**EDITOR'S NOTE: Please be mindful of the page length guidelines of 35 pages inclusive of everything except tables and figures. We are fine with going over this a bit because reference sections are longer in meta-analyses.**

*We have removed duplicated references to reduce the page limit.*

Reviewers' comments:  
  
Reviewer 1: Thank you for the opportunity to review this paper. I appreciated the  
revisions that the authors undertook, and I continue to see this as a  
useful paper. I have a few quite minor comments or edits to suggest,  
but otherwise will be pleased to see this in print.

*We would like to thank the reviewer for their helpful comments.*

Fig. 3: The caption (and others) still refers to colors, here orange  
and blue. (I think having colors would be fine -- I only thought it  
would be better if the colors also worked without color. I'm on  
board with varying both luminance and hue.)

*We have updated the captions referring to the correct colors, which are now yellow and dark blue, which should contrast with one another when printed in black and white.*

I also thought the plots could be a little taller, and the titles  
("Object Recognition", ...) less huge.

*Figure 3 has been made as tall as possible. The full panel title has been removed from Figure 3, as this was deemed unnecessary, and the font here reduced in size.*

pp. 26-27 "This could lead to an increase in significant results and  
even alter the developmental trajectory of mispronunciation  
sensitivity." It wouldn't really alter the developmental trajectory  
as much as it would alter the \_apparent\_ developmental trajectory.

*This sentence on page 25, lines 633-635 has been changed to read:*

“This could lead to an increase in significant results and even alter the measured developmental trajectory of mispronunciation sensitivity measured in experiments.”

p.31 discusses why children do not show stronger effects of  
mispronunciations when the distracter picture is an unfamiliar object.  
The paper indicates that "no studies have directly examined this  
assertion", which is true. This is probably because when children  
hear a minor mispronunciation, they do not think that they are hearing  
a novel word at all, or at least not a word for another object. This  
has been shown repeatedly (Dautriche, Swingley, & Christophe, 2015;  
Swingley & Aslin, 2007; Swingley, 2016; perhaps also White & Morgan,  
2008). Given this work, one would not expect that a word-learning  
effect would boost the size of mispronunciation effects when a novel  
object is available.

*We have added the following sentence to the end of this paragraph on page 31, lines 789-793:*

“One possible explanation is that when the size of the mispronunciation is small (e.g. 1-feature change), infants are unlikely to map this label onto a novel object and even seem to be biased against doing so (for evidence from infant word learning see Dautriche, Swingley, & Christophe, 2015; Swingley, 2016; Swingley & Aslin, 2007).”

p. 33 says: "Looks to the target in response to mispronunciations  
may be slower than in response to correct pronunciations in infants  
(as predicted by TRACE, Mayor & Plunkett, 2014)..." This fact is  
certainly consistent with TRACE modeling, but it's strange to call it  
a prediction, since it was already shown by Swingley & Aslin in 2000.

*We have reworded the beginning of this sentence on page 33, lines 829-831 to read:*

“Looks to the target in response to mispronunciations may be slower than in response to correct pronunciations in infants (Mayor & Plunkett, 2014; Swingley & Aslin, 2000)..."

p. 34 "A lack of a field standards can have serious consequences," --  
the first part of this sentence is not grammatical. Isn't this also  
more of a technique than a field? I guess I think of psychology as a  
field. Or maybe developmental psychology. How about something like  
"Variation in measurement standards can ..."

*We appreciate this suggestion and have used it to rewrite the beginning of this sentence on page 34, line 859:*

“Variation in measurement standards can have serious consequences…”

p.34 recommends growth curve analysis as a methodological  
improvement. I haven't followed this closely, but I have heard a few  
discussions that this method has a high false positive rate because it  
doesn't model the non-independence of sequential data points. I don't  
know the truth of the matter about this, but I suppose if making  
recommendations, the authors would want to have a view of this. (and  
-- perhaps they do, and have reassured themselves this isn't an  
issue.)

