Response to Reviewers

Editor:

- *One point made by both reviewers is that you need to more clearly motivate your hypotheses. I agree that you can do more to explicitly spell out the links to theory, particularly for the predicted differences between different types of bilinguals.*

Thank you for this feedback. In this revision, we created stronger links between existing theories and our study. In this context, we clarified how our study can inform several hypotheses that predict a bilingual word learning advantage (for example, the ones explained in Bogulski et al., 2019; see pages 6 and 7). We now argue that our main goal is to test the predictions of what we term the learning adaptation hypothesis, i.e. that bilinguals adapt their mutual exclusivity bias to facilitate word learning; such changes in learning behaviour should specifically benefit when acquiring more complex mappings (Poepsel & Weiss, 2016). We also clarified how trial-by-trial analyses are relevant for the current experiment and we describe them in more detail and with more connection to the past literature (see pages 9 and 10). As described in more detail later, we no longer plan to recruit two separate groups of bilinguals. Instead, we plan to calculate language entropy to measure participants' language balancedness and motivate that decision in the text.

- *Sample plan*

  In the text you tell us that you that you planed on effect size "based on the effect size reported by Poepsel and Weiss (2016) and Escudero et al., (2016)". As the reviewers point out, this is underspecified. Is this an effect size for the main effect of language group, or language group by trial type interaction? What actually is this effects size? What analysis did you do with this value?

  In addition, you currently do not have separate justifications of the sample required for each of your hypotheses, but this is a requirement for being accepted as a Stage 1 RR. That is: every row in the big table at the end has to have its own sample size justification. (If there are places where you really aren't able to do that, you can still do the analysis as an exploratory analysis, but they won't be part of your pre-registered analysis plan and this means they can't feature in the abstract and if you talk about them in the discussion they can't be the main focus.) [Note: we were provided with additional help on how to do this; these comments are left out for brevity]

Based on your feedback, we revised our power analyses. As suggested, we completed five separate power analyses, one for each of our hypotheses (see summary in Table 2; page 18 but also reproduced below). We used our pilot experiment for the power analyses of four hypotheses (H1, H2, H4 and H5). However, since we did not have a significant result in our pilot regarding the comparison between monolinguals and bilinguals (H3 and H4), we used Poepsel and Weiss' (2016) data to calculate the power needed for H3. For H4, this was not possible because Poepsel and Weiss (2016) did not assess learning continuously. Overall, we found that we need a minimum of 150 participants to have a power of .8 for all hypotheses but H4. Since we are able (money and time wise) to collect up to 200 participants, we decided to use the so-called random stopping practice (Rouder,

2015) to calculate Bayesian factors in intervals of ten participants periodically. Therefore, we propose to collect between 150 and 200 participants and decide when to stop in this window based on the Bayesian factor for H4 (see pg. 18 for more details).

Hypothesis	Model	Effect of interest	Data	Participants needed for an effect size of >.8
H1: Is it easier t to learn simple (1:1, one word maps onto one object) or complex (1:2, one word maps onto two objects) mappings?	DV= Accuracy; IV= Block, Mapping Type and Language Group	Main effect of mapping type	Pilot data	50 participants (.999)
H2: Do bilinguals learn words more easily than monolinguals under some circumstances?	DV= Accuracy; IV= Test, Mapping Type and Language Group	Interaction between Mapping Type and Language Group	Poepsel and Weiss (2016)	150 participants (0.842)
H3: Does accuracy on a previous trial and target count (how often a word has been encountered) impact accuracy on a current trial?	DV= Accuracy; IV= Mapping Type, Language Group, Last-target-accuracy and target count	Main effect of Last-target-accuracy	Pilot data	50 participants (1.000)
		Main effect of Target count		50 participants (1.000)
		Interaction between Last-target-accuracy and Target count		50 participants (.894)
H4: Will the use of mutual exclusivity bias be reduced for bilinguals?	DV= Accuracy; IV= Mapping Type, Language Group, Last-competitor-accuracy and target count	Interaction between Last-competitor-accuracy and Language group	Pilot data	200 participants (0.584)
H5: Does the use of mutual exclusivity change between mapping types?	DV= Accuracy; IV= Mapping Type, Language Group, Last-competitor-accuracy and target count	Interaction between Last-competitor-accuracy and Mapping type	Pilot data	100 participants (.934)

