Dear Reviewers and Editorial Board,

Thank you for your thoughtful comments on my manuscript. Below please find a section outlining more general changes to the paper, followed by sections to each reviewer’s comments. I trust that you will find this version of the manuscript more fluid and coherent, largely thanks to your suggestions, but also due to some changes that I have made to improve flow. I have summarized all of these changes in black. In the following sections, reviewer’s comments appear in black, and my responses to these comments appear in red. Finally, changes to the manuscript are highlighted in red. Thank you for the opportunity to collaborate with LAB and I look forward to moving towards publication.

**General Revisions**

***General changes:***

* There is now an analysis section that describes the plan for data analysis.
* The discussion has been significantly restructured to be more concise. The major change is that I now present the results with respect to each research question, followed by general theoretical comments. I conclude with limitations.
* I have attempted to justify the use of the volitional subjunctive as the structure of interest in this paper, as well as to explain that reassembly is not the only result of Putnam and Sánchez’s (2013) model. Rather, there could be variability in mapping functions onto form (e.g., the uninterpretable mood feature onto inflectional morphology). These contributions contribute significantly to the coherence of the theoretical arguments advanced in the manuscript.
* I have described how Putnam and Sánchez’s (2013) predictions may potentially be adjusted to account for the protracted acquisition of features in heritage languages. This aligns with recent research that I have published and argues that feature reassembly can be a bidirectional process.
* All sections of the paper have been made more concise.

***Changes to statistical analysis:***

* I have adjusted the statistical analyses based upon the reviewers’ comments and have also made some changes to better address the research questions. For instance, properly testing the hypotheses require additional interactions, such as the interaction between frequency of use and task, that were not integrated into the previous modeling. In order to address all of these variables, nested model comparisons were necessary through pairwise comparisons, and only those models whose effects were considered significant were included in the final model.
* The third GLMM now evaluates lexical frequency with a slightly different predictor than the other two models. This is because this model is specifically testing the frequency of subordinate verbs, and sometimes participants produced grammatical sentences using periphrastic constructions. In these instances, the expected subordinate verbs were not inflected, so it would not be appropriate to determine the influence of frequency using these verbs. Note that my research questions did not center around the role of lexical frequency, so I have limited my discussion of these data.
* I have removed the forest plot figures and have reported the statistical models in the prose.

**Reviewer #1**

Please be sure to cite the Putnam & Sánchez (2013) piece in your references. It’s quite surprising that it’s not there...

Corrected.

The treatment/discussion of “features” in footnote 1 requires further elaboration, even if the primary focus of this manuscript isn’t directly involved with the advancement of any particular theoretical claims. One reference that is worth consulting in this respect is Lohndal & Putnam (2021; hereafter, L&P), who provides a more detailed treatment of the conceptualization of “features” that is largely compatible with P&S’s (2013) proposals. As such, the notion of “feature” seems to function as the fundamental unit of linguistic structure in this paper, hence, perhaps it should not appear in a footnote, but rather in the main body of prose.

Thank you for this comment. I have described the role of features and their mapping onto lexical items in greater detail. I have also cited a recent article that I published in LAB supporting this framework with Spanish heritage speakers.

Page 10, lines 22-28: Two points are in order here: First, building upon my previous comment, a more detailed treatment of “features” in this paper would allow A to say something that the features responsible for subjunctive mood (in connection with the morphology that ‘realizes’ these features). Second, it sounds a bit awkward to say something along the lines of a grammatical feature (or set of features) being “less like to reassemble”; rather, I think that something along the lines of "would be less likely to be impacted”, since ultimately what’s going on here is a resistance to reassemble (or, perhaps, different associations with allomorphy).

I have rephrased the predictions of the research question to align with the notion of being impacted by reassembly (or not), as the reviewer mentions. Furthermore, I have described how the notion of feature is relevant here for the study of the subjunctive because the uninterpretable mood feature must be mapped onto the appropriate morphology that generates lexical items (e.g., inflected verbs with subjunctive inflections).

