**Peer Review and Communication History**

**MS Title**: Conducting Language Production Research Online: A Web-based Study of Semantic Context and Name Agreement Effects in Multi-Word Production

**Author Names**: Jieying He, Antje S. Meyer Ava Creemers, Laurel Brehm

**Submitted:** May 5, 2021

**Editor First Decision**: Revise & Resubmit

July 21, 2021

Dear Jieying He,

I have now received two reviews of your manuscript, “How to conduct language production research online: A web-based study of semantic context and name agreement effects in multi-word production” from qualified researchers. I also independently read the manuscript before consulting these reviews. The reviewers and I agree that your manuscript has some strengths, but also that there are some issues that need to be addressed. I therefore encourage you to submit a revised version for further consideration at Collabra: Psychology.

One point in partiular that stood out to me (also raised both both reviewers), is that the manuscript makes a strong claim that online onset latencies are problematic, despite presenting data showing that they are (at least in this context) pretty reasonable. It seems to me that it might be better to focus on the broader usefulness of including multiple different kinds of measures (which is, I think, an excellent point to make) rather than the potential limitations of onset latency.

Both reviewers offer several other useful comments and suggestions on framing and analysis that I won’t reiterate here, but these should certainly be addressed should you decide to submit a revision. In your resubmission, please include a document with a point-by-point response to these comments, outlining each change made in your manuscript or providing a suitable rebuttal.

In summary, I think this is a promising manuscript and, I hope you will revise it for further consideration at Collabra: Psychology. I look forward to receiving your revision.

Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This is the last opportunity for major editing, therefore please fully check your file prior to re-submission.

If you have any questions or difficulties during this process, please contact the editorial office at [editorialoffice@collabra.org](mailto:editorialoffice@collabra.org).

We hope you can submit your revision within the next six weeks. If you cannot make this deadline, please let us know as early as possible.

Sincerely,

Bob Slevc

# Reviewer 1

##### Open response questions

### Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

The paper entitled “A web-based study of semantic context and name agreement effects in multi-word production” introduces and reports data from a modified blocked-cyclic naming paradigm that can be applied in web-based settings. This paradigm allows to circumvent potential problems regarding the audiovisual synchrony by investigating outcome variables measured within a single trial, i.e., within a single audio recording. The authors find evidence for an influence of name agreement, but only weak evidence for semantic context effects on the various outcome variables.  
Without doubt, the approach applied by the authors is of high interest for speech production research and opens several new perspectives both in lab-based as well as in online settings and should therefore be appreciated by the research community. However, I have concerns regarding methodological and conceptual issues that should be addressed or clarified prior to publication:

1. **Data analysis:** For all outcome variables, the authors apply Bayesian mixed-effects models and provide two parameters that allow to evaluate the evidence for or against specific effects, Credible Intervals (Cr.I.) around slope parameters and Bayes Factors (BF). While I generally appreciate the statistical analyses, I have concerns about the author’s approach:  
   **a. Prior selection:** The authors applied “weak, widely spread” priors for all variables. How were these priors chosen? Where any prior or posterior predictive checks performed?  
   **b. Priors and iterations in BF analyses:** While the effect of priors on Bayesian modeling is fortunately limited, they can have tremendous effects on BF, especially if weakly informative. Therefore, the authors should at least perform prior predictive checks or, even better, perform the BF analyses on different priors (sensitivity analyses) as inappropriate priors may bias the BF towards the null hypothesis. Potentially, inappropriate priors may even partly explain the contradicting results of BF and Cr.I. analysis.  
   Furthermore, I was surprised that the authors seem to have conducted their BF analyses on the same number of iterations as the analyses for slope and Cr.I. estimations (4.000-7.000). To my knowledge, stable BF estimations usually need far higher numbers of iterations (e.g., around 20.000-30.000). The authors should therefore elaborate on the appropriateness of their models and priors and the possible influences on BF and Cr.I. analyses (see e.g., Nicenboim et al., 2021; book draft “An introduction to Bayesian data analysis for cognitive science).  
   **c. Model selection:** A minor aspect is the selection of the random structure: How did the authors define their reduced random structure? Did they start with a maximal fully-crossed random-effects structure and reduced the structure due to convergence issues or was the decision taken a priori? The approach should be made clear in the manuscript and/or the analysis scripts.  
   **d. “Warm-up phase”:** The authors state that “… each chain had 4000 to 7000 iterations depending on model convergence […], with the first half representing a burn-in period” (p.16). From the analyses scripts I understand that the warm-up phase was only 1000 iterations in each model.
2. **Lack of semantic context effects:** Belke and Meyer (2007) in their very similar paradigm found small, but significant semantic context effects on pause rate. The authors of the study do not find evidence for such hypothesized effects, but also only anecdotal to weak evidence against them. The reader may want to wish for a discussion whether these differences may be due to a lack of power and what effect size may have been reliably detected, especially because BF were relatively low. The reliability of the results may also affect the theoretical considerations regarding theories that do (not) assume competition at the lexical level (p. 8/9 and p. 30). Therefore, the authors should discuss the reliability and the inference they make about semantic context effects.
3. **Effect over time:** The experiment consisted of only 32 different pictures that were named repeatedly within and across trials (~16 times each). As briefly touched by the authors, this may have affected the results, e.g., through repetition priming. I think this objection should receive further attention. One way could be to analyze whether the effects (for a given stimulus) changed over time, within and across trials. Were the effects in the first block different from those in the third block?  
   This kind of analysis would also be interesting in light of the expected lack of semantic context effects in the first repetition: While the authors did not analyze voice onset effects from the second row for obvious reasons (even though there might also be some smart work arounds allowing to analyze onset latencies from the second row), analyzing them from the second repetition would be easily feasible, highly interesting, and in line with the research tradition.

***Minor issues***

1. I would appreciate if the authors made it clearer which analyses were planned a priori and which post-hoc. E.g., were the analyses from the second row only planned before data collection? Were the participant exclusion criteria set before or after data collection?
2. Inferences regarding the audio-visual synchrony: In their introduction (p.3), the authors summarize results from Bridges et al., (2020) by concluding that “web-based experiments currently cannot offer the audiovisual synchrony (i.e. simultaneous presentation of visual stimuli and audio recording) needed to accurately record onset latencies in language production”. I think this claim is not fully warranted as Bridges et al. (2020) did not test the synchronization of visual stimulus presentation and onset of audio recordings, but rather the synchronization of visual and auditory stimuli.  
   In the same vein, in their interpretation of speech onset latency effects, the authors conclude “in spite of the encouraging results of this particular study, we advise against planning web-based experiments relying exclusively on speech onset latencies unless precise timing of stimulus presentation”. Based on the focus of the study and the study results, this conclusion is not fully warranted. The authors, like other recent studies, overall find evidence for a precise-enough audiovisual synchrony to replicate lab-based language production effects (e.g., Fairs & Strijkers, 2021; Stark et al., 2021; Vogt et al., 2021) and Bridges et al. (2020), who report limited accuracy, tested the synchronization of visual stimuli and sounds presented via speakers instead of accessing the microphone.  
   (However, to additionally test the reliability of the speech onset measurements, it would be interesting, to analyze whether the speech onset effects in the current study were consistent within a single participant.)
3. Sample size estimation (p. 9): The selection of the sample size seems rather arbitrary. It would be great if the authors at least gave an example of the previous semantic context experiments they based their estimations on.
4. The apparatus section (p.9/10) is very short. As next to presenting a modified paradigm, the authors add to the still very little and recent evidence on browser-based speech production research, I think it would be valuable to give additional detail on the programming part of the experiment: E.g., where can the reader find additional information about the platform FRINEX? How did the voice recording work? This would give readers seeking to conduct browser-based speech production a head-start. Furthermore, only additional technical information, in combination with other, recently published speech production experiments, would allow to make informed decisions about the audio-visual synchrony and (computational) aspects that may influence it. This is a personal preference the authors are welcome to disagree with.
5. Materials (p.10).  
   a. The stimulus set was composed of 32 pictures “with one- or two-syllable names”. The stimulus list in the Appendix suggests that this was not the case for all stimuli with low name agreement (e.g., “inktvis / octopus”). The authors may want to clarify.  
   b. Where do the name agreement ratings (p.10) come from (e.g., from the norms/the experiment/pretests)?
6. Materials and design (p. 10-12): I had difficulties understanding the trial structure. I very much appreciate that the authors provide a link to an exemplary experiment, but I think the description could be more straightforward by making a clear differentiation between the terms “picture”, “picture grid”, and “items”. Providing the exact number of trials, the number of times each participant saw each picture grid and the number of times each participant named each individual picture would also help. Furthermore, I think the manuscript would benefit if Figure 1 contained an example of a heterogenous picture grid, too.
7. Analysis + Results: What’s the difference between silent intervals and pauses?
8. The authors switched the APA citation style on p. 32
9. The authors could further streamline their introduction and discussion. Although the manuscript is well written and a nice-read, the authors could do a better job in carving out their study’s major contribution: The novel paradigm? The web-based setup? The novel response variables? Then they may want to provide somewhat more detail on the background of their major focus (e.g., regarding recent attempts of web-based speech production research and technical details that may prove necessary for conceptual replications, relating the paradigm to other paradigms, an evaluation of the usefulness of the different response variables). This is an advice the authors may or may not want to follow, but that could further improve the reading flow of the manuscript.