- *Assessing evidence for H0*

One of the reviewers asks you to consider including a measure to assess the strength of evidence for H0/H1. I agree with this and would strongly encourage you to consider using either Bayes factors or equivalence tests (Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving inferences about null effects with Bayes factors and equivalence tests. The Journals of Gerontology: Series B, 75(1), 45-57.)

This isn't a requirement for stage 1 acceptance, however if you don't have any method to assess evidence for the null, then you cannot make statements such as the following:

> *"Thus, if we do not observe a performance difference between language groups (either as an interaction with mapping type or as a main effect), we will conclude that bilinguals are not better at learning words statistically than monolinguals.-- " [Note: we were provided with additional help on how to do this; these comments are left out for brevity]*

We decided to add the calculation of the Bayesian Factor to each of our analyses. We will calculate it as you suggested (with the same procedure as Silvey et al., 2021). For this reason, we added the null hypothesis for each hypothesis in the text (see pages 26 and 28).

- *Random structures: is the mapping of labels and words randomized across participants? If not, I think you might need a nested random effects structure for these? (If it is randomized and I missed this, maybe remind the remind the reader of this when you talk about the random effects)*

We created five different lists of possible combinations of mappings of words and objects. Each participant will be randomly assigned to one of the five lists. Therefore, as you suggested, we added "list" as a nested random structure (within which labels and words are nested; see page 26).

- *Target count (H4) as a fixed effect: you say this will be log transformed, will it also be centered around 0 as you do for your other fixed effects?*

Yes, it will be centered. All the variables will be centered, which is now explicit throughout the entire manuscript (see, for example, page 28).

- *"We will exclude any participants with a linear slope of the accuracy of −1 SD deviation from average (slope will be calculated across bins of 36 trials; c.f., Roembke et al., 2018)." this needs justification beyond a reference to another paper*

We decided to remove this sentence. As you pointed out, we did not have enough theoretical explanation for this decision. We will, instead, exclude participants who do not reach 40% accuracy in block 5 from the trial-by-trial analyses only. We decided to ensure we had enough usable trials for these analyses (see page 29).

- *I agree with the reviewer that it is odd to use terminology of "balanced" and "unbalanced" bilinguals where your criteria seem to be more about onset of acquisition than about proficiency*

Thank you for this helpful feedback. In this revision, we modified how we capture differences in proficiency across bilinguals. As you and the two reviewers suggested, our previous idea of using the age of acquisition to divide balanced and unbalanced bilinguals was not optimal. We therefore decided to have just one group of bilinguals in which we continuously measure language balancedness. To do so, we will compute language entropy (as done by Gullifer and Titone, 2020). We decided to use this measure for several reasons: it has been found to relate to different measures of bilingualism (Gullifer & Titone, 2020), it is a continuous measure, it is used in more and more papers (e.g., DeLuca et al., 2020; Kang et al., 2023; Salig et al., 2023), and it is possible to compute it from a slightly modified version of the language questionnaire LEAP-Q (Kaushanskaya et al., 2020). In the main analysis, we will just compare the monolingual and the bilingual groups. We will then do an exploratory analysis considering language entropy as a covariate. We hope this solution will address your and the reviewers' concerns about the previously suggested division between bilinguals while allowing us to explore whether language history affects cross-situational language performance.

