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

November 21, 2023

To the editors of *Psychological Review*:

We have enclosed our revised manuscript entitled *Feature and order manipulations in a free recall task affect memory for current and future lists* (manuscript submission ID: REV-2023-0784). We appreciate the reviewers' insightful comments on our original submission. We provide detailed point-by-point responses on the following pages. The reviewers' comments are italicized and our responses are in **bold**.

One issue that we wish to highlight is our manuscript's theoretical contributions. As noted in your decision letter for our previous revision, all three reviewers asked for clarification regarding how our study advances current theory. In summary, our work advances theory in several important ways, all related to characterizing **how memory organization changes over time (controlling for the to-be-remembered content of experience), and how the way earlier events are experienced affects how later events are experienced and remembered**.

To provide some background, nearly all prior list-learning studies of episodic memory (e.g., using free recall and other similar tasks) average participants' behaviors across a series of similarly structured lists (e.g., a sequence of random word lists, a sequence of blocked categorized lists, etc.). A small number of recent papers have begun to explore across-list effects in these tasks (e.g., Lohnas et al., 2015, Psych Review; also reviewed by Kahana, 2020, Annual Review of Psychology). However, the focus of this latter work has been on explaining intra-list errors (e.g., mistakenly recalling a word from a previously studied list rather than correctly recalling a word from the just-studied list). Our study is aimed at systematically teasing apart the interactions between three aspects of the stimulus (list order across lists, word order within lists, and item appearance) and different aspects of behavior (we focus on recall clustering, but we also examine serial position effects and output order effects). We ask: **how do the ways participants experience (and remember) early lists affect how they experience and remember later lists?** Just as memory researchers have historically conceptualized list learning tasks like free recall as experimental analogs of real-world episodic memory, our work helps to elucidate how earlier experiences might affect how people experience and remember later experiences.

The specific factors we chose to highlight in our study were intended to build direct links with more "naturalistic" tasks (e.g., memory for real-world events, memory for narratives and movies, etc.) that contain rich temporal and semantic structure at multiple scales. An important aspect of memory in these scenarios is that earlier experiences can impact how later experiences are perceived and remembered.

The order and appearance manipulations within and across lists in our study were intended to evoke analogous changes in memory for the stimuli in our experimental paradigm. Specifically, lists in different conditions of our experiment contain structure at different timescales (akin to event boundaries) and across different modalities (semantic, lexicographic, and visual).

Reviewers 2 and 3 also remarked on the "exploratory" nature of our work. Our study is (by design) somewhat exploratory, because although the basic free recall paradigm has been used by psychologists for roughly a century, our approach and manipulations are new. In turn, this enables us to explore new across-list phenomena and interaction effects whose complexity lies somewhere between that of traditional random list learning and natural experience.

One set of advances to come from our study pertains to several novel behavioral findings that we uncovered by incorporating appearance and order manipulations into a free recall paradigm:

- Even though we didn't test participants' memories for the words' visual appearances, randomly varying the words' font colors and onscreen locations within individual lists led participants to more strongly temporally, semantically, lexicographically, and visually cluster their recalls— a proxy for estimating how much participants relied on these stimulus features in organizing their memories. As we discuss in our revised manuscript, this result has **important theoretical implications for how experience and memory interact with task goals and attention**.
- When we changed the lists' visual properties *across* lists (e.g., altering on different lists whether all of the words were presented in black at the center of the screen, or in random colors and locations that changed with each new word), the moment of change seemed to induce a sort of "event boundary" in participants' experiences. Just as transitioning from one situation to another seems to reset how new experiences are processed and remembered, we found that visually induced event boundaries seemed to reset how the participants processed and remembered pre- versus post-boundary lists of words. Again, as discussed in our revised manuscript, this finding shows that **even aspects of experience that are incidental to the current task, situation, or focus of attention can induce event boundaries**, including corresponding effects on memory.
- A basic tenet of most modern theories of memory is that prior experiences guide future expectations, behaviors, and memories. When we sorted early lists semantically, we found that participants recalled later (randomly sorted) lists in a semantically clustered order as well. Other order manipulations (lexicographic, visual) did not show this pattern. **This suggests that certain feature dimensions of our experiences are "stickier" than others, such that structuring earlier experiences along those dimensions can have lingering effects on how we experience and remember later experiences.**

A second important set of advances from our study relates to the **experimental paradigm** itself. We show how three different aspects of stimuli on a list-learning task (semantic, lexicographic, and visual features) may be systematically and (largely) independently varied to study the resulting impact on memory. We also show how **these manipulations may be carried out in real time** to home in on individual differences. We expect that our dataset of 491 participants (who collectively studied and recalled

hundreds of thousands of words across 11 experimental conditions) will also serve as a **valuable resource for testing future models and theories of how the structure of experience affects how we perceive and remember ongoing and future experiences**.

Thank you for considering our revised manuscript.

Sincerely,

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

Reviewer #1: Manning and colleagues present an interesting and innovative study of recall organization effects in the free recall task. Across several conditions, the features of study items are manipulated, and the presentation order of the items is manipulated, in order to examine how these manipulations affect memorability of the studied items, and several organizational measures of recall dynamics. The paper includes a technically sophisticated and innovative manipulation that leverages real-time analysis of a participant's recall behavior to alter subsequent lists such that their structure is either consistent or inconsistent with that participant's recall behavior to that point. The authors make appropriately measured claims relative to the results that they report, and the analyses were well-designed and convincing.

We thank the reviewer for their positive comments.

My major concern about this paper is that, as it is currently structured, the theoretical stakes of the study are unclear. The abstract sets up the scope and mission of the paper, with a central point being that a manipulation of the list structure affects how participants recall that list, as well as later lists. However, the results aren't clearly tied to a theoretical perspective, making it difficult to assess the importance of particular results.

This important point was raised by both other reviewers as well. We have addressed this issue in three (main) ways. First, we have unpacked some of the core intuitions behind our study design in the introduction (pages 5–6) to help clarify the overarching logic. Second, we have added additional citations of relevant literature throughout our results section (pages 27 – 53) to tie our specific results to relevant findings in the literature. Third, we have expanded our discussion of the theoretical implications of our work, with a focus on clarifying connections between the questions raised in the introduction, our present study's findings, and the relevant literature. Of note, we have added several new sections to the discussion: *Memory consequences of feature variability* (pages 57–59), *Free recall of blocked versus random categorized word lists* (pages 61–62), and *How do different types of clustering relate to each other, and to memory performance?* (pages 65–-66).

For example, one of the central results is that if early lists are sorted by category, some participants will continue to organize their recall responses by category on later lists. This result seems reasonable and plausible, and generally consistent with the idea that people can strategically alter the sorts of internal cues they use to guide memory search. But without a clearly stated theoretical framework guiding the manuscript, it was hard to determine the importance of the result.

Nearly all published studies of free recall treat each list as independent of all other lists, and most use randomized list orders to average away order-specific effects on memory. These designs typically preclude detailed examinations of how studying earlier lists might impact memory for later lists– or, more broadly, how early experiences might affect how we process and remember later experiences. For this reason, although modern formal (mathematical) accounts of episodic memory (which have been primarily developed using free recall data) *imply* that prior experiences will affect future experiences, there is little direct empirical evidence for this phenomenon.

Only a few studies (e.g., Lohnas et al., 2015, Psych Review; also reviewed by Kahana, 2020, Annual Review of Psychology) have attempted to directly examine across-list effects in free recall. These studies used randomized word lists, and they were largely focused on explaining prior list intrusion

errors– i.e., incorrect recalls of words that participants studied on prior lists, rather than on the most recently studied list that they were being tested on. That prior work was important because it showed how formal (mathematical) accounts of free recall could scale to multiple lists. Multi-list effects were a challenge for earlier models, which had trouble reconciling memory effects at wide ranges of timescales (e.g., nearby items on a single list versus items from temporally separated lists). This "range of timescales" issue is also a key challenge for relating accounts of list learning (which typically unfolds over seconds to minutes) to real-world episodic memory (which can unfold over anywhere from seconds to decades).

