General comments:
We added a new figure on page 8 (Figure 1) that we think will help readers understand the logic behind the trial-by-trial analyses. Therefore, the other figures have shifted numbers by one. We also decided to change the term “limited advantage” to “specific advantage” based on the literature.

Editor:

- *One concern is the disconnect between the fact that questions about bilingual balance are now exploratory, and the framing in the introduction. I agree with this. For example, The Current Study section (p. 15) talks about looking at effects of extent of balancedness in the list of “critical components” of the experiment. It is ok to mention it, but it should be made clear that this is an exploratory element, mentioned after the factors related to the key hypotheses*.

Thank you for this feedback. We now explicitly state that the effect of balancedness on statistical word learning is only an exploratory question (see pages 13 and 14). Consistent with this, we no longer mention it as part of the critical components in The Current Study section (but instead in the final paragraph of that section). We also removed any references to distinctions between balanced and unbalanced bilinguals when introducing existing studies on word learning during the introduction (note, though, that we still define balancedness in brackets on pages 3 and 4 to introduce the general concept).

- *I also got quite confused where you talk about the “Hypotheses 1-4” (p 7) as these don’t seem to match hypotheses later in the paper.*

We are sorry for the confusion we created between our hypotheses and the ones listed on page 7. We added Hypotheses 1-4 (as introduced by Bogulski et al., 2019) to create a stronger link between the theory and the current manuscript. However, since this addition created confusion and the manuscript concentrates on the predictions of one hypothesis, the learning adaptation hypothesis, we decided to cut them out of the current version. Instead, we now concentrate our discussion on the learning adaptation hypothesis and its predictions. We hope that this change will make the manuscript more readable and coherent.

- *My other concern regards the Bayes Factors. First, can you clarify that you are moving to use Bayes Factors throughout as your statistic for inference, rather than significance? There are places where the word significance is used (e.g. p. 28) and these will need to be changed. If you are planning to mix analyses that needs to be very explicit and well justified (it generally isn't recommended -see Dienes, Z. (2023, August 14). Use one system for all results to avoid contradiction: Advice for using significance tests, equivalence tests, and Bayes factors. https://doi.org/10.31234/osf.io/2dtwv).*

Thank you for this suggestion. We read the papers you suggested and decided to use Bayes Factor throughout the manuscript. We deleted any place in which the word significant was used.

- *Second, while I applaud the move to these analyses, they are currently under specified. The author information for PCIRR says*
  “For studies involving analyses with Bayes factors, the predictions of the theory must be specified so that a Bayes factor can be calculated. Authors should indicate what distribution will be used to represent the predictions of the theory and how its parameters will be specified. For example, will you use a uniform distribution up to

some specified maximum, or a normal/half-normal distribution to represent a likely effect size, or a JZS/Cauchy distribution with a specified scaling constant?"
You say you will follow a similar approach to Silvey et al. In that paper, we represented H1 as a half normal distribution with a mean of zero and SD set to x, where x is a rough estimate of predicted effect size. That may be reasonable here too (for one sided predictions assuming small effects are more likely) but you need to explicit say what distribution you are using and why. Most critically, for each hypothesis you need to specify the parameter x – i.e. what size of effect you predict. For where to get these values- I think for this study you have values from pilot data / previous study. If not, Silvey et al suggest some possible ways to estimate main effects and interactions which may (or may not) be appropriate and there are more suggestions in these papers by Dienes:
Dienes, Z. (2021). How to use and report Bayesian hypothesis tests. Psychology of Consciousness: Theory, Research, and Practice, 8(1), 9.
https://osf.io/preprints/psyarxiv/bua5n
Dienes, Z. (2021). Obtaining evidence for no effect. Collabra: Psychology, 7(1), 28202. https://osf.io/preprints/psyarxiv/yc7s5

We wanted to thank you for pushing us to use the Bayes Factor. We apologise for not being specific enough in the previous version. We added more details on how we will conduct the analyses (see footnote 3 on page 16). We will use a half-normal distribution with a mean of zero and SD set to x, where x is a rough estimate of the predicted effect size. As you suggested, we will use the effect size from our pilot and data from Poepsel and Weiss (2016). We hope we have now provided enough details as suggested by the PCI:RR guidelines.

