Dear reviewers,

We would first like to express gratitude for your thoughtful consideration in providing feedback on this manuscript. The reviews provided have allowed us to carefully consider nuances in this work that may have initially been overlooked, and we are extremely appreciative of this opportunity to implement these changes. We have carefully reviewed each comment and are confident that we have addressed each one appropriately. We appreciate the opportunity to submit a new version of the revised manuscript for your review. On the following page, we have included each comment provided by the reviewers with a small explanation as to how we have addressed the issue in the revised manuscript. We look forward to hearing from you in due course.

All the best,

Michaela Ritchie

Dr. Loaiza-Kois - Comments	
The first effect is defined more concretely in terms of the concrete effect on performance whereas this is a bit more vague (p.2)	This has been reworded in the abstract: "The proposed study will investigate whether the generation effect, a memory advantage for self-generated verbal information, is enhanced in multisensory conditions. Such a finding would be consistent with the multisensory facilitation effect, a phenomenon wherein multiple sensory inputs may reduce the cognitive load required to process and respond to co-occurring stimuli from multiple senses"
I would say more on this. Explain one of these studies. At the moment it's not clear (p.14)	A brief summary of McCurdy et al.'s (2020) meta-analytic findings are now included.
I am not catching this distinction. It needs to be explained better or another example is needed. I am also not sure why this point is important to the rationale at this stage either. (p.16)	The structural/organizational changes to this version of the manuscript should make this more clear. For example, p.8-9 make a contribution to clarifying why it is important to consider. Redundant multisensory information is a component of overt generation tasks, and so it is important to consider whether it contributes to the magnitude of the effect. "Although congruent stimuli from multiple senses may allow for more efficient processing of information, redundant multisensory information is also useful (MacLeod, 2010; Wallace, 2009). When the same information is presented to multiple senses at one time (e.g., the visual presentation of a word accompanied by its auditory articulation), the presence of the information from multiple sources increases the probability that the stimuli will be recognized and responded to. Redundant multisensory information may serve to enhance the distinctiveness of the target stimuli from either sense, a process that has been thought to underlie the generation effect (Kinoshita, 1989). Therefore, the simultaneous presentation of cue and target stimuli in a typical generation task may

	promote its distinctiveness for later recall (Kinoshita, 1989)."
I do not understand the difference between these hypotheses. If one just adds "of target items" to H1b, it seems like it is simply just H1 as explained previously. I do not see the point of making a distinction like this. I would start with this hypothesis first. So: H1: There will be a generation effect overall in target recognition and confidence ratings (i.e., generate > read conditions). H2: This generation effect predicted in H1 will be greater in multisensory conditions compared to unisensory conditions, such that there will be higher recognition of target items when they were generated in the multisensory (audiovisual) condition compared to the unisensory (auditory or visual only) condition. This seems like a simpler setup and then there is no need for the a/b thing. (p.17)	The hypotheses have been restructured to read as follows: Hypothesis 1 (H1): We predict a generation effect, with participants performing better in the generate condition compared to targets that were read, as reflected by better cued recall performance for target items that were generated. Hypothesis 2 (H2): The generation effect hypothesized in H1 will be more pronounced in multisensory conditions than in unisensory conditions. We expect greater recall of target items when they are generated in the multisensory condition compared to when they are generated in the auditory or visual only condition.
What is the effect size, specifically? Is this for the predicted overall main effect of task type or the interaction? (p.18)	The specific effect size used for our power calculation is now included (d = 0.37). This is the main effect size anticipated for task (generate vs. read), as we anticipate at least a medium effect, if not a larger effect size for the interaction between task type and sensory modality. Observed power will also be collected when the data are analyzed.
Now that I have read the whole study, I do not know how the task could be "incorrectly completed." I think some concrete examples of what is here meant would be helpful/more explicit. (p.18)	This is now clarified with examples of incorrect task completion on pages 13-14.
Who are the participants? Are they students taking part for course credit or something else? Is there any age restriction or similar? (p.18)	These details are now included under the heading "Participants". Participants will be recruited from the researchers' university population o fundergraduate students enrolled in a Psychology course. In compensation for their participation, they will be awarded one bonus credit toward their Psychology course. Participants must be of at least 19 years of age, and will be asked to report their age, gender, and first language.

