope your experiences with Collabra: Psychology have been positive and that you will continue to consider it as an outlet for your work.

As there are no further reviewer revisions to make, you do not have to complete any tasks at this point.

You will be receiving separate correspondence regarding any production and technical comments, data deposits, as well as publication charges. We work with the Copyright Clearance Center to process any applicable APC charges. Please note that your APC transaction must be completed before your article gets published.

You will have an opportunity to check the page proofs before we publish your article. Thank you again for publishing in Collabra: Psychology.

Sincerely,
Nathan Evans

# Reviewer 1

##### Rating scale questions

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  | ✔ |  |

##### Open response questions

### Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

Overall the authors have addressed all y concerns that I previously mentioned. I am happy for this manuscript to be published as is.