*Since the last revision of our manuscript, we have also become aware of the criticism proposed against growth curve analysis. We have therefore added the following information to this sentence on page 35, lines 874-877:*

“Both Growth Curve Analysis (Mirman, Dixon, & Magnuson, 2008) and Cluster Permutation Analysis (Maris & Oostenveld, 2007; Von Holzen & Mani, 2012) offer

potential solutions to analyze the full time course (although Growth Curve Analyses are not without criticism, see Huang & Snedeker, 2020).”

Bibliography: the entire bibliography is given twice, at least in my  
copy.

*We thank the reviewer for pointing this out and have removed the duplicated bibliography.*

Reviewer 3: This manuscript presents a well-designed meta-analysis on infants' sensitivity to mispronunciations (e.g., Swingley & Aslin, 2000), an oft-studied and theoretically relevant topic in the field of language development. The authors included records from 32 studies that measured infants' mispronunciation sensitivity and we believe that the field will be interested in the findings of this meta-analysis. We found the meta-analysis to be well-executed and the authors were responsive to the previous reviews. Overall, we believe that with some minor revisions (as suggested below), this work will make a valuable contribution to both the literature on mispronunciation sensitivity and to the growing call for best practices in open science.

*We would like to thank the reviewer for their helpful comments.*

General Comments:  
We thought that the authors satisfactorily responded to the majority of the reviewer's comments. We have a few additional considerations and points of clarification, some of which were mentioned by the previous reviewer, that we believe would strengthen the manuscript.  
  
On Page 6 on the manuscript (Lines 109-111): We do not understand what the authors mean here by developmental trajectory. Their literature review suggests that both infants and adults show graded sensitivity to mispronunciation size, so it's unclear how this sensitivity changes over development, as a trajectory would imply.  
Similarly, on Page 24, starting on Line 591, the authors state that:  
  
"The developmental trajectory of mispronunciation sensitivity was influenced by type of mispronunciation and overlap between target and distractor labels, but mispronunciation size, mispronunciation position, and distractor familiarity were found to have no influence."

Their size of mispronunciation analysis showed graded sensitivity, in line with previous infant and adult studies, but there was no moderator of age.  
There are a few more instances where the authors talk about developmental trajectory that are vague. On Page 26, Line 623, the authors say that "… could lead to an increase in significant results and even alter the developmental trajectory of mispronunciation sensitivity." This wording is not precise as it could be—data analysis decisions could influence reported results, not developmental trends (if a true developmental effect exists). We do feel as if the authors do a nice job of talking about how their analyses can test for developmental trajectories in the General Discussion (Page 29, Lines 707-715), and their summary here treats development with the proper precision. We would like to see this level of consideration when using this term in other contexts in the manuscript.

*We appreciate the reviewer encouraging us to be more precise in our references to developmental trajectory. We have now edited these sections referencing developmental trajectory as follows:*

*Page 6, lines 120 – 122, we have replaced the last sentence with:*

“By aggregating studies testing infants of different ages on mispronunciations of varying size, this also has consequences for identifying any graded sensitivity changes over development.”

*Page 24, lines 603 – 606:*

“When age was added as a moderator, mispronunciation sensitivity was found to vary by type of mispronunciation and overlap between the target and distractor labels over development, but age did not influence sensitivity to mispronunciation size, mispronunciation position, and distractor familiarity.”

*Page 25, lines 633-635:*

“This could lead to an increase in significant results and even alter the apparent developmental trajectory of mispronunciation sensitivity measured in experiments.”

On page 29, in agreement with the original reviewer, we have a few further comments regarding their discussion of vocabulary size measures. On Lines 718-722, in addition to researchers' lack of interest in vocabulary size, another reason that only 13 papers reported vocabulary correlations is that previous researchers may have not reported null results. Would a funnel plot including only the 13 papers that report vocabulary be helpful to probe if there's a publication bias for significant vocabulary correlations? We think that non-significant vocabulary correlations in file drawer are just as likely an explanation as a reduction of interest by researchers.  
  
To the extent that the authors believe that vocabulary size is of theoretical interest, which we believe it is, it might be worthwhile to suggest that researchers include infants' vocabulary sizes in all publications, regardless of their significance. Further, they might suggest that future work includes vocabulary size as a continuous variable in statistical mixed-models, which would be a more robust test of the relationship between vocabulary size and mispronunciation sensitivity.

