Jeremy R. Manning
Dartmouth College
Department of Psychological & Brain Sciences
HB 6207 Moore Hall
Hanover, NH 03755

October 16, 2020

Dear Dr. Schiffer:

We have enclosed our revised manuscript entitled *Geometric models reveal behavioral and neural signatures of transforming naturalistic experiences into episodic memories* (submission NATHUMBEHAV-18094982A).

We appreciate the reviewers' additional comments, and have included point-by-point responses to each of the three reviewers below. The reviewers' comments are shown in *italics* and our responses are shown in **bold**.

In addition to addressing the reviewers' concerns, to the best of our ability we have also edited our manuscript to comply with the Nature Human Behavior formatting guidelines as requested. However, we note that, with the additions the reviewers requested, our manuscript length exceeds the length limit specified in those guidelines. We believe the revised manuscript we have submitted represents the clearest framing of the paper. If it is necessary to substantially cut back on space, we see at least two possible options:

- We could de-emphasize some of the complexities of our methods in the results section. However, we worry this may substantially impact readability.
- Another option would be to move our "classic analyses" (Fig. 3) back to the supplement. In the previous round of revisions, Reviewer #2 had suggested promoting that figure to the main text and placing additional emphasis on the corresponding findings. However, we could revert to something more similar to our initial submission's framing in order to save space.

We would also welcome your suggestions with respect to what to cut, move (e.g., to the methods section or supplement), or de-emphasize as needed.

Thank you for considering our revised manuscript.

Sincerely,

Jeremy R. Manning
Jeremy.R.Manning@Dartmouth.edu

Reviewer #2:

Remarks to the Author:

Thank you to the authors for their substantial revision - I was happy to see that this paper is still under consideration! My enthusiasm about this work has only increased after seeing these changes.

Thank you for the positive feedback!

The revisions have largely addressed my concerns, but I do have some remaining (minor) questions:

1) The authors' reasoning about why they defined the topic space solely on the episode annotations makes sense. However I do think that the results need to be interpreted in light of this decision. For example, the revised manuscript still states that topic proportions for the recalls "are not as sparse as those for the episode" which seems to be a direct consequence of the training corpus choice and not a fact about the recall data itself. Since the topics were fit to the episode annotations, making comparisons about their test performance on the episode vs. recalls seems misleading (since in one case this is testing on the training data, and in the other it is generalizing to a new dataset).

We have added a note to this effect (p. 7, emphasis added, references expanded out for convenience):

> **"The second clear pattern present in every individual participant's recall correlation matrix was that, unlike in the episode correlation matrix, there were substantial off-diagonal correlations. *One potential explanation for this finding is that the topic models, trained only on episode annotations, do not capture topic proportions in participants' "held-out" recalls as accurately.* A second possibility is that, whereas each event in the original episode was (largely) separable from the others (Fig. 2B), in transforming those separable events into memory, participants appeared to be integrating across multiple events, blending elements of previously recalled and not-yet-recalled content into each newly recalled event (Figs. 2E, Supp. Fig. 5; also see Howard et al., 2012; Manning, 2019; Manning et al., 2011)."**

We explore the off-diagonal entries of these recall correlation matrices in more depth in another manuscript (Manning, 2019; https://psyarxiv.com/6zjwb). One relevant finding in that paper is that the off-diagonal entries appear to exhibit meaningful structure (e.g. see Fig. 3 in that paper, which highlights additional meaningful structure that participants impose on an event when they recount it).

