Dear Dr. Sreekumar,

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

We would also like to note two minor corrections that we have introduced to the method section after a careful inspection:

- In materials on p. 10, we clarified that in order to create 16 lists of 50 words each, we will remove 5 words with lowest UK prevalence scores of all will be removed, rather than removing 5 words at random.
- In the procedure on p. 14, we fixed an error regarding the feedback message for practice trials in the lexical decision task (study phase), where feedback will comprise symbols (green tick for correct vs. red cross for incorrect) rather than words ("correct" vs. "incorrect").

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

Reviewer 1

> *1. The scientific question seems valid, but somewhat underwhelming. Maybe the authors could give some details about the size of this effect and how it interacts with other important cognitive processes. Currently, the question seems somewhat isolated and I am unsure what we really gain from answering it. I am sure there is an appealing reason to study this effect, but currently the authors do not include this rational into the stage 1 report.*

Semantic effects on word memory were generally considered to be straightforward (more imagery, concrete sensory information, etc. in the representation leads to better memorability) until we found a rather counterintuitive effect (Dymarska et al., in press). We found that for concepts involving bodily experience (i.e., interoception, hand and foot action, haptic experience etc.), participants were producing slightly higher hit rates (due to richness of representation) but also higher false alarms that ultimately impaired discrimination of the old and new items.

Semantic richness theory is perhaps the most detailed account of precisely why semantic information affects word memorability, and predicts that richer semantic representations = stronger semantic activation = stronger memory trace = better memorability (Hargreaves et al., 2012; Pexman et al., 2003; Sidhu & Pexman, 2016; though please see our response to Reviewer 2, point 9, for nuance). Since our results did not fit these predictions, we put forward two other explanations (adaptive attention account and somatic attention account), and suggested that semantic richness theory needs to be expanded to accommodate the process underlying this result. However, we could not distinguish between the two possibilities in that paper, due to the nature of the task that was employed in the dataset we analysed: an expected memory paradigm where participants may have used strategies to elaborate on the word's meaning (see response to Reviewer 2, point 11). We think that employing a surprise memory task minimises the likelihood participants will employ such strategies, which leads to different predictions for the two accounts of the mechanism underlying Body effects. We believe that this study will go a long way towards determining how and why some forms of sensorimotor information in word meaning can hinder rather than help memorability.

Because of such clear predictions that allow us to answer a specific question, we considered this study to be a suitable candidate for the registered report format, where we already have a set of stimuli, some idea of what effect sizes are elicited (though we account for the possibility that they will be smaller than in an and expected memory task: see point 5 below), and what effects the predictor variables elicit on lexical decision (i.e., the task in the study phase). We therefore can pursue the

specific unknown with confidence, hoping to adapt the semantic richness theory to these findings in a way that will be beneficial for future research on semantic effects on word memory.

> *2. For me it remains somewhat unclear, if Hypothesis 3 really takes care of all scenarios. A potential other explanation for increased false alarms is greater similarity of word meaning between lures and targets for the items with higher Body component scores. Can the researchers rule that out? I would also like them to summarize the other alternative explanations that may be construed and how their design takes care of them.*

We agree with the reviewer's suggestion that greater similarity of word meaning can influence the participants' ability to discriminate between old and new items and thus inflate FA. Indeed, in Dymarska et al. (in press) we ruled out low distinctiveness of Body items as the reason for inflated FA, since most items with high Body scores have a distinct meaning that is not easily confusable with others (e.g., *bathe, climb, dance, massage*). However, it's worth noting that inadvertent similarity of meaning between lures and targets is a greater concern in factorial designs where each item list is constructed to be high vs. low on a given variable of interest. By contrast, in our megastudy regression design (same as that analysed in Dymarska et al., in press), all 6 predictor variables apply continuously to all items, meaning that any similarity of meaning between lures and targets would have to apply simultaneously to all 6 predictor components rather than just the Body component. Nonetheless, it is possible in principle that assigning words to target and lure lists at random could inadvertently lead to an increased similarity of meaning between targets and lures in some of the target-lure list pairs, which could confound the results if it coincided with higher-than-average Body scores in those lists.