##### Rating scale questions

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  | ✔ |  |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  | ✔ |  |

# Reviewer 2

##### Open response questions

### Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

The premise of the manuscript is overly pessimistic and does not seem to be based on sound empirical evidence. As a consequence, the authors’ main recommendation to avoid using onset latencies seem to be unwarranted. The authors claim that onset latencies “cannot offer the audiovisual synchrony needed to accurately record onset latencies in production experiments.” This is probably true, but the question is how much this matters. Literally all measurements have errors, so the real question is not whether there is a noise in measurement, but how big the noise is and if that noise is biased (not equal across conditions). As long as the noise generated by hardware is random, the challenge can be overcome (and has been overcome multiple times, indeed by the authors themselves for the naming agreement effect) relatively easily by collecting larger number of participants. In fact, recent studies by Strijkers’ group (<https://psyarxiv.com/2bu4c>) show that the classic effects in onset latencies in single word naming is comparable between lab-based and online studies. Our lab got 20-40ms semantic interference effect in picture-word-interference-like task about ten times in online studies. So I think the recommendation to avoid using onset latencies is simply empirically unfounded. In addition, onset latencies associated with multi-word utterances certainly have larger variability (= larger amount of noise) than those associated with single word utterances, because there are more processes involved in generating complex utterances (and this problem is in addition to the problem associated with simultaneous picture presentations discussed by the authors). Thus, to make the recommendation to use multi-word utterances, the author must demonstrate that the amount of noise introduced by hardware variabilities in single word utterances is larger than the amount of noise introduced by making the target utterances multi-words. For those reasons, I do not believe the authors’ recommendation to use multi-word utterances is justified, and it has the potential to mislead researchers in the field.

When talking about how speech duration/pause rate can be used to study planning processes, I think the authors should cite many previous studies that used the duration measure to study planning processes, including: Ferreira & Swets (2005, a book chapter on resumptive pronouns), Kandel et al (CUNY: <https://www.cuny2021.io/wp-content/uploads/2021/02/CUNY_2021_abstract_131.pdf>) Momma & Ferreira (2019, Cognitive Psychology) among others.

The authors cited Wheeldon & Lahiri, 1997 in the context of justifying that planning of multi-object naming utterances is incremental. I think this is an odd choice; Wheeldon & Lahiri I believe showed that phonological planning is incremental. The authors were specifically concerned about the incrementality at the level of lemma selection (non-phonological planning), so I am not sure if W&L is relevant in the context. (I believe it should be cited elsewhere in the paper, but not in p. 8). Perhaps the more relevant citation is Griffin (2001), which showed that naming agreement effect in multi-word utterance is absent in onset latency measures. More generally, I think the authors should be careful about what levels of processing they are talking about when making a claim about ‘incrementality.’ Relatedly, I think the authors at least should mentions previous studies showing that lemma/conceptual planning in multi-word object naming may not be as incremental (e.g., Meyer, 1996, JML).

The authors wrote “A power simulation for determining sample size before the present study was not possible, as no comparable studies were available. However, the data reported here can now be used for a power simulation to estimate the sample size needed to observe effects of interest in future work.” If this is the claim, I think the authors might as well provide do the power calculations and include the results in the manuscript?

The methods used in this study seems to be mostly fine. I think there is an issue of how simultaneous picture presentations affect interference effects, but I believe the issue is adequately discussed.

##### Rating scale questions

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  | ✔ |  |  |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  | ✔ |  |  |  |

**Author Response**  
Sep 6, 2021

Dear Dr. Slevc,

Thank you very much for your letter dated July. 21th, 2021, and the reviewers’ valuable comments on our paper. We do agree that we should be careful when making the claim about the reliability of onset latencies. This work was done in parallel with work in other labs also addressing the feasibility of online language production work. We were not aware of this work when planning the study reported here. We explain this in the Introduction. In the Discussion we review the implications of our results together with those of the other studies now brought to our attention by Reviewer 2 and claim that onset latencies are much more reliable than we thought. Moreover, we have clarified the usefulness of multiple dependent measures and revised analyses following the reviewers’ suggestions.

We have submitted two versions of our revised manuscript: one version with only the new text printed in red (i.e., clear version), followed by the other version with all the change highlighted by using the track changes mode in MS Word (i.e., changed version). In our *Response to Reviewers* we reproduce the reviewers’ comments and explain how we addressed each comment or concern. We hope this manuscript will be acceptable to be published on *Collabra: Psychology*. We are looking forward to hearing from you.

Sincerely,

Jieying He

Max Planck Institute for Psycholinguistics

**Response to Reviewers**

Reviewer 1

The paper entitled “A web-based study of semantic context and name agreement effects in multi-word production” introduces and reports data from a modified blocked-cyclic naming paradigm that can be applied in web-based settings. This paradigm allows to circumvent potential problems regarding the audiovisual synchrony by investigating outcome variables measured within a single trial, i.e., within a single audio recording. The authors find evidence for an influence of name agreement, but only weak evidence for semantic context effects on the various outcome variables.

Without doubt, the approach applied by the authors is of high interest for speech production research and opens several new perspectives both in lab-based as well as in online settings and should therefore be appreciated by the research community. However, I have concerns regarding methodological and conceptual issues that should be addressed or clarified prior to publication:

1. Data analysis: For all outcome variables, the authors apply Bayesian mixed-effects models and provide two parameters that allow to evaluate the evidence for or against specific effects, Credible Intervals (Cr.I.) around slope parameters and Bayes Factors (BF). While I generally appreciate the statistical analyses, I have concerns about the author’s approach:  
a. Prior selection: The authors applied “weak, widely spread” priors for all variables. How were these priors chosen? Where any prior or posterior predictive checks performed?

Thank you for your question. We selected these weakly informative priors for all variables based on calculations (see Nicenboim & Vasishth, 2016), differences between the priors and posterior distributions, and comparisons between three models with different sets of priors on each dependent measure. For instance, we started the analysis by setting *Normal (0, 0.2)* as the prior for the name agreement effect on log-transformed utterance duration, which means that we assumed that the difference between high and low name agreement conditions would be around 4259 ms[[1]](#footnote-1) and can be positive or negative. In reality, the range is likely to be much smaller (602 ms in our study), suggesting that the prior *Normal (0, 0.2)* was weakly informative. We then plotted the difference between the prior and posterior distributions of the name agreement effect on the log-transformed utterance duration (see Figure 1, left), which shows that the prior *Normal(0, 0.2)* does not have a large influence on posterior distribution, suggesting this prior is suitable. Next, we fitted the models with another two sets of less weakly informative priors: *Normal (0, 4)* and *Normal (0, 10),* and then compared the results with the prior *Normal (0, 0.2)*. Figure 1 and Table 1 below show that the three different sets of priors do not influence the posterior distribution/estimation, which further implies that the prior *Normal(0, 0.2)* is reasonable.