*Although we appreciate the reviewer’s suggestion to conduct a funnel plot analysis on papers that reported vocabulary scores, considering both funnel plots already show evidence of publication bias, this reduction is unlikely to show a different pattern. We are unfortunately unaware of a metric for comparing the asymmetry of funnel plots. All of our data and analysis scripts will be publicly available after publication, and the interested reader could generate this plot.*

*In a previous revision of this manuscript, we were asked to tone down our assumptions of what authors are doing in regards to vocabulary measurements. We do fully agree that the lack of reported vocabulary measures and/or correlations may be due to a file drawer issue. We have therefore added these sentences on page 29, lines 729 – 739:*

“On the one hand, this may reflect a decreasing interest in the relationship between mispronunciation sensitivity and vocabulary size and/or to invest in data collection that is not expected to yield significant outcomes. On the other hand, non-significant correlations between mispronunciation sensitivity and vocabulary size may be more likely to not be reported, reducing our ability to uncover the true relationship (Rosenthal, 1979; Simonsohn et al., 2014). Considering the theoretical importance of infants’ vocabulary size, however, more experimental work investigating and reporting the relationship between mispronunciation sensitivity and vocabulary size, whether the relationship is significant or not, is needed if this link is to be evaluated. We encourage researchers to measure and report infants' vocabulary size in future studies.”

*And page 35, lines 880 – 883:*

“As mentioned previously, one example of standardization would be for all experimenters to measure and report vocabulary size.”

Similarly, to the reviewer's comments about the colors of the figures on the previous draft, we had a difficult time reading a few of the figures in greyscale. If there is a way to create more contrast, it would make their figures easier to interpret. This is especially the case in Figure 3—the mispronunciation line blends in with the error visualization. Also, in the figure caption for Figure 3, the colors are still labeled as "orange" and "blue". The greyscale contrast concern and the color naming in the captions persists in some of the other figures, too (e.g., Figure 4).

*We have updated the captions referring to the correct colors, which are now yellow and dark blue, which should contrast with one another when printed in black and white.*

Proofing comments:  
Line 48: "infants' initially episodic representations" ◊ "initial" might sound better! We get the gist; it just reads a little strange to us.

*This has been changed.*  
  
Lines 187-188: There seems to be a problem with numbers here or how the authors came up with 299 items needs to be made more explicit. I think that the authors removed 99 duplicated between the expert list and the google search list, but this isn't totally clear.

*We have now reformulated this section to make our selection process more explicit, as well as updated an error in Figure 1:*

“We first generated a list of potentially relevant items to be included in our meta-analysis by creating an expert list (see Figure 1 for an overview of the selection process). This process yielded 110 items. We then used the Google Scholar search engine to search for papers citing the original Swingley and Aslin (2000) publication. This search was conducted on 22 September, 2017 and yielded 288 results. From this combined list of 398 records we removed 99 duplicate items and screened the remaining 299 items for their title and abstract to determine whether each met the following inclusion criteria…”

Line 255: typo, maybe remove 'are' in sentence "…. studies [are] typically examine…"

*This has been changed.*  
  
Line 483: missing period

*We were unable to find what the reviewer is referencing in this comment and hope that the typesetters spot any missing punctuation we missed.*

Line 576/577: The authors claim that studies with younger children did not control for overlap between target label and distractor label. However, we believe that most of these studies intentionally did not have overlap. These seems like different things.

*The information given in the referenced sentence on page 24, lines 586-589 refer to the results of Fischer’s exact test, which evaluates the independence of infants’ age group and whether they received stimuli in which the target and distractor labels overlapped or did not. We cannot infer intentionality of experimenters using this test.*

*However, we have removed the word “control” from the mention of the results of this analysis in the Discussion, on page 32, lines 808 – 810:*

“Contrary to predictions made from the literature, our meta-analysis revealed that studies which include target and distractor images that overlap in their onset elicit greater mispronunciation sensitivity than studies in which these labels do not overlap.”

Line 747: factors is spelled incorrectly (facturs)

*This has been changed.*