Editor:  
  
The reviewer provided a thorough and comprehensive review, so I will note belabor or reiterate it.  Nevertheless, I do think it may be useful to highlight some areas of consistent concern between my own reading of the manuscript and that of the reviewer.

1. I think the most important of these issues is actually one of tone - there are a few points in the manuscript where you go beyond describing your data to imputing motives on the part of the researchers whose work you are reviewing.  I think it's useful to lay out possible hypotheses about why your data look as they do, but I think it's also important to note that you cannot make claims about researcher motivation with any kind of certainty.  And, as the reviewer notes, there are potentially multiple explanations for many of these phenomena, so it may be more useful to discuss a range of possibilities, rather than subscribing to a particular favorite or preferred interpretation.

*We appreciate the Editor underscoring the importance of tone highlighted by Reviewer 1. We have made changes to our manuscript to address issues in tone, specifically related to post-naming analysis windows (which is supported by the asymmetric funnel plots for correct pronunciations in Figure 2) and decreasing number of studies measuring vocabulary.*

1. A second critical conceptual concern, and one a bit more closely tied to the theory (rather than the sociology) of the literature you are discussing is that your characterization of the literature was not always as precise as it might be.  The reviewer provides two in depth examples of questionable characterizations of prior methods, though I think it would do to give consideration to this concern throughout the introduction.  From my own perspective, this relates to an interesting theme that I would have liked to see discussed more thoroughly (though I can understand why it was not the focus of this manuscript) which is differences among the methods and stimuli employed by various labs to address these questions about phonological and lexical information.  It seems plausible that there is a significant amount of variance to be explained by these kinds of procedural differences.

*We have edited the Introduction carefully with the suggestions of the Editor in mind. In addition to the suggestions made by Reviewer 1, we have tried to add in more details about the different sources of variance that we examine in our moderator analysis. For example, to address the specific request made by the Editor here, we have edited the paragraph introducing the goal of the moderator analysis in the Introduction:*

“But, this may not account for all variability found in the literature. Indeed, different laboratories may vary in their approach to creating a mispronunciation sensitivity experiment, using different types of stimuli and methodologies. Many studies pose more nuanced questions, such as examining the impact of number of phonological features changed (mispronunciation size) or the location of the mispronunciation. Some studies may differ in their experimental design, presenting a distractor image that is either familiar or completely novel. In our meta-analysis we code for features of the experiment that are often reported but vary across studies and include an analysis of these features to shed further light on early phono-lexical representations and their maturation.”

1. On a methodological level, the reviewer raised several questions about the nature and assumptions of your analyses that seem like important questions to think about and discuss a bit more extensively.  The analyses you have conducted appear appropriate to the questions you are asking, but I believe some additional discussion about the inferences that can be drawn from the results, given the limitations of your dataset, and perhaps some additional caution in your claims, may be warranted.  
     
   *We respond to this question in detail below, specifically regarding the tone of our claims of why authors may choose a specific post-naming analysis window and why vocabulary information may not be collected. We point out the limitations in our dataset due to a small sample size, for example regarding productive vocabulary (page 18) and coda mispronunciations (page 22), as well as our inability to address the relationship between consonant/vowel mispronunciation sensitivity, mispronunciation position, and vocabulary size (Page 32). In our Discussion, we call for more studies to examine vocabulary size (Page 29), the impact of mispronunciation size across different ages (Page 30), differences between mispronunciations of different positions (Page 31), and the role of target-distractor overlap at different ages (Page 32).*

Reviewers' comments:  
  
Reviewer 1: This paper is a competently done meta-analysis of the 30 or so studies  
that have tested toddlers' knowledge of the sounds of words by placing  
children in an image-fixation task where words are pronounced normally  
or are mispronounced (Swingley & Aslin, 2000). When children's  
performance is impaired by strange pronunciations, relative to normal  
pronunciations, it is evidence that children's lexical knowledge is  
correct, at least with respect to the features that contrast between  
the correct pronunciation and the mispronunciation. Given the fairly  
large number of studies that have used this technique to address  
various questions about children's lexical knowledge, and given the  
clear utility and reliability of the task for future use, a  
meta-analysis is timely and useful.  
  