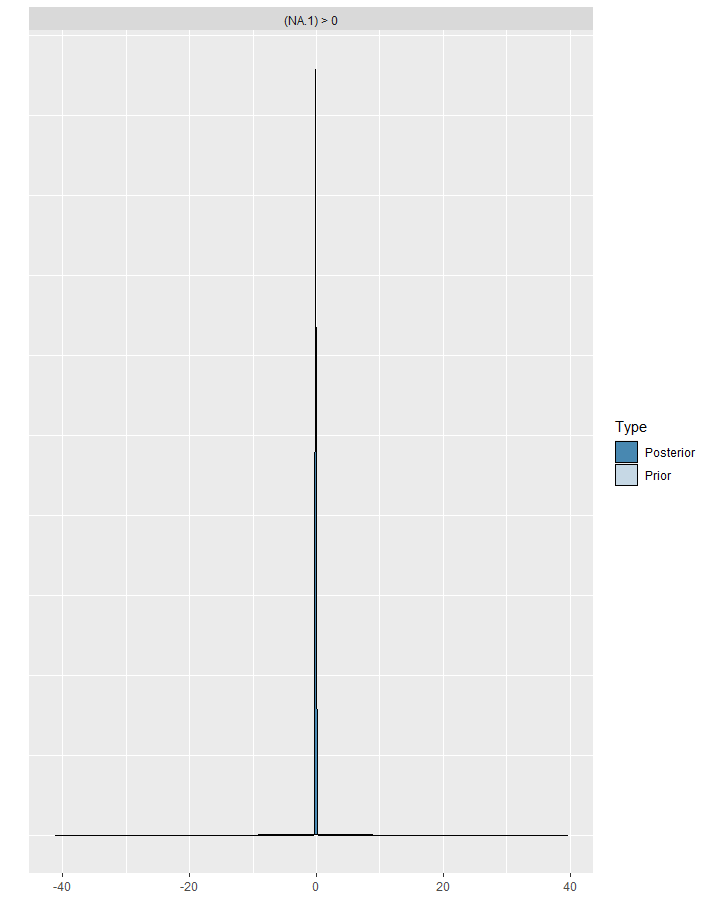
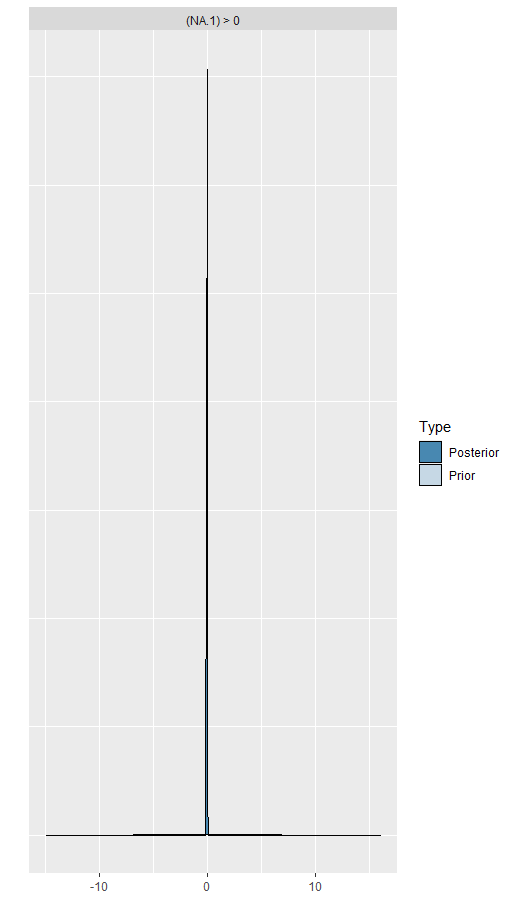
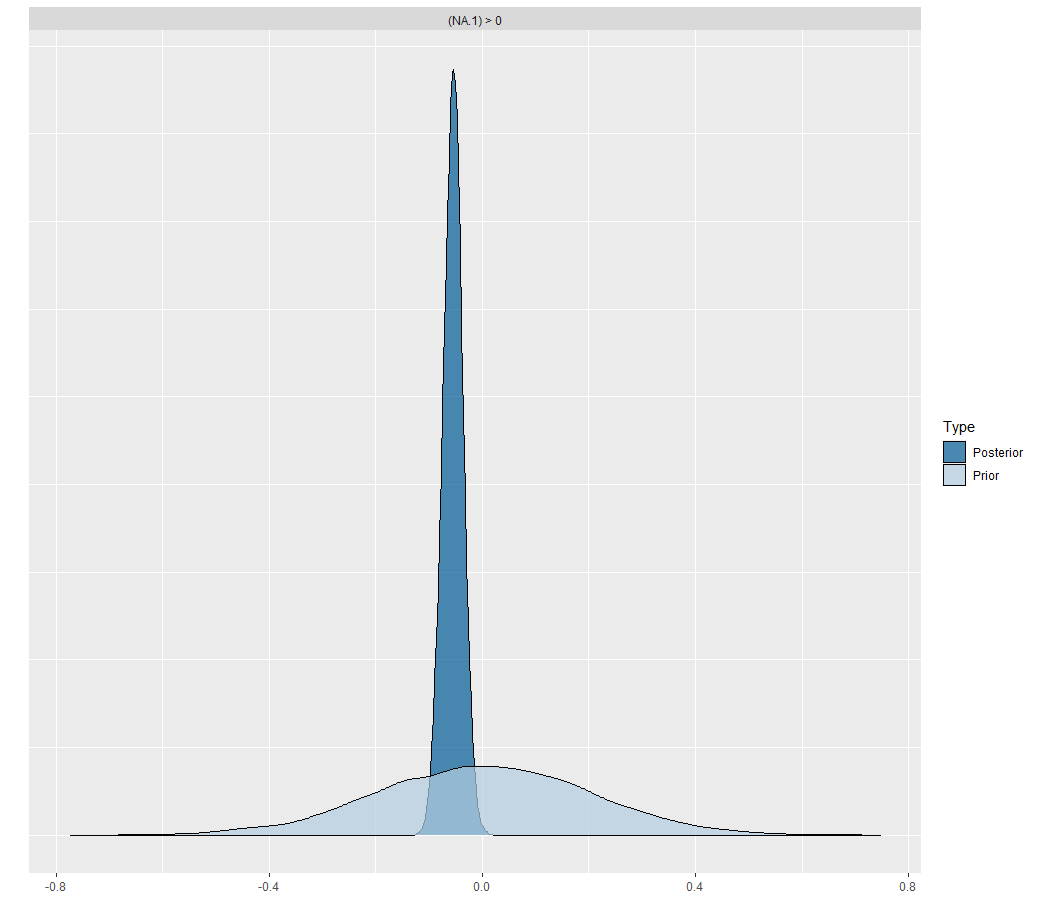


Figure 1 *Prior and posterior distributions of name agreement effects on log-transformed utterance duration given three sets of priors (left: Normal (0, 0.2); middle: Normal (0, 4); right: Normal (0, 10)). NA refers to name agreement.*

*Table 1* Summary of posterior estimation for the name agreement effects on the log-transformed utterance duration given three sets of priors.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Prior | Estimate | Est.error | 95% Cr.I | |
| Normal (0, 0.2) | -0.06 | 0.02 | -0.09 | -0.02 |
| Normal (0, 4) | -0.06 | 0.02 | -0.09 | -0.02 |
| Normal (0, 10) | -0.06 | 0.02 | -0.09 | -0.02 |

Similar to the log-transformed utterance duration, we also checked the suitability of priors for each effect on each dependent measure by calculating the difference between conditions, by plotting the difference between prior and posterior distributions, and by comparing the posterior estimations with different sets of priors (see Table 2). We found that the choice between these priors had no influence on the posterior probabilities, suggesting that the priors (i.e., Prior 1 in Table 2) we used in our study are reasonable. We have addressed this in our manuscript (see page 17, lines 321-323), and also uploaded all these figures/estimations to an OSF repository (https://osf.io/6jg4p/).

Table 2 Three sets of priors for each dependent measure. Log- refers to log-transformed.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Prior 1 | Prior 2 | Prior 3 |
| Accuracy | Student\_t(1, 0, 2.5) | Student\_t(1, 0, 5) | Student\_t(1, 0, 10) |
| Log-utterance duration | Normal (0, 0.2) | Normal (0, 4) | Normal (0, 10) |
| Log-total pause time | Normal (0, 1) | Normal (0, 10) | Normal (0, 25) |
| Total chunk number | Normal (0, 3) | Normal (0, 10) | Normal (0, 40) |
| First chunk length | Normal (0, 3) | Normal (0, 10) | Normal (0, 40) |
| Log-onset latency | Normal (0, 0.2) | Normal (0, 4) | Normal (0, 10) |

b. Priors and iterations in BF analyses: While the effect of priors on Bayesian modeling is fortunately limited, they can have tremendous effects on BF, especially if weakly informative. Therefore, the authors should at least perform prior predictive checks or, even better, perform the BF analyses on different priors (sensitivity analyses) as inappropriate priors may bias the BF towards the null hypothesis. Potentially, inappropriate priors may even partly explain the contradicting results of BF and Cr.I. analysis.

Furthermore, I was surprised that the authors seem to have conducted their BF analyses on the same number of iterations as the analyses for slope and Cr.I. estimations (4.000-7.000). To my knowledge, stable BF estimations usually need far higher numbers of iterations (e.g., around 20.000-30.000). The authors should therefore elaborate on the appropriateness of their models and priors and the possible influences on BF and Cr.I. analyses (see e.g., Nicenboim et al., 2021; book draft “An introduction to Bayesian data analysis for cognitive science).

Following your comments, we re-conducted the BF analyses with far higher number of iterations (i.e., 20000). As Table 3 shows, the results were highly comparable to those using fewer iterations (i.e., 4000-7000): although the BF values slightly differ from those of the previous analyses, the overall patterns of both analyses were the same. We updated these stats inthe *Methods* and *Results* (see page 18, lines 339-341; Table 3 on page 28; Table B2 in Appendix B).

Table 3. Bayes factors given different numbers of iterations.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | NA effect | SC effect | Null interaction |
| Accuracy | Iteration=4000 | 1.76 | 3.71 | 0.82 |
| Iteration=20000 | 1.75 | 3.64 | 0.86 |
| Log-transformed utterance duration | Iteration=7000 | 7.19 | 0.22 | 5,25 |
| Iteration=20000 | 7.60 | 0.22 | 5.49 |
| Log-transformed total pause time | Iteration=5000 | 334.44 | 0.37 | 7.04 |
| Iteration=20000 | 343.85 | 0.40 | 7.85 |
| Total chunk number | Iteration=5000 | 6.55 | 0.03 | 38.83 |
| Iteration=20000 | 6.34 | 0.03 | 38.32 |
| First chunk length | Iteration=4000 | 1.57 | 0.02 | 24.72 |
| Iteration=20000 | 1.55 | 0.02 | 24.34 |
| Log-transformed  Onset latency | Iteration=4000 | 338.06 | 0.44 | 7.22 |
| Iteration=20000 | 340.22 | 0.47 | 7.36 |

While prior checks (i.e., Responses to *a. Prior Selection*) have showed that the priors we set were suitable, we also conducted sensitivity analyses, following your suggestion (e.g., Nicenboim et al., 2021), to explore whether prior specification influenced BF results, especially for the models with log-transformed utterance duration and log-transformed total pause time. For each effect on each dependent measure, we tried out 10 priors with a range of standard deviations, from 0.1 to 10 (i.e., 0.1, 0.2, 0.4, 0.6, 1, 2, 4, 6, 8, 10), resulting in 10 models. Each model involved 20000 iterations.