A question concerning the interpretation of these experimental results involves whether or not features have been “reassembled”, or perhaps, their associate with allomorphy that expresses subjunctive mood has become more variable. Although the P&S (2013) model would support both, an important contribution that A could make here, along the lines of what L&P propose on p. 11 (6) of their article, is that the “loss” or “reconfiguration” of features is one of several outcomes. What some of these groups/individuals in this study may be doing is exhibiting variable feature-morphology associations. Again, although the larger focus and purpose of this paper is not a theoretical contribution per se, it’s worth mentioning this, especially since this is in line with the general position offered by A in their interpretation of their own results.

This is an insightful question that deserves greater attention in the manuscript. Features can become reassembled entirely, but it is also very possible that what happens due to decreases in exposure is that there is an increase in difficulty mapping synsem features onto forms, as indicated in the Lohndal and Putnam (2021) article. I have incorporated this into the literature review, predictions, and discussion section. In particular, I have attempted to highlight that these two outcomes are situated at different points along a continuum, whereby increasing difficulty of mapping forms to functions is part of the reassembly process (e.g., the early stages of the P&S framework), while “full-fledged” feature reassembly is the outcome of this process (and is actually quite rare in heritage grammars, as Perez-Cortes et al., 2019 discuss).

Page 25, lines 4-24: The question raised by A re: whether or not the P&S (2013) model is adaptable to children, or at the very least requires some deeper thought. I would suggest also taking a look at Putnam et al. (2018) for an expansion of this topic. Surely, this is likely something that will receive short shrift in this piece, but it’s worth acknowledging how this could contribute to a larger a discussion on these and related matters.

While I agree that this article *could* be more theoretical in nature, I think that one of the main messages I hope to communicate is that Putnam and Sánchez’s (2013) predictions might run in reverse when talking about initial HL acquisition. For instance, children may begin to exhibit more stable receptive knowledge before they develop consistent form-function mappings in production, contra the reassembly process postulated for HL attrition and reassembly. The rate of this process may be driven by patterns of exposure and language dominance, as Putnam et al. (2018) describe.

**Reviewer #2**

I would argue more generally that we need follow-up studies with larger samples to confirm all of the findings of the present study.

I have added a sentence to the final paragraph of the discussion section to discuss this.

I recommend including an analysis section in between the methods and results. I am not completely clear on the dependent variables (see specific comment below for p. 18). And if the models are using multiple observations per person (as opposed to a tallied score for one or both tasks), I wonder if would be more appropriate to nest observations (e.g., responses on a task) within individuals. Given the relatively small sample size, an analyses section could also explain the suitability of the statistical models for this number of participants and describe any preliminary tests that may have been conducted to determine the suitability of these models.

Thank you for these comments. I have opted to restructure the statistical analyses completely. I provide more detail about the changes to this analysis in the first section of this review document.

From the abstract, lit review and the research questions, it becomes clear that the study focused on the volitional subjunctive in nominal clauses, which is perfectly reasonable. However, I think that it would help readers—especially those who do not have a background in Spanish—to explain \*why\* that subjunctive.

This is an insightful comment that I have incorporated into both the introduction of the manuscript and the description of the volitional subjunctive mood. It provides readers with a clearer justification of why I have chosen this element of the subjunctive, defended by citations in recent research (as well as Blake’s work). I am confident that this apportion has strengthened the reasoning for using this component of the subjunctive mood in this project. Finally, I have mentioned that the volitional subjunctive is interface-free, along with the fact that it’s an uninterpretable feature (two ways of saying the same thing) and have said that “non-interface” structures are generally more resilient in heritage grammars.

I understand the choice to hold morphological regularity constant. However, it is important to acknowledge that verbs that have a stem change in the subjunctive are likely to be more salient to students in both oral and written language. For example, the 3s present subjunctive form tenga (indicative counterpart tiene) may stand out more than the 3s present subjunctive form hable (indicative counterpart habla). See Collentine’s work, in particular his 1997 piece in Spanish Applied Linguistics. Collentine’s participants are traditional FL learners, but the difference in salience is relevant to heritage learners as well.

This is indeed a key finding, and a paper that came out during the process of data collection for this process actually supports this assertion based upon research with heritage speakers (Giancaspro et al., 2022). I have addressed this in the limitations section, as well as through a footnote in the methods section.

Regarding the MLE school, many English-language middle schools in the US offer Spanish as an elective (typically designed for traditional FL students). Did the school that the MLE 7th and 8th graders attended offer such courses? If so, were any of the participants enrolled in those courses?