Are these taken from previous research to establish their appropriateness for use? For example, is the synonym strength sufficient to allow people to generate it within the time allowed for study? I imagine that such norms already exist given how replicable the generation effect is, so it seems prudent to just use what has already been done.	Unfortunately, despite the pervasive use of word pairs in this area of research, our attempts to access existing materials have been unsuccessful. We have reached out to authors and searched for open source resources and believe that we have made every reasonable attempt to access established materials with no luck. We are open to the idea of pilot testing our word pairs on a different set of participants if this is a suitable solution for the reviewers.
The way this was written made it seem like the generate/read conditions were manipulated between-subjects, so I added "first" to make it clearer in each case	Thank you!
I do not understand why this task is not completed on a computer. There are a LOT of things that could influence the task besides these other risks, such as an inconsistent tone of voice or inconsistent timing/pacing, merely due to human error. If the issue is that there are restrictions around access to software or programming, I highly recommend lab.js, which is free and also includes an already programmed paired-associates learning task that could be easily adapted for this.	We have opted to program this study on PsychoPy, a similar software for conducting psychological studies, as we are more familiar with its functions. Our original intent was to use flash cards and a pen-and-paper approach because we are familiar with the materials used for this approach, as we have just completed data collection for another study which used similar materials. This comment allowed us to re-evaluate our plans for this research—thank you very much!
I can see why it would be more practical to manipulate sensory condition between-subjects in this flashcard method, but if you do the task on the computer, then this would not be necessary. You could do all of it within-subjects and block the conditions so that they are counterbalanced with the order of read/generate at the top level, and then each of the 3 sensory conditions counterbalanced within those tasks. That would give you 12 possible counterbalances (which 72 is divisible by), such that the permutation of the bottom 3 sensory conditions (6 possible) is duplicated for the read/generate conditions.	Also addressed by using PsychoPy, thank you!
This can also be programmed in lab.js, or another task could be used that is already programmed in their suite of options since the	Also addressed by using PsychoPy.

nature of the distraction likely does not matter nor is the task important to the design/results	
Once again, this could all be programmed in lab.js or one of their tasks could be adapted. Another issue I see is that I do not understand how the lure options will be created to be plausible alternatives to the target item. For example, in the case of chilly-cold, are there lures of sufficient synonym strength as "cold"? How is this determined (e.g., has this established done in prior work)? If they are not all related lures, then participants could simply choose the target that is related which is not really the point. Overall, it seems like it would be easier to simply give a cued recall test -- isn't this what is typically done in this literature? What is the advantage of this 3AFC?	Although free recall, recognition, and cued recall tests are all commonly used in the literature surrounding the generation effect, we recognize that the lure options will not have been of sufficient semantic strength to be a feasible choice for many of the target words. In consideration of your thoughtful feedback, we have opted for a cued recall test in our design, which is now reflected in the method section of this manuscript.
This language is quite a bit different to the language used in the introduction. I interpreted the rationale as multisensory facilitating the effect, not accounting for it. This brings up a deeper point, too: If there is a generation effect no matter what but it's just stronger in the audiovisual condition, does that really account for the generation effect? It seems like it could "account" for it only if the generation effect was significant in the audiovisual condition and not the other conditions. Otherwise it is simply that the multisensory presentation enhances the effect.	It is our hope that some changes in wording have clarified this ambiguity. Specifically, we anticipate an interaction between task type (generate, read) and sensory modality of encoding (visual, auditory, multisensory). The procedure and interpretation of an ordinal interaction is now clearly defined in our proposed analysis and we are grateful for your guidance toward relevant literature on the approach.
I would have liked to see more detail in exactly what pattern of results would qualify as support for/against each hypothesis. From earlier on, I surmise the following: 1. You will conduct a 2 (task) x 3 (sensory modality) ANOVA and expect an overall main effect of task type, such that generate condition shows greater target recall and confidence ratings compared to the read condition, in line with H1 and consistent with prior literature. (So you don't need a separate	The hypotheses, proposed analysis, and appended table have all been revised with the aim of making the expected results, and how they would align with (or disconfirm) the outlined hypotheses, more transparent. Additionally, we are appreciative of your guidance toward relevant literature in interpreting ordinal interactions and have implemented them in our Proposed Analysis section.

t-test; this will already be evident in the main effect). 2. Furthermore, this generation effect should be qualified by an interaction, such that the effect is larger in the audiovisual condition compared to the auditory and visual conditions. That is, the generate condition should show greater target recall the audiovisual conditions compared to the other sensory modalities. This would be consistent with H2 that the generation effect is more pronounced in multisensory conditions. Please note that H2 is an ordinal interaction. This means that the expected result is that there is a generation effect for each sensory modality, but it's larger for one than the other. This is problematic for unambiguous interpretation (see Loftus, 1978; Wagenmakers et al., 2012). Thus, you should also transform the data and repeat the analysis to verify that the interaction still occurs even when the data are transformed to a different scale, thus reducing the chance that the ordinal interaction is due to a mere artifact of scale. See Labaronne et al. (2023) Journal of Cognition for an example of how this was done in another context (and it's also a registered report)	
This does not seem relevant given that you have done a power analysis, and Bayesian inference does not absolve issues of power either.	This has been removed and the observed power for our analyses will be reported once the actual data are obtained.
Overall an interesting proposed study, and I also like the idea of this table! I think that it may need updating based on the comments I've suggested, though.	The table has also been updated!
Dr. Sharon Bertsch - Comments	
On page 14, the authors discuss the difference between redundancy and congruency. I'm not sure I understand the difference as it's explained. It appears that the current experimental design uses redundant (same words read/written and spoken), not congruent (semantically similar words read and spoken), stimuli. This should be clarified,	This issue has been clarified throughout the manuscript. Additionally, the experimental goal has been clarified. Regardless of the stimuli being redundant, it is imperative to understand why its concurrent presentation through an auditory and visual medium yields a larger generation effect than covert generation tasks. We hope that the