In order to rule out this possibility in the present RR, we will use a binned sampling method when dividing the stimuli into word lists to ensure that all lists contain items that span from low to high scores in all 6 components. We will sample from bins (set as component score quartiles) of each orthogonal component, such that every list will include 3 words from each quartile of each component (i.e., 3 words x 4 quartiles x 4 components = 48 words), plus 2 words selected at random from different bins to bring the total to 50 words. As a result of this sampling method, the extent of the distribution from low to high across all 106 stimuli lists will be the same for every component, which means that there can be no greater similarity of lures and targets for the Body component than for any other component. If after this sampling method is employed, the Body component still elicits the effects that we outline (i.e., being the only component to increase FA), then these effects are not due to target-lure similarity, because if they were, then the same pattern would be present for all components. We outlined this sampling plan on p. 10.

Apart from target-lure similarity of meaning (and its related issue of item distinctiveness), and the adaptive advantage and somatic attention accounts which are covered in the current paper, we have not identified other alternative explanations that would lead only one out of six simultaneous predictors to increase false alarms.

> *3. The table at the end of the document was very hard for me to understand. I think it should include all hypotheses and their operationalisations as well the interpretations.*

We have elaborated on the hypotheses, the sampling plan, and the theory for which the outcome would be relevant in the Study Design Table at the end of the manuscript, and hope it is now sufficiently clear.

4. The sequential sampling plan is unclear to me. Is it not possible that a BF that has crossed the treshold can return to the undecisive region again, if much evidence against or for an effect is collected? How do you deal with this, since there are five BF that need to be across the treshold at the same time. So as far as I see it, one could cross the treshold but be undecisive again by the time the others have also crossed the treshold. But maybe I misunderstand the procedure

Yes, it is possible for Bayes Factor (BF) values of a given predictor in Bayesian regression to move in and out of the equivocal zone (i.e., crossing back and forth over the threshold) as the dataset changes with successive participants. However, with more data, evidence typically begins to accumulate consistently in favour of either the null or alternative hypothesis, and the BF value consistently clears the threshold with no return. When there are multiple predictors to consider in a Bayesian regression, the BF values for some effects may stabilise earlier than others, but all effects will accumulate evidence at their own rate until there is sufficient data for them to stabilise clear of the threshold. The only question remaining is whether one has sufficient resources to keep collecting data until all variables have a stable level of evidence out of the equivocal zone. For this reason, sequential hypothesis testing plans specify not only the threshold for inferencing, but also the maximum sample size (Nmax) that one has the resources to test (and the minimum sample size, Nmin, at which one will start analysing data).

The stopping rule in our sequential hypothesis testing plan requires all sensorimotor predictors for all DVs to be simultaneously out of the equivocal zone in order to stop testing. This means that if one predictor drops below the threshold of $BF_{10} = 6$ (and remains above the reciprocal $BF_{01} = 1/6$), even if it previously cleared it at a smaller sample, we will continue testing because its accumulated evidence is not yet stable, and we are looking for a stable effect which will provide a robust estimate of the true effect of each sensorimotor component on memory for words. We will begin analysing data at Nmin = 20 participants per list (total Nmin = 2120), which is the sample size in Cortese and colleagues' dataset (for an expected recognition memory task) that allowed us to detect the key Body effect on FA (Dymarska et al., in press). However, in case effect sizes are smaller in our proposed surprise recognition memory task (see also response to point 5 below), we have the resources to triple the sample size up to Nmax = 60 participants per list (total Nmax = 6360). We expect that eventually the effects will stabilise and will provide clear evidence for or against the effect we are interested in, but in the unlikely event that any of the predictors remain in the equivocal range for any of the DVs after testing Nmax participants, we will conclude that those particular effects are too small to be reasonably detected.

5. How were the borders of the sequential sampling plan determined? Was any formal power analysis performed? Why not?

The number of participants specified in the minimum sample size (Nmin) is based on the Cortese et al (2010; 2015) and Khanna and Cortese (2021) studies, from which the stimuli come. We found that this sample size was adequate to detect the effects in an expected memory task, when we reanalysed the Cortese et al. data (see Dymarska et al, in press). Nonetheless, it is possible that the effect sizes in the surprise memory task require a larger sample size to be detected, and so we set the Nmax to be three times the size of Nmin (see point 4 above). As there is already a good chance of detecting all effects at Nmin or close to it, we expect to reach the stopping rule long before reaching Nmax.