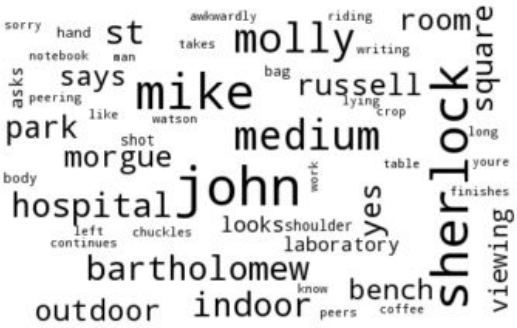
2) The changes to the text preprocessing pipeline all make sense, but I didn't see my concern about the "Music Presence" addressed - if this is being encoded as "Yes"/"No" then it is hard to

see how this will generalize to recall descriptions (since this just means that e.g. the word "Yes" would appear in topics related to the presence of music).

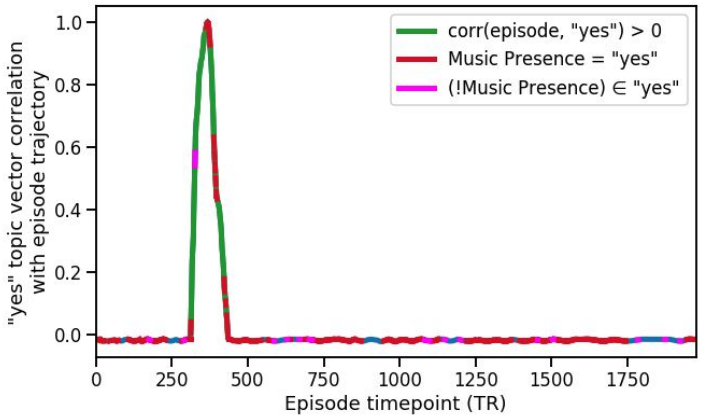
We agree that the words “yes” and “no” are not specifically related to music in their colloquial uses, nor are participants likely to use those words to refer to the presence or absence of music in their recalls. In fact, the word “no” is filtered out at the start of our analysis pipeline, since it appears so frequently throughout the annotations and recalls that it cannot provide reliable event-specific information.

In contrast, the word “yes” is a comparatively rare word (both in the episode annotations and participants’ recall transcripts). The rarity of the word “yes” in the annotations makes it a potentially informative feature for the topic models. For example, the presence of music often co-occurs with the presence of specific characters, actions, locations, and plot elements. There are several ways of exploring these associations; we view them as beyond the scope of our current manuscript, but we have included a few example analyses here to give a sense of how we approach this question.

One way of gaining an intuition about which themes the word “yes” weights on is to use wordles analogous to those in Fig. 7. To create the word cloud for a given word (in this case, “yes”), we can compute the average topic weights assigned to different words, weighted by how much each topic weights the target word (“yes”). Words weighted heavily by topics that also weight the word “yes” heavily appear in a larger font in the wordle:

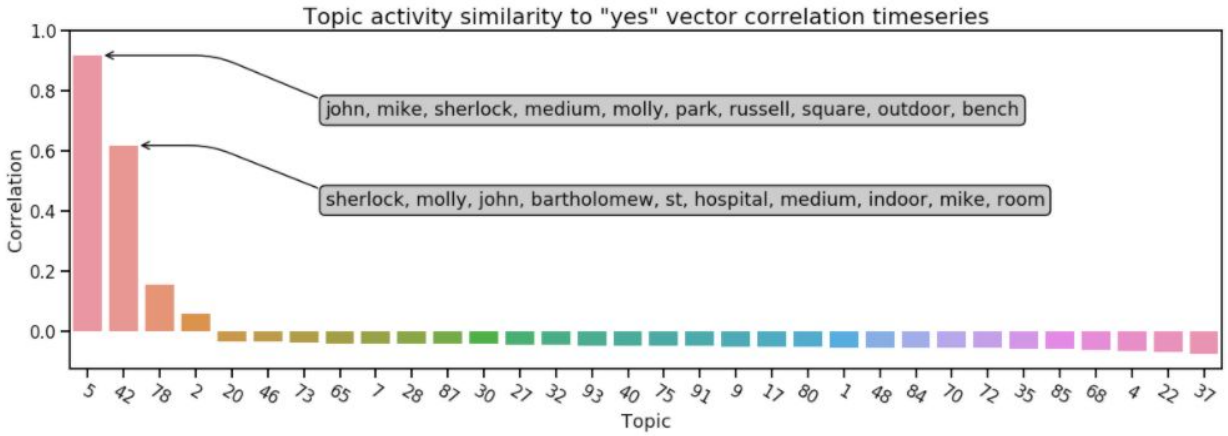


We can also examine how the topic vector (created by normalizing each topic’s weight on the word “yes”) correlates with the topic vectors of each timepoint in the episode:



