Dear Dr. Sreekumar,

Thank you for giving us the opportunity to revise the manuscript of the Stage 1 Registered Report again. We have addressed the reviewers' comments. Please find our responses below.

Changes to the text in the manuscript are highlighted in yellow for ease of reviewing.

Reviewer 1

1. *All in all, I have the impression that the authors put a lot of work into the revisions and that the Stage 1 manuscript has significantly improved.*

We thank the reviewer for these kind words.

2. *Concerning my point 4: No changes were made to the manuscript. While it is interesting to read how the authors plan to perform sequential sampling in their rebuttal, they must make sure the manuscript is clear on this point, too.*

We elaborated on the sequential sampling plan in the Participants section (p. 10) in order to clarify the criteria to stop testing.

3. *Concerning my point 5: I do not entirely agree with the authors. While they are correct that their Bayesian approach does not require a frequentist power analysis, it does require justifications for choosing borders of the sampling plan. This was the reason for my question and the authors should include such justifications into the manuscript. (Why is 6360 the upper limit?)*

The upper limit of Nmax = 6360 (i.e., 60 participants per word list) represents a tripling of the lower limit Nmin = 2120 (i.e., 20 participants per word list) and constitutes the practical extent of resources we have available for this RR. We have therefore set Nmax to meet two important criteria – it is high enough that the hypothesised effects are likely to be detected even if they are much smaller than in the expected memory task, and yet it is not so high that it requires an unreasonable amount of resources to detect effects. While alternative Nmax limits are of course possible, we feel the one we selected offers the best balance between these two criteria. In other words, if a particular effect is not detected in the planned study by Nmax, it is likely too small to be of interest.

We have now included a more detailed justification of Nmin and Nmax in the manuscript (p. 10).

4. *Concerning my point 6: Please explicitly state in the manuscript that no outliers will be excluded and that no outlier correction will be performed (e.g., Windsorising) beyond the exclusion criteria.*

We have included this statement in the manuscript (p. 18).

5. *Concerning my point 7: I do not agree with the authors. Since the results from the published study were not collected in an unbiased environment, effects such as regression to the mean makes it likely that the RR will differ from the original study. Any comparison the authors draw between the two datasets will be affected by this. Therefore, I would strongly urge to include the explicit condition. But I would leave it up to the editor to decide, whether my concern is crucial to Stage 1 acceptance.*

We acknowledge the reviewer's concern about the comparison of results from different sources, and we sought the Editor's view on this issue. The Editor's opinion is below, together with our response:

Editor's correspondence:

> *As long as the experimental protocol here closely matches your prior work (with the explicit memory tests) in all aspects except the surprise memory test, I am happy to accept a comparison of your new results to the old ones without adding the explicit condition to your new study. Therefore, I'd encourage you to write explicitly about how the design ensures that any differences in relative FA increase or decrease you see between the current and previous studies can be attributed solely to the memory test being a surprise one and that there are no other potential confounds that could explain the difference. Reviewer 1 was probably (rightly) concerned about exactly this, so if you choose not to include an explicit memory condition, you should include a strong justification. For instance, reviewer 2 in the earlier round asked if 20 participants per word would be sufficient. Are all these details similar to the previous study? Would item variability due to having only 20 participants per word (or a much higher number of participants if your sampling plan ends up leading you there) lead to a difference in results between the two studies by chance? This is, of course, just one possibility, but I am sure you could think of others. Acceptance would be contingent on this justification for leaving out the explicit condition being convincing*

We would like to note that our RR hypotheses rely on the results of the surprise memory task only; none of our registered hypotheses involve a comparison between the surprise and expected memory tasks. That is, the effects of the Body component scores on hit rate and false alarms in a surprise memory task will indicate whether adaptive advantage or somatic attention drive the memory for body-related words (as outlined in Table 1, p. 7).

Nonetheless, we do plan to discuss the results of the surprise memory tasks in relation to previous results for an expected memory task by Dymarska et al. (2023) in order to obtain a broader understanding of semantic richness theory. We agree that for such a discussion we need to ensure that it is clear whether any differences stem from the difference between the tasks (i.e., surprise or expected) rather than other potential confounds or differences in design, procedure, data handling or analysis. We are confident that such discussion of task differences can be reasonably carried out with respect to the existing megastudy data from Cortese and colleagues' expected memory task (i.e., as analysed in Dymarska et al., 2023) and outline our justifications below. Please forgive the length of the following response, but felt it was necessary to be as thorough as possible.