As shown in Figure 2 and 3, the evidence from BFs for name agreement effects are slightly influenced by different priors: the evidence becomes weaker as the prior standard deviation increases on log-transformed utterance duration (Figure 2, left), while the evidence is really strong for a range of different values for the prior standard deviation on log-transformed total pause time (Figure 3, left). However, the evidence for semantic context effects is overall weak or null on the two time measures (Figure 2, right; Figure 3, right), which implies that the BFs results for semantic context effects are less likely influenced by different priors. We therefore think that the priors we set were reasonable, but so that readers can decide for themselves, we have also uploaded all these materials to an OSF repository (https://osf.io/6jg4p/). As you suggested, we have elaborated on the appropriateness of our models and priors in the manuscript, see page 18, lines 341-344.

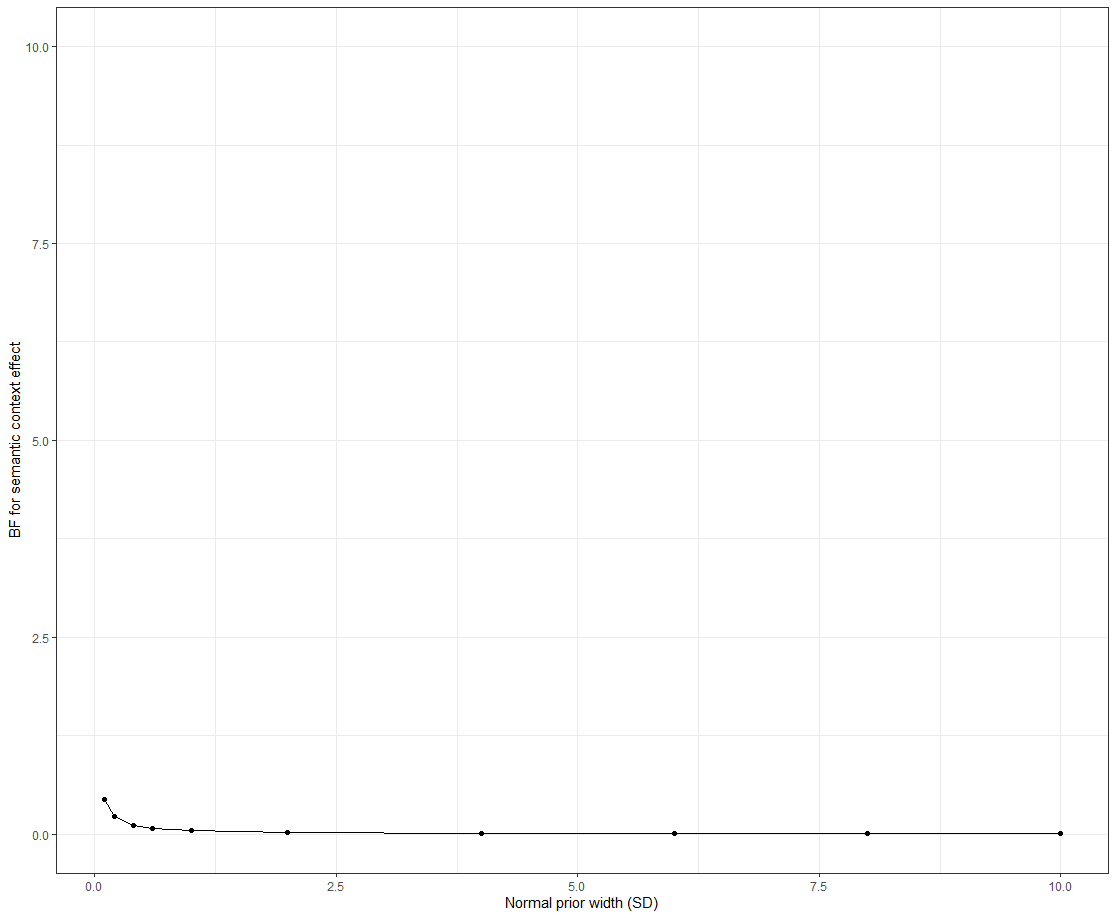
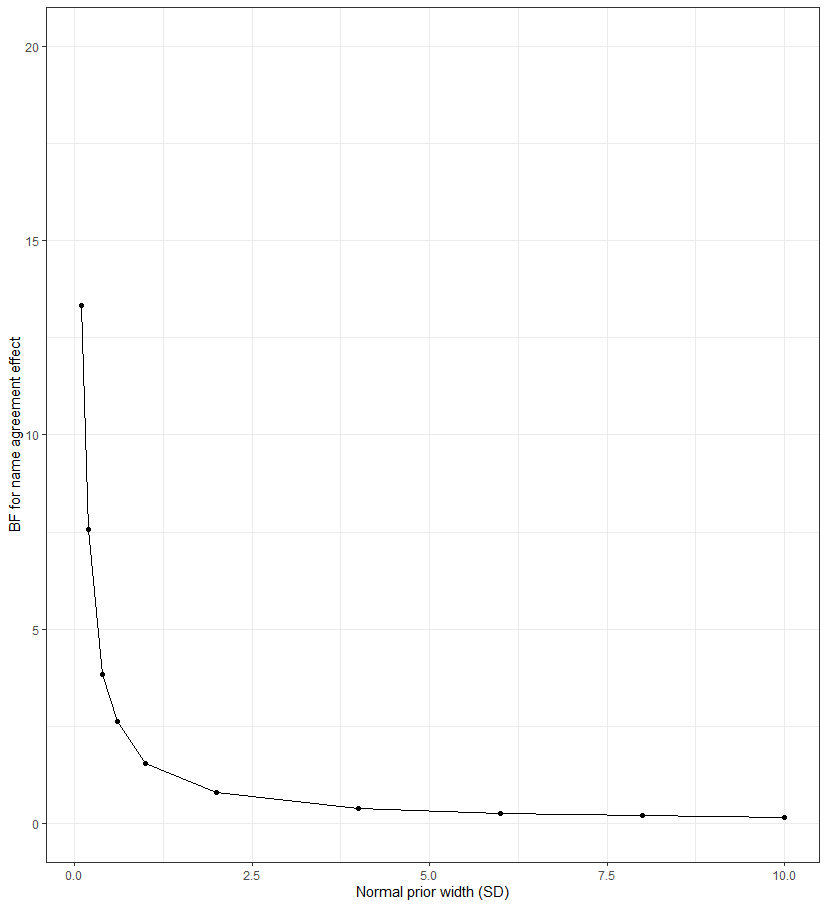


Figure 2. Prior sensitivity analyses for the Bayes factors for the models with log-transformed utterance duration, left figure for name agreement effect, right figure for semantic context effect.