I have clarified in the manuscript that the MLE students did not receive any Spanish instruction at any stage of their education.

Along the same lines, I noticed on p. 24 that it says that the DLI group had their input decrease to just 1 class a day. Was this class a Spanish as a foreign/world language class or was it more a of a language arts (reading and composition) class in Spanish? Were they attending schools designated as DLI where the proportion of Spanish instruction shifted, or were these children who attended a DLI elementary but moved into an MLE middle school? These details have important implications for Spanish use at school, so I would like to see more information on this.

I have clarified in the manuscript that this was a Spanish for heritage speakers course, and have also added a reference that indicates that this is relatively typical for DLI programs.

Did Gathercole (2002) really use experimental methods? Unless she randomly assigned students to conditions (which seems unlikely), she didn’t. Maybe you mean quasi-experimental (e.g., comparison of 2 groups of people that are not randomly assigned to conditions). Maybe you just mean longitudinal and/or comparative.

This is a good point of clarification. I have reduced this portion of the text to abide by word counts, so I no longer use the term “experimental” to describe this experiment.

I tried to look up the specific Gathercole article to verify the methodology used, but the reference is not in the bibliography.

Corrected.

Also, I’m not sure why it’s ironic that only Gathercole has done this. Is it that we have a gap in the research that only she has sought to fill? This should be clarified.

I have removed the word *ironic* from the text for two reasons. Firstly, I have recently published an article that adopts this method in assessing differential object marking, and secondly, I agree that it’s not ironic but rather an opportunity for scholarship.

p. 9: SDBA= Spanish-Dominant Bilingual Adults, correct? Please spell out the abbreviation upon its first use.

Corrected.

p. 12: MLE-7/8 n = 11, as shown in the table. It says n = 25 in the text, but the table shows that it is the total for MLE at all grade levels.

Corrected.

p. 13: “The children enrolled in the monolingual school were matched for age, socioeconomic status, and family background with those in DLI.” How so? This doesn’t look like a study with matched pairs because family background with those in DLI. How so? This doesn’t look like a study with matched pairs because Table 1 show us different numbers of participants in each cell. Was it one-to-many matching? Reading the rest of the paragraph, my guess is that the research was conducted at MLE and DLI schools that had similar demographic profiles (i.e., similar socioeconomic status (SES) and family backgrounds), though I’m still unclear on the age part beyond the fact that recruiting from the same grade (e.g., 5th at MLE and DLI schools) will yield participants with similar ages. Choosing schools with similar profiles—while commendable—is not “matching” as it is understood in statistical analyses. I’d like to see a clearer explanation of this matching.

Indeed, this is not matching in the statistical sense, so I have removed this terminology from the manuscript. The manuscript now indicates that the schools were matched based upon their populations, but there is no mention of participant matching.

p. 13: “While only some children’s parents spoke English, all spoke Spanish, and all participants were predominantly exposed to Spanish at home (see Table 2).” With regard to “some children’s” and “all”, are we talking about the children at the school generally, or the participating children? It seems like the latter, but this should be clarified.

Corrected.

p. 13: “The SDBAs... represent a source of input for the HS groups.” How so? Where they recruited from the same community where the schools were located? Are some of them parents/family members of child participants? Do some of them teach at the DLI school? Please elaborate.

Yes, all of the SDBAs were from the broader community and most were from the same town; I have clarified this in the manuscript.

p. 13: Please explain how frequency of use of Spanish is being measured. I think that you get to this later, so it may be a matter of moving text up or telling the reader that these measures are described later on.

Corrected.

p. 14: Please explain the proficiency measure used in Table 2. It is not the full DELE, given that it only has a maximum of 18 points. Is this a subset of the DELE? If so, what types of items were included? (It seems like a lot of DELE points for volitional subjunctive only. Is it for all subjunctives or some other combination of items?) Reading on further, it looks like it was the BESA. Again, this is a matter of moving text up or telling the reader that these measures are described later on.

The DELE is not used to assess proficiency in this study. While I report the SDBA’s DELE results descriptively, I have added clarification that the data represented in Table 2 are from the BESA Spanish morphosyntax sentence completion subsection.

Also, if Table 2 is reporting the BESA results and 4 items were not reported, shouldn’t the maximum score on the table be 14?