Additionally, the definition of power in Bayesian analysis is not the same as in frequentist analysis (Krushke, 2015, doi.org/10.1016/B978-0-12-405888-0.00001-5), and the sequential hypothesis testing plan with Bayes Factors replaces the need for a power analysis in the current study (Schönbrodt et al., 2017). Our plan is to employ a stopping rule when we find evidence for *or against* each hypothesis, which means that if the sample size is too small to detect the effects, Bayes Factors will remain in the

equivocal range and we will continue testing to obtain a larger sample. At the same time, when there is enough evidence for or against the hypothesis and collecting additional data would not change the outcome, we will avoid wasting resources on an inflated sample size. One of the advantages of sequential hypothesis testing with BF is that one does not have to guess a priori what the true effect size might be, unlike in formal power analysis which depends on a correct effect size guess in order to accurately estimate the required sample size. Thanks to the flexibility of sequential hypothesis testing, if detecting the effect in a surprise memory task requires a larger sample than in the expected task, we are able to expand our sample to meet that requirement. However, if the effect is detectable at the same sample size, we are not forced to collect additional data to control for the possibility of a smaller effect size. The approach we take is overall more efficient than power analysis, particularly for small effect sizes (Schönbrodt et al., 2017).

> *6. This study relies on prolific, so some additional data quality checks may be good. For example, they could exclude participants with a d-prime below 0.1. In general, some info on how they will treat outliers could be helpful.*

We agree with this point, and are already planning to include data quality checks. First, as outlined in the Ethics and Consent section and in line with Prolific policy, participants will be informed that they will not be paid, and their data not considered, if they do not provide any meaningful responses or if they fail two attention checks. Additionally, we will only recruit participants who have had at least 95% approval rate on Prolific, and have been successfully completing studies as required. We have now specified this criterion in the Participants section on p. 9.

In order to further ensure quality of the data, we are going to exclude participants who time out on more than 30% of trials in the study or test phases, as outlined in the Data Analysis Plan, in the Data exclusion section (p.16). Additionally, in line with Cortese et al. (2010; 2015), we are excluding participants who do not meet the 60% overall accuracy threshold on the memory task. This overall accuracy calculation takes into account both old words (targets) where participants correctly respond "old" (i.e., hit rates) and new words (foils/lures) where participants correctly respond "new" (i.e., correct rejections, aka 1 – false alarms). Employing an overall accuracy threshold actually subsumes the suggested d' threshold, as it ensures that participants with d' < 0.1 are excluded. We have clarified these points in the manuscript (p. 16).

In terms of outliers, we will exclude trials with RT below 300ms as motor errors, as indicated in the data exclusion section. Since each trial will time out after 3000ms, we do not anticipate the need to further exclude slow responses, which is also motivated by our intent to replicate the methods of Cortese et al. as closely as possible, to allow for comparison of the effects of the components on the expected memory task. Finally, as we are not analysing RT as a dependent variable, further outlier controls are not required.

> *7. The authors should consider introducing an explicit memory condition into the study. Currently, they are relying on comparing the results structure from their study to existing data, but a direct comparison within one study would be preferable. This would strengthen the study a lot.*

On this point, we disagree with the reviewer and have opted not to include an explicit condition (expected memory task) for several reasons. The most important reason is that none of our hypotheses relate to an expected memory task: all are based on the potential outcome of a surprise memory task, so it is unclear what the role of an additional task would be in this RR.

The second reason is that, even if we wanted to add the task for exploratory interest, its benefits do not outweigh its costs. We designed the current RR study to provide a direct comparison with our analysis of the expected memory task (obtained from Cortese et al., 2010; 2015), which we presented in Dymarska et al. (in press). This design includes using the same set of words, setting the minimum number of participants at the same level as Cortese's sample size, and replicating the procedure as closely as possible. Because all analyses are run at the item level, it means the present RR and the existing Cortese dataset are directly comparable as a within-item manipulation of task (i.e., surprise vs. expected, respectively). In order to add an explicit condition to the current RR, we could run it either as a parallel task condition (i.e., between-participants task manipulation) or as a follow-up condition after the surprise task (i.e., within-participants manipulation). The former option would require recruiting a different sample of at least 2120 participants, thus doubling the cost of data collection, but would add nothing new to what has already been achieved by the Cortese et al. dataset, and would not change anything about the comparison that one can make between tasks. The latter option would also double the duration of the experimental portion of the study, which increases risk of participant fatigue and potentially decreases data quality, and – given that participants on Prolific are paid by duration – would almost double the associated costs of data collection, which is difficult to justify for the purposes of an exploratory analysis. In both cases, the additional data collection for an expected task would severely limit the maximum sample size Nmax we have the resources to test, and thus potentially impair our ability to detect the hypothesised effects of the RR surprise memory task. In sum, we consider the existing dataset from Cortese et al. to be a valuable resource, and we think that repeating it is not currently warranted given the resources that would be required.