- *You also need to explain more clearly how you will use the information from the mixed models to get the data summary which is fed to the calculator etc. I know that information is in Silvey et al, but you can't assume the reader will look at that and it isn't a "standard approach".*

Again, we are sorry for not explaining the information needed in detail. We added more information about this in footnote 5 on page 23 and specified that readers can see all the functions we will use in the OSF folder.

- *A related point is that you talk about doing "power analysis", but that implies that what you are going to be interpreting is the p value. If instead you are going to compute Bayes factors you need to do analyses that BOTH (1) show that if H1 is true, you have a reasonable chance of finding your threshold BF (e.g. BF>3) AND (2) show that if H0 is true you have reasonable chance of finding BF<0 (which btw is generally harder). I think that you used simulation to do your power analyses, in which case you should be able to adapt these to do the Bayesian version. (For looking at 1, you can run the same simulation but see how often you get BF>3 (or whatever threshold you are choosing) rather than how often you get a p <.05; For looking at (2) you will have to run versions of the model where the relevant coefficient in the mixed model is set to 0, and then you can see how often you find BF below your threshold fo 1/3). I would be happy to share some scripts with you where I did something similar if that is helpful (feel free to email me). I would also request that you share your own scripts for power analyses on OSF as it will make it easier for me – and future readers- to evaluate the approach you are taking.*

Thank you for the suggestion. Based on the additional reading we did, we decided not to redo all the power analyses using the BF. Instead, we decided to keep the stopping rule as the most relevant practice in the power analyses (following Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. Psychological Methods, 22(2), 322–339. https://doi.org/10.1037/met0000061). We kept the power analyses with the p-value, but just to determine a minimum for the range between which we can stop. We moved most of the information on our original power analyses into the supplementary materials since they are now less relevant and to streamline the paper. We uploaded all the scripts in the OSF folder (https://osf.io/4e5nq/).

Reviewer 1:

- *My main concern is that the introduction still reads as if it the study will ask a question about whether bilinguals' language balance relates to CSSL, but this has now become only exploratory. Throughout the introduction, whether bilinguals were balanced or not in previous studies is mentioned inconsistently, which leads the reader to think that this study might look at balanced bilingualism as a reason for possible incongruent findings to date. But now it is only an exploratory analysis, and only for one of the hypotheses (though I could imagine that potential effects of last competitor accuracy on mutual exclusivity could be related to language balance). To be clear, I think looking at whether balance across languages matters should be exploratory, since it isn't clear what the theory behind why it would matter is, but it feels like the introduction was only minimally edited in response to previous concerns. Below are more comments related to this issue.*

We are sorry that the previous version of the introduction might have been misleading. Following your suggestion, we deleted all the references to "balancedness" and "unbalancedness" in the main parts of the introduction. Instead, we now briefly introduce the exploratory question on how bilinguals' balancedness may relate to statistical word learning at the end of The Current Study section (see pages 13 and 14). We hope this solution will make the introduction more coherent with our hypotheses.

- *Pg 3-4, why split out that the bilingual advantage for word learning was found in balanced and unbalanced bilinguals, seems like it is just found in bilinguals.*

As we specified in the previous response, we modified the introduction following your comment. However, we left a reference to balancedness on pages 3 and 4 to introduce the concept of balancedness (i.e., a definition) at the beginning of the manuscript. We put the definition in brackets to underline that it is important but not central.

- *Pg 4 – "the goal of the study is twofold – understand if there is a bilingual advantage and to look trial by trial", so why mention whether participants are balanced bilinguals or not sporadically?*

As mentioned previously, we no longer make references to bilinguals' balancedness in the main parts of the introduction.

- *Pg 7 – new paragraph starts with "Hypotheses 1-4", I don't think I necessarily agree that more balanced bilinguals = better for the regulation hypothesis, or the learning adaptation hypothesis. I'm not sure what the mechanism would be? If the L1 regulation hypothesis has to do with more regulatory experience in their L1, I'm not sure that has anything to do with their "balance" in experience/exposure.*
    - *Relatedly, this paragraph calls these "Hypotheses 1-4" but they are not called that in text when they are described, and then the study itself has H1—H5, so I would recommend not calling them this.*

We agree with you and the editor that this part of the introduction was confusing. Based on your feedback, we decided to remove this section (and also to shorten the manuscript, which was starting to be too long). Instead, we now focus our theoretical discussions on what we termed the learning adaptation hypothesis, which's prediction we are specifically addressing in our study and which is therefore most relevant for our study.