Yes; corrected.

p. 16: I’m a bit uneasy about the items with creer being used to test indicative mood, just because we can see some mood variation in negated epistemics like no creer que, which could carry over into the affirmative counterparts if the speaker wants to emphasize doubt. A safer bet would have been testing after *tener que*, *saber que*, or *ver que*. I suggest acknowledging this in the limitations.

I have added a sentence to the limitations section; however, multiple published articles (Giancaspro, 2019, 2020) have used this matrix verb as a control with minimal to no use of subjunctive by any of the target populations.

p. 17: “In the HSs’ production data, the children produced the subjunctive in a total of 131/426 instances (30.7%), and alternative forms in the remaining 295 instances (69.2%).” If this is the case, why does every child group in Figure 1 have production percentages over 30.7%? I understand that the 30.7% is for all children and Figure 1 shows the subsets, but I don't see how every group doing better than 30.7% can mathematically also work out to 30.7% for all children.

Thank you for this concern. The 30.7% included the SDBA’s data and represented the total count of non-subjunctive forms, rather than the amount of subjunctive forms. I have corrected this section of the manuscript by including only HSs’ data in the table and by reporting the correct percentage of subjunctive production (45.4%).

p. 18: “In both models, the suppliance of the expected mood inflection was incorporated as the dependent variable...” I’d like to see the phrase clarified. For any participant, one can calculate a preference (receptive) score and a productive score. So, is the DV a combined receptive + productive score? Or is this more of an odds ratio, that is, the odds that participants will answer correctly? This is one thing that could be included in an analysis section.

This has been taken care of by incorporating an analysis section that describes the process of data coding more precisely, per the recommendations of other reviewers and the editor.

p. 21: “some who showed great variability on the proficiency test” Do you mean that these 6 students showed a wide range of scores on the proficiency test?

I have clarified this through clearer wording.

Footnote 9: Do you mean to refer the reader to footnote 8? Both notes may work better as notes on tables rather than footnotes to the text.

I have moved footnotes 8 and 9 to the table description per your recommendations; however, I did not intend to refer readers to footnote 8 in Table 5. I apologize if I have misunderstood this recommendation.

p. 26: Like my comment on Gathercole (2002) above, I wonder if “experiments” is the most accurate term here.

I have changed this to “these tasks.”

p. 26: Although every verb in the present study was a regular –ar verb, the authors could look at frequency of those verbs. That is, did students perform better on the items featuring more frequent verbs?

When developing this project, lexical frequency was taken into consideration. An analysis of the frequency data is complicated because there were instances in which participants used different verbs (e.g., verb phrases; instead of *amar*, *tenga que amar*) that were inflected for the subjunctive. Truly calculating lexical frequency would require discarding a large amount of data (that is, all responses where the inflected verb is not the expected verb selected for the study), which could further skew results. Anecdotally, as can be verified by viewing the GitHub repository and downloading the coding, there was no main effect for lexical frequency in the modeling. I am aware that this repository is blinded for the purposes of peer review, but I am comfortable with providing access to it if the editor deems it appropriate. While I aspire to be maximally transparent, and will be able to do so once the unblinded repository is made available after review, I also recognize that analyzing only those responses that contain inflected forms of the expected verbs would be problematic because it would also require discarding large amounts of data.

p. 26: Regarding the discrepancy between the findings of the present study and those of Potowski (2007a), how many of the subjunctive items in the latter had stem changes? This could be another factor in addition to the amount of Spanish used for instruction at the DLI school.

This is an excellent question. Potowski’s tasks were open-ended, so they may have favored subjunctive use because they could choose their own verbs. However, due to space limitations in the revised manuscript, I opted to remove this paragraph, because it does not contribute to the central arguments of the paper.

p. 28: I agree that we need to consider the importance of output in Spanish (which a DLI setting does not necessarily guarantee) and the role of input quality when considering why we didn't see differences between MLE and DLI students. Something else to consider is the nature of language use questionnaires. Language use at home can vary depending on the interlocutor(s) and the topic of conversation, and it’s difficult for a brief questionnaire to capture all of that complexity.

I have added this very accurate observation into the paragraph.

I see other minor errors and recommend a thorough proofreading of the manuscript.

This has been part of the revisions.