Reviewer 2

> *1. My overall impression of this submission for a Registered Report is fairly positive. I think the authors are considering a high powered study with some relevant controls for their variables of interest. My recommendations for revision are fairly minor.*

We thank the reviewer for these kind words.

> *2. My largest criticism is whether the hierarchical regression analysis they have chosen is appropriate here. I think the standard for investigating lexical variables is instead to use mixed effects models that not only allow for variability across subjects but can additionally allow for variability across items and even variability in the effect size across items (e.g., varying slopes). It has been the standard to use ever since the influential Clark (1973) article and can be seen in the following investigations of item level effects in recognition memory:*
>
> *Cox, G. E., Hemmer, P., Aue, W. R., & Criss, A. H. (2018). Information and processes underlying semantic and episodic memory across tasks, items, and individuals. Journal of Experimental Psychology: General, 147(4), 545–590.*
>
> *Freeman, E., Heathcote, A., Chalmers, K., & Hockley, W. (2010). Item effects in recognition memory for words. Journal of Memory and Language, 62, 1-18.*

We entirely agree that item-level variability affects word memory, and with the findings of Cox et al. and Freeman et al. showing that word characteristics (lexical and semantic variables) account for most variance in recognition memory performance. Indeed, the importance of such item-level variability is precisely what motivated the item-level analysis of Cortese et al.'s (2010, 2015) memory studies across a much larger sample of words than used in the Freeman and Cox studies (i.e., the "megastudy" approach to recognition memory). By collating a very large sample of words (typically

in the thousands), measuring participant performance for each word in an experimental task, and then calculating mean performance scores for each word, megastudies provide item-level databases that have the capacity to detect small but theoretically-important effects via large-scale regressions (Balota et al., 2012). Indeed, item-level analysis of megastudy datasets has become increasingly popular in studies of word recognition memory (e.g., Dymarska et al., 2023 preprint; Khanna & Cortese, 2021; Johns, 2022; Lau et al., 2018), mirroring similar trends in psycholinguistic studies of word reading.

While it would be possible in the present RR to conduct trial-level analysis in a mixed effects model, we have several reasons for employing item-level analysis instead.

First is that the theoretical accounts we plan to test (i.e., adaptive advantage vs. somatic attention) make predictions about different classes of items, which can be appropriately tested at the item level (Balota et al., 2012).

Second, one aim of the present study is to enable us to compare our results for a surprise memory task with previously-found effects of sensorimotor components in an expected memory task (Dymarska et al., in press) by replicating the same regression model structure. Since the dataset of the expected memory task (developed by Cortese et al., 2010, 2015) does not provide trial-level performance, all analyses in Dymarska et al. are at the item level, and so item-level analysis is also required in the planned surprise memory task to allow direct comparison.

Third is that some of the DVs that form the basis of our RR hypotheses, specifically HR-FA and d', cannot be analysed at the trial level in mixed effects analysis, and *must* be calculated by summarising over multiple responses. To avoid any confusion, we reiterate here that while HR-FA and d' are sometimes calculated at the participant level (i.e., summarising performance over all items seen by each participant), in our analyses HR-FA and d' are calculated at the *item level* (i.e., summarising performance over all participants that saw each word).

Nonetheless, if the Reviewer and the Editor feel strongly about it, we could run mixed effect models on the trial-level data as an exploratory analysis, in addition to the confirmatory item-level analysis. We would only be able to analyse Hits (i.e., logistic mixed effect regression on responses to old/target items) and False Alarms (i.e., logistic mixed effect regression on responses to new/lure items), but we could explore various random effect structures where convergence allows. In order to ensure that any difference in the results between this surprise memory task and the previous expected memory task cannot be attributed to the difference in the analysis, we would still base our inferences on the confirmatory item-level analysis, but the trial-level analysis could be made available in supplemental materials for anyone who wishes to inspect the effects.