First, we plan to use the same items and list design as Cortese et al.'s expected memory task. That is, we will present participants with the same target items (i.e., all words in the present RR were also used in Cortese et al.'s studies), we will divide them into lists of the same size as used by Cortese et al. (50 words), and we will balance presentation of study and test lists in the same way (i.e., lists are paired so that list A serves as distractors when testing memory of list B, and vice versa – this is now made clearer on p.12). In this way, the 5300 words that we will test for sensorimotor effects in a surprise memory task in the present RR have already – using the same design – produced these sensorimotor effects in an expected memory task (Dymarska et al., 2023), so **any differences in results between tasks cannot result from by-item variability.**

Second, we also plan to use the same experimental procedure as Cortese et al. (apart from the critical task manipulation in the study phase). We will implement the same distractor task between study and test phases as that used by Cortese et al. (i.e., verification of 18 simple addition and subtraction problems), and in the test phase itself we will run the recognition memory task as per Cortese et al. (i.e., target items presented in a different random order to the study phase, intermixed randomly with distractor words, where participants respond by pressing a "new" key on the left or an "old" key on the right). While Cortese and colleagues did not include a time limit in their original test phase procedure, such a timeout is required to make online testing feasible; we therefore opted to implement a limit of 3000 ms based on a similar task in Dymarska (2023), where 3000 ms was more than 4 SD above the mean response time in a surprise memory task. As so few responses are likely to be affected by this time limit, **it does not constitute a plausible reason for any systematic differences in results between tasks**.

Third, we have ensured that the Nmin sample size of 20 participants per word list matches the number of participants who saw each word list in Cortese et al. (2010). That is, the sample size of 20 data points per word that was sufficient to detect critical effects in an expected memory task (Dymarska et al., 2023) is our starting point for testing effects in the present surprise memory task (i.e., we may end up collecting data from more participants due to sequential hypothesis testing). While it is possible that by-participant variability could introduce some volatility to the item-level scores (e.g., a low-performing participant could drag down HR, HR-FA and d' while increasing FA), such a risk is mitigated by the sheer size of the megastudy sample involved. That is, having 20 participants per word list would be a concern if memory performance were being examined for a small number of words because each participant would have a relatively large effect on the total dataset. For instance, if a study examined only (say) 100 words in the present design (i.e., 20 participants study a list of 50 target words and are tested on 50 targets + 50 distractors), then each participant would contribute scores to all words and comprise 2.5% of the total dataset, meaning that participant variability could easily influence results. However, because our planned study uses a very large number of words (5300), the present design means that each participant (at Nmin) would contribute scores to only 100 out of 5300 words and comprise only 0.05% of the total dataset, meaning that **by-participant variability cannot plausibly influence results in a way that produces task differences**.

Fourth, our planned data exclusions follow those of Cortese et al. (e.g., we employ the same participant replacement and outlier removal criteria) and the analysis plan and predictors are identical to our earlier work on the expected memory task (i.e., regression analysis with 2 lexical and 4 sensorimotor predictors derived from PCA: Dymarska et al., 2023). As previously noted, this consistency ensures that **any difference in results between tasks cannot be attributed to differences in statistical models.**

Finally, we note that the reviewer was concerned about regression to the mean (RTM) producing differences in results between the present RR and the previous analysis in Dymarska et al. (2023) of the Cortese megastudy dataset. While the large samples of both items and participants in the Cortese megastudies reduces the risk of subsequent samples exhibiting RTM (i.e., because large samples are less likely to draw asymmetrically from the distribution), the nature of our hypotheses mean that – even if RTM is observed – it will not confound the interpretation of results. If memory performance in the present RR regresses towards the mean, then *all* hypothesised sensorimotor effects will be reduced: that is, if a null Body effect on FA is accompanied by null effects of Food and Object components (against hypotheses), then it indicates RTM and will be interpreted as such (see Study Design Table on p. 25). However, if a null Body effect is accompanied by the hypothesised facilitatory effects of Food and Object components, then RTM is not occurring, and the critical Body effect can be attributed to the adaptive advantage account. (Or likewise, if an actual Body effect on FA is accompanied by Food and Object effects, then it can be attributed to the somatic account). In other words, **regardless of whether the present RR surprise memory task exhibits RTM relative to the analysis reported in Dymarska et al. (2023), our design allows us to test and interpret our hypotheses appropriately.**

Reviewer 2

1. *This is a good revision that addresses many of my concerns.*

We thank the reviewer for these kind words.

2. *My biggest point that I would like to make is that I maintain my position that mixed effects analyses would be useful here. I understand the authors' rationale for wanting to maintain some consistency with a previously published analysis. However, I would recommend doing mixed effects analyses *in addition* to the current analyses to see whether the results are comparable. I would also like to mention that Jeff Rouder and colleagues have developed techniques for performing hierarchical Bayesian mixed effects models that can calculate d' at an item level. The following papers may be useful:*

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application to the theory of signal detection. Psychonomic Bulletin & Review, 12, 573-604.

Pratte, M. S., Rouder, J. N., & Morey, R. D. (2010). Separating mnemonic process from participant and item effects in the assessment of ROC asymmetries. Journal of Experimental Psychology: Learning, Memory & Cognition, 36, 224-232.

We agree that it would be interesting to run additional mixed-effects analyses and we hope to include the suggested Bayesian mixed-models as exploratory analyses in the Stage 2 submission.

3. *The introduction gives a much clearer perspective on the theories of interest. However, I think it would also be useful to include an additional paragraph on some of the methodological issues, such as controls for word frequency and other things that their analyses consider. Otherwise the mentions about these word level controls come in out of nowhere in the Methods section.*

We have amended the introduction to introduce the word level controls earlier in the manuscript (p. 2-3).