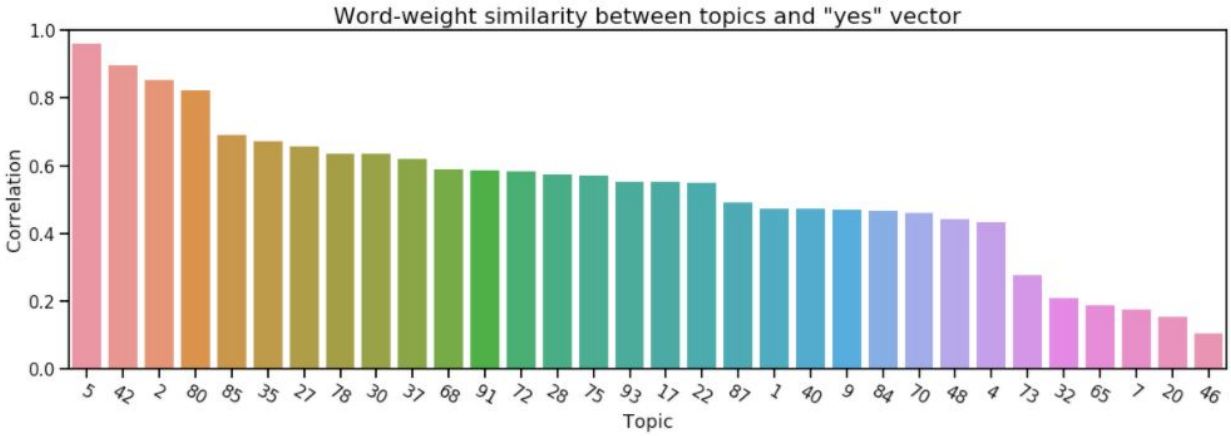
Here green reflects moments when the correlation between the topic vector for "yes" and the episode timepoint's topic vector is positive. Red indicates moments when background music was present. Magenta indicates moments when the episode annotations contained the word "yes" but when no music was present. This reveals that the music-related themes our model picked up on seem to be most strongly associated with one scene in the episode.

We can also ask which *other* topics' timeseries of correlations with each moment of the episode were similar to the timeseries of correlations for yes's topic vector:



Here, the top-weighted words of the two topics with the most similar timecourses are displayed in the gray boxes.

Finally, we can ask which topic vectors' weightings on different words were most similar to the topic vector for the word "yes" (the two most similar topics are identical to those in the previous figure):



Taken together, our interpretation of these analyses is that the "Music Presence" annotations provide a (likely relatively weak) signal for when these associated themes appear. Further, when participants reference those themes in their recalls through their uses of *other* words associated with those themes (e.g., even if they don't specifically use the word "yes"), our modeling framework will still "match" references to music-related themes (i.e., semantic features or topics in the episode that co-occur with the presence or absence of music) with other words associated with those semantic features or topics.

We also note that the above point—that annotations and recalls don't need to use the same words (or in the same ways)—is a central feature of our modeling framework. We think solving the "matching problem" (i.e., labeling specific things participants say with specific events they experienced) in a way that is robust to wording differences is an important advance in studying naturalistic memory behaviors.