Reviewer 1

- *My main comment concerns the question of whether balanced bilinguals would differ from unbalanced bilinguals, which seems to be a main question of interest for this RR, but is not at all motivated in the manuscript. What is the reason they would differ? Further, the definition of a balanced bilingual is not what I would expect – I expected a balanced bilingual to be one who is reasonably equally proficient in both languages. The authors seem to define balance based on when exposure to both languages occurred, but use a very early cut-off for that. For example, the authors state "balanced bilinguals learned both languages simultaneously, whilst unbalanced bilinguals learned English as their mother tongue and German later in life". Is this a criterion that will be imposed on participant groups (e.g. will participants who do not meet these criteria be excluded before/after data collection), or are the authors expecting this to be true? I can imagine many circumstances where simultaneously learning two languages could result in unbalanced bilingualism and vice versa. Further, the definition for unbalanced says "they will only be included if their LexTALE score exceeds 70% for English and 50% for German" but this means that they could both be equal and exceed 70% so then they would be balanced in proficiency but be sequential bilinguals. I would encourage the authors to motivate this question further, and then to clarify the definition of each group – if it's about balance in proficiency then is it fair to impose an Age of Acquisition cut-off? Maybe renaming the groups would also be helpful.*

Thank you for your thoughtful feedback. We made two changes to address these concerns: First, we better motivated in the text why we are interested in language balancedness as a potential moderator variable of the bilingual word learning advantage (see pages 7 and 14). More specifically, we describe that several existing hypotheses (e.g., Bogulski et al., 2019) predict that a word learning advantage should be higher if participants are more balanced.

Second, we agree with you that the previous plan to separate bilinguals into two groups based on age of acquisition was not optimal (and may have led to the exclusion of many participants). Instead, we now propose to calculate language entropy for bilingual participants (as done by Gullifer and Titone, 2020). We decided to use this measure for several reasons: it has been found to relate to different measures of bilingualism (Gullifer & Titone, 2020), it is a continuous measure, it is used in more and more papers (e.g., DeLuca et al., 2020; Kang et al., 2023; Salig et al., 2023), and it is possible to compute it from a slightly modified version of the language questionnaire LEAP-Q (Kaushanskaya et al., 2020). In the main analysis, we will just compare the monolingual and the bilingual groups. We will then do an exploratory analysis considering language entropy as a covariate. We hope this solution will address your concerns about the previously suggested division between bilinguals while allowing us to explore whether language history affects cross-situational language performance. (We will address the point you made about the LexTALE cut-off points in a response below.)

- *The power analysis doesn't really seem like a power analysis? What does it mean that "based on the effect size reported by Poepsel and Weiss (2016) and Escudero et al., (2016), we concluded that a sample size of 50 should be sufficient to detect the effect of the language group if it exists". What are the effect sizes reported in those studies, and what power does a sample size of 50 allow you to have to detect that effect size?*

Based on the feedback we received, we revised the power analyses significantly. Please see our detailed response to the editor at the beginning of this document (and specifically Table 2). (We deleted the specific sentence you quote.)

- *Why is only one word presented on each trial? I appreciate using orthographic instances so as to not cue a specific language, but in the Escudero et al., 2023 study I believe they presented the same number of words as pictures on the screen. The proposed methods are a deviation from more traditional statistical learning paradigms, and change the statistics that can be computed on any given trial. This seems like it could have meaningful differences in outcome. I can see how showing three pictures and providing words on different consecutive trials would sequentially narrow the choices participants could make (e.g. if they saw objects A, B, and C, and heard nonword1 and clicked on A, then the next time they hear nonword2 they would most likely to say it mapped to object B or C, etc.), but could all three words be displayed at the bottom and dragged to the image they think it labels? If not, this is another limitation of the comparison of this study to previous ones if there end up being differences in findings.*