- *While Hypotheses 1, 2, and 3 (for the RR) are clear, H4 and H5 are not particularly well described.*
- *For one, it isn't clear how last-competitor accuracy is calculated. Since there are always two competitors per trial, is it averaged accuracy across the two from the last*

time each one was the target? And for 1:2 mappings, how is accuracy on the other mapping considered in this, because it seems like it could be a competitor right?

We are sorry for not being clear. Following your suggestion, we decided to add more information about how last-competitor accuracy is calculated (see page 25). We also added a figure (new Figure 1) as part of the introduction that illustrates how last-trial accuracy, last-trial accuracy, and last-competitor accuracy are calculated. We hope that these additions help clarify the trial-by-trial analyses.

As you correctly described, the last-competitor accuracy is calculated as the average between the two competitors' accuracy when they were last encountered as targets (so it can be 0, 0.5 or 1). Last-competitor accuracy is thought to be a proxy (operationalisation) for mutual exclusivity, so it is important to consider what is present on that specific trial – T2 (the other target) is not present on that trial, so while it may compete internally (and could slow down RTs, for example), it is not a feasible target on the current trial because it is simply not present.

- *Another concern is that I don't fully understand how these are different from finding an effect of mapping type in H1 or an interaction between mapping type and language group in H2. If the interpretation is about mutual exclusivity, then finding that bilinguals do better on the 1:2 mappings would already tell you that they rely less on mutual exclusivity. What will this analysis add?*

Whilst H1 represents the main effect of mappings on performance, H2 (as you pointed out) is about whether there is an interaction between language group and mapping type. If there is evidence for an interaction between language group and mapping type in the predicted direction, this could be consistent with bilinguals using the mutual exclusivity assumption less than monolinguals. H1 could be considered a sanity check (we want to be sure that 1:1 mappings are easier to learn than 1:2 mappings). In general, this approach (H1 and H2) replicates the typical analytic strategy that has been used in the existing statistical word learning literature (e.g., Poepsel & Weiss 2016 or Benitez 2016). Having said that, if we do find evidence for an interaction between language group and mapping type (H2), this does not definitively show that bilinguals rely less on mutual exclusivity. In fact, it can also be that they are just successfully understanding the two labels for the same object. Our trial-by-trial analyses, specifically H4, test whether an interaction in H2 can be (at least partly) explained by differences in mutual exclusivity use.

- *It's also not clear how your effects can be interpreted. I think that for 1:2 mappings maybe the expectation is that there is no effect of last-competitor accuracy, because it shouldn't matter whether you got that right or wrong if it needs to be remapped, but lack of clarity on how that's calculated for 1:2 trials which are not blocked and could be the target or the competitor on other trials makes it hard to understand. The language in the table at the end says something like "if bilinguals show less of an effect of last-competitor accuracy" but what is less? Just not significant, not as significant, smaller betas?*

We agree that interpreting the effect of last-competitor accuracy for 1:2 mappings (one-word maps onto two different objects) is more complex than for 1:1 mappings. Last-competitor accuracy is calculated as the mean of accuracy for both competitors (it can, therefore, be 0 if the participant was not correct the last time they encountered competitor 1 and competitor 2, 0.5 if they were correct for only one of them or 1 if they were correct both times), and we now clarify this in the manuscript (see page 25). Thus, in general, if participants make use of the mutual exclusivity assumption and have knowledge of competitors' targets, they should be more likely to be correct on a current trial, independently of mapping type. For 1:2 mappings, as described previously, it is important to consider what is present on that specific trial – T2 (the other target) is not present on that trial, so while it may compete internally, it is not a feasible target on the current trial because it is simply not present. Having said that, how mutual exclusivity is exactly implemented is less clear for 1:2 mappings than for 1:1 ones. This is the case because participants may have correctly selected an object when it had the other meaning.