Although I hope the authors can implement a few suggestions and minor  
wording changes described below, I think this contribution, though a  
modest advance, is nevertheless solid and informative. I also found  
it fully appropriate that the authors did not portray their study as  
an attempt to resolve theoretical disputes in the literature or as an  
attempt to show that one loose theoretical description of the  
developmental process is better than another. I appreciated their  
sticking to the data. Below I have a small number of narrative  
questions and some proofreading comments.  
  
Questions  
  
p. 5 "In a particularly demanding task, such as when the target and  
distractor image share the same onset (e.g. doggie and doll),  
infants' ability to access the phonological detail of familiar words  
may be restricted (Swingley, Pinto, & Fernald, 1999)."  
  
I thought that study showed the opposite. Children were very similar  
to adults in their ability to use word-initial and word-medial  
phonetic features appropriately to drive their looking behavior.  
  
This study is also referred to on page 7 in the context of children's  
responses to mispronunciations. But that study did not use  
mispronunciations; it took advantage of the fact that children knew a  
few similar-sounding words, to address similar questions to those that  
were later addressed using mispronunciations.  
  
I think it is useful to refer to that study (the 1999 one) as a  
counterexample to the idea that overlapping picture names impairs  
performance in some generic way. It seems like the default  
expectation should be that overlap would \*delay\* finding the target a  
little, but this would be true both for children and adults (as  
studies testing this tend to show, starting from the 1999 study).

*The results of Swingley et al. (1999) show that word recognition is delayed if the target and distractor images share the same onset (e.g. doggie and doll) in comparison to rhyming target-distractor pairs (e.g. ball-doll). We thank the reviewer for this thoughtful comment and we have now found a better example to support our arguments on page 5:*

“The ability to access and use this detail, however, is governed by the task or developmental demands probed in a particular experiment. For example, if infants are tested on a more subtle mispronunciation that changes only one phonological feature, they may be less likely to identify the change in comparison to a mispronunciation that changes two or three phonological features (White & Morgan, 2008). If older infants are more likely to be tested using a more demanding mispronunciation sensitivity task, this may attenuate developmental effects across studies.”

*And we now refer to a mispronunciation study to illustrate our point on page 7:*

“For example, when examining sensitivity to a vowel mispronunciation of the target word "ball", the image of a ball would be paired with a distractor image that shares onset overlap, such as "bed", as opposed to a distractor image that does not share onset overlap, such as “truck”. This ensures that infants cannot use the onset of the word to differentiate between the target and distractor images (Mani & Plunkett, 2007).”

p. 17: "there might be no relationship between age and target looks in  
response to a correctly pronounced or mispronounced label." Not in  
this dataset. But studies using correct pronunciations exclusively  
have repeatedly found that older children recognize a given set of  
words faster than younger children (Fernald et al., 1998). I think it  
would be a mistake to leave this statement as-is without comment. I  
suspect that what is going on is that researchers expand the range of  
words they can use as they begin asking the increasingly intricate  
questions that are possible with older children. So an experiment  
with 30 month olds can use the lower-frequency, later-learned,  
"harder" word "monkey", which would be a poor choice in a study with  
15 month olds.

*Although we did not wish to speculate on this lack of an effect, as it was not related to our main question, the reviewer is correct to point this out. We have now added the following explanation for our results on page 17:*

“ …indicates that there was no relationship between age and target looks in response to a correctly pronounced or mispronounced label. However, experimental studies (e.g. Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998) and a recent meta-analysis (Frank, Lewis, MacDonald, 2016) have found that children's speed and accuracy in recognition of correctly pronounced words increases with age. Perhaps older children are more likely to be tested on less-frequent, later learned words than younger children, which could lead to a lack of a relationship between age and target looks in response to correct pronunciations in the current meta-analysis.”

p. 21. I had a confusion about the moderator analyses. Does the  
order of entry matter? This is why I wondered:  
  
Table 2 introduces the moderators in this order: size, position, type,  
distractor-overlap, distractor-familiarity.  
  