As you correctly pointed out, there is a methodological difference between Escudero et al. (2023) and this manuscript. The reason why it is essential that participants are only presented with one word per trial is that it allows us to perform trial-by-trial analyses that are consistent with previous literature (this information is now present in the text as a footnote on page 26). As you suggested, we also highlight this difference between this paper and the previous literature on CSSL with bilinguals on page 15. While, to our knowledge, this study would be the first to use this specific version of the CSSL paradigm when comparing monolinguals and bilinguals, it has been used repeatedly when studying CSSL in monolinguals (e.g., Roembke & McMurray, 2016; Yurovsky & Frank, 2015). Moreover, it should be highlighted that there is not just one "true" version of the CSSL paradigm (c.f., Roembke et al., 2023); instead, studies often differ in the exact parameters they select (e.g., number of words, how nonwords are created, how many objects are presented on the screen, whether objects are novel or not, whether participants are entirely naïve to the presence of co-occurrence statistics or not)—our goal here was to select the specific parameters that can best answer our scientific question. There is no a priori reason to believe that changing the exact statistics on each trial should change whether a bilingual word learning advantage can be observed or not.

- *Main analyses – it doesn't seem like the main analysis accounts for repeated measures on each item, right? Yes, there is a block main effect, but if I understood the methods correctly each word will be the target in multiple trials within a single block. So yes, accuracy should increase across blocks, but correct accuracy on the first trial is based entirely on chance, and on later trials can be due to statistical learning. I understand that the trial–by-trial analyses will account for some of this, but it seems like maybe the main analysis needs to account for trial as well? Or calculate accuracy on the last instance in the block for each word? Or an average across the block?*

You correctly understood our design; there are indeed repeated measures for each item (each word is seen six times per block). As you pointed out, we consider only the block as a fixed effect in hypotheses H1 and H2, not the trial. This is a common practice in the literature (see, for example, Roembke et al., 2016), and, as you said, we can still account for more precise changes using trial-by-trial analyses (as part of hypotheses H3-H5). Like with any experimental design, performance on any trial can be due to several factors: chance, momentary (in)attention, actual learning, or a combination of those. One advantage of linear mixed-effects models (over, for example, an ANOVA) is that they do not require we average performance before entering them for analysis; this allows us to evaluate the impact of several random effects in more detail. Across blocks, we can then capture the learning slope.

Minor comments:

- *Pg 4 – "Fourth, the bilingual cognitive control hypothesis proposes that bilinguals' advantage on monolinguals is due to bilinguals' better cognitive control on executive function tasks. Another possible explanation is that bilinguals have a higher executive control than monolinguals…" aren't the 4th and the "another possible explanation" essentially the same?*

Yes, as you pointed out, they are the same. For this reason, we deleted "another possible explanation" and connected the two sentences.

- *Pg 7 – description of the second/testing phase in a lot of these tasks – "participants' memory can be tested through different methods" is memory the key outcome here, or learning? In order to be tested on anything, even at an immediate time point, there is some memory involved, but that's different from testing memory separate from learning.*

Memory plays a role, but it is not the critical outcome here. We, therefore, exchanged the word *memory* with *learning*, which is our key outcome.

- *Pg 11 – "whilst the interaction between languages and mappings was not significant" – maybe edit languages to "language background"*

As suggested, we changed the word *languages* to *language background*.

- *Pg 12 – "opposite effects" are referred to a lot, but this is hard to keep track of because what the opposite is referring to keeps changing. Maybe mixed results would be clearer?*

"*Opposite effects*" was indeed confusing. We changed the sentence, and now it says: *but results were mixed in how exactly groups differed*, as suggested (see page 12).

- *Also, the Crespo 2023 study finding is a little more complicated than is alluded to here – maybe they are less 'hurt' by overlapping cues? It's not really that they found a bilingual advantage in 1:1 mappings overall, just under this one specific condition.*

We are sorry that our description of Crespo et al. (2023) was misleading. We clarified our description by expanding that paragraph (see page 12).

- *What are typical ranges of scores for LexTALE? Is 70% not too low of a cutoff for native monolingual English speakers?*

Based on our findings in other studies (e.g., when assessing LexTALE in student populations), LexTALE scores for participants' mother tongue range from 0.7 to 1, though averages are typically well above 0.8. Thanks to your comment, we again read Lemhöfer and Broersma (2012) carefully and decided to change the scores according to Table 9 of their paper. A score of 80% is now requested for their mother tongue (consistent with an upper intermediate/advanced user).