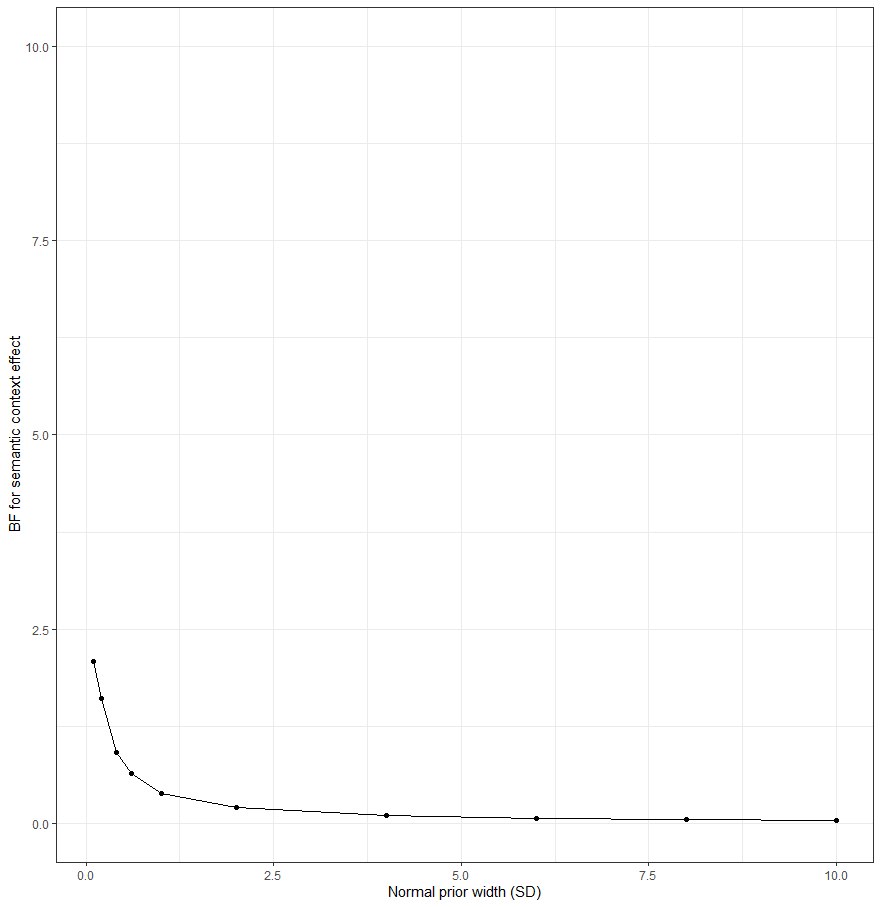
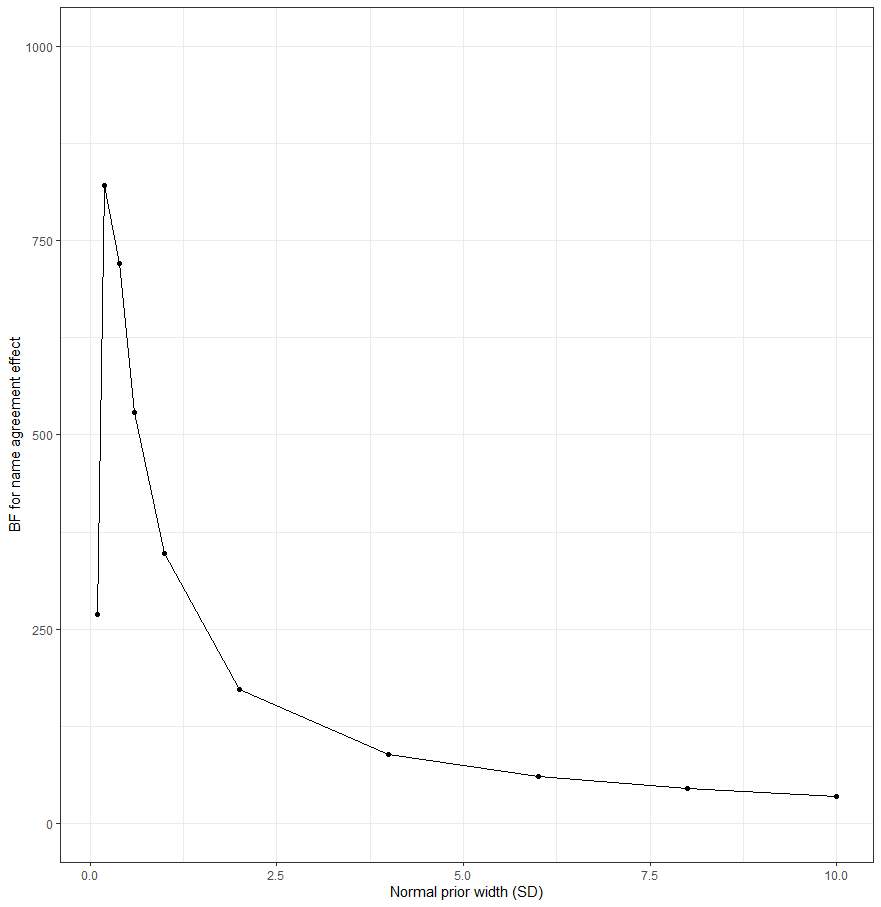


Figure 3. Prior sensitivity analyses for the Bayes factors for the models with log-transformed total pause time, left figure for name agreement effect, right figure for semantic context effect.

c. Model selection: A minor aspect is the selection of the random structure: How did the authors define their reduced random structure? Did they start with a maximal fully-crossed random-effects structure and reduced the structure due to convergence issues or was the decision taken a priori? The approach should be made clear in the manuscript and/or the analysis scripts.

Thank you for your question. We defined our reduced random effects structure based on both data-driven and design-driven considerations. In terms of data-driven consideration, we first checked what the results would be like by using *linear mixed-effect models* for time measures and *ordinal mixed models* for chunk measures. To define the random structures for these models, we indeed started with a maximal fully-crossed random-effects structure, and then simplified the structure by successively removing the by-item or by-participant slopes that accounted for the least amount of variance, until model convergence. We found that the best fitting models were the ones with random intercepts for participants and items, but without random slopes. Given that we were uncertain about noise in our web-based experiment, we used Bayesian data-analytic methods, rather than the frequentist approaches, which would allow us to focus on the best estimate we can obtain of the effects (Nicenboim & Vasishth, 2016, 2018). We thus conducted the Bayesian mixed-effect models and Bayes factors by using the “best fitting” models identified initially (i.e., with random intercepts for participants and items, but without random slopes).

In terms of design-driven consideration, as name agreement and semantic context are both between-item factors, a random intercept for items is usually sufficient (see Barr et al., 2013), which means no random slope was included for items. As we only have four observations for each condition per block for each participant, the random slope variance would be confounded with trial-level error (see Barr et al., 2013), thus a random intercept of participant would be sufficient. Combined, the random effects structure in our model only contained random intercepts for participants and items. Following your suggestion, we have clarified how we define the random structure for all models in Methods, see page 16, lines 303-304.

d. “Warm-up phase”: The authors state that “… each chain had 4000 to 7000 iterations depending on model convergence […], with the first half representing a burn-in period” (p.16). From the analyses scripts I understand that the warm-up phase was only 1000 iterations in each model.

Yes, we did not explain this correctly. The warm-up phase indeed was 1000 iterations in each model. We have changed the paper accordingly, see page 16, lines 306-309.

2. Lack of semantic context effects: Belke and Meyer (2007) in their very similar paradigm found small, but significant semantic context effects on pause rate. The authors of the study do not find evidence for such hypothesized effects, but also only anecdotal to weak evidence against them. The reader may want to wish for a discussion whether these differences may be due to a lack of power and what effect size may have been reliably detected, especially because BF were relatively low. The reliability of the results may also affect the theoretical considerations regarding theories that do (not) assume competition at the lexical level (p. 8/9 and p. 30). Therefore, the authors should discuss the reliability and the inference they make about semantic context effects.

Thank you for your suggestion. Belke and Meyer (2007) found that the semantic context effects on pause rate was only present for the older speakers (aged 52-68 years) but not for the young speakers (18-21 years). This is actually consistent with our results that we observed a moderate semantic context effect on total pause time (β = 0.108, SE = 0.057, 95% Cr.I = [-0.005, 0.22]) but such an effect was argued against by only weak evidence (BF = 0.40). This pattern points to a relatively weak or null semantic context effects on total pause time for the young speakers (aged 19-26 years) in our study. We have now provided more details when discussing Belke and Meyer’s (2007) work in the Introduction (see page 6, lines 89-90) and Discussion (see page 35, lines 588-591).

3. Effect over time: The experiment consisted of only 32 different pictures that were named repeatedly within and across trials (~16 times each). As briefly touched by the authors, this may have affected the results, e.g., through repetition priming. I think this objection should receive further attention. One way could be to analyze whether the effects (for a given stimulus) changed over time, within and across trials. Were the effects in the first block different from those in the third block? This kind of analysis would also be interesting in light of the expected lack of semantic context effects in the first repetition: While the authors did not analyze voice onset effects from the second row for obvious reasons (even though there might also be some smart work arounds allowing to analyze onset latencies from the second row), analyzing them from the second repetition would be easily feasible, highly interesting, and in line with the research tradition.

As you suggested, we explored how the effects changed over time by adding “Block” as a fixed factor in our model. As shown in Table 6, name agreement interacted with blocks on the log-transformed utterance duration, showing that name agreement effects presented in the first block (β= 0.10, 95% Cr.I = [-0.145, -0.051]), but disappeared in the last two blocks. This finding suggests that repetition across trials may eliminate the name agreement effect on the measure of utterance duration. However, we did not find any interaction between semantic context and block, which suggests that semantic context effects are not influenced by repetition across trials.