Table 1 shows that "position" mostly means Onset versus Medial, as  
there are few studies testing Coda, and perhaps only two testing onset  
and coda separately as a controlled design feature (Ren & Morgan in  
press, and Swingley 2009). But Medial almost always means it's a  
Vowel being tested. Thus, the features of position and type are not  
fully independent of one another.  
  
Does the fact that "position" is entered before "type" mean anything,  
in helping us to interpret that there seems to be a (somewhat  
inconsistent) effect of position, and not an effect of type?

*The previous explanation for Table 2 was unclear and it is not surprising that the reviewer was confused about the moderator analyses. Each moderator analysis was conducted separately. For example, the moderator analysis examining Misp. position did not include the factor of Misp. type and vice versa. We agree that, for the reasons given by the reviewer, combining the analyses of Misp. position and type would be problematic. The order of moderators in Table 2 corresponds to the order they are introduced throughout the manuscript, the results of one moderator analysis do not influence the results of another. We introduce Table 2 as follows in the manuscript on page 20:*

“In this section, we consider each moderator individually and investigate its influence on mispronunciation sensitivity. For most moderators (except mispronunciation size), we combine the correct and mispronounced datasets and include the moderator of condition, to study mispronunciation sensitivity as opposed to object recognition. To better understand the impact of these moderators on developmental change, we include age as subsequent moderator. Results of the 5 main moderator tests (mispronunciation size, mispronunciation position, mispronunciation type, distractor overlap, distractor familiarity) as well as the individual effects for each moderator interaction are reported in Table 2. The statistic that tests whether a specific moderator explains a significant proportion of variance in the data, QM, was significant for all moderators and subsequent significant interactions of critical terms are interpreted.”

*We have edited Table 2 to emphasize that these moderator tests were separate from one another and added more description to the caption:*

“Summary of the 5 moderator tests, including effect estimates for effects and critical interactions.”

p. 25 This argument will be scrutinized carefully, as it lays out  
how unscrupulous researchers might cheat in their data analyses. I  
looked at it carefully too. I think there is a problem with the  
argument. The argument is that researchers do studies with young  
children, fail to find an effect of mispronunciation, and then  
desperately tweak the analyses until an effect of mispronunciation is  
found. But why would they do that? A failure to find an effect of  
mispronunciation is not a failed, hard to publish study. It would be  
consistent with a major theoretical point of view, indeed a dominant  
point of view before the early 2000s, that young children have vague  
phonetic knowledge of words.  
  
The failed, hard to publish study is one in which children recognize  
neither the correct pronunciations nor the mispronunciations. The  
most plausible cause of this would be the materials: either the words  
were too hard, or the pictures were not sufficiently represenative in  
the minds of the participants.  
  
Now, it is possible that it has happened before that a study would use  
words that were too hard or pictures that were too obscure,  
nevertheless accidentally find recognition of correct pronunciations  
as a Type 1 error, and thereby produce an apparent mispronunciation  
effect where there is none. And indeed this could happen because of  
researcher analysis-window tweaking. But this isn't quite what the  
present authors are suggesting.

*The reviewer’s suggestion is more logical and plausible that our original position. Indeed, this fits well with the asymmetrical funnel plots found in our analysis of publication bias, especially for correct pronunciations. As a result, we have changed our wording in this paragraph on page 25 to read as follows:*

“For example, consider the possibility that a particular study does not find that infants looked to the target object upon hearing a correct pronunciation. With this pattern of behavior, interpreting an effect of mispronunciation sensitivity becomes difficult; how can infants notice a phoneme change when they do not even show recognition of the correct pronunciation? A lack of recognition or a small effect for correct pronunciations would be more difficult to publish (Ferguson & Heene, 2012). In order to have publishable results, adjustments to the analysis approach could be made until a significant effect of recognition for correct pronunciations is found. But, these adjustments would also need to be made for the analysis of mispronunciations, which may impact the