We think the trial-by-trial analyses with last-competitor accuracy will still be informative since the participant has some information related to the competitor when they were correct, and this is better than having none. The pilot data are going in the direction we just explained: we found a significant main effect of last-competitor accuracy and interaction with mapping type, with a less pronounced effect of last-competitor accuracy for 1:2 than 1:1 mappings. Regarding the table, we are no longer using the term "significance" (consistent with our focus on BF analyses), but we specified in the column "Interpretation given different outcomes" that we are expecting a BF greater than 3 for 1:1, but not for 1:2, mappings.

Some additional comments

- *Interlingual homographs like pie in English and Spanish don't seem super related to what is described at this point (in spoken language they sound different) but seem more related to this specific paradigm, but that isn't clear.*

We are sorry if the example created some confusion. We gave this example to show that there are different mapping types that map onto different existing semantic concepts. We chose "pie" because it is indeed relevant in our experiment since all the words were written, and we did not focus on how words sounded. To increase clarity, we exchanged the original example now with "cola" (see page 5), which also has the same pronunciation in English as in Spanish.

- *Edit on page 6 – "advance over monolinguals is due to having more regulatory skills in their L1 (first acquired and often most proficiency language)…" – it seems like a citation would be necessary here, there are so many paths of bilingualism (which was brought up in the first draft), and first language is not necessarily the most proficient for a lot of people, particularly depending on age of acquisition of the second language and the context (e.g. schooling).*

We agree with you, and we cut this part out of the manuscript to avoid confusion.

- *Pg 6 – "therefore, bilinguals may have better cognitive control than monolinguals, which may in turn facilitate word learning" – what is the link between cognitive control and word learning? There is also an example "(e.g. higher inhibition may reduce interference from competitor referents)", but that seems like maybe it's just for bilinguals? So is it the better cognitive control or is it the experience of being bilingual?*

Thank you for your comment. Again, this section is part of the manuscript that was cut.

- *Are participant exclusions included in the sample size? E.g., will you recruit 150 and then exclude participants, or is 150 the non-excluded number?*

Your second example is correct. Following your suggestion, we explicitly spell out in the manuscript that we want 150 usable/non-excluded participants (see page 16).

- *Bayes Factor – on Page 27 is says a BF larger than 3 is moderate support for the alternative hypothesis but a BF smaller than 1/3 is moderate support for the null, but on page 19 it said you would use 6 or 1/6? Is the 6 or 1/6 specific to H4?*

We are using two different criteria on purpose. The 6 or 1/6 BF criterion is only used in the context of the stopping rule for recruitment, while the 3 or 1/3 BF criterion is used to evaluate whether there is support for hypotheses or not. We wanted to be more conservative in the stopping practice than in the analyses, consistent with Rouder (2014). We now explicitly clarify in the text that we use this criterion to be conservative (see page 16).

- *Can the authors spell out how low/high entropy maps to balance? I assume that high entropy scores = high diversity of outcomes = more use of both languages (e.g., less predictable use of each language), but I am not sure this is correct*

Yes, you are correct. High entropy signifies more use of both languages. Following your suggestion, we added this information in the text (see page 25). We spell out that participants with a lower entropy score have less integrated language usage and, therefore, can be considered more unbalanced bilinguals. In contrast, participants with higher scores have more integrated language usage and can, therefore, be considered more balanced bilinguals.

- *Pg 30 – how does only including participants that reach 40% accuracy during the last block ensure that you have enough usable trials?*

We apologise for the confusion we created. We decided to add this criterion because the pilot study had many non-learners, and including them in the trial-by-trial analyses made them noisier. We now explicitly say when we specify the goals of our study that we will only do trial-by-trial analyses with learners (see page 4).

Reviewer 2:

- *I only have a very minor suggestion that the BFs should be included in the final section when considering conclusions/implications.*

- *Typos:*
  p. 10 'In another recent study (Aguasvivas et al.)...' – verb missing, and Basque misspelled
  p. 13 ..Singaporean balanced bilinguals adults - balanced bilingual adults
  p. 16 first line: with 'A word and three objects'
  p. 18 'The bilingual group will be proficient' (remove 'speak')

Thank you for the suggestions and for taking the time to review this manuscript! As suggested, we added the BFs in the final section. We also corrected all the typos.