as the stated experimental goal is to test congruent stimuli.	experimental goal is more clearly stated in this version of the manuscript.
H1a "Multisensory engagement will enhance recognition of target items during generation tasks." Perhaps add 'compared to targets that are read' H2 refers to confidence ratings, but these aren't reviewed anywhere in the introduction. In H2b , why should confidence increase under audiovisual conditions compared to visual or audio alone?	Per the guidance of our other reviewer, we have refined our hypotheses so that our experimental goals are more transparent. Additionally, we have opted for a cued-recall test without the confidence ratings, and thus this discrepancy in our earlier draft is no longer relevant.
Where will the word pairs come from? Will the lures used in the recognition test be matched to the target in terms of word frequency? I doubt the Kucera & Francis (1967) norms are current, but there should be something relevant to today's undergraduate students.	We feel that we have made every reasonable attempt to access the relevant materials from earlier research. Beyond a thorough search of the literature and supplementary materials, we have also contacted multiple authors who have conducted work using similar materials, with no response. For this reason, we have opted to develop and use our own list, and again anticipate that performance issues would be consistent across sensory conditions and therefore not affect the analysis. In additional attempt to navigate the issue of using new materials in the proposed study, we aim to include only participants who have successfully generated all anticipated target words. With this said, any resources that you may have access to or know of would be greatly appreciated.
Perhaps the audio condition should play words pre-recorded by the researcher to improve standardization.	This consideration has now been implemented in our Method section.
I don't know enough about auditory processing to speculate, but in the auditory generation test condition, could the fact that the first letter of the target word spoken by the researcher as a cue might match the sound of the target first phoneme differently? E.g., if the target is 'chilly' would saying the letter 'c' as an auditory cue create error variance compared to a target word of 'cat'?	We appreciate this thought, and also recognize that this concern may not be solely tied to the auditory condition, given that the visual perception of certain letters may also promote the processing of certain sounding words over others. With that said, we are confident that the semantic constraints provided by task will outweigh this concern. To extend the example given, if the cue word "chilly" is given, as: "chilly-c", the phonological sound of the letter c, may

	promote the thought of words that being with a soft-sounding c (e.g., cinnamon). However, the semantic constraint of the task requires that participants think of a response that is a synonym of the cue word given, which we expect would reduce the likelihood that a participant would be inclined to respond with the incorrect target word. Contrarily, a participant might see the word pair "chilly-c_____" in the visual processing condition, and be more inclined to think of words that begin with a hard "C" instead of considering a word that begins with C but produces a "ch" sound or even a soft "c" sound. In this example too, we are confident that the semantic constraint provided by the task rule (to generate a synonym to the cue word given) will restrict responses to relevant options, and thus rule out the possibility that the phonological (or visual) processing of a letter will influence responses. Beyond all of this, we also recognize that because any concern of this issue would not be unique to the auditory condition, we should have no reason to expect systematic differences in performance across sensory conditions attributable to this.
In the Procedure subsection, assuming the within subject design described, will the read/generate items be displayed in blocked form or randomized form? It seems to matter. Additionally, the design is described in this section is as Task type manipulated within subjects, but in the Proposed Analysis section is described as between subjects. This matters also in terms of the sample size needed, as the effect sizes for these two types of manipulations are very different.	Although not included in the original draft, the procedure will follow a blocked design, which is now transparently reported in the procedure section of this draft. In the Proposed Analysis section, there was an oversight which led to a typographical error that task type would measured between subjects. This is not consistent with our plans, as task type is intended to be measured within subjects, and this factor was recorded as such in our power analysis as well.
Please also plan to include the effect sizes that you find in your analyses, and of course, report the power for any non-statistically significant findings.	These will be reported in the final manuscript once data have been collected!

In addition, looking forward to a full manuscript submission down the road, the literature review was much too heavy on the background and multiple theoretical explanations (and problems) of multisensory facilitation, and too brief on the ins and outs of the generation effect.	We feel that we have adequately addressed this imbalance in this version of the manuscript.