Dear Dr Diederich, Editor in Chief, *Journal of Mathematical Psychology*

Dr Colonius, Action Editor and Guest Editor of the special issue

Thank you for considering our revised manuscript “Assessing cross-modal interference in the detection response task”, for publication in the special issue of the *Journal of Mathematical Psychology* on computational and mathematical approaches to multisensory integration. We thank the Reviewers again for their thoughtful comments and provide our detailed response below. Edits to latest round of reviews are marked YELLOW in the revised manuscript. For completeness, we kept changes from the previous round marked as well (GREEN). We believe the manuscript is now ready for publication in the *Journal of Mathematical Psychology* and hope you agree.

Sincerely,

Alex Thorpe,

The University of Newcastle

Response to Editor and Reviewer comments:   
  
Reviewer #1: One issue: The current manuscript could benefit from integrating the relevant recent research on the topic of workload capacity and the role of distractors (Little, Eidels, Fifić, M., & Wang, 2015; Little, Eidels, Fifić, & Wang, 2018) that are highly relevant for the theoretical considerations on the page 11, second to the last paragraph).

>>> We thank the reviewer for this suggestion. Information regarding workload capacity in the presence of distractors and the resilience function has been added to the Introduction section (p. 10).

The authors responded to all of my suggestions in a satisfactory manner. Overall, I think that the manuscript describes an interesting research project that could further improve our understanding about the perceptual mechanism underlying the detection response task (DRT and the system's capacity in the cognitive operations. There is a notable integrative effort to combine previously separated approaches, such as the measures of workload capacity (as provided by the system's factorial technology) and the classical detection task, thus increasing the validity of the findings. These crossovers between different technologies can be expected to have a facilitatory effect of the future work on this matter.

>>> Thank you.  
  
  
  
Reviewer #2: I cannot support publication of this study. I will not go into details again (I already did this in my first review), I will just list a number of aspects that illustrate the overall impression I got from the study.

>>> We believe we have addressed the Reviewer’s concerns and implemented her/his suggestions in the last round. For the remaining points, please see our response below.

1. "Mean response times (RT) and hit rates (HR) were compared using repeated-measures ANOVA and Bayesian repeated-measures ANOVA. Bayesian tests were applied only in cases of non-significant or ambiguous results, to present evidence in favour of no difference; this cannot be achieved using frequentist tests alone." (p. 19)

If the F-test is not significant, the authors look further what the Bayesian test says. This two-stage procedure is statistically flawed. In statistics, "the more the merrier" does not hold. The analysis should follow from the hypotheses.

>>> There are many arguments in favour of Bayesian statistical analysis (articulated clearly in work by EJ Wagenmakers, Richard Morrey, Jeff Rouder, John Kruschke, Dani Navarro, and others), and in theory we could have reported only Bayesian tests. Yet despite its appeal, Bayes Factors have not yet become the ‘industry standard’ in reporting psychological results, and it is therefore not quite clear whether it is sufficient to report Bayes Factors alone, while the majority of students worldwide are still taught to interpret frequentists tests. For this reason, we see merit in reporting frequentist tests at baseline, and supplementing them with Bayes Factors when we are seeking evidence in favour of the null hypothesis -- which Freq’ tests cannot provide. This is the approach taken in the present manuscript. We could have discarded all frequentists tests and reported only Bayes Factors, but we feel this change would be too radical and would not serve the best interest of the JMP readership. The approach we have taken is quite different from, say, running two sets of analyses and choosing to report the test-outcome that best supports our hypothesis, which we whole-heartedly agree would have been a flawed process.

2. Pseudo-RTs for the tracking data: In my first review I have demonstrated that this measure is flawed and I have shown that there is no monotonical relation to performance. Instead of a convincing counterargument, the authors kept the analysis of this flawed measure and "fixed" the problem by running a few simulations and sensitivity analyses and copying my comment into the paper.

We thank Reviewer 2 for reiterating their concern re- pseudo-RTs for the tracking data. In the lack of more concrete text in this round of revisions we went back to her/his original review, and will assume the relevant concern is this:

“*Regarding the pseudo-RT transformation: If RT is used as a performance measure, better performance should be associated with lower RTs. Consider a very good participant that shows perfect tracking behavior. The needle of this participant would never leave the yellow area, and you would end up with no data. Now consider a participant that randomly moves the needle up and down very fast. This participant would have very fast responses, although "performance" is actually very poor. The first example already illustrate that the pseudo-RT-transformation does not work. The second example may not be applicable to the present experiment because the speed of the needle is probably kept constant by the program, but the "idea" of a tracking task (i.e., with a proper response device) would actually allow different speed.*

>>> The same point was indeed raised by the Reviewer in the first round. We already agreed in our first revision that the more lenient the threshold, the fewer ‘pseudo trials’ we record, and acknowledged the issue in the text. We undertook an analysis exercise to assess how much of an impact this issue might have on the outcome of capacity analyses. Appendix B (p. 48) and Figure 14 show relatively stable capacity estimates across different data-set sizes. But, that is not the main point with respect to the Reviewer’s current concern. The main point, rather, is that capacity estimates were stable irrespective of the tracking error \*tolerance-level\*. That is, we repeated the analysis of the same tracking-error data reported in Results, for different levels of error-tolerance (epsilon). And, when doing so, the overall pattern of results (namely, limited capacity) changed very little, as can be seen in Figure 14.