*Table 6.* Results of Bayesian mixed-effects models for time measures across blocks.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Estimate | Est.error | 95% CI | | Effective samples |
|  |  | lower | upper |
| **Log-transformed speech duration** | | | | | | |
| *Population-level effects* | Intercept | 9.241 | 0.032 | 9.179 | 9.305 | 1223 |
| Name Agreement | **-0.055** | **0.014** | **-0.082** | **-0.027** | **5321** |
| Semantic Context | 0.024 | 0.014 | -0.003 | 0.051 | 5677 |
| Block1 (1 vs. 2&3) | **0.106** | **0.02** | **0.067** | **0.144** | **5261** |
| Block2 (2 vs. 3) | 0.02 | 0.017 | -0.013 | 0.054 | 5314 |
| NA × SC | 0.009 | 0.028 | -0.046 | 0.065 | 5440 |
| NA × Block1 | **-0.09** | **0.039** | **-0.165** | **-0.012** | **5627** |
| NA × Block2 | -0.017 | 0.034 | -0.083 | 0.05 | 5157 |
| SC × Block1 | 0.02 | 0.038 | -0.056 | 0.095 | 5159 |
| SC × Block2 | 0 | 0.034 | -0.067 | 0.068 | 4815 |
| NA × SC × Block1 | 0.012 | 0.073 | -0.133 | 0.153 | 5498 |
| NA × SC × Block2 | -0.029 | 0.064 | -0.157 | 0.097 | 6039 |
|  |  |  |  |  |  |  |
| *Group-level effects* | Participant\_sd(Intercept) | 0.189 | 0.023 | 0.150 | 0.241 | 1809 |
| Item\_sd(Intercept) | 0.043 | 0.001 | 0.032 | 0.058 | 3492 |
|  | | | | | | |
| **Log-transformed total pause time** | | | | | | |
|  | Intercept | 7.644 | 0.1 | 7.448 | 7.839 | 1388 |
|  | Name Agreement | **-0.254** | **0.054** | **-0.361** | **-0.145** | **6954** |
|  | Semantic Context | **0.111** | **0.054** | **0.004** | **0.219** | **7165** |
|  | Block1 (1 vs. 2&3) | **0.206** | **0.077** | **0.054** | **0.358** | **6991** |
|  | Block2 (2 vs. 3) | 0.041 | 0.067 | -0.088 | 0.172 | 6535 |
|  | NA × SC | 0.061 | 0.109 | -0.154 | 0.275 | 6978 |
|  | NA × Block1 | *-0.302* | *0.154* | *-0.608* | *0.006* | *7121* |
|  | NA × Block2 | -0.027 | 0.134 | -0.289 | 0.236 | 7287 |
| *Population-level effects* | SC × Block1 | 0.018 | 0.15 | -0.273 | 0.315 | 6521 |
| SC × Block2 | 0.008 | 0.132 | -0.247 | 0.271 | 6401 |
| NA × SC × Block1 | 0.068 | 0.297 | -0.525 | 0.645 | 6979 |
| NA × SC × Block2 | -0.076 | 0.261 | -0.585 | 0.436 | 6617 |
|  |  |  |  |  |  |  |
| *Group-level effects* | Participant\_sd(Intercept) | 0.590 | 0.074 | 0.469 | 0.753 | 1951 |
| Item\_sd(Intercept) | 0.166 | 0.026 | 0.121 | 0.224 | 3594 |

Note: we defined two contrasts for the factor of Block: one was the 1st block versus 2nd and 3rd blocks (1 vs. 2&3), the other was the 2nd block versus 3rd block (2 vs. 3).

4. Minor issues

1) I would appreciate if the authors made it clearer which analyses were planned a priori and which post-hoc. E.g., were the analyses from the second row only planned before data collection? Were the participant exclusion criteria set before or after data collection?

Following your suggestion, we have described whether the analyses were planned prior or post-doc in more details, see page 18, line 345; page 19, line 359.

2) Inferences regarding the audio-visual synchrony: In their introduction (p.3), the authors summarize results from Bridges et al., (2020) by concluding that “web-based experiments currently cannot offer the audiovisual synchrony (i.e. simultaneous presentation of visual stimuli and audio recording) needed to accurately record onset latencies in language production”. I think this claim is not fully warranted as Bridges et al. (2020) did not test the synchronization of visual stimulus presentation and onset of audio recordings, but rather the synchronization of visual and auditory stimuli.  
In the same vein, in their interpretation of speech onset latency effects, the authors conclude “in spite of the encouraging results of this particular study, we advise against planning web-based experiments relying exclusively on speech onset latencies unless precise timing of stimulus presentation”. Based on the focus of the study and the study results, this conclusion is not fully warranted. The authors, like other recent studies, overall find evidence for a precise-enough audiovisual synchrony to replicate lab-based language production effects (e.g., Fairs & Strijkers, 2021; Stark et al., 2021; Vogt et al., 2021) and Bridges et al. (2020), who report limited accuracy, tested the synchronization of visual stimuli and sounds presented via speakers instead of accessing the microphone.

(However, to additionally test the reliability of the speech onset measurements, it would be interesting, to analyze whether the speech onset effects in the current study were consistent within a single participant.)

Thank you for your comment. We do agree with you that our inference regarding the audio-visual synchrony is not fully warranted any more. Since there was no literature available showing precise-enough audiovisual synchrony in web-based language production study when we planned our study, we assumed that we might encounter similar audio-visual synchronization issues to Bridges et al. (2020). However, based on the studies that have come out since we began this research (e.g., Fairs & Strijkers, 2021; Stark et al., 2021; Vogt et al., 2021) and our finding of strong name agreement effects on onset latencies, we agree that we can now be more optimistic about the reliability of onset latencies in web-based studies. Thus, we revised the first two paragraphs of Introduction by highlighting the historical context of planning our study, i.e., there were concerns about the feasibility of onset latencies because little relevant work was available at that time (see page 3, lines 29-45). We also revised the conclusion on page 29 (now page 33) by claiming that onset latencies are much more reliable than we thought (see lines 547-560).

You asked about the consistency of the effects \*within\* participants. We were not sure how to address this comment. The below graphs show the effects across participants and there is quite some variation (range: -181~376 ms for name agreement effect; -200~573 ms for semantic context effect). To determine within-participant consistency we could determine the effects per participant and block. This will surely also show substantial variation. However, we have no data from parallel lab-experiments, and therefore do not know how to evaluate the variation in our experiment. We have therefore not incorporated such analyses in the paper.