Our manuscript presents an attempt to provide a richer account of *how* past experiences can influence how we perceive and remember future experiences. By studying this phenomenon using free recall, we connect prior work on modeling episodic memory with the notion that, in everyday life, our past experiences influence how we perceive and remember our future experiences. Further, because we reproduce this phenomenon in a free recall laboratory experiment, we were able to measure how different aspects of experience, and different dimensions of variation in experience, affect memory for future experiences. The example the reviewer identified– that studying early lists sorted semantically leads participants to semantically cluster later (randomly sorted) lists shows that prior experience *can* influence future memory. But it is not universal. For example, sorting early lists lexicographically or visually did not show this "carryover" effect. Taken together, this set of findings tells us that different aspects of our experience appear to have different effects on how we perceive and remember future experiences (e.g., as opposed to each feature dimension exerting equal influence, or influence in proportion to its task relevance or attentional salience, etc.). We have added a discussion to this effect on pages 57–59.

The finding that adding irrelevant features to a list boosts memorability and boosts several organizational measures is also interesting. But the conclusions drawn from this result are limited, as there are many potential reasons why it may be true.

We agree that this finding is interesting! In addition to the theoretical implications, we also see potential practical implications of our finding that participants show improved memory when we manipulate incidental features in certain ways. For example, how might this be further optimized? Could we use other incidental features, or a different set of manipulations, to further enhance memory performance? Is there a global maximum to this effect? And might analogous strategies be leveraged in other domains, such as classroom learning? We discuss these ideas on page 68 of our revised manuscript.

On the theoretical side, the core questions our finding gets at relate to how our memory systems are affected by "incidental" (task-irrelevant) stimulus features. We agree with the reviewer that there could be several possible explanations for this finding. Nonetheless, we do have some ideas about the "how and why" of this improvement (and we have added additional discussion to this effect on pages 57–59).

One question is about how we "treat" task-irrelevant features. For example, do we simply ignore those incidental features, since they are uninformative with respect to the current goal, as might follow from normative accounts of episodic memory? Since participants are never asked to recount the colors or locations, one could imagine that any space or mental effort allocated to encoding or attending to those features might detract from mental resources that could have instead allocated to encoding the words themselves (e.g., "stripped" of their font colors and onscreen locations). Or, alternatively, do we (passively or actively) incorporate those incidental features into our mental context (as predicted by recency-weighted average models of episodic context, e.g., Howard and Kahana, 2002)? In other words, how are participants internally representing their experience of studying each word, and which aspects of those representations affect how the words are organized and retrieved from memory?

A second question relates to how we might be able to leverage task-irrelevant features to affect memory performance. We found that varying the font color and presentation locations on later lists (after holding the font colors and locations fixed on earlier lists) boosts memory performance for those later lists. This might be an event boundary effect– for example, whereby participants intuit that the "situation" or "set of rules" has changed once the visual structure of the lists changes, with corresponding effects on memory. Or this could reflect an attention effect, whereby participants' attention systems start to flag after repeated study/test trials, but then are "reinvigorated" when new information appears to become available.

This theoretical agnosticism regarding the implications of the results makes it difficult for the reader to understand the importance of many of the results. It isn't that the paper is theory free; it contains clear discussions of theories of event cognition and computational models of memory search. The paper sets up a key mystery on page 4, regarding whether the situation models and schemas that organize memories of naturalistic experiences map on to organizational phenomena seen in word list studies. But the results of the paper aren't connected to that mystery in a clear way.

On one hand, there are clearly (major) differences between any list learning paradigm and the sorts of experiences that are more typical of everyday life. Therefore there are always going to be gaps between list learning and more naturalistic or realistic scenarios. Nevertheless, our list learning paradigm does incorporate some important features of more realistic experiences. For example, participants appear to treat feature changes as "boundary like," both within lists (e.g., especially visible in the purple curves in Fig. 3) and across lists (e.g., early vs. late lists in the Reduced (early) and Reduced (late) conditions, as shown in Fig. S8). These putative "event boundaries" have consequences for memory that are analogous to event boundaries in more realistic settings, such as enhanced memory for post-boundary content. Further, although real-world schemas are likely more complex than the learned representations in our paradigm, there do seem to be some analogues. For example, when early lists are sorted semantically, participants continue to semantically cluster their recalls of words from later lists, even when they are sorted randomly. One interpretation is that participants learn to pick up on the statistical regularities present in early lists (i.e., semantic sorting), and they organize their memories to take advantage of those regularities, similarly to how real-world schema

representations enable us to more efficiently process and remember our real-world experiences. Even when participants later encounter new items that do not conform to their current schema, they seem to interpret new (randomly structured) content through the lens of the current schema. Interestingly, however, this phenomenon only appears to occur for a subset of feature types. Specifically, sorting early lists semantically (by word category or referent size) produces this schema carryover effect on later lists, but sorting early lists lexicographically (by first letter or word length) or visually (by word color or presentation location) does *not* lead to carryover effects on later randomly sorted lists. This shows that not all "structure" in our experiences is equally meaningful or equally influential on how we interpret subsequent experiences. We have added a note to this effect on pages 57–59 of our revised manuscript.

Specific comments

When lists are sorted by a particular feature, clustering due to that feature increases. For the lists sorted by taxonomic category, this result could be better connected with the categorized free recall literature comparing performance on lists where categorized items are presented in blocks vs randomly intermixed. I was curious how much of the increased feature clustering was likely due to the temporal proximity of similar items after the list is sorted.

In general, it is difficult to fully tease apart feature clustering from temporal clustering when both are related. In the extreme, for example, if changes in a given feature were perfectly correlated with item position, any clustering measure would show exactly the same degree of temporal clustering and feature clustering (for that feature). For this reason, our ability to distinguish putative "feature clustering" versus "temporal clustering" effects on early (sorted) lists in the order manipulation conditions is limited. We have added a clarifying note to this effect on pages 23–24.

The more interesting question is what happens on *late* (unsorted) lists, which do *not* conflate feature values and temporal positions. We found that when we sorted early lists by semantic features (category and size), participants also showed an increase in semantic clustering on unsorted late lists (relative to late lists in conditions where early lists were also unsorted). This cannot be an artifact of sorting, since late lists in the feature rich (control) condition and in each of the order manipulation conditions are all unsorted.

Another interesting finding is that there appear to be some interaction effects between temporal and semantic clustering. When lists are sorted semantically, participants show an increase in "temporal" clustering, relative to their temporal clustering on control conditions. This would seem to reflect an interaction between "true" temporal clustering (which occurs on all lists and conditions in our experiment) and semantic clustering that *looks* like temporal clustering because semantic order and temporal order are conflated on those lists.

That said, this is an important point and we have added an additional analysis and discussion to attempt to address the question. In most cases, feature values and temporal position are *not* perfectly

correlated, and it is possible to estimate the contributions of feature clustering over and above temporal clustering alone in those cases.

In our previous submission, we described a permutation-based correction procedure for detangling observed clustering scores (i.e., the average similarity rank of successive recalls, for the given measure) from the position and feature labels of the recalled items. We gave the example of a list whose items were all "large" (i.e., had the same referent size label). In that (artificial) scenario, any items the participant recalled would show an uncorrected size clustering score of 1, since all adjacent recalls would have the same "size" label. However, this clustering score would not provide "real" evidence that the participant organized the items by size, since *any* set of recalled items from the list would yield the same clustering score. Our correction procedure randomly permuted the order of the recalled items many times. We can then use the average percentile rank of the observed clustering score, in the distribution of clustering scores obtained from permuted recalls, as the "corrected" clustering score. In our example (where any set of recalls would yield a clustering score of 1), the average percentile rank of the observed clustering score would be 0.5 (i.e., the expected "chance" level of clustering if the recalls were made in a purely random order).

Following the reviewer's comment, we have developed an additional correction procedure (described on pages 23–24) to specifically distinguish temporal versus feature clustering. For a given set of recalled items (whose presentation positions are given by x1, x2, x3, …, xN), we can circularly shift the presentation positions by a randomly chosen amount (between 1 and the list length) to obtain a new set of items. Since the new set of items will have the same (average) temporal distances between successive recalls, the temporal clustering score for the new set of items is equal (on average) to the temporal clustering score for the original recalls. However, we can then re-compute the feature clustering score for those new items. Finally, we can compute a "temporally corrected" feature clustering score by computing the average percentile rank of the observed (raw) feature clustering score within the distribution of circularly shifted feature clustering scores. This new temporally corrected score provides an estimate of the observed degree of feature clustering *over and above* what could be accounted for by temporal clustering alone.