> *3. Another important point concerns the sample size. 2,120 participants is really admirable for their standard. However, 20 participants per word means about 10 target and 10 lure trials per word, which is not extensive and can still lead to a lot of variability at the item level. The authors could consider using longer lists of words and/or more participants to up this.*

Our sampling plan sets Nmin at 20 participants per *list*, which means that 20 participants will view each word as a target, and then another set of 20 participants will view the same word as a lure. We realise our previous phrasing was ambiguous and have clarified it in the method section (p. 9).

Additionally, 20 participants per list is only the minimum sample size that we will start with. If it is not sufficient to detect effects at the desired level of evidence, we will continue testing up to a maximum sample size Nmax of 60 participants per list, as per the sequential hypothesis testing plan (see also response to Reviewer 1, points 4-5). As a result, our design will produce between 20-60

target trials *and* 20-60 lure trials per word, which should be more than sufficient to produce consistent measures of performance per item.

> *4. Finally, it's very clear that one of the big problems with this line of work is the correlations between the different lexical variables. While I admire the fact that the authors are including word frequency in their comparison, one of the current gold standards is actually contextual variability. Have the authors considered using this measure? One of the leaders in developing newer measures of context variability is Brendan Johns, and he had a recent paper demonstrating the advantages of these measures:*
>
> *Johns, B. T. (2022). Accounting for item-level variance in recognition memory: Comparing word frequency and contextual diversity. Memory & Cognition, 50, 1013-1032.*

There are two related issues here that we will address separately. First, we agree that intercorrelations between item-level variables is a major concern in investigations of the sort we propose. Hence, our predictors are orthogonal (uncorrelated) components derived from principal components analysis (PCA: see p. 12-13) in Dymarska et al. (2023, JML). That is, the components have been rotated to ensure that each captures unique variance withing the large set of lexical and semantic variables entered into the PCA (see Table 2 on p. 13), and so there are no potential intercorrelations to influence our results.

Second, we also agree that contextual variability is an important variable in word recognition memory, which is why we included it in the PCA that produced the component predictors. That is, we do not include lexical word frequency as a predictor in the planned regression analyses, but rather the predictor variable that we refer to as Frequency is a PCA component. Variables that loaded on this component were: lexical word frequency in US and UK English, contextual diversity, prevalence, subjective familiarity, and (negatively) Age of Acquisition and linguistic distributional distance (see Table 2 on p. 13). The Frequency component thus represents the common variance shared by all these lexical variables, that is not shared with the other components. Two of the loading variables are related to context variability: contextual diversity (i.e., log contextual diversity across documents in the SUBTLEX corpus, as incorporated in the Elexicon Project) and linguistic distributional distance (i.e., mean distance to 20 nearest neighbours in distributional space, representing the diversity of words that tend to occur in a similar context to the target word).

In particular, contextual diversity's loading on the Frequency component was 0.94, the highest of all variables, which suggests that contextual diversity is most closely correlated with the Frequency component. The loading for linguistic distributional distance was -0.89, still strong but less so than that of contextual diversity or the lexical frequency variables. We chose to name the component "Frequency" because it is subjectively the simplest label for the loading pattern, but the component incorporates multiple variables that span contextual variability, lexical frequency, and other subjective judgements. The full PCA loading table and path diagram, originally from Dymarska et al. (2023, JML), is included in supplemental materials for reference.

> *5. One of the points he also emphasizes in this paper is that there is actually a quadratic relationship between frequency and/or context variability and performance, which the authors may want to consider.*

We considered this point and decided to check whether it may also be the case for our orthogonal Frequency component (as opposed to a specific contextual variability or lexical frequency variable). Taking word memory performance from the Cortese et al. dataset we analysed in Dymarska et al. (in

press; expected memory task), we calculated the correlations between the Frequency component and the DVs of hit rate (HR), false alarms (FA), HR-FA, and d'. We also created a quadratic version of the Frequency component as per Johns (2022), and calculated its correlation with the same DVs. We found that the correlation coefficients were roughly the same for the linear and quadratic Frequency functions, with neither offering a systematic advantage across DVs. The results can be found in supplemental materials.

This result suggests that while there may be a quadratic relationship between word frequency / context variability and word memory performance when analysing raw lexical variables, it does not necessarily extend to the PCA-derived Frequency component we plan to use in this RR. Alternatively, since Johns found that the quadratic relationship offered no clear advantage over the linear relationship when analysing disyllabic words, it may be the case that we found a similar pattern because disyllabic words comprise just over half of our item set. In any case, a quadratic relationship is unlikely to occur in our planned surprise memory task on this same item set, and therefore we opt not to apply it to our Frequency component.