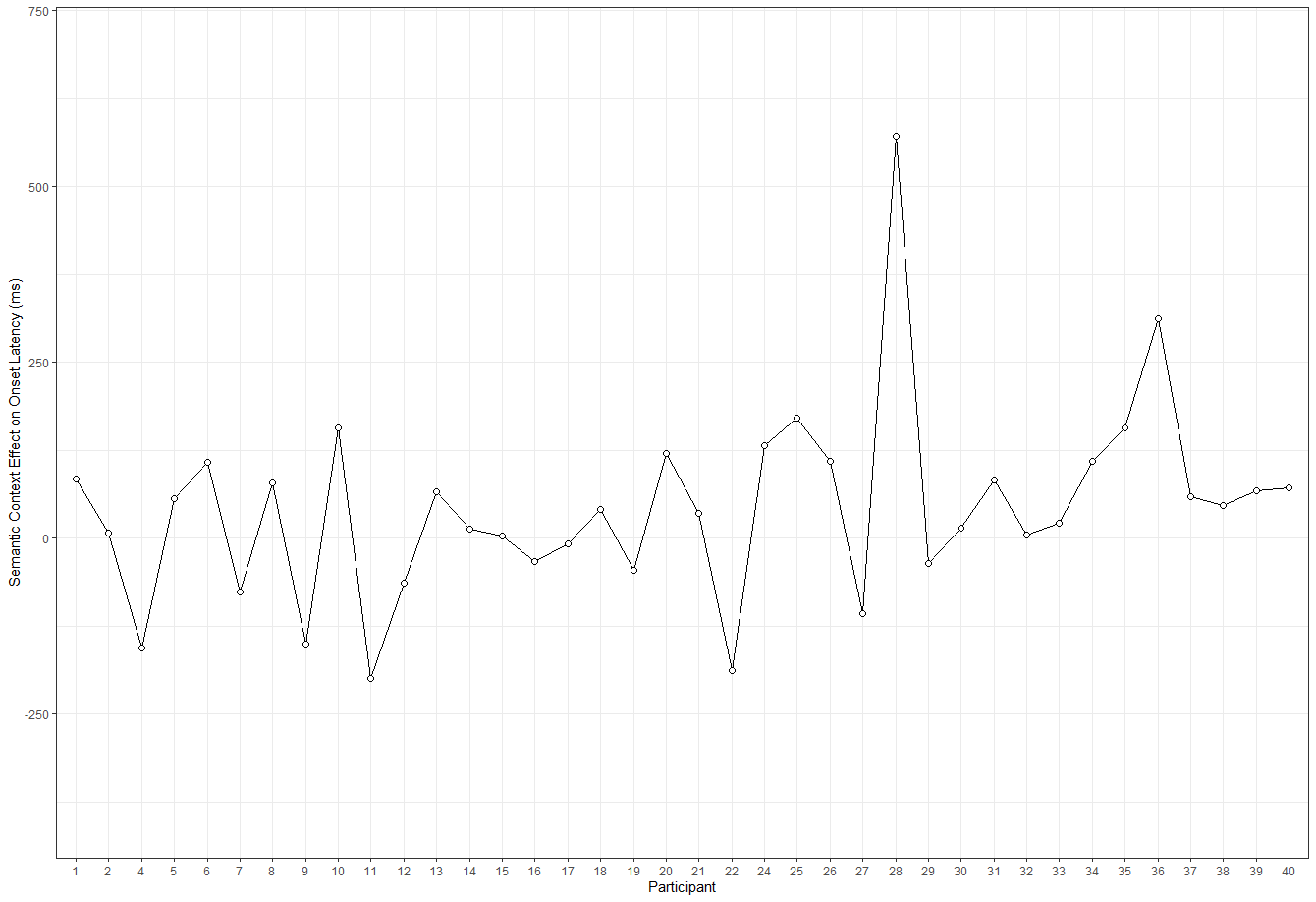
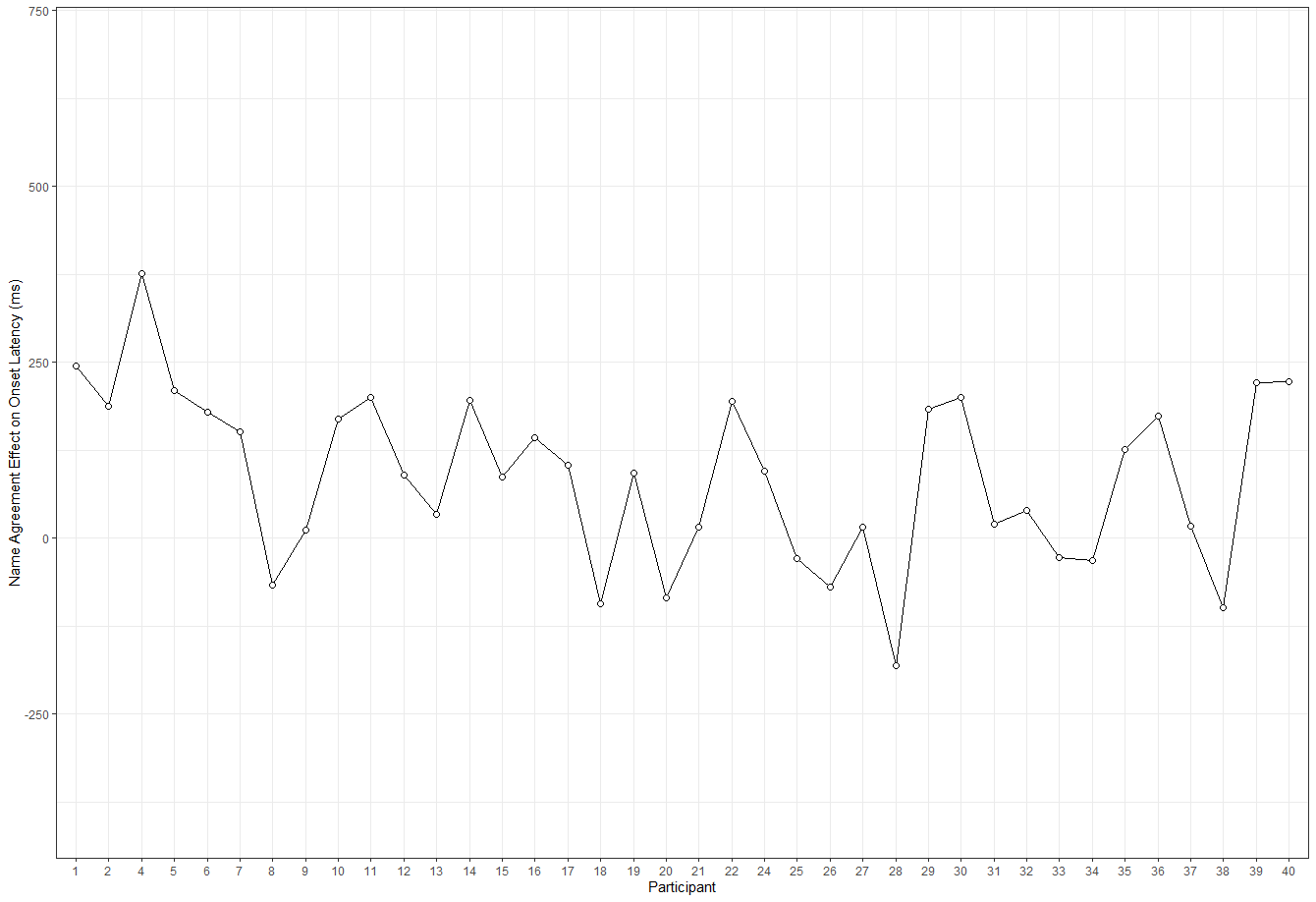


Figure 4. *Effects of name agreement (low - high; upper) and semantic context (homogeneous - heterogeneous; below) on onset latencies across participants*.

3) Sample size estimation (p. 9): The selection of the sample size seems rather arbitrary. It would be great if the authors at least gave an example of the previous semantic context experiments they based their estimations on.

Thank you for your suggestion, we have provided examples of previous semantic context experiments for the selection of participants, see page 9, line 171.

4) The apparatus section (p.9/10) is very short. As next to presenting a modified paradigm, the authors add to the still very little and recent evidence on browser-based speech production research, I think it would be valuable to give additional detail on the programming part of the experiment: E.g., where can the reader find additional information about the platform FRINEX? How did the voice recording work? This would give readers seeking to conduct browser-based speech production a head-start. Furthermore, only additional technical information, in combination with other, recently published speech production experiments, would allow to make informed decisions about the audio-visual synchrony and (computational) aspects that may influence it. This is a personal preference the authors are welcome to disagree with.

Thank you for your suggestion. we have added a citation for FRINEX which provides details for programming on the FRINEX web-based platform, see page 10, line 180.

5) Materials (p.10).  
a. The stimulus set was composed of 32 pictures “with one- or two-syllable names”. The stimulus list in the Appendix suggests that this was not the case for all stimuli with low name agreement (e.g., “inktvis / octopus”). The authors may want to clarify.  
b. Where do the name agreement ratings (p.10) come from (e.g., from the norms/the experiment/pretests)?

We have clarified these two issues in the Methods section, see page 10, line 189, line 191.

6) Materials and design (p. 10-12): I had difficulties understanding the trial structure. I very much appreciate that the authors provide a link to an exemplary experiment, but I think the description could be more straightforward by making a clear differentiation between the terms “picture”, “picture grid”, and “items”. Providing the exact number of trials, the number of times each participant saw each picture grid and the number of times each participant named each individual picture would also help. Furthermore, I think the manuscript would benefit if Figure 1 contained an example of a heterogenous picture grid, too.

Thank you for your suggestion. We have made a clear differentiation between these terms and also provided the exact numbers, see Materials section on pages 11-12.

We have re-plotted the trial structure by including the four conditions of our study, see Figure 1 on page 14.

7) Analysis + Results: What’s the difference between silent intervals and pauses?

The two terms refer to the same concept so we have replaced *silent intervals* with *pauses*, see page 15, line 280.

8) The authors switched the APA citation style on p. 32

We have corrected this, see page 36, lines 622.

9) The authors could further streamline their introduction and discussion. Although the manuscript is well written and a nice-read, the authors could do a better job in carving out their study’s major contribution: The novel paradigm? The web-based setup? The novel response variables? Then they may want to provide somewhat more detail on the background of their major focus (e.g., regarding recent attempts of web-based speech production research and technical details that may prove necessary for conceptual replications, relating the paradigm to other paradigms, an evaluation of the usefulness of the different response variables). This is an advice the authors may or may not want to follow, but that could further improve the reading flow of the manuscript.

Following your suggestion, we have slightly revised the Introduction in the light of these comments.

**Reviewer 2**

1. The premise of the manuscript is overly pessimistic and does not seem to be based on sound empirical evidence. As a consequence, the authors’ main recommendation to avoid using onset latencies seem to be unwarranted. The authors claim that onset latencies “cannot offer the audiovisual synchrony needed to accurately record onset latencies in production experiments.” This is probably true, but the question is how much this matters. Literally all measurements have errors, so the real question is not whether there is a noise in measurement, but how big the noise is and if that noise is biased (not equal across conditions). As long as the noise generated by hardware is random, the challenge can be overcome (and has been overcome multiple times, indeed by the authors themselves for the naming agreement effect) relatively easily by collecting larger number of participants. In fact, recent studies by Strijkers’ group (<https://psyarxiv.com/2bu4c>) show that the classic effects in onset latencies in single word naming is comparable between lab-based and online studies. Our lab got 20-40ms semantic interference effect in picture-word-interference-like task about ten times in online studies. So I think the recommendation to avoid using onset latencies is simply empirically unfounded.

Thank you very much for your comments. We agree that we can now be more optimistic about the reliability of onset latencies in our web-based studies. The paper submitted for reviewer feedback reflects our state of knowledge at the time when the study was planned. We thank you for the references to studies that were conducted in parallel with our work, and which we did not know of. In the Introduction and Discussion, we now place our study in its historical context, remove the claim about avoiding using onset latencies, and highlight that similar work was conducted in parallel with ours (see page 3, lines 29-45). We then discuss the conclusions to be drawn from this body of work, see page 33, lines 547-560.

In addition, onset latencies associated with multi-word utterances certainly have larger variability (= larger amount of noise) than those associated with single word utterances, because there are more processes involved in generating complex utterances (and this problem is in addition to the problem associated with simultaneous picture presentations discussed by the authors). Thus, to make the recommendation to use multi-word utterances, the author must demonstrate that the amount of noise introduced by hardware variabilities in single word utterances is larger than the amount of noise introduced by making the target utterances multi-words. For those reasons, I do not believe the authors’ recommendation to use multi-word utterances is justified, and it has the potential to mislead researchers in the field.