On sorted lists, our temporal correction procedure "over-corrects" by removing potentially *meaningful* feature clustering effects, e.g., to the extent that presentation position and feature values are correlated. Therefore it is difficult to interpret "null" effects observed for these corrected fingerprints. We found that participants show that participants *semantically* cluster their recalls more on semantically sorted lists and later randomly sorted lists that follow semantically sorted lists, as compared with clustering on randomly sorted early and late lists (Fig. 5D). For the other order manipulation conditions, however, we found no systematic changes in feature clustering that survived this very conservative control analysis and our multiple comparisons correction procedure. In summary, we were able to detangle temporal and semantic clustering, but our examinations of other forms of clustering using this approach were inconclusive.

We also appreciate the reviewer's suggestion to flesh out connections between our findings and prior work on memory for blocked versus randomized categorized lists. We have added a discussion to that

effect on pages 61–62.

On page 42 the authors describe an ambiguity of interpretation. Participants might be malleable with regard to how they organize incoming information. Alternatively, they might have a preferred way of organizing information. This motivates the adaptive experimental conditions. The stabilize vs destabilize lists were not reliably different in terms of overall memory performance, but they were reliably different in terms of degree of temporal organization. I wasn't clear whether it is important that the stability manipulation affects overall performance. The stabilize lists order the items by some characteristic feature, and the destabilize lists actively avoid this. So it seemed that the differences in temporal organization could be due to the tendency for items in neighboring study positions to have some similar features on the stabilize lists. I wasn't sure if this interpretation was consistent with the authors interpretation or not.

As the reviewer notes, our goal in running the "adaptive" condition was to attempt to detangle several plausible interpretations of the results from the other experimental conditions.

If participants essentially pick up on whatever patterns appear in the presented items, then "adapting" how the items are presented should have no consequences on memory (with respect to overall performance *or* other aspects of memory, like temporal clustering). On the other hand, if participants impose their own preferred "fingerprint" on incoming information, then matching (or not matching) those preferences should have consequences on memory. Our "adaptive" results are consistent with this latter interpretation.

As the reviewer suggests, sorting lists according to participants' fingerprints will tend to result in neighboring list items having similar features, along particular dimensions. In turn, this appears to lead to increased temporal clustering. More interesting, though, is that the particular *dimensions* that the "stabilize" condition sorts on are different across participants. In other words, the weightings of features (with respect to how people organize their memories) tend to be different across people. So it's not simply that *any* feature dimension will be equally effective at influencing memory performance or temporal organization for everyone. Rather, people tend to be affected differently by different combinations of features, which we can get at through their memory fingerprints.

In addition to evidence from the adaptive condition, we can also get at this idea by comparing memory fingerprints within versus across people, in the feature rich condition (where lists are ordered randomly). We designed an analysis to compare the similarity (correlation) between the fingerprint from a single list (from one participant) and (a) the average fingerprint from all other lists from the same participant versus (b) the average fingerprint from each other participant (across all of their lists). We found that participants' fingerprints on a held-out list are reliably more similar to the same participant's fingerprints on other lists than to other participants' fingerprints (pages 29–30).

Reviewer #2: Summary

The authors present results from a feature-rich free-recall paradigm, wherein 16 lists of 16 items each are presented

with variation in six features reflecting semantic, lexicographic, and visual properties of words. They find that the addition of variable visual features leads to increases in the observed organization based on other features, including temporal organization. They also examined lists that change from reduced to feature rich, or vice-versa, halfway through the lists. They find that recall is higher when switching from reduced to feature rich lists, compared to the fully reduced and fully feature-rich conditions, potentially due to the switch serving as an event boundary. The authors next present a series of manipulations designed to examine effects of sorting lists by each of the different features. In each of these conditions, lists are only sorted on the early lists. They find that clustering based on the sorted feature are increased in each case in the early lists, and temporal clustering is also increased in most conditions. However, they find that feature clustering generally did not transfer to new lists on average, as clustering to the relevant feature was not greater than late lists in the feature rich condition.

Note: this summary is mostly correct. We did, however, find that sorting early lists semantically (by category or referent size) led to increased semantic clustering on early *and* late (randomly sorted) lists.

They argue, however, that individual differences in feature clustering modulate whether the ordering manipulation on the early lists transfers to the late lists. Finally, the authors present results from an adaptive paradigm, which attempts to detect the clustering tendencies of individual participants and use this "fingerprint" to adaptively organize lists to either match or mismatch their fingerprint. They find a trend toward an increase in recall in the adaptive condition, and an increase in temporal organization.

To clarify, in the "adaptive" condition we see the (weak/trending) increase in memory performance and the increase in temporal clustering only on "stabilize" lists; the increases are computed relative to participants' performance on the "destabilize" and "random" lists.

Evaluation

The feature-rich free-recall paradigm provides a rich set of data that may be useful in examining the relationship between different types of organization in free recall. The strategy of adaptively organizing lists based on earlier recall performance is innovative and may have interesting applications for future work. The effect of rich visual features on organization by other features is interesting, and should be followed up further in revisions or future work. The use of multiple organizational measures with permutation-based correction is also a strength, as it provides a set of useful measures of different aspects of recall behavior.

We thank the reviewer for their positive feedback.

The work is largely exploratory in nature, which isn't necessarily a problem, as benchmark datasets with many conditions have historically often proven useful for theory development. However, connections to prior work are not always clear. For example, is a decrease in organization across early and late lists predicted based on existing data? How do the present results relate to the rich literature on ordering effects in categorized free recall (e.g., Puff 1974, Borges & Mandler 1972)? Also, one of the major conclusions in the manuscript is that there are "carryover" effects of organized lists that are encountered early on. I have concerns about the interpretation of the individual difference measures that are cited as evidence of carryover effects, as I believe they can be plausibly explained without there necessarily being any true carryover effects. Assuming that these effects can be validated, the manuscript could

benefit from some speculation on how these effects might occur, or on what future work could be done to determine the underlying mechanisms; currently, the hypothesized mechanism for transfer effects is unclear. Similarly, it is unclear why adding randomly varying visual features would affect clustering along other item dimensions, and the manuscript would benefit from more discussion of why that might be. Finally, there is a lot of discussion of clustering "fingerprints", but the justification of this idea is currently unclear. Evidence of individual reliability and diversity across individuals of clustering profiles would help justify the idea of a "fingerprint".

The reviewer raises some important questions and observations here. We agree that our work here is largely exploratory, by design. Although the basic free recall paradigm has been used by psychologists for roughly a century, our approach and manipulations are new, as the reviewer notes above. This enables us to explore new across-list phenomena and interaction effects whose complexity lies somewhere between that of traditional random list learning and natural experience. In doing so, we hope to build new connections between the list learning and naturalistic memory literatures.

Nevertheless, the reviewer's point that we need to do a better job of connecting with prior work is well taken. To that end, we have expanded our introduction (pages 5–6), discussion (pages 57–59, 61–62, and 65–66), and parts of our results section (pages 27 — 53) to clarify and discuss additional connections with prior work. We also appreciate the pointer to Puff (1974) and Borges and Mandler (1972), and we have added citations of those papers in our discussion section (page 62).

The reviewer also asked about evidence of whether the clustering effects we observe are truly like "fingerprints." As the reviewer notes (and we agree), we see two important aspects of the "fingerprint" notion. First, within a participant, patterns of clustering scores should be stable across lists. Second, across participants, we should see some diversity in these patterns. We carried out an additional analysis to test both of these properties.