> *6. Another point – it wasn't clear whether any of the measures that the authors considered were correlated with word frequency or any other predictor they used. The Introduction should make this clear – it would be very easy to report and discuss a correlation matrix.*

Please see our response to point 4 above. In brief, all 6 predictors (Frequency, Length, Body, Communication Food, Object) are components extracted from a PCA that were rotated to be orthogonal, uncorrelated predictors (see p. 11). Removing potential variable intercorrelations was one of the motivations behind conducting a PCA rather than using raw variables as predictors.

In this RR, we are analysing a slightly smaller set of words than what was used to create the PCA (see Dymarska et al., 2023, JML), which means that the intercorrelations between components are not precisely 0. Nonetheless, all intercorrelations are still extremely small with less than 1% shared variance between components. We have included the correlation matrix in supplemental materials to illustrate this point and reinforced in the introduction that the components are orthogonal (p. 3).

> *7. I found the description of the various theoretical mechanisms somewhat puzzling. They seem to pop up at various points as explanations for relevant phenomena. I think it might make more sense to describe some of the underlying theory and/or competing theories initially and then describe the perplexing and contradictory effects reported in the literature.*

Our initial outline of the paper intro used this suggested theory-first structure (i.e., semantic richness, then competing adaptive vs. somatic accounts, then present the perplexing effects) but found it was less clear than the present structure (i.e., semantic richness, then perplexing effects, then competing adaptive vs. somatic explanations). We feel that the current structure works better because the adaptive advantage and somatic attention accounts are not currently integrated in semantic richness theory, and are only relevant in how they can potentially explain the perplexing Body effect; hence, they make better sense when presented *after* the Body effects are explained.

Nonetheless, to make the theoretical aims and structure clearer, we have added some text on p. 2 and p. 6 to elaborate how semantic richness theory relates to memory, what are its predictions, and how it may need to be adapted given our findings (see also response to point 9 below). Depending on what effects emerge in the present RR, we will finally be in a position to integrate either the adaptive or somatic account with semantic richness theory in order to explain why semantic information relating

to sensorimotor experience of the Body affects word memory differently to other forms of semantic and sensorimotor information.

> *8. "semantically-rich, distinctive words tend to facilitate recognition memory in the classic mirror pattern…" (p2). I'm not sure what the authors are referring to here, whether this is the advantage for low frequency words or for concrete words. Regardless, I don't think it's at all clear that the advantages reported were because the words are "semantically rich." The causes of the word frequency effect are still debated in theoretical models today! For instance, Dennis and Humphreys (2001) argued that word frequency effects are just because of frequency – higher frequency words were experienced in more contexts and thus produce more interference. This says nothing about there being differences in the words' semantic content.*

Semantic richness theory is not concerned with frequency effects; it is restricted instead to how semantics – that is, information relating to word meaning, aka the representations of concepts to which words refer – affects lexical processing and memory. Semantic richness has been shown to influence word memorability independently of word frequency effects, where words with richer semantic representations are remembered better even when the analysis controls for frequency (Cortese et al., 2010; 2015; Hargreaves et al., 2012; Lau et al., 2018; Sidhu & Pexman, 2016). Concreteness is one possible variable that can be used to probe the richness of semantic representations, although it does not elicit very strong effects on memory compared to other semantic variables such as imageability (Khanna & Cortese, 2021). Indeed, a wide range of semantic variables have now been shown to elicit semantic richness effects on memory that are independent of frequency, including body-object interaction (Sidhu & Pexman, 2016), higher animacy and perceived threat (Bonin et al., 2014; Leding, 2020), and sensorimotor experience relating to food and objects (Dymarska et al., in press). Most of these variables facilitate word recognition memory by increasing hit rates while decreasing false alarms (i.e., the mirror pattern of effects we refer to in the quoted text above).

In the current RR, we are focusing on semantic variables that capture different aspects of sensorimotor experience, in order to determine how semantic richness theory can accommodate the unusual effects of Body experience (i.e., via the adaptive or somatic accounts). None of our hypotheses refer to frequency because it is not a semantic variable and semantic richness theory centres around semantic effects. Of course, we agree that word frequency has strong effects on word memory, and have previously found strong effects of the Frequency component on performance in an expected memory task (Dymarska et al., in press), which is why we are including Frequency as a control predictor in the planned regression before we evaluate semantic effects (see point 4 above for detail on how the Frequency component was obtained). However, we are not making any predictions about the effects on Frequency on word memory in the current study, as it is not relevant to the theories we are testing.