Reviewer 2

- *The authors are encouraged to consider including a measure to assess the strength of evidence for H0/H1.*

We decided to implement Bayesian Factors for linear mixed effect models to assess the strength of H0. We will calculate it as the editor suggested (with the same procedure as Silvey et al., 2021). For this reason, we added the null hypothesis for each hypothesis in the text (see pages 26 and 28).

- *One area where the ms. can be strengthened is in being more explicit about the links between the hypotheses and theory. For instance, why will the differences between 1:1 and 1:2 mappings increase over the course of training/exposure, or why would the differences between 1:1 and 1:2 mappings be more pronounced in balanced than unbalance bilinguals? It would be helpful to be more explicit about the possible processes that may be underpinning these differences. Additionally, it would be helpful for the authors to elaborate on the links between the specific measures used and the specific cognitive processes they are thought to index (e.g., last trial accuracy).*

Thank you very much for this comment. We address all of your points in response to your more detailed comments below.

- *It'd be helpful to provide a brief operational definition of balanced vs. unbalanced bilinguals.*

We now explicitly define what we mean by balanced and unbalanced bilinguals on pages 3 and 4. In addition, we no longer plan to recruit two separate groups of bilinguals. Instead, we plan to calculate language entropy (Gullifer & Titone, 2020) as a continuous measure of each person's language balancedness. Please see our answer to the editor at the beginning of the document for an explanation of why we decided to measure differences in language balancedness this way.

- *For hypothesis 4, please clarify the distinction between 'better cognitive control' and 'higher executive control'.*

There is no distinction between the two; we therefore deleted “another possible explanation” and connected the two sentences.

- *It'd be helpful to state somewhere here explicitly which of the hypotheses listed on p. 6 the current study focuses on (i.e. the current study focuses on mutual exclusivity, and it won't be able to distinguish between the other hypotheses).*

Following your comment, we added a paragraph on page 7 explaining what this study can say regarding the hypotheses listed. To our understanding, the first four hypotheses (listed in Bogulski, 2019) predict that the more balanced someone is between languages, the better language learners they are. For this reason, if the results of the exploratory analysis finds that more balanced bilinguals are better word learners than less balanced ones, we would interpret this as evidence in line with all hypotheses. While our analyses cannot distinguish between these hypotheses, they would lend further support to the general idea that how proficient someone is in their second language has an impact on how good of a word learner they are. The last hypothesis we consider is what we now term the learning adaptation hypothesis: bilinguals have adapted their learning assumptions to facilitate the acquisition of multiple mappings. For the learning adaptation hypothesis, it is also true that higher balancedness should be associated with better language learning, but it additionally predicts a specific advantage when what has to be learned is more complex (i.e., multiple mappings). Our paper focuses on testing this specific prediction (we now explicitly state this in the manuscript; see page 14). For this reason, participants learn two different mapping types that differ in complexity. If we find a difference between monolinguals' and bilinguals’ performance for more complex mapping types (as reported by Poepsel & Weiss, 2016), we will interpret this as evidence for the learning adaptation hypothesis.

- *It'd be helpful to provide some more information here on the cognitive mechanisms these measures are thought to index.*

We added more details regarding the trial-by-trial analyses and the cognitive mechanisms they map onto. On page 10, you will find additional information on how these measures were used in the past and how they can be directly connected to different cognitive mechanisms. For us, three indexes are particularly interesting: target count, last-target accuracy, and last-competitor accuracy. The effect of target count has been argued to be a measure of statistical or more implicit learning processes, while the effect of last-target accuracy has been seen as indicator of more explicit learning (Trueswell et al., 2013; Roembke & McMurray, 2021). Moreover, the use of mutual exclusivity can be estimated by how accurate participants are on a current trial based on whether they had selected the correct referents for the competitor objects the last time they were the target object (last-competitor accuracy; Roembke & McMurray, 2016; Roembke et al., 2018). That is, are participants able to rule out competitor objects as potential targets if they had correctly mapped them on a word before?