We reasoned that the "feature rich" condition should give us the "cleanest" estimates of clustering (i.e., without potential biases introduced by the manipulation conditions in our study). Each participant in the feature rich condition studied and recalled a total of 16 lists, yielding 16 sets of "fingerprints" for that participant. We asked: holding out one of these fingerprints at a time, could we "match up" which participant it belonged to? Specifically, we created two distributions of correlations. The first distribution comprised "within-participant" correlations between the fingerprint from a held-out list and the average fingerprint from all remaining lists. Each participant contributed a total of 16 correlations to this distribution. The second distribution comprised "across-participant" correlations between one the fingerprint from one held-out list from one participant, and the average fingerprints (across all lists) for each other participant. We repeated these across-participant comparisons for each pairing of lists (from the "template" participant) and other participants. Therefore each participant contributed 16 x (N - 1) correlations to this second distribution (one per list, times N - 1– i.e., the number of participants excluding the template participant). When we compared these distributions we found that the "within-participant" correlations were reliably higher than the "across-participant" correlations. This new analysis helps to justify our framing of participants' clustering scores as "fingerprints." We describe this new analysis on pages 29–30.

Major issues

1. Page 16: 4 lists seems pretty quick for establishing a fingerprint. Do you have any reliability measures on this? Also, do you have evidence of reliable diversity in the fingerprint? Does everyone cluster most by the semantic features, or is there evidence that some participants really tend to cluster more by one of the other features in a reliable way? These are important things to demonstrate to justify the idea of a "fingerprint".

These are great questions! We have added several additional analyses to the supplement to address these issues:

How quickly fingerprints stabilize is difficult to answer objectively. For example, one challenge is that participants' fingerprints may change over time, e.g., across different order manipulations (e.g., Figs. 6, S8). Nevertheless, we have added two analyses to get at this question. First, we can ask: using the average fingerprint from lists 1…*N*, what is the average Euclidean distance to the fingerprint for list N + 1? We plot the distances as a function of N in our revised Figures 6 and S8. In most conditions the fingerprints appear to stabilize after just a few lists. We also show in Figure 9 how the memory fingerprints change across lists, for each list type in the *adaptive* condition (stabilize, destabilize, and random). Specifically, the fingerprints tend to become more consistent across lists in the stabilize condition, and more variable across lists in the destabilize condition.

How "reliable" across-participant fingerprint differences are is also difficult to answer, again because fingerprints *within* an individual might not be stable over time. That said, one way of examining this question is to compare within-individual versus across-individual fingerprint variability. If participants' fingerprints are more similar to their *own* fingerprints (on other lists) than *other participants'* fingerprints on other lists, this would indicate that there are reliable individual differences in people's memory fingerprints. We have added a new analysis to look at this, as described above (pages 29–30). In brief, we confirmed that individual participants' held-out fingerprints from a single list are more similar to the same participant's average fingerprint from their other lists, than to any other participant's average fingerprint (across all lists).

Although our individual differences analysis indicated that individual differences in memory fingerprints are reliable (i.e., somewhat stable across lists for a given individual), this does not mean that everyone's fingerprint is "random." Rather, we found that participants are more likely to cluster their recall dimensions along some dimensions versus others. This can be seen most directly by looking at the memory fingerprint bar plots (Figures 4, S5, S6). We see two general patterns that seem to hold across experimental conditions. First, there are substantial differences in clustering scores across different feature dimensions that tend to hold across participants. For example, as the reviewer notes, participants in most conditions were more likely to show semantic clustering (by category or referent size) than clustering along the other feature dimensions. This can be seen by the relative heights of the bars in those plots. Second, however, for any given feature dimension, we see substantial variability across participants (this can be seen by the sizes of the error bars, which reflect bootstrap-estimated 95% confidence intervals).

2. Page 21: I wonder if this is the most effective way to match organizational tendencies of individual participants. The relatively small sample size for determining the fingerprint of each participant means that this estimate will be noisy. As a result, organizational factors that the participant isn't really using (despite corrected organization that is greater than zero) will sometimes factor into the organization of new lists. Matching based on the correlation with the fingerprint will also mean that pretty different designs might be selected with equal probability. For example, if I'm understanding your adaptive selection procedure correctly, a candidate item ordering with perfect category clustering and chance clustering on other features would have the same correlation with the template fingerprint as another candidate that had moderate category clustering and slightly lower, but still above chance, clustering on the other dimensions. I wonder if another algorithm wouldn't be better suited to adaptive organization, like just selecting the feature with the strongest clustering in the template and ordering to maximize adjacent item similarity along that. Do you have any evidence that participants reliably cluster recall along multiple feature dimensions in a stable ways that differ between individuals (e.g., one participant might reliably cluster by both category and size, while another clusters by length, color, and location)?

There are a few important points to unpack here. One question is about whether correlating fingerprints (versus some other approach) is the most effective way to match organizational tendencies of individual participants. The reviewer proposes using just a single feature dimension and using that dimension to order the lists. This would be similar to our approach in the early lists in the "order manipulation" conditions, where we sorted those early lists in each condition along a single feature dimension. The key difference is that the reviewer is proposing that we choose which dimension to sort along according to participants' prior recalls. Although we cannot carry out this manipulation using our existing data (i.e., without running an additional condition using the proposed approach), we can get some insights into what we'd expect, based on the existing data in the order manipulation conditions.

When we sorted early lists along a single feature dimension, we found that participants varied in how much they clustered their recalls along that dimension (e.g., the error bars in Figure 4 are relatively spread out across a wide range of clustering scores). Participants who clustered their recalls along their condition's feature also tended to remember more items (Fig. 7A) and showed stronger temporal clustering scores (Fig. 7E). A possible interpretation is that participants who showed the strongest clustering scores might have come into the experiment with biases in how to organize incoming information that happened to match their assigned condition (e.g., as though they were participating in the "adaptive" variant the reviewer is proposing). Consistent with this view, participants in the order manipulation conditions who showed strong feature clustering on early (sorted) lists also showed strong feature clustering along their condition's feature dimension on late (random) lists (across all participants and conditions: $r(179) = 0.592$, $p < 0.001$). Another possible explanation for this result is that sorting early lists along a given feature dimension induces some participants to organize incoming information on those lists, and on later lists, according to that feature dimension. The adaptive condition we ran was designed to help distinguish between these interpretations.

The reviewers' question about whether participants reliably cluster their recalls along *multiple* feature dimensions simultaneously, e.g., as opposed to just a single dimension at a time, is a good one. One way of examining this issue is to ask: what do individual participants' average fingerprints actually look like? Do they put all of their weight on a single dimension, or is substantial weight also spread along other dimensions? We reasoned that the "feature rich" condition (i.e., including visual features but sorting every list randomly) would provide the cleanest (i.e., least biased) insight into participants' clustering tendencies without potential confounds of sorting along one or more dimensions, or of changing the list statistics across lists. We ran a simple analysis whereby, for each participant in the feature rich condition, we ranked their average fingerprint scores from largest to smallest. We then used across-participant t-tests to discover which ranks' distributions of scores were reliably larger than the chance value of 0.5. We found that the top 3 most highly ranked feature dimensions were consistently (reliably) above chance, indicating that participants reliably cluster along (at least) 3 feature dimensions (including temporal clustering, which was often ranked highest). We report these results on pages 52–53 (also Tab. 23).

3. I also wonder about potential effects of serial position on organization. Will recency and contiguity, combined with a relatively smoothly varying feature dimension, lead to greater apparent organization along that dimension? Or does the correction procedure correct for that? I don't think it will, as the correction procedure only accounts for the items recalled, but not temporal organization. I assume that temporal organization in a list with temporally organized features will lead to apparent clustering based on those features. That leads to some questions about your individual difference measures (see below).

Reviewer 1 raised a similar concern about potential confounds between temporal and feature clustering on sorted lists, and we developed a new correction procedure to disentangle the two forms of clustering. For convenience, here is our description of the new analysis, taken from our response to Reviewer 1:

> ***Following the reviewer's comment, we have developed an additional correction procedure (described on pages 23–24) to specifically distinguish temporal versus feature clustering. For a given set of recalled items (whose presentation positions are given by x1, x2, x3, …, xN), we can circularly shift the presentation positions by a randomly chosen amount (between 1 and the list length) to obtain a new set of items. Since the new set of items will have the same (average) temporal distances between successive recalls, the temporal clustering score for the new set of items is equal (on average) to the temporal clustering score for the original recalls. However, we can then re-compute the feature clustering score for those new items. Finally, we can compute a "temporally corrected" feature clustering score by computing the average percentile rank of the observed (raw) feature clustering score within the distribution of circularly shifted feature clustering scores. This new temporally corrected score provides an estimate of the observed degree of feature clustering over and above what could be accounted for by temporal clustering alone.***

Here, the reviewer is raising an additional concern about additional potential effects of recency. Since the null distribution in our "temporal correction" procedure above retains the positions of recalled

items (i.e., while circularly shifting the presentation positions of the items), the same correction procedure also corrects for serial position effects (e.g., primacy and recency), in addition to more subtle potential temporal or position-related effects. In summary, it is not always possible to fully detangle temporal or serial position effects from feature clustering effects on the order manipulated lists, since feature values and temporal positions are often strongly correlated on those lists. However, we do see at least *some* evidence that participants are truly exhibiting feature clustering over and above what would be expected by temporal clustering (combined with order manipulations, primacy, and/or recency alone). For example, participants show reliable semantic (category and size) clustering even after factoring out the potential effects of temporal order and serial position (Fig. 5D).