Now, consider a very ‘good’ subject, as in the example provided by the Reviewer. This subject can track the moving target accurately and stay close to the central needle. Yet, even a very good subject cannot track the needle perfectly across the entire course of the trial (this is not an assumption, but rather an empirical fact). If we ‘zoom in’ the analysis, and minimize the error tolerance level, then at some points in time we will discover deviations of the tracking marker from the target location. If the subject is as good as suggested, they should be able to quickly get closer to the target and will manifest speedy pseudo response times, as expected. But the more interesting test-case is this: What if a very good subject can follow the target very closely, but never quite be on top of it? In that case, with a relatively large epsilon we might record very long response times for a subject who is actually performing quite well, theoretically confounding the analysis. This is where the analysis reported in Figure 14 becomes useful: if that were the case, then systemically varying the tolerance level should, at some critical point, be affected by the constant yet small error. As soon as the error-tolerance index, epsilon, gets smaller than the subject’s error, these deviations should manifest as errors, captured as frequent and quick pseudo response times, and affect the capacity analysis. Our dedicated analysis, however, shows this does not happen (see Figure 14).

We suspect, based on the Reviewer’s comment that this explanation did not quite come through in the original text. We apologise and have now improved the clarity of exposition [p. 35]. There could be other possible solutions for this issue. For example, one could normalise the amount of time spent out of the error-tolerance range by overall task duration (so that good subjects score low, as they spend little normalised time out of the ‘permitted’ range, and poor subjects score high). Or, for another example, our own team member, Rachel Heath, is exploring methods that subject tracking error to sophisticated fractal analysis based on her non-linear dynamics approach. But, these and other techniques offer completely different analyses, and importantly, do not produce the desired pseudo-response times that allow us to connect tracking data with response-time analysis tools.

3. Summary of the main findings: Participants were slower in condition A than in condition B. The modality of the distractor did not matter. Instead of honestly reporting this modest finding, the reader is flooded with poorly motivated models (Wald) and extra analyses of capacity coefficients that are nothing more than log transforms of RT distributions. Since the log transform is a monotonical transformation, we have K\_A(t) < K\_B(t) whenever we have the same relation for the survivors S\_A(t) < S\_B(t) and vice-versa, and from S\_A(t) < S\_B(t) follows the same relation for mean RT. As a side note, the many numerical problems for t = 0 still exist in Figure 11, 14, 17, 18. In my last review I illustrated the problem (basically, log S(t) = -inf for t -> 0).

The Reviewer pointed out the numerical problems associated with computing a ratio of log transform (e.g., integrated hazard functions) in the last round of revisions. In the Reviewer’s own words “*C(t) suffers from an in-built design flaw because it is defined as the ratio (!) of two logarithms C(t) := log S\_AB(t) / [log S\_A(t) + log S\_B(t)] instead of the difference, C\*(t) := log S\_AB(t) MINUS [log S\_A(t) + S\_B(t)], which would be much more natural to use with logs anyway… I tend to consider this as a minimum requirement for a revision. In fact, the authors noted this problem themselves at a later page.”*

>>> In response, we modified in R1 (the previous revision) the entire analysis to a difference form, as suggested. This change was also implemented in Figures 10, 13, 16, and 17. We believe this response should have satisfied the reviewer’s concerns, and in fact find it a bit unfair to bring it up again after the requested changes to the analysis had been made. As a side note, any numerical problem for t=0 might be important mathematically, but practically inconsequential for behavioural data; the common assumption is that humans simply cannot respond so quickly and such responses are typically censored anyway.

4. Cohen's d, eta-square, z-standardization of capacity coefficients: The inclusion of dimensionless (and therefore meaningless) effect measures did not in any way improve the paper. In fact, response time has a nice SI unit, milliseconds, which is easily communicated. Translating meaningful measures to standard deviations of an arbitrarily chosen experimental condition does not add any information. Did anyone ever enter a grocery store to buy half a standard deviation of milk?

>>> We have included eta-square values at the specific request of Reviewer 1. As for Cz, the measure developed by Houpt and Townsend in 2012 was added to aid inference. By now Cz is considered common practice in SFT analysis (e.g., cited in Google Scholar more than 120 times, 24/4/2020). Its main contribution is supplementing inference from capacity analysis, which used to be based only on visual inspection of the capacity coefficient plots. It appears to now be accepted as a legitimate and useful statistical tool.

5. Figure 1 presents a linear ballistic accumulator, whereas a diffusion model is used in the text. I am again wondering why none of the 6 coauthors noticed such inconsistencies.

>>> We thank the reviewer for noting the model presented in Figure 1 is an LBA and not the Wald model, yet at the same time note the figure is said to provide a general illustration of “the process of making a decision as described by a sequential sampling model” [SIC; and we don’t mean it here as the acronym for Survivor Interaction Contrast]. Due to its simplicity the LBA provides a convenient platform for illustrating Evidence Accumulation and the associated parameters. This was not an inconsistency, but rather a deliberate choice to enhance the clarity of the exposition. It may work for some but perhaps not for all, and we understand the Reviewer’s concern. To avoid confusion, we removed this figure from the manuscript, as all relevant information it depicted is described in-text.

6. The purpose of the study is not clear. There's no real theory other than replication of Wicken's findings, and the methodological contribution is questionable (see 2 and 3). In some sense, the authors acknowledge this in the last sentence, "Our findings also present novel methods for applying cognitive models to more real-world data sets in the future."

>>> That is a fair point, and we now have improved the exposition of the goals in the first page of the Introduction to ensure the purpose is stated clearly and early.