**Reviewer #3**

*The reviewer provided multiple corrections on the manuscript itself. Most of these corrections are simple, so I have not included them below. Only those comments that were left on the manuscript and that require further elaboration are included.*

Please carefully proofread the paper again; a number of (grammatical) errors and errors in Figures have been pointed out in the PDF. In addition, please introduce all abbreviations at first mention.

Thank you for your comments. The manuscript has been largely rewritten and has been reread multiple times.

Some sentences are very long, run over multiple lines, and are separated by numerous citations within – this makes reading and understanding somewhat cumbersome. Perhaps consider rephrasing some of these to produce as smoother writing style.

Thank you for your recommendations. I have used this writing style in recent publications, including one in LAB. I have attempted to make some sentences shorter, although I do find that there are some areas in which it would not be appropriate to move citations to the end of the sentence. In these instances, the citations refer to the specific part of the sentence. Where possible, I have broken these into separate sentences. The other reviewers indicated that they found the manuscript to be clear, so I trust that this revision strikes a balance between the reviewer’s concern and adequately representing the ideas of different authors who are cited.

I think some of the contextualization needs to be a bit more in depth. For example, who are your SDBAs? And please provide more context to the type of schooling – this information was bit scattered here and there and appeared a bit inconsistent. Please check and provide a more coherent picture.

I have added information about the SDBA’s community. I have also incorporated information about who these participants were (undergraduate or graduate students, which describes their occupation and level of education).

Whereas your analysis seems solid and the individual steps you took can easily be followed, I encourage the author(s) to also briefly describe tables/figures to point out what the most relevant information is or what the reader is supposed to see from these tables and figures. However, some important information is missing, and this should be added: you mention that you have a random structure in both regression models. However, nothing is said about the impact of the random effects. Please add. In addition, the model fit of both models needs to be reported.

I have added information about Figures 1 and 2 to the text in a descriptive statistics section before the statistical modeling. I have also conducted nested model comparisons and revised my statistical modeling accordingly. These comparisons first compare models with different fixed effects that contain participant and item as random intercepts. The model of best fit included the predictors school, age group, BESA proficiency, frequency of use, and task. Then, I attempted to carry out a second nested model comparison between two GLMMs that evaluated these same predictors, with one model containing random intercepts only, (1 | Participant) + (1 + Item), for participant and item, and one with random intercepts *and* random slopes. The model with random intercepts and random slopes failed to converge, so the final model included for the statistical analysis was the one with the lowest AIC in the *anova* comparing the models with random intercepts only. I referred to Cunning’s (2012) paper here and have cited that in the manuscript. I trust that this significantly improves the transparency of my data reporting, all of which is traceable on a public repository.

Specific comment to Figure 4/Model 2: Note that if you have an interaction in your model (which you have, namely 3!), you shall interpret the interactions instead of the main effects! This might actually have an effect on your interpretation here, but since this is not given in your results section, it is a bit hard to tell how you interpret the findings. Please add this and also pay attention to how to interpret such regression models. In addition to this, I would suggest considering a step-wise model building process (backward, or forward). You could then, if statistically not significant, remove/or not add predictors/interactions, rather than keeping them. Take a look, for example at Gries 2021. One possible way could be to start out with the maximum model (all main effects, all two-way interactions) and then to step-wise remove non-significant predictors (either main effects, if not in an interaction, or interactions) via drop1 in R.

Thank you for this comment. I have described the extensive changes to the statistical modeling in the first section of this document. Please note that I have used nested model comparisons to construct the model of best fit through pairwise comparisons. Based upon the results of the nested model comparisons, the models of best fit never included any interactions, so this is also now resolved through the revised analysis.

I am a bit skeptical about the discussion, simply because of the interpretation of Model 2 (see my comment above). Perhaps once this has been taken care off, the discussion needs to be adjusted.

I have restructured the statistical analyses, as addressed in the first subsection of this document, and have adjusted the discussion section according to the results obtained.

Perhaps consider having a somewhat more introductory start into the topic by having a shorter introduction and then, as a separate section, to systematically provide an overview of HSs language development.

Thank you for this comment. I have ultimately decided to leave the introduction as it is because I feel that all paragraphs of this section flow together and creating a different section would affect flow. I trust that this is not a significant obstacle for publication.

Is this still up-to-date? 2008, when this was published, is quite a while ago..