4. There's a lot of reliance on "trending" (.05 > p > .1) statistical results, most notably for the important comparison of recall in the stabalize adaptive condition with the other conditions. There are also a lot of comparisons in the manuscript, with no apparent control for multiple comparisons that I noted.

Although we report many "trending" statistical results, as the reviewer notes, we have tried to temper and tone down our claims about what those results *mean* whenever possible (e.g., by using language like "participants *tended* to X" or "X was *numerically larger* than Y," and so on). None of our major claims depend on these "trending" effects. For example, with respect to the reviewer's comment about the "adaptive" results, we don't treat the trending effect for the "stabilize" versus "random/destabilize" as evidence for (or against) our core claims; our intention was simply to report what we found. "What we found" was sometimes highly significant, and in other cases the statistical results were "trending"or "not significant." Although our field tends to treat 0.05 as a critical threshold that separates "significant" from "non-significant," we see the "truth" as less binary. We suspect that, as in most psychological studies, neither our data nor our tests are sensitive enough to meaningfully distinguish between a *p*-value of 0.049 versus 0.051, for example. In the interest of transparency, we simply report all tests and *p*-values, share our code and data (including a complete history of every analysis we ran since collecting the data, available in our GitHub repository's history), and describe what we found as accurately as we are able.

The main effects, on which we *do* base our core claims, are almost all very highly significant (we report those *p*-values as "$p < 0.001$," and in many cases the *p*-values were *substantially* less than 0.001; we made all raw statistical results available in our shared analysis code in our linked-to GitHub repository). These results would pass even a very conservative correction for multiple-comparisons, like Bonferroni correction (e.g., treating every test as completely independent). That said, it is also not clear to us how we would appropriately control for multiple comparisons to yield an *accurate* estimate of the *p*-values we report. A simple Bonferroni correction (e.g., multiplying all *p*-values by the raw number of tests we report in the manuscript and upper-bounding at 1) would yield *p*-values at or near 1 for most of the "trending" effects the reviewer is referring to. However, we feel that would mask a lot of the important nuance in our results, e.g., by removing any information about the relative reliability of different effects that readers may find useful. For analyses where we simply carry out "all possible comparisons" (e.g., Fig. 5, Tabs. 1–23), we use the Benjamini-Hochberg procedure to "correct" *p*-values and reduce the false discovery rate.

Because many of our analyses are exploratory, our goal here is to "report what we found and interpret as best we can," rather than attempting to bolster support for one particular theory over another. To be clear, we do see our work as making important theoretical contributions, as outlined above and in our discussion section (pages 66–68). But our analyses were primarily *designed* to clarify "what was going on in the dataset" rather than to test a specific hypothesis or theory.

5. I'm confused about some apparent differences between conditions. In Figure S8, why aren't the early lists of the reduced (early) condition near the early lists of the reduced condition in the component plots? Aren't those early lists effectively the same, and so wouldn't you expect similar clustering behavior? Perhaps relatedly, why is there apparent organization by color and location in the early lists of the reduced (early) condition (Figure S5)? Similarly, why is there apparent organization by color and location in the late lists of the reduced (late) condition? Aren't those lists shown in black, with no variation in location?

These issues relate to how we computed "corrected" fingerprint values. In our previous submission, we defined the "corrected" fingerprint as the proportion of shuffled fingerprints that are less than or equal to the observed fingerprint. By this definition, if all list items have the same value along some feature dimension (e.g., word color), then the observed and shuffled fingerprints will have the same value, and therefore the corrected fingerprint will be 1.

We appreciate the reviewer's point that this behavior is counter-intuitive, particularly with respect to this "edge case" that occurs when the feature values along some dimension are constant across all words on a given list. Therefore in our revised manuscript, we have redefined the fingerprint correction procedure to be the average value across all shuffled fingerprints, where each shuffled fingerprint is given a value of 0, 0.5, or 1 according to the following criteria:

- **If the shuffled fingerprint is strictly greater than the observed fingerprint, it receives a value of 0**
- **If the shuffled fingerprint is *equal* to the observed fingerprint, it receives a value of 0.5**
- **If the shuffled fingerprint is strictly less than the observed fingerprint, it receives a value of 1**

This revised definition results in the behavior we think the reviewer is "expecting"-- specifically, when all feature values are the same across all words on a list, the corrected clustering score will be 0.5 (i.e., exactly at "chance"). We have updated all figures and statistics to use this revised definition.

We note that we do continue to see some (unexpected) differences across conditions, e.g. on early "reduced" versus early "reduced (early)" lists. We don't have a specific theoretical explanation for this. We hope that by publishing our code and data, others might dig into these sorts of mysteries further!

6. I'm not convinced that the correlations presented in Figure 7 provide evidence for persistent effects of the ordering manipulation of the early lists. Assume that (1) there is individual variability in how well participants tend use a given feature to organize their memory and recall, (2) there is individual variability in temporal organization, and (3) participants who tend to temporally cluster their recalls also tend to recall more items (Sederberg et al. 2010, Memory & Cognition). Given the ordering of the early lists, a tendency to cluster by the relevant feature will lead to increased temporal organization, which may also lead to increased recall. I believe each panel of Figure 7 can be

explained using these assumptions, without resorting to any carryover effects. I focus on the panels that correlate features of late list recall will behavior on early lists. (C) Feature clustering is strongly correlated with temporal clustering on early lists (see panel E); temporal clustering has previously been shown to correlate with recall (SE10), and recall may correlate across early and late lists for any number of reasons. (D) If individual differences in feature clustering tend to increase temporal clustering and recall on early lists, that tendency to cluster by the relevant feature on early lists should also be correlated with feature clustering on late lists. (G) Feature clustering on the early lists is confounded with temporal clustering (see panel E), and individual differences in the tendency to temporally cluster recalls will likely persist between early and late lists. (H) Similarly, temporal clustering on early lists is confounded with feature clustering on early lists, and individual differences in feature clustering are likely to be somewhat stable between early and late lists. More broadly, if the ordering manipulations actually modified subsequent behavior on late lists, then why are there not clear carryover effects on average? Even if the manipulation only affected a subset of participants, an average effect should still be measurable. The conclusion that ordering manipulations may modulate later recall performance and organization is a key claim in this manuscript, so this is an important concern to address.

The reviewer is raising several important issues here. First, we agree that the literature supports the first 3 "assumptions" proposed by the reviewer– i.e., that feature and temporal clustering vary across individuals, and that temporal clustering and recall performance are positively correlated. The key question is whether, as the reviewer suggests, these three assumptions alone can account for the "carryover effects" we report.

As the reviewer notes, it is difficult to distinguish feature vs. temporal clustering on early lists, since both measures are strongly correlated (Fig. 7E). This could potentially be explained by the fact that early lists are sorted by the given condition's feature, so temporal position and feature values are correlated on those lists. On later lists, however, the association between temporal position and feature values in the stimulus itself is broken, since those lists' items are presented in a random order. Nonetheless, we continue to see correlations between feature clustering and recall probability (Fig. 7B) and between feature and temporal clustering (Fig. 7F). Because features and time are *not* conflated on these later lists, the associations (and correlations) in those panels cannot be explained by temporal clustering effects alone.