3) I better understand the "recall template" fMRI now, but am still quite confused by Fig 8B. The caption says that they "align each participant's recall timeseries to the TR timeseries of the episode" but the x/y axes read recall/viewing and then recall/recall in the 2nd and 3rd stages - should these both be viewing/viewing if the data is warped to viewing TRs? What is going on between the 2nd and 3rd stages - is this just showing the proximal masking, and if so why are some of the near-diagonal entries changing (maybe a colorscale change)? I also am still not sure I correctly understand the logic of this analysis and what the warped matrix means. What I think the final result is showing is this: Topics associated with TR X during viewing are re-appearing when subjects are recalling information associated with TR Y from viewing. Is that the right interpretation of the warped matrix? Also the authors should note that this relationship is not symmetrical (swapping X and Y in that statement can give different values, as shown in Fig 8B) and it is unclear to me whether the upper or lower part of this matrix is what we want.

The final "stage" shown for the episode template pipeline (Fig. 8A) displays the correlations between pairs of episode topic vectors at nearby timepoints, while the final stage shown for the recall template pipeline (Fig. 8B) displays the correlations between pairs of recall topic vectors at nearby timepoints. The "proximal correlation matrices" we

mention in the figure and text denote that we have taken temporal autocorrelation matrices and masked out everything above the *n*th diagonal (where *n* is the duration, in TRs, of the longest episode event). When we say "correlations at nearby timepoints," we are referring to the unmasked parts of these proximal correlation matrices.

To compute proximal correlation matrices for the episode, or for a given searchlight's activity patterns, we simply correlate the topic vectors (or voxel responses) from every pair of timepoints, and then mask out anything beyond the *n*th diagonal. In this way, the RSA analysis that we designed to identify searchlights whose responses show a similar (proximal) correlation structure to the episode's topics (Fig. 8C) does not require any further temporal alignment, since both the topic timeseries and voxel responses are computed at the same timepoints.

However, computing the proximal correlation matrix for participants' recalls requires an additional "warping" step to bring the behavioral data (during recall) and neural data (while *viewing* the episode) into temporal alignment. For example, different participants take different amounts of time to recount the episode, and their transcripts have different lengths. These differences occur at the "episode" level (i.e., with respect to total recall duration and/or transcript length) and also at the "event" level (i.e., a given participant's recounting of a particular event may be more or less detailed, and take more or less time, than another participant's recounting of the same event). The purpose of the warping step is to temporally stretch or compress the topic timecourse of each participants' recounting so that it is temporally aligned with the voxel responses recorded as the participants were watching the episode. That warping step is what we are highlighting in the two rightmost matrices in Figure 8B (i.e., to use the reviewer's terminology, the 1st and 2nd stages of that analysis pipeline).

The transition between the 1st and 2nd stage of the pipeline shown in Fig. 8B shows the effect of using dynamic time warping to temporally align an example participant's recall topic proportions matrix with the episode's topic proportions matrix. The rightmost matrix (stage 1) shows the correlation between the topic vectors for each *un*warped recall timepoint (i.e., sliding text window) and each episode timepoint (TR). The middle matrix (stage 2) shows the correlation between the topic vectors for each *warped* recall timepoint (i.e., TR) and each episode timepoint (TR). In both matrices, the row (episode timepoint) matched to each column (recall timepoint) by the dynamic time warping algorithm is indicated in yellow. We chose to visualize this step by correlating the example recall trajectory with the episode (rather than with itself, as in the other matrices) before versus after warping. We felt this would help to illustrate how the diagonal "straightens" as a result of the non-linear "stretching" the algorithm performs to align the two timeseries. An important feature of the warping algorithm is that it is a monotonic transformation-- i.e., it does not change the relative orders of the recalled timepoints; it only stretches or compresses different parts of the recall topic proportions matrix while preserving its temporal order.