> *9. "higher scores in this component made no difference to either hits or false alarms, which Dymarska et al suggest may be due to lack of distinctiveness in communication-related words."(p3) How do we know there was a lack of distinctiveness? Even if the effect was found, how would we know that this was specifically due to distinctiveness? I don't think that's necessarily clear without an independent definition or theoretical conception of distinctiveness. I'm not saying there isn't one – it's possible to define distinctiveness as isolation in some type of representational space – but you cannot conclude that performance advantages are necessarily due to distinctiveness. It's possible that other factors – such as just having more features or stronger encoding of said features – could be responsible.*

The original formulation of semantic richness theory for memory assumed that richer semantic representations = stronger semantic activation = stronger encoding of the memory trace = better memorability (Hargreaves et al., 2012; Pexman et al., 2003; Sidhu & Pexman, 2016). However, Lau et al. (2018) found that it was not quite so simple, and that sometimes richer semantic representations (e.g., having a higher number of senses) led to worse memorability, specifically by inflating false alarms, which they concluded was due to a lack of distinctiveness at the semantic level. Usually, richer semantic representations are also more semantically distinctive, which leads to better memorability because of clear and specific overlap between the retrieval cue and memory trace. However, words with more senses are semantically richer but also more ambiguous (i.e., less distinctive), which does not translate to better memorability because of the decreased overlap between the retrieval cue and memory trace. The semantic richness theory for memory was thus updated to allow distinctiveness to act as a constraint:

> (a) richer *AND more distinctive* semantic representations = stronger semantic activation = stronger *AND more distinctive* memory trace = better memorability due to increased hit rates and decreased false alarms.
>
> (b) richer *BUT less distinctive* semantic representations = stronger semantic activation = stronger *BUT less distinctive* memory trace = worse memorability due to inflated false alarms (null effects also possible where distinctiveness is not low enough to cause confusion).

In Dymarska et al., (in press), we found that this distinctiveness-constrained variant of semantic richness theory can adequately explain the effects of sensorimotor experience relating to Food, Object, and Communication. That is, the memory facilitation effects of Food and Object are consistent with (a), and the null effects of Communication are overall consistent with (b) where the representations of high-scoring words are not distinctive enough to facilitate memory but not so indistinct as to be regularly confused with other items. When we inspected words that were high in Communication scores, they appeared tentatively consistent with this idea. Many words that relate strongly to Communication experience appeared to cluster with other words of rather similar meanings that could easily be confused with one other (e.g., *scream, yell, shout*; *chat, talk, speak*). However, other strongly-Communication words seem relatively distinct in meaning (e.g., *song, pun, lecture, sneeze*), meaning that low distinctiveness is not endemic amongst high Communication scores. That is, if stronger Communication experience does not systematically increase distinctiveness of a word's representation, it could explain why semantically richer representations (in terms of Communication experience) do not necessarily facilitate word memory.

We agree that having more features indeed predicts a performance advantage, but it is not due to another, separate factor to those outlined above. It is entirely consistent with (a) above, because more features increase the richness of a semantic representation *and* also make the representation more distinctive. As such, more semantic features lead to higher hit rates and lower false alarms. If it did not make the representation more distinctive at a semantic level, then it would not facilitate word memory because a richer representation alone – and the corresponding stronger encoding it produces in the memory trace – is insufficient for a performance advantage if the representation itself is easily confusable with other items.

We clarified this issue in the manuscript by adding a definition of distinctiveness on p. 2.

> *10. "Instead of producing the semantic richness pattern of increased hits and fewer false alarms…" (p3) This is just the mirror effect, not a "semantic richness" effect.*

Yes, it is indeed the mirror effect, which is how we first describe it on p. 2 of the intro. However, as we explain at that point, this mirror pattern is precisely what is predicted by semantic richness theory:

"semantically-rich, distinctive words tend to facilitate recognition memory in the classic mirror pattern (Glanzer & Adams, 1985) of increasing hit rates and reducing false alarms."