The "carryover" effects we report relate specifically to comparisons between early and late lists. In Figure 7, we report four such comparisons:

- **Panel C: Feature clustering on early lists is correlated with recall probability on late lists. As the reviewer notes, this could potentially be explained by assuming that participants with better memories (on both early and late lists) are the same participants who show strong temporal clustering (on both early and late lists). Specifically, what we report in the panel as "feature clustering" could actually be an artifact of temporal clustering alone.**
- **Panel D: Feature clustering on late lists is correlated with recall probability on early lists. Because items on late lists are presented in a randomized order, feature clustering scores on late lists cannot be explained by temporal clustering alone. Therefore explaining this**

association would seem to require some additional factor(s) outside of the three assumptions mentioned by the reviewer.

- **Panel G: Feature clustering on early lists is correlated with temporal clustering on late lists. As the reviewer notes, this could potentially be explained by assuming that some participants tend to exhibit stronger temporal clustering in general (i.e., across lists). If so, what we report in the panel as "feature clustering" on early lists could actually be an artifact of temporal clustering alone.**
- **Panel H: Feature clustering on late lists is correlated with temporal clustering on early lists. Again (as with Panel D), because temporal and feature clustering on late lists are not conflated, the "artifact" explanation (i.e., that what we call temporal and feature clustering are actually both reflections of temporal clustering alone) no longer holds. Explaining this association would (again) seem to require some additional factor(s) outside of the three assumptions mentioned by the reviewer.**

In other words, the primary evidence (in Fig. 7) for what we are interpreting as "carryover" effects is that the same participants who remember more words and show strong temporal clustering on early lists also continue to show strong *feature* clustering on late lists. Because feature and temporal clustering on late lists are decoupled, feature clustering on late lists cannot be explained by temporal clustering effects (on early *or* late lists). Further, we note that the specific features reflected in the figure (in each color/condition) are determined by the experimental condition– e.g., the dark purple dots reflect category clustering in the category manipulation condition, the dark green dots reflect color clustering in the color manipulation condition, and so on. It's not simply that some participants are clustering *all* features more on late lists. Rather, participants who show strong feature clustering on late lists *for their condition's feature* also tend to recall more words and show stronger *temporal* clustering on late lists, even though those late lists' items are presented in a randomized order.

7. In the abstract: "Inferring that we are in one type of situation versus another can lead us to interpret the same physical experience differently." - It doesn't seem like this necessarily applies here; participants will likely view organized early lists and randomized late lists quite differently. Even if you do see carryover effects of the early list order manipulations (see above for questions on this point), it doesn't follow that participants are necessarily interpreting the later lists differently.

We agree that, all else being equal, one would expect participants to experience sorted (early) versus randomized (late) lists differently. That's one reason we think that it's particularly interesting when participants assigned to one of the order manipulation conditions cluster late lists along their condition's feature dimension. It implies that the early lists *may* be influencing how they remember late lists. This effect (what we're calling a "carryover" effect) is strongest for participants with better memories (Fig. 7C, D) and who show the strongest temporal clustering effects on early and late lists (Fig. 7G, H). We agree that it's not clear whether participants subjectively "interpret" early lists versus late lists differently. All we can measure directly is whether participants who were randomly assigned to different conditions exhibit different behavioral tendencies on those later lists. We include a clarification to this effect on pages 49–50.

8. More discussion of theoretical implications is needed. Why would adding incidental visual features lead to increased clustering based on temporal, category, size, length, and first letter features? Why and how specifically would viewing ordered lists early on affect later organization? The discussion focuses mostly on what results were observed, without much engagement with what mechanisms that might be involved in producing these results.

We have substantially expanded our discussion of the theoretical implications of our study (pages 57–59, 61–62, and 65–66). That said, we do not have definitive answers for all of the interesting questions the reviewer raises here!

In brief:

- **We do not know why adding incidental visual features leads to increased clustering along other feature dimensions. We speculate (pages 57–59) that when ongoing experiences vary along more dimensions, this provides a sort of "richness" signal that leads participants to increase their attention and effort on the task. However, we also clarify that our current study does not provide direct evidence of any particular mechanism for this finding; our intention with respect to that particular finding was simply to report what we found. Our main takeaway is that changing the content of our ongoing experiences, even in ways that are incidental to our current task or focus of attention, can affect how we organize memories of our experiences.**
- **The reviewer also asks how "specifically" viewing ordered lists early on might affect later organization. We interpret the carryover effects in our study along the same lines as prior work showing that priming participants with one context or task (versus another) can affect how future information is experienced, learned, and remembered (e.g., pages 2, 50). We speculate that participants might be "primed" (to use the term loosely) by early ordered lists to adopt a particular schema for interpreting or organizing information on those lists, e.g., by forming expectations about how the lists are organized or sorted along different feature dimensions. If the same schema remains "active" when the participant encounters later (random) lists, it could explain why those random lists are sometimes clustered according to the same feature dimension that was used to sort early lists (page 63–64).**

Minor issues

Page 4: "As formalized by models like the Context Maintenance and Retrieval Model (Polyn et al., 2009), the stimulus features associated with each word (e.g. the word's meaning, size of the object the word represents, the letters that make up the word, font size, font color, location on the screen, etc.) are incorporated into the participant's mental context representation (Manning, 2020; Manning et al., 2015, 2011, 2012; Smith and Vela, 2001). During a memory test, any of these features may serve as a memory cue, which in turn leads the participant to recall in succession words that share stimulus features." - This makes it sound like the influence of mental context is a fact, rather than the theoretical construct that it is.

Good catch. We have modified that sentence to read "*According to* models like the Context Maintenance and Retrieval Model…" to clarify that we are reporting on *theory* as opposed to

observable fact.

Page 13: You note that some features are correlated, but seeing some quantification of this would be helpful. Most relevant is the correlation between features in your 16 selected lists, as that reflects your list assignment procedure. Plotting a correlation matrix of some kind would be helpful. It's the variation in similarity across items that seems relevant here, so I'd suggest doing something like calculating the pairwise distance between items within each list based on each feature, using the same distance metrics you used for list sorting of each feature, to obtain a distance matrix of interitem distances for each feature. Then, for each pair of features, what is the rank correlation (or Kendall's tau, or a similar measure) between their interitem distances? For example, to what extent does two items being in different categories correlate with them also being of a different size?

We have added a supplementary figure with the requested correlation matrix (Fig. S9).

Page 14: What are the units for location distances? Inches?

We have added a clarification that location distances were computed as percentages of the width of the viewable portion of the display (page 15).

Page 16: Will this selection procedure lead to lists that have greater similarity of adjacent items early in the list? Will the last item often have low similarity to the prior item, and have been selected just becase you ran out of items? A plot similar to figure 3B, but showing some sort of relative similarity of each adjacent pair of items as a function of serial position (perhaps normalized within each list and feature by rescaling such that similarity of adjacent items varies between 0 and 1, and then averaged across lists and participants), would be helpful. The boundary probability is related to this, but has been thresholded and will be somewhat less similar to serial position differences. Already it looks like the non-semantic features show an uptick in boundary probability at the end of the list; I'm guessing that normalized adjacent similarity will show a more obvious uptick. Any variability in the similarity of adjacent items with serial position is potentially relevant for interpreting your results.

Great suggestion! We have added an additional panel to Figure 3 (Fig. 3B) with the requested figure. In brief, we do see a slight "uptick" for the final items in the lexicographic and visual order manipulation conditions, but the magnitude of the uptick is within less than 1 standard deviation of the variability in distances we see across items at other positions.

We also note that we corrected a bug in our "even boundary" analysis (now Fig. 3C), whereby we were computing distances incorrectly for the "length" condition in our previous implementation, which led that curve to look (unexpectedly) different from the others. Essentially, due to a labeling error in that event boundary analysis, we were treating length as a discrete variable (i.e., distance = 0 if lengths are equal, and 1 otherwise), but we've corrected those computations to treat it as a continuous variable (i.e., distance = the absolute difference between words' lengths).

Page 21: Are "stablize" and "destabalize" the best terms here? The "destabalize" lists won't match the template determined based on the initialization lists, but they should be pretty stable in their organization, shouldn't they? It seems like different terms like "matching" or "non-matching" might fit better here. However, I might be

misunderstanding how the adaptive procedure worked; were fingerprints updated with each list, and was this used to select matching or non-matching lists? The methods text makes it seem like there is only one template calculated on the initialization lists, but the plotting of trends over lists in Figure 9 makes me think that maybe the fingerprint is updated with subsequent lists. Is that the case? If not, why plot the trend of correlation over list as in Figure 9D?