It is up to date because the date was mis-listed; I have corrected this to Montrul (2018) and added this to the citations section. Thank you for pointing this out.

How is "morphosyntactic proficiency" a way to operationalize exposure?

Following the researchers cited in the paper, proficiency (both lexical and morphosyntactic) has been operationalized as a proxy to language exposure (see Giancaspro & Sánchez, 2021; López Otero et al., 2023a, 2023b). The argument is that bilinguals who are more exposed to a language develop higher proficiency in it. While I agree that this argument seems to render proficiency a bit redundant with patterns of exposure, most manuscripts continue to explore this variable. Since this variable was included in Dracos and Requena’s (2022) study, which is cited at length throughout my manuscript, I have opted to use a similar proficiency measure to maintain consistency across results.

The setting is not entirely clear - which age cohorts are you referring to here?

Clarified.

So DLI does not only mean that content subjects are taught in Spanish and English, but that English is also taught as a separate subject? It would perhaps be useful to provide more information on DLI programs.

Due to space limitations, I have referred readers to another reference.

At this point, I am wondering, if it made sense to provide some actual sentence-level examples (contrasting subjunctive, with indicative and imperative?) to make it even more transparent for the reader what it is you are investigating (especially since you are not investigating all but only a special case of subjunctive uses?).

This is a logical suggestion, but due the incorporation of additional analyses and the increased word count, I have chosen to leave this as-is due to space limitations.

What does this mean?

These are syntactic categories from Rizzi’s cartographic theory that have become standard in literature on the subjunctive. I have included a brief description in the manuscript.

Why did you treat 7th and 8h grade as one group. Wouldn't you expect differences also between grade 7 and grade 8 students?

Ideally, I would have incorporated grade as a continuous variable, but there weren’t enough participants. I have addressed this as a limitation.

50% vs daily - so in the former, Spanish was not used daily, but in the latter, if daily, what is the % between Spanish and English?

I have clarified that all groups get daily Spanish instruction (e.g., 50% each day versus just one course each day).

How was SES operationalized?

The school reports do not discuss how SES is operationalized, and this is a comment addressed in my dissertation on a similar topic. However, in consistency with the conventions of educational research, the school reports have been anonymized here to protect the identity of the schools and their students. Unfortunately, this is an insurmountable limitation that I have addressed in a footnote.

What is their age?

I have added the average age and SD in the prose of the manuscript.

I find it a bit hard to believe that these children managed to rate their Spanish use on a 5-point Likert scale. Have you piloted and tested, how well this worked with children completing it independently?

I did not pilot this instrument. Firstly, previous research has adopted similar methods with children who were considerably younger than those in my study (Castilla-Earls et al., 2022). Specifically, the authors demonstrate that their instrument, which attempts to rate exposure to Spanish, has high psychometric validity in children between ages four and eight. Furthermore, children are better judges of their own interactions with their peers while at school or in the community, since their parents are not there. I have included the Castilla-Earls et al. (2022) reference in the citations.

To me, it is still unclear what kind of test it was. Oral? written? complete gaps? Could you perhaps specify this?

It was written; this is now included in the manuscript.

You keep calling the second test differently - preference task, selection --- perhaps better use one term only?

Prior to reading this comment, I had come to a similar conclusion in a reread of this comment. I have addressed this throughout the manuscript, including in the figures, and now refer to the second task as the “selection task.”

What is currently missing is the overall fit of the model, i.e., the predictive power. Please add this for both models as this is important in assessing the quality of the models!

AIC added. If this is not a sufficient measure of model fit, it would be beneficial for the reviewer to provide additional information about what s/he/they identifies as necessary to demonstrate model fitness.

What does "standardized prior to analysis" mean? This should be mentioned and explain in the methods section!

I trust that the inclusion of an analysis section that describes the computation of the variables has addressed this comment sufficiently.

I am surprised - wouldn't this be something you'd check before running the regression models, as part of the data inspection? I'd consider moving this part before the regression models.

Thank you for your comment. Recent articles in heritage languages (including my own, as well as Giancaspro et al., 2023) have incorporated individual analyses after the statistical analyses. I believe that it provides a natural segue into the discussion, so I have opted to be consistent with previous articles and have left this section where it is.