Hence, our intention on p. 3 was to highlight that instead of producing a pattern of effects that would be consistent with semantic richness theory, the Body effects are entirely different. We have amended the text on p. 3 to make this intention clear:

"Instead of producing the mirror pattern predicted by semantic richness theory of increased hits and fewer false alarms, higher Body scores unexpectedly had little effect on hits but led to more false alarms…."

> *11. …when participants are not aware they will be later tested on their memory for presented words… such elaboration is far less likely, and offers us an opportunity to adjudicate between theoretical accounts." (p5-6) I am a fan of the approach that the authors are taking and I like surprise memory tests. However, I didn't think this statement made a lot of sense. How do the different theoretical accounts require this manipulation? How does elaboration change their predictions? I think this statement should be made more clear.*

As outlined on p. 8, the two competing explanations for the Body effects on word memory make different predictions about what will happen in a surprise memory task because only one of them (the adaptive advantage account) relies on elaboration at encoding.

In brief, participants in an expected memory task are likely to use elaboration as a strategy to make the word memorable, such as by placing a concept in a particular scenario or context. According to the adaptive advantage explanation, such elaboration will trigger survival-relevant words to spread activation to a network of other, related concepts, which will make the memory trace for those words less distinctive and prone to false alarms. However, participants are unlikely to use elaboration as a strategy in a surprise memory task because they do not know that they will be tested on memory for the words. Therefore, without such elaboration, spreading activation to related concepts will not take place for survival-relevant words, and they will no longer be prone to false alarms. It is possible that the pattern of the Body component effects in Dymarska et al.'s (in press) expected task was due to such an elaboration strategy, but if so, the same pattern will not emerge in the surprise memory task.

By contrast, the somatic attention account does not rely on elaboration at encoding and so is unaffected by the task manipulation. In both expected and surprise memory tasks, when word meaning is automatically accessed on reading during the study phase, any representations relating to bodily states will extend attention to touch and other irrelevant modalities, which will make the memory trace for those words less distinctive and prone to false alarms. It is possible that the pattern of the Body component effects in Dymarska et al.'s (in press) expected task was due to such automatic processes, and if so, the same pattern of effects will emerge in the surprise memory task.

We have added a clarification to p. 6, where we signposted our detailed predictions for clarity and included a table which outlines these two predictions (see Table 1). We also included the text above with the detailed predictions on p. 8.

> *12. It's also important to note that the usage of a lexical decision task can still produce strategies. In fact, the nature of the encoding task has a huge effect on performance and can even change the nature of the word frequency effect – see the following paper: Criss, A. H., & Shiffrin, R. M. (2004). Interactions Between Study Task, Study Time, and the Low-Frequency*

Hit Rate Advantage in Recognition Memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 30(4), 778–786.

We certainly agree that different encoding tasks may produce different strategies, which needs to be taken into consideration when designing a surprise memory task. In the current study we aimed to select a task which would eliminate the kind of encoding strategies that participants were likely to have used in the expected memory task analysed in Dymarska et al. (in press). Specifically, we wanted to eliminate the motivation to elaborate on the meaning of the words during encoding, in order to disentangle two possible explanations of the pattern of effects of the Body component (see also response to point 11 above). We therefore chose lexical decision as the encoding task because the objective does not require such elaboration on word meaning: that is, quickly judging whether a string of letters is a word or nonword is unlikely to lead to participants to make a conscious effort to represent the kind of elaborate contexts/scenarios for the word that are important to the adaptive advantage account. Indeed, psycholinguistic research shows that lexical decision generally involves only automatic activation of word meaning (e.g., Neely, 1977, JEPG) that lacks semantic detail compared to more deliberate conceptual processing (Pexman et al., 2008, Cognition), but – critically – *does* include the kind of sensorimotor information that is important to the somatic encoding account (Banks et al., 2018, Royal Soc PTB; Dymarska et al., 2023, JML). In this way, we expect that using lexical decision as the encoding task will allow us to adjudicate between the accounts.

While it is possible that other encoding tasks may also produce the effect we are looking for, semantic activation in lexical decision is well understood and thus represents a relatively safe choice in our study. In particular, all four sensorimotor components that we plan to examine in this RR elicit semantic activation during lexical decision (as shown by Dymarska et al., 2023, JML), which allows us to turn our attention to the memory task and the effects that occur there, and to outline our predictions regarding the possible underlying mechanisms with confidence.