To clarify, the "fingerprint estimates" were always updated to reflect the average across *all lists up until that point in the experiment* (see *Online "fingerprint" analysis*, page 19)-- not solely the four "initialization" lists. Regarding our plotting decision in Figure 9D, there are a few points to our reasoning:

- **Aside from the initialization lists (which were always the first 4 lists– i.e., lists 0 – 3), participants experienced blocks of (4 lists each of) random, stabilize, and destabilize lists in a block-randomized order. In other words, for some participants, lists 4–7 were "random" lists, for others lists 4–7 were "stabilize" lists, and so on.**
- **The key trend we're trying to highlight in Figure 9D is that on stabilize lists, participants' fingerprints on the most recent list tend to become more and more similar to the average fingerprint (across all lists up to that point). On destabilize lists, by contrast, participants' fingerprints on the most recent list tend to be come *less* and *less* similar to the average fingerprint across all lists up to that point. In other words, people tend to organize words on stabilize lists in ways that get increasingly *more* similar to their average fingerprint, as they see more stabilize lists. On the other hand, people tend to organize words on destabilize lists in ways that get increasingly *less* similar to their average fingerprint, as they see more destabilize lists.**
- **This basic pattern also reflects our choice of terminology: stabilize lists tend to "stabilize" people's fingerprints (i.e., memory recall ordering tendencies), whereas destabilize lists tend to "destabilize" people's fingerprints.**

Page 22: Did you take item availability into account for your Nth recall curves? For example, if a participant recalled item 1, then item 2, then if taking item availability into account, you shouldn't place a zero into position 1 when calculating the probability of 2nd recall curve, because recalling the item in serial position 1 wasn't possible by the time you got to the 2nd recall. Instead, that list and position would be marked as missing data, as you were unable to assess the probability of recalling serial position 1 at output position 2 on that list. If the analysis is not conditional on availability, then this should be noted and justified.

Our "probability of nth recall" curves (Fig. S7) simply report the proportions of trials on which participants recalled the item presented at each serial position, at output position n. Our approach follows several prior studies' implementations of this analysis (e.g., Howard and Kahana, 1999, Figure 1; Polyn et al., 2009, Figure 6; Zhang et al., 2022, Figure 6E; etc.). This "unnormalized" approach also closely matches standard ways of computing serial position curves (i.e., the unnormalized proportion of trials on which item x was correctly recalled) and probability of first recall curves (i.e., the unnormalized proportion of trials on which item x was recalled *first*).

The reviewer's suggestion to "normalize" these curves based on item availability is *also* reasonable, and reflects a similar intuition to how the lag-CRP curves are computed. It also follows Farrell (2010)'s suggested approach to examining output position (although there is some debate about whether whether this is appropriate; e.g., see Moran and Goshen-Gottstein, 2013, but also Farrell, 2014's response).

In general, we see pros and cons to each approach, and there is precedent in the literature for both approaches. We appreciate Moran and Goshen-Gottstein (2013)'s argument that the "unnormalized" variant of this analysis might provide a "purer" estimate, and also Farrell (2014)'s simulation-based argument showing that the unnormalized curves may underestimate the "true" magnitude of recency effects.

Our view is that which is preferable comes down to what one is trying to learn from these curves. For example, as the reviewer implies, participants rarely make repeated recalls. So given that the participant had previously recalled an item (prior to their nth recall), we would expect the normalized curve to provide a better estimate of something akin to "the conditional probability that the participant would next recall item x, given that it is their nth recall." On the other hand, this approach implicitly "assumes" that participants will not repeat recalls. Although repetitions are rare, they do occur on occasion. The "unnormalized" approach to computing these curves reflects something more analogous to traditional serial position curves– i.e., the "raw" probability that a given item, x, will be recalled at output position n. We concede that this simpler variant might "mask" some underlying mechanism or process (e.g., as argued by Ferrell 2014), but we see that detangling as a job for *models* to explain, as opposed to something we can address in a theory-free way here.

Given the above, we ultimately decided to stick with our original "unnormalized" framing of the probability of nth recall analyses. We did add a brief note on these issues to our revised manuscript (pages 24–25) to help justify and explain our approach.

Page 23: For the lag-CRP analysis, how did you determine what lags are "possible"? Are lags considered not possible if the item at that lag has already been recalled, or did you only account for whether or not there was an item presented at that position? You mention that sucessive repetitions are excluded, but for this type of analysis, typically all repetitions are excluded. This allows repetitions to be excluded as "possible" lags. In other words, the lag-CRP is generally conditional on the next recall not being a repeat of any previous recall, but here I'm not sure if that was the case. If the analysis wasn't conditional on the recall not being a repeat, I don't immediately see an issue with that, but it should be noted given the difference from many previous implementations of lag-CRP analysis. I think it will result in overall lower conditional probabilities in the lag-CRP curve, given that more transitions will be considered "possible". If that is indeed the case, this difference should be noted to aid comparison to prior work.

The reviewer is correct; in the standard lag-CRP computation, *all* repetitions are typically excluded, as opposed to solely *successive* repetitions. This was a typo in our prior draft; we have corrected the methods accordingly (page 25). Thank you for catching this!

Page 25: Effect sizes would be very useful in general, and especially here to back up your statement that clustering was "substantially" greater in the feature rich condition. The mean clustering scores would also be useful to see in the text or in a plot.

We have added effect sizes (Cohen's d) to all reported statistical results, along with bootstrap-estimated 95% confidence intervals. We have also added annotations to each "clustering" bar graph to report the mean clustering score along each feature dimension, in each experimental condition (Figs. 4, S5, S6).

Page 27: "Given our above findings that (a) participants tended to remember more words and exhibit stronger clustering effects on feature rich (versus reduced) lists..." - I thought you found that there was no difference in recall for feature rich and reduced lists (page 25)?

Another good catch. We found no difference in overall recall (aggregating across all lists) for the feature rich vs. reduced conditions. We have updated the text accordingly (pages 31–32).

Ppage 41: "We found that participants who exhibited larger carryover in feature clustering (i.e., continued to show strong feature clustering on late lists) for the semantic order manipulations (but not other manipulations) also tended to show a larger improvement in recall...on late lists, relative to early lists." - Should this be a smaller decrease in recall, rather than a larger improvement in recall? Figure S3 shows a substantial decrease in recall probability for late lists, at least in the category condition.

The reviewer is correct; we have updated the text accordingly (p. 49).

Violin plots of clustering should be cut off between 0 and 1. Currently, the distributions go beyond the allowable bounds of the measure, presumably due to kernel smoothing, and this may be confusing to readers.

This is due to kernel smoothing, as the reviewer notes. In our revised manuscript we have replaced violin plots with bar plots, which we think are overall easier to read.

Figure S4: Diverging color scales (i.e., scales tha are designed to have a clear midpoint) should be used for all of these heatmaps. The color mapping should then be set so that the midpoint is at zero (e.g., -1 to 1 correlation for the top two rows). The RdBu map in matplotlib is one good choice for this. Currently, it's hard to tell especially for the difference in correlation whether the difference is positive or negative. This also applies to the difference heatmaps in Figure S7.

We have modified the "difference" plots in Figures S4 and S7 to use the "vlag" colormap (a diverging palette similar to the recommended RdBu colormap), and we have also centered the color scale on 0 so that red shading denotes positive values and blue shading denotes negative values.

Figure S8: The legend is pretty hard to read. Please make the text labels larger.

We have increased the font size for the Figure S8 legend.

Optional suggestions

The finding of increased clustering in the feature-rich condition is surprising and interesting; I would like to see a figure showing that effect in detail, and a figure would help highlight that finding.

We agree! We unpack this finding in more detail in Figures S1 (middle column) and S5.

Page 11: Rather than RGB, a perceptually more uniform space like CIECAM02-UCS may be better for selecting colors and estimating their distances. Of course, a high degree of accuracy is not necessary in this setting, so RGB is probably sufficient. But it might still be worth considering using a more perceptually informed color space in future work. In particular, I'd recommend choosing a system where hue can be modulated independently from saturation and value, and only varying hue, to avoid having very light colors like "CUP" in Figure 1, which can barely be seen on a white background.