I’d even try if a correlation can be constructed here - there seems to be such a trend. Moreover, why are some points not on the lines? How can, for example the lowest red point be inbetween 1 and 2 (x-axis) if this is the absolute number of of subjunctives produced? how can 1.5 sentences be produced? Please explain.

Thank you for the comment. The purposes of individual analyses are to provide descriptive support for correlations, but this would not be appropriate. The graph was generated using the *jitter* layer using *ggplot* in R, which displaces the points slightly along a graph to enhance their visibility. Otherwise, many data points would gather along the same point (e.g., x1, y2) and only one point would be visible. I have two published articles with this format and would prefer to be consistent. To address your concern, I have added a footnote.

This is what I've commented on before (and you also mentioned yourself) - this might cast some doubt on the measure of the questionnaire in eliciting this kind of (truthful) information...

Thank you for your comment. I hope that in the revised version of the manuscript, it becomes clear that students often overhear Spanish, but do not respond using this language. This provides a natural explanation for why students reported “0” on the language questionnaire. In fact, it’s a more honest measure than what parents may provide. There are many studies (Babino & Stewart, 2017; Ballinger & Lyster, 2011; Hamman, 2018; Potowski, 2004, just to name a few) that have found this to be the case. Therefore, this finding is indeed in line with previous studies that have used multiple methodologies and that have identified that oftentimes, children do not use Spanish in immersion programs even though they are exposed to it. I have dedicated a considerable amount of the discussion section to this, so I do not feel that making further changes to this in the manuscript would be beneficial.

Perhaps instead of a footnote, this could be either directly mentioned in the text (also indicating that this concerns both Tables 5 and 6) or added as a note below both tables.

Thank you for your comment. I feel that including this in the description of the tables would be cumbersome, because the description is intended to be brief. I have chosen to leave the footnotes intact as the best possible solution.

But Freq. of use and parental languages does not necessarily suggest this? What is your thought here?

Heritage language research has consistently shown that Spanish HSs often respond in English, even when input is in Spanish (see Babino and Stewart, 2017 for a recent example in the context of bilingual schools or Flores et al., 2017 for a study in the acquisition literature. So, these ratings are in clear alignment of these findings.

This really makes me wonder again who these SDBAs are – this information is, I think, absent from the paper...

Corrected; this is a very insightful comment and I have made a concerted effort to show that there is no evidence of attrition in the caregiver input.

Where was this shown? I believe in Fig. 4? note my comment above (and in my response letter) about interpreting main effects and interactions.

This is not shown in a figure, as I have removed the forest plots (visual model outputs). They seemed to cause more confusion than clarification. Regardless, the model of best fit did not include any interactions, so this comment should now be taken care of. None of the data represented in the first draft or in the current manuscript involve interactions, but I was wondering if the reviewer would be willing to clarify why main effects should be ignored in the event that interactions are significant. I have not seen this in any previous research on Spanish as a second or heritage language, and it is also contra my statistical training. I would value this information for future manuscripts, but I would also be very surprised if main effects should be altogether discarded simply because an interaction involving one of these effects is also significant. For instance, it is completely plausible that, in a comparison of HS and L2 learners, that there is an overall effect for proficiency, as well as a group-by-proficiency interaction (see, for example, Montrul & Perpiñán, 2011). It would not be proper to discard the proficiency effect simply because one group is more affected by proficiency than the other; discarding main effects when they also are involved in interactions would be empirically problematic.

Same as before, please check your model again!

Thank you for your comment. I am not quite clear on what the reviewer would like for me to change here. The effect for age group indicates that older children were more likely to use the subjunctive than younger children. This does not involve any interaction and is clearly stated in the outputs of model #2.

This substantial information also needs to be added above (see one of my earlier comments).

This information is available in the participants section.

Preposition missing?

The preposition here is “contra,” which is standard in academic writing.

Given that you have this information in your regression model, I'd suggest to visually inspect the interaction rather than purely checking % for each group. A multifactorial perspective is to be preferred, I think.

Thank you for this comment. I am not clear as to which interaction the reviewer is referencing here. In the first draft of the manuscript, there were no significant interactions. In the revised version, due to nested model comparisons, none of these interactions were included in the models, so it is unclear how this comment relates to the information presented here.

Def. article missing.

“On one hand” is a standard phrase in American English.