The 3rd (leftmost) stage shown in Fig. 8B then displays the proximal correlation matrix for the example participant's recalls. In other words, we computed the correlation matrix for that participant's warped recall topic proportions (i.e., after they have been temporally aligned to the episode), and then masked out everything beyond the *n*th diagonal. As the reviewer correctly points out, the near-diagonal entries shown in the 3rd stage differ from those in the 2nd stage. Whereas the entries in the stage 2 matrix show correlations between (warped) recall topics and episode annotation topics, the entries in the stage 3 matrix show correlations between (warped) recall topics at different timepoints. We chose to label axes corresponding to the warped recall trajectory as "Recall time (TR)" rather than "Viewing time (TR)" to help differentiate them from axes corresponding to the episode trajectory, as well as maintain consistency with how we labeled video-recall and recall-recall correlation matrices in other figures.

With respect to the reviewer's question about what the warped matrix "means," there are a few aspects of the analysis to unpack. The neuroscientific question we are exploring in that analysis is: "when participants are watching the episode, which brain regions respond in a way that predicts the distorted way they will recount the episode later?" One way of conceptualizing this analysis is as a "naturalistic extension" of Paller and Wagner (2002)'s classic "subsequent memory effect" analysis.

From a methodological standpoint, as described above, the dynamic time warping procedure enables us to align "recall time" (which varies by participant) with "episode time" (which is the same for every participant and also matches the fMRI timing). We can then ask: while *viewing* the episode, which searchlights' (neural) temporal correlation matrices are similar to each participant's warped (behavioral) recall temporal correlation matrices? Because every participant has a different behavioral recall temporal correlation matrix, every participant's searchlight "template" (i.e., their recall temporal correlation matrix) is unique. For this reason, our analysis picks out searchlights whose responses while viewing the episode are specifically related to the idiosyncratic way each participant will later recount the episode.

Reviewer #3:

Remarks to the Author:

While I have remaining concern about whether the paper's contribution really extends beyond the methods (this was the main theme in my initial review), I believe the revised manuscript does a better job making the case for why the methods advance is important, in and of itself. I do not have any new concerns.

We appreciate the reviewer's feedback.

Reviewer #4:

Remarks to the Author:

#General comments:

The authors should be commended for their impressive revision. I found the revised version much improved. Especially, I really like how the authors now emphasize how their approach can help to understand how the brain preserve and distort our ongoing experiences to encode them into episodic memories. I also really appreciate the authors' efforts to link some of their findings to classic effects in the recall literature.

Thank you for the positive feedback! We are excited about these findings as well!

Having said that I am still slightly disappointed by the lack of discussed relationships between this new method and current theoretical and neurobiological frameworks of human memories. I might misinterpret the authors but they seem to assume in their responses to the referees that their method constitutes a theoretical advance that suffices in itself. I agree that the current framework is remarkable and will help to study more naturalistic memories, something that is currently greatly lacking. However, if we cannot link their approach to existing theories or concepts, then the current methodological breakthrough loses some of its meaning. Helping researchers to understand the scope of the work and how they can use it, seems really important in this context. My point is that the authors should discuss in general more what their method can bring to understand the brain and how memories are implemented in the brain. This will mostly impact the discussion. I list some relevant frameworks below. I'm not expecting that the authors discuss extensively their results in light of these frameworks (although I raise some questions) but that they briefly relate their method to these frameworks and how it may help to test new questions.

We take the reviewer's point and have added references to the studies below to our revised discussion section.

1) Schema (Giboa and Marlatte, TICS, 2017): how can their method help to study memory schema? How does the temporal unfolding of the narrative structure relates to the notion of schema (e.g. the notion of gist) ? Does the involvement of the vmPFC during both encoding and recall speak to this issue?

