1. Reviewer 1

> **There is one major limitation for the broader application of this approach and that's demonstrating its effectiveness in fMRI data. The majority of neuroimaging studies are done using fMRI (and I do not say this to discount the other valuable tools of EEG, MEG, NIRS, etc.). To get other neuroscientists to grab onto this approach, I recommend the authors perform another analysis using fMRI data. My suggestion for the authors would be to use data from the Human Connectome Project (http://www.humanconnectomeproject.org/). The outcome variable used with this approach is not important as so much the demonstration that their technique is effective for fMRI data. Without including fMRI data, the approach comes off as useful but speculative for fMRI data, which drastically limits its impact potential.**

We thank Reviewer 1 for this remark. We now apply the same analyses to fMRI recordings of 100 subjects performing a reading experiment similar to the MEG study described in original manuscript. We kept the exact same methods and parameters, and used the fmriprep package to perform the preprocessing.

Overall, the results consistently show that B2B compares favorably to other methods. Furthermore, these novel analyses complements the temporal results obtained with MEG. Specifically, the effects of Word Length are predominantly observed over the early visual cortices, and the Word Frequency are predominantly observed over the fronto-temporal areas.

2. Reviewer 2

> **First, the authors justify their approach of using backward modeling by the ability to estimate common noise sources. However, this would also be possible using multivariate forward modeling. For that reason, I'm wondering how multivariate forward modeling would fare under these circumstances. I would find it highly informative to include a (similarly cross-validated) multivariate GLM to the analyses and results.**

We thank Reviewer 2 for this remark. The multivariate GLM / MANOVA is indeed a relevant analysis. We now indicate in supplementary materials the following analyses:

> "Following the recommendations of one of our reviewers, we implemented a multivariate variant of the forward model, i.e. a MANOVA, using the statsmodels implementation. MANOVA is primarily used as a inferential statistics, and does not trivially convert to a predicting method. Consequently, we did not find a way to compare MANOVA against B2B with the ΔR evaluation. However, the effects of MANOVA are generally summarized with the Wilk's Lambda stastistics or its transformation into an F-value. For each searchlight, we thus use the F-values of the Wilk's Lambda statistics as proxy for $\hat{S}$ and feeds it to a second-level Wilcoxon signed-rank tests across subjects, like we did for the other models.

The results show that all factors, including the Dummy variable, are systematically above chance level in all recorded brain regions (Supp Fig.). This result can be explained by the large dimensionality of Y. Indeed, limiting the searchlight to a 1mm radius did not lead to these spurious effects but provided results similar to the Forward model.

Nonetheless, MANOVA does appear to capture some plausible effects. Indeed, the F-values obtained for both Word Length and Word Frequency were weakly but significantly higher than those obtained with the Dummy variable in the occipital and temporal brain areas (Supp Fig.). This results suggests that the effect size of the MANOVA can be biased and, thus, is not valid for second-level statistics.

Overall, this suggests that MANOVA (1) can lead to positively biased estimates of $\hat{S}$ (2) appears weaker than B2B in terms of second-level analysis across subjects (3) misses the effect of Word Function detected with the Forward model, and (4) does not trivially translate into a prediction tool. Together these elements thus suggest that MANOVA is less suitable to the present objective than B2B."

Likewise, while I applaud the authors efforts in conducting these analyses, I am wondering whether the impact of this method could not be made even stronger by adding more experimental findings. Are there any other openly available datasets that could be used to demonstrate the ability to disentangle known factors?

We agree with Reviewer 2. As indicated to Reviewer 1, we now successfully apply the same analyses to the fMRI recordings of 100 subjects reading word sequences.

Rather than using the feature importance delta R, is it possible to just report the parameter estimate E hat at each time point? What would be the downside of this?

It is possible to report $\hat{E}$ (now relabelled $\hat{S}$), however they are interpretable only if the second regression H is not regularized. We now added a Supplementary Figure to that effect.

Note, however, that $\hat{E}$ statistics are not in the same unit as the forward $\hat{H}$ coefficients (a.k.a betas values in GLM). Consequently, it is not straightforward to compare these two metrics directly. As a proxy, we thus report their respective second-level p-values across subjects. The associated figure shows that B2B's $\hat{S}$ coefficients compare favorably to the Forward $\hat{H}$ coefficient and the MANOVA coefficients.

The authors describe their method as unbiased but then introduce two regularization parameters, which will bias the estimated coefficients towards zero. While the statement "B2B leads to unbiased (i.e. zeros-centered) scalar coefficients for non-causal features" is correct for the mean of those coefficients, the variance will also be biased. I think these are important criteria to consider.

We agree with this remark. We now clarify the meaning and implication of the bias in the problem statement:

"Note that $\hat{S}$ is unbiased, in the sense that it is centered around zero when there is no effect, only if the second regression H is not regularized. Second-level statistics testing whether $\hat{S}$ is superior to 0 are thus only valid if H is not regularized."

> **The results of backward regression are not shown in Figure 4, but it would be nice to see those results. Is there a possibility to add them to Figure 4 in a separate panel?**

We now added the decoding performance to a supplementary figure. The latter show that approximately the same cortical network can be used to decode Word Length and Word Frequency. It also shows that the Dummy Variable can be decoded from various brain areas, and from the visual cortex in particular.

> **I think the authors could better explain the dimensions of the data (dy referring to the number of measurement channels, and dx to the number of independent variables). Perhaps the authors could add this nomenclature to their graphical illustration? Likewise, for the regression of Y onto X, it is unclear if this is a mass-univariate regression or a multivariate regression (i.e. taking the error covariance between Ys into account or not). This might be obvious to them, but is not necessarily obvious to the reader.**

We amended the text in multiple places to clarify this issue.

> **I find the choice of the letter E for one of the causal factors unfortunate since it is typically reserved for the true error component of a multivariate model. Likewise, the choice of N is unfortunate since the letter n has multiple meanings in the text: one particular instance of n, or the number of samples. This is a very minor comment though.**

We now changed E into S for Selection matrix. We also changed the number of samples into m, and keep N as the Noise matrix. Thank you for these suggestions

> **Typo line 28: "were" should read "where"?**

Yes this is now corrected.

> **Line 181: do the authors refer to the Wilcoxon signed-rank test?**

Yes, this is now corrected.