Thank you for your comment. As noted, we now don’t advise against the use of speech onset latencies any more, and we also acknowledge that measures derived from utterance durations and pauses may be noisier than speech onset latencies.

see pages 32-33, lines 535-544.

2. When talking about how speech duration/pause rate can be used to study planning processes, I think the authors should cite many previous studies that used the duration measure to study planning processes, including: Ferreira & Swets (2005, a book chapter on resumptive pronouns), Kandel et al (CUNY: <https://www.cuny2021.io/wp-content/uploads/2021/02/CUNY_2021_abstract_131.pdf>) Momma & Ferreira (2019, Cognitive Psychology) among others.

Thank you for your suggestion, we have cited the literature you suggested, see page 4, line 56-57.

3. The authors cited Wheeldon & Lahiri, 1997 in the context of justifying that planning of multi-object naming utterances is incremental. I think this is an odd choice; Wheeldon & Lahiri I believe showed that phonological planning is incremental. The authors were specifically concerned about the incrementality at the level of lemma selection (non-phonological planning), so I am not sure if W&L is relevant in the context. (I believe it should be cited elsewhere in the paper, but not in p. 8). Perhaps the more relevant citation is Griffin (2001), which showed that naming agreement effect in multi-word utterance is absent in onset latency measures. More generally, I think the authors should be careful about what levels of processing they are talking about when making a claim about ‘incrementality.’ Relatedly, I think the authors at least should mentions previous studies showing that lemma/conceptual planning in multi-word object naming may not be as incremental (e.g., Meyer, 1996, JML).

Thank you for your suggestions. We have removed the reference to Wheeldon and Lahiri’s work and now provide a slightly longer discussion of advance planning in the current paradigm (see pages 4-5, lines 57-68). We acknowledge that speakers can flexibly adjust their span of advance planning (e.g., He et al., 2021). However, it is also clear from the literature that when speakers have to name more than two or three objects they are very unlikely to fully prepare the utterance before speech onset. Hence, utterance durations may provide information about the time required for speech planning.

4. The authors wrote “A power simulation for determining sample size before the present study was not possible, as no comparable studies were available. However, the data reported here can now be used for a power simulation to estimate the sample size needed to observe effects of interest in future work.” If this is the claim, I think the authors might as well provide do the power calculations and include the results in the manuscript?

Following your suggestion, we have reported the results of post-hoc power calculations for the semantic context effects and the interaction between name agreement and semantic context on time measures in Results section, see pages 29-30 lines 484-503 , Figure 5; also see Appendix C. We also rephrased the corresponding text in our manuscript, see page 36, lines 611-616.

5. The methods used in this study seems to be mostly fine. I think there is an issue of how simultaneous picture presentations affect interference effects, but I believe the issue is adequately discussed.

Thank you very much.

**Editor Final Decision: Revise & Resubmit**

Nov 1, 2021

Dear Jieying He,

I am writing regarding your revised manuscript you submitted to Collabra: Psychology. Thank you for your careful attention to the concerns the reviewers and I raised. I am happy to provisionally accept your manuscript for submission.

Before final acceptance, I encourage you to consider a couple minor suggestions raised by Reviewer 1 in a (very) minor revision. In particular, I agree that a few details of Figure 1 could be clarified and also that you may want to consider toning down the title a little bit (as you are presumably suggesting *one* way to conduct language production research online, not the *only* way).

In any case, I look forward to receiving your final revision and accepting it for publication in Collabra: Psychology.

Please ensure that your revised files adhere to our author guidelines, and that the files are fully copyedited/proofed prior to upload. Please also ensure that all copyright permissions have been obtained. This is the last opportunity for major editing, therefore please fully check your file prior to re-submission.

If you have any questions or difficulties during this process, please contact the editorial office at [editorialoffice@collabra.org](mailto:editorialoffice@collabra.org).

We hope you can submit your revision within the next six weeks. If you cannot make this deadline, please let us know as early as possible.

Sincerely,

Bob Slevc

**Reviewer 1**

**Open response questions**

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

The authors have done a good job in revising the manuscript and have satisfactorily responded to my comments and suggestions. Especially, the additional explanations regarding the trial structure makes the methods and results much clearer. The additional information provided regarding the procedure of their analyses makes them easier to evaluate. Only very minor questions came up while going through the revisions.

Unfortunately, the authors did not copy the changes they made to the manuscript to the response to reviewers and the line and page indications mostly do not correspond to neither the word nor the pdf version of the manuscript, making it unnecessarily difficult to review their changes. I would highly recommend to take care of correct references in the revisions of their next papers as it would make reviewer’s work much easier.

Furthermore, in the response to reviewers, the additionally provided Figure 1 is difficult to interpret (response 1a.). I believe the authors that the figure suggests the priors are suitable (which is also suggested by Table 1), but the different x-axis makes it impossible to compare the different distributions.

The adapted Figure 1 of the manuscript is very helpful (minor comments, point 6). But the figure suggests that the order of the stimuli was always 1234 – 2341 – 3412 – 4123 in homogenous blocks and 1234 – 4123 – 3412 – 2341 in heterogenous blocks. If this was the case, this should be noted somewhere in the manuscript as this “rhythmic” presentation might affect chunking.  
In the figure note of Figure 1, I don’t understand the info provided in the parentheses (“mouse and vulture in the examples”). As far as I understand it, the other stimuli also come from homogenous sets of other categories.

I appreciate that the authors now embed their findings in the context of recent studies providing evidence that the assessment of vocal onset latencies in web-based settings is reliable enough to measure classic speech production effects (minor comments, point 2). However, I think that the authors now need to make it clearer what the additional contribution of their findings is and in which contexts such multi-word production paradigms might be more (or less) suitable than classic picture naming paradigms.  
I also think that, given the reduced claim of the paper, the title is a slight overstatement as it suggests that multi-word paradigms are the way to go if language production research is to be conducted online. It might be a matter of taste but I think the point of the paper is not that this is the online way to conduct such research in browser-based settings but rather one possibility (and a smart way to circumvent some problems, especially in online settings).

p.3, l. 39 – did not depend on  
p.12, l. 228 – during the experiment  
p.17, l. 324 – with 1 degree of freedom  
p. 21, l. 398 – was significantly longer  
p. 49, l. 938 (and other figure notes) – context effects (left) (space missing)

**Rating scale questions**

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  | ✔ |  |  |

**Reviewer 2**

**Open response questions**

Please write your review here. The author(s) will see this review. Your identity will not be revealed to the authors unless you also include your name (i.e., sign your review) in this box. It is up to you whether to reveal your identity or not, either is fine.

I believe that my major concerns in my previous review (espeically the authors’ description of the reliability of onset latency measures in web-based studies) is adequately addressed. The manuscript also now include the discussion/citation of the recent web-based studies replicating classic effects, so it provides a good overview of the current state of the field. The post-hoc power analysis also provides a (somewhat sobering) guidance on the adequate sample size for detecting the semantic context effect and its modulation in web-based studies, which would be useful for future web-based studies. Just like in my previous review, I do not have much concerns about their analysis procedures and experimental design. Overall, I recommend that this manuscript be accepted for publication.

**Rating scale questions**

|  | Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree |
| --- | --- | --- | --- | --- | --- |
| The study/studies in this manuscript have strong construct validity (good measures and/or manipulations of the constructs the authors wish to study). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong statistical validity (appropriate statistical tests, assumptions are clear and reasonable, no statistical errors, appropriate statistical inferences, etc.). (Choose “Neutral” if this is not an empirical manuscript) |  |  |  |  | ✔ |
| The study/studies in this manuscript have strong internal validity (any causal claims or implications are well-justified, alternative explanations are thoroughly considered, etc.). (Choose “Neutral” if this is not an empirical manuscript, or no causal claims are made or even vaguely implied.) |  |  |  | ✔ |  |
| The study/studies in this manuscript have strong external validity (authors appropriately constrain their conclusions based on the limits of the generalizability of their findings to other contexts (including from lab to real world), other populations, other stimuli or measures, etc.) |  |  |  | ✔ |  |

**Author Response**

Nov 10, 2021

Dear Dr. Slevc,

Thank you very much for your letter dated November 1st, 2021, and the reviewers’ valuable comments on our paper. We have made modification of the manuscript by revising the title, providing more details of the order of stimuli, and discussing additional contribution of our findings following your comments. We have fully checked our files following the author guidelines.

We have submitted two versions of our revised manuscript (one with the new text printed in red; the other is the final version) and a letter of response to reviewers. We are looking forward to hearing from you.

Sincerely,

Jieying He

Max Planck Institute for Psycholinguistics

**Editor Final Decision: Accept**

Nov 15, 2021

1. The mean utterance duration is 10577 ms, and because we assumed a lognormal distribution, to calculate the ms, we need to find out exp(log(10577)+0.2)-exp(log(10577)-0.2), which is 4259 ms. [↑](#footnote-ref-1)