We have added a note to our discussion section to describe how our work might related to schema learning and/or knowledge, as well as a potential speculative role of the vmPFC (p. 17; emphasis added, and references are expanded below for convenience):

> **"Our work also provides a potential framework for modeling and elucidating *memory schemas*– i.e., cognitive abstractions that may be applied to multiple related experiences (e.g., Baldassano et al., 2018; Gilboa and Marlatte, 2017). For example, the event-level geometric scaffolding of an experience (e.g., Fig. 6A) might reflect its underlying schema,**

and experiences that share similar schemas might have similar shapes. This could also help explain how brain structures including the ventromedial prefrontal cortex (Fig. 8; also see Gilboa and Marlatte, 2017) might acquire or apply schema knowledge across different experiences (i.e., by learning patterns in the schema's shape)."

2) Collective memory (Gagnepain et al., NHB, 2020): Thanks to their geometric space embedding conceptual contents, the authors address the non-trivial problem of matching up descriptions of a shared experience despite differences in verbal recall. How can this framework help to study collective memory and the recall of shared experiences and memories ?

This is a very interesting question space, and one we are hoping to elaborate on in follow-up work. For example, one could use properties of the average recall trajectory (e.g. Fig. 6B, black lines) and/or variability in individual participants' recalls (e.g., Fig. 6B, arrows; also Fig. 6C) to determine when a "sufficiently complete" recollection has been generated by a group of participants or witnesses. Unfortunately we lack the space to delve into these ideas in depth in the current manuscript, but we have added a note emphasizing the utility of our framework at matching up different people's recountings of a shared experience, along with a reference to the Gagnepain et al. paper (p. 17; reference is expanded for convenience):

"Building an explicit model of these dynamics also enables us to match up different people's recountings of a common shared experience, despite individual differences (also see Gagnepain et al., 2020)."

We also elaborate on challenges related to how word embedding frameworks like the one we develop here can be used to solve the "matching problem" in one of our preprints (Manning, 2019; section entitled "The matching problem"; https://psyarxiv.com/6zjwb), which addresses and discusses some of these ideas as well.

3) Memory in space (Bellmund et al., Science, 2018 & NHB, 2020). Several works suggest that neural codes guiding navigation in space also shape cognition and memories. How can this new method (and perhaps the current findings, especially in the hippocampus and vmPFC, see Constantinescu et al., Science, 2016) featuring the embedding of memories into a geometrical conceptual space can be related to these existing notions ?

The connections between semantic versus spatial navigation is a very deep question, and one that we feel is somewhat beyond the scope of this paper. Therefore we have added only a brief note with references to this work (p. 26; references expanded for convenience):

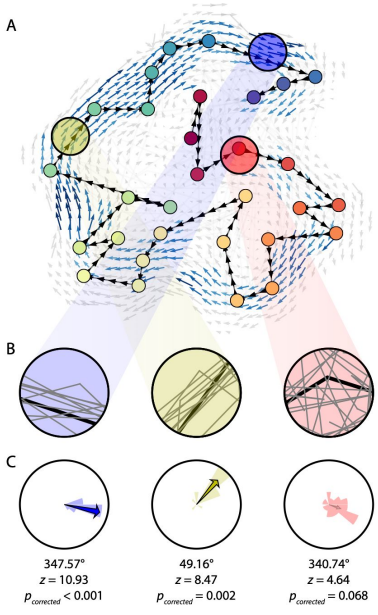
"Our work casts remembering as reproducing (behaviorally and neurally) the topic trajectory, or "shape," of an experience, thereby drawing implicit analogies between mentally navigating through word embedding spaces and physically navigating through spatial environments (e.g., Bellmund et al., 2020, 2018; Constantinescu et al., 2016)."