- *It'd be helpful here to provide broader context for the paradigm here -- why would CSSL be a useful paradigm to shed light on bilingual word learning? It would be also useful to consider the extent to which the specific paradigm that includes an explicit selection of the target object in every trial might influence the cognitive processes that underpin the learning (e.g., enhance the contribution of explicit processes, including specific metacognitive strategies), and if so the implications for models of word learning.*

We added some additional information at the end of page 10 to address your comment. We now explicitly state that CSSL is a mechanism used by adults and children and that since bilinguals are exposed to more complex contexts, it is likely that they will have to make use of implicit word learning mechanisms (like CSSL).

- There is an inconsistency between Table 1 and the text about L1/L2 for unbalanced bilinguals.

Since we no longer use two separate groups of unbalanced and balanced bilinguals, the table was redone with just one line for bilinguals (and the numbers in Table 1 for bilinguals are now correct).

- *Please clarify whether the choice of sample size (50 ppts per group) was based on a power analysis using the effect size reported by Poepsel & Wiess (2016).*

We redid our power analyses based on the editor's and reviewers' comments. Please see our detailed descriptions when responding to the editor at the beginning of this document. We now explicitly clarify in the text which data were used for which power analysis in the manuscript (see Table 2).

- *It'd be useful to provide the English/German-like ratings for the stimuli somewhere (Appendix, OSF).*

The ratings are now available in Appendix B.

- *Please provide a rationale for the choice of objects, and clarify whether they were all artifacts, and how they were sourced.*

Following this comment, we decided to use a database of objects frequently used in the field. Therefore, we chose the NOUN database (Horst & Hout, 2016). The objects were chosen to be the ones with the highest novelty score (see page 20/21).

- *Larger images/figures would be really helpful.*

We made all the images bigger.

- *It'd be useful to be more explicit about what type of strategies will be taken as evidence of 'cheating'.*

We do not plan to distinguish between different cheating strategies; we will exclude any participant who indicated that they did not try their best (Did you try to the best of your abilities to learn the words? Yes/**no**) or that they used a cheating strategy (Did you use any "cheating" tactics to learn the words, such as taking notes, during the experiment? **Yes**/no). We made this more explicit in the text on page 27.

- *It'd be helpful to state explicitly the possible reasons for the overall advantage of the 1:1 mapping -- this could be due to the inherent difficulty of learning 1:2 vs. 1:1 mappings, but it could also be due to the frequency of exposure (if I understood the design correctly, the 1:1 mappings will be presented twice as often as the 1:2 mappings).*

You understood the design correctly, and as you pointed out, the frequency of exposure can play a role, as well as the complexity of the mappings; we now state this explicitly on page 29. There are three main reasons why we plan to teach participants eight 1:1 and four 1:2 mappings: to be consistent with Poepsel and Weiss' design (2016) that similarly used a 2/3 and 1/3 split between 1:1 and 1:2 mappings, not to overstress the presence of 1:2 mappings (if 1:2 mappings were as frequent as 1:1 mappings, this may fundamentally change how participants learn); 1:1 mappings may overall be more frequent than 1:2 mappings in real life. Having said that, we agree that it would be very interesting to explore experimentally how the ratio of 1:1 to 1:2 mappings changes CSSL—this is indeed the goal of another ongoing project of ours.

- *For H2, the authors may want to consider the possibility of no significant two-way interaction between mapping type and language group, but a significant three-way interaction between mapping type, language group, and block that might indicate that the bilingual advantage only emerges with increased levels of training.*

Thank you for the idea. We decided to add the possibility of a significant three-way interaction between mapping type, language group and block in favour of our hypothesis in the table (pg. 37).

- *For H6, it would be useful to consider whether the interaction may only be expected to emerge after some learning has occurred, i.e. in later blocks.*

As before, we decided to add the possibility of a significant three-way interaction between mapping type, language group and target count in favour of our hypothesis in the table (pg. 45).