Agreed. We also noticed this issue after we had already started the data collection process. We decided, as the reviewer also notes, that the issue was not sufficiently critical that it would be necessary to halt data collection, re-work the color selection criteria, and scrap our previously collected data. In practice, from piloting the experiment many times ourselves, we did find that "hard to see" colors occurred quite rarely. That said, for future work we agree that an alternative color space (e.g., CIECAM02-UCS, as suggested) would be better.

Page 15: The picture shown in Figure 2 makes sense, but I wonder if it wouldn't be simpler to show a bar graph of selection probability for a set of sample items, rather than illustrating laying out "sticks" and then selecting a point along the sticks. In words, you randomly select an item, with the probability of selecting a given item set to the normalized similarity of that item. Optionally, you could show a bar graph of distance and a bar graph of normalized similarity/selection probability for the same set of items.

We appreciate the reviewer's suggestion. We chose this particular figure format to most closely match our actual selection procedure. It also has the benefit of being particularly visually compact (e.g., relative to a bar plot), which we prefer aesthetically.

Figure 3: I suggest plotting presentation position with 1 indexing to match prior work.

We have updated all plots to use 1-indexing rather than 0-indexing.

Figure 3: It looks like the sawtooth pattern that is clear in the early lists in the category condition persists to some extent in the late lists. A block structure isn't apparent in recall initiation, but it does seem like there might be an influence on later recalls, even though the category block structure is no longer present on the late lists. This might be worth discussing as a post-hoc test, if you can quantify the effect and it seems reliable.

We agree that there's something interesting going on for the late lists in the category order manipulation condition. Informally, what seems to be happening is that participants in that condition learn to "chunk" their recalls into groups of roughly 4 items on the early lists. Even though the category ordering goes away on the late lists, participants still seem to carry over that same 4-item

chunking strategy on late lists. We attempted a few different ways of quantifying the effect, but we were unable to come up with a simple approach that yielded a clear/reliable interpretation. Given that this was somewhat of a tangent from the main thrust of our already complicated paper, we decided to leave this analysis to future work rather than to dig into it further. We did publish all of our data, however, so the chunking behavior is "hiding" there for someone else to explore if they wish!

Figure S7: I suggest making all of these plots be squares for consistency.

Done.

Reviewer #3: Manning et al. present a novel study which leverages the free recall paradigm to reveal ways in which manipulations of visual, lexicographical, semantic and temporal features influence memory for current and subsequent lists. The experiment design serves as a bridge between list-learning studies meant to minimize and control relationships between stimuli, with more naturalistic studies using more dynamic and correlated stimuli. They use established and novel analysis approaches to characterize mean performance and individual variability. They found that manipulating item features, even when irrelevant to the memory task, influenced recall within a list and could influence the manipulations of feature dimensions in subsequent lists. At the same time, sensitivity to the feature manipulation varied across participants. Their research design also involved optimizing study order during the experiment itself, leading to improved memory and temporal clustering for those participants.

Overall I think this manuscript provides an interesting set of results which elucidates organizational factors in free recall and has the potential to inform memory for more naturalistic stimuli. As the manuscript currently stands, however, I am less clear how it fulfills the scope of Psychological Review to "make important theoretical contributions to any area of scientific psychology". This is not because I am confident that the manuscript could not do so, but rather because I think the authors need to make explicit how their study develops or tests psychology theories on any of the major topics that they raise: episodic memory, event segmentation, situation models/schema. It may also be appropriate to relate the present study to the literature on prospective/future episodic thinking. Although the authors introduce some theories in the introduction, I think the authors need to be more explicit about how they aim to bridge and develop theories in the introduction, use this to motivate the results, and finally to discuss the theoretical importance of their results. In the results, I would suggest that the authors reframe their analyses, especially the half dozen or so analyses motivated by "we wondered", to stand on firmer theoretical grounding. In the Discussion, the "Theoretical Implications" section explains primarily the empirical importance of how the present paradigm is meant "to help bridge" studies with more naturalistic versus more simplistic stimuli. This section could expand greatly on the theoretical importance, including what the results mean for theories in each type of paradigm.

We appreciate the reviewer's comments. The core points here, regarding the theoretical implications of our work and connections with the relevant literature, were raised by both other reviewers as well. We have addressed this issue in three (main) ways. First, we have unpacked some of the core intuitions behind our study design in the introduction (page 5) to help clarify the overarching logic. Second, we have added additional commentary and citations of relevant literature throughout our

results section (pages 27 – 53) to tie our specific results to relevant findings in the literature. Third, we have substantially expanded our discussion of the theoretical implications of our work, with a focus on clarifying connections between the questions raised in the introduction, our present study's findings, and the relevant literature (page 57–59, 61–62, and 65–66).

I also think a bit more work has to be done to unpack the interactions between the clustering scores.
a. For feature-sorted lists, if the lists are constructed so that items from a specific feature dimension (e.g. onscreen location) are more likely to be presented in close temporal proximity, wouldn't feature clustering be correlated with temporal clustering? For instance, perhaps a participant generally exhibits strong temporal clustering and ignores screen location. For lists in which words studied in nearby screen location are more likely to be neighbors, then wouldn't the participant exhibit greater spatial clustering simply due to the temporal proximity of items with greater spatial similarity?

This point was also raised by both other reviewers. In brief, if a list's items are sorted by a given feature, then clustering along that feature dimension will be positively correlated with temporal clustering. The critical question is whether (and under which circumstances) participants exhibit feature clustering *over and above* what would be expected by temporal clustering alone, on a feature-sorted list. To examine this issue, we designed a new correction procedure to disentangle the temporal and feature clustering. For convenience, here is our description of the new analysis, taken from our response to Reviewer 1:

> ***Following the reviewer's comment, we have developed an additional correction procedure (described on pages 23–24) to specifically distinguish temporal versus feature clustering. For a given set of recalled items (whose presentation positions are given by x1, x2, x3, …, xN), we can circularly shift the presentation positions by a randomly chosen amount (between 1 and the list length) to obtain a new set of items. Since the new set of items will have the same (average) temporal distances between successive recalls, the temporal clustering score for the new set of items is equal (on average) to the temporal clustering score for the original recalls. However, we can then re-compute the feature clustering score for those new items. Finally, we can compute a "temporally corrected" feature clustering score by computing the average percentile rank of the observed (raw) feature clustering score within the distribution of circularly shifted feature clustering scores. This new temporally corrected score provides an estimate of the observed degree of feature clustering over and above what could be accounted for by temporal clustering alone.***

In summary, it is not always possible to fully detangle temporal or serial position effects from feature clustering effects on the order manipulated lists, since feature values and temporal positions are often strongly correlated on those lists. However, we do see at least *some* evidence that participants are truly exhibiting feature clustering over and above what would be expected by temporal clustering (combined with order manipulations, primacy, and/or recency alone). For example, participants show reliable semantic (category and size) clustering even after factoring out the potential effects of temporal order and serial position (Fig. 5D).

b. It would be ideal if the authors could relate their findings to prior research examining the tension between different types of clustering in free recall (such as temporal, semantic and spatial), and their relation to recall probability (e.g. Howard and Kahana, 2002b; Healey et al., 2019; Unsworth, 2009; Sederberg et al., 2010).

We have substantially expanded the scope of our discussion section in our revised manuscript. We have added some discussion specifically about the tension between different types of clustering, and how clustering and recall performance are related, including citations of the suggested papers (pages 65–66).

On a more minor note, I think much of the information presented in the last two paragraphs of the Participants section would be easier to compare and contrast if it were in table format.

We struggled with how to present this information succinctly. The general challenge is that each aspect of participants' demographics we report has its own set of values. For example, ages are reported as integers ranging from 17 to 31. Genders included male, female, and not reported. Ethnicity and race each have several response types as well, and so on. Converting this into table format would require a separate table for each demographic dimension (11 separate tables in all). We thought the simplest solution would be to summarize the demographics in paragraph form. We also provide (as part of the associated GitHub archive containing our code and data) a Jupyter notebook that summarizes each demographic area, using a series of tables. We also share (as an Excel spreadsheet) the complete table of all participant demographic responses.