More generally, although we do see similarities between semantic and spatial navigation, our more nuanced view is that we also see important differences. First, we cannot teleport in real space, whereas our thoughts can exhibit rapid jumps between different non-contiguous regions of word embedding space (e.g., at event boundaries). Second, when two people visit the same fixed physical landmark, they necessarily visit the same physical location. When two people describe a shared experience, their idiosyncratic thoughts, goals, prior experiences, etc. can lead them to process and remember the experience differently. Third, when we revisit a fixed physical landmark, we (by definition) revisit the same spatial location. In other words, it is possible to visit the same spatial location several times. However, we cannot revisit our prior experiences in this way. We only get to experience each moment once; subsequent "visits" to a prior experience must occur through a different medium than the original experience (e.g., our memory, another person's memory, a physical recording, a written account, etc.). Fourth, we can revisit different aspects of the same prior experience on different "visits"-- e.g., when we remember a specific event we can focus on the physical occurrence in one remembering, the emotional occurence in another remembering, the social implications in another remembering, etc. We can also revisit the same experience in different levels of detail or depth. There are no deep analogs of these phenomena in spatial navigation. Finally, *physically*, we are always located at a single moment and in a single location, whereas our mental state can be "spread" over many concepts, times, and/or locations.

#Minor comments:

- I appreciate the authors' effort to describe their method to produce Figure 6. I think I understand it better. If possible, a supplementary figure that describes the main steps of this method would really help.

We have added Supplementary Figure 4 to help illustrate this analysis and unpack it a bit more:



The figure caption (Supp. Fig. 4) explains what is shown in the figure, and we have added a reference to this new figure in our main text.

- I don't think the authors have addressed my comment about the lack of baseline in their searchlight analyses. My point was not about their segmentation procedure. My point was to use an episode model that is not embedded in the topic space (for instance a categorical model based on annotation). The idea was then to compare the similarity of the topic model (with neural temporal matrices) to the similarity of the categorical model (to see if the topic model explains more of neural activity than a basic categorical model)

The reviewer's suggestion to develop categorical models is an interesting one, but implementing such an approach would be non-trivial for several reasons. The simplest categorical model we can think of would be to assign each scene a unique category. However, depending on the implementation details, this would result in either a block diagonal matrix (nearly identical to the "episode correlation matrix" in Fig. 8A) or the identity matrix (which would not serve as an effective RSA searchlight matrix). More sophisticated semantic models, whereby (for example) overlapping *mixtures* of categories could be active in each scene (or timepoint) start to approach the topic modeling framework we implemented. In our approach, for example, each moment's topic proportions reflect a weighted blend of automatically derived themes. Generating unbiased manual category annotations would require running additional experiments (e.g., one to define the specific categories and a second to assign category labels to each timepoint or scene); we view this approach as beyond the scope of our current manuscript. Further, prior work (e.g. Griffiths et al., 2007, Psychological Review) suggests that topic models like those we leverage in our manuscript perform well at predicting people's semantic judgements and behaviors.

The analyses in Figure 8 use a permutation-based procedure to identify temporal correlations in the neural data that are *specifically* matched to the temporal correlations in the topic proportions matrices (p. 37). For each searchlight, we correlate the temporal correlations in topic proportions with temporal correlations in neural patterns to obtain an "observed correlation." To estimate a "baseline" for these analyses, we then repeat this procedure for 100 phase-shifted topic proportions matrices (i.e., circularly shifting the rows of the topic proportions matrices) while holding the neural data fixed. This results in a "null" distribution of 100 correlation values. We report the percentile rank of the observed correlation relative to the null distribution.

In other words, the circularly shifted topic proportions matrices provide alternative timecourses that share the same autocorrelation structure as the unshifted matrices, but break the temporal alignment between the stimulus (or behavioral data) and the neural data. In our view, this provides the most conservative test of whether the neural responses are specifically best explained by the topic timecourses, or whether they also match alternative models with similar temporal structure.

We do agree with the reviewer's point that we cannot specifically rule out potential alternative semantic models. We discuss several such models in our discussion section (p. 18–19). Our claim is not that topic models provide the "one true description" of participants' thoughts or behaviors, but rather that our framework (which leverages topic models as one of many possible word embedding models) provides a way of characterizing thoughts and behaviors sufficiently well that we are able to match up different participants' recountings of their shared experiences, gain meaningful insights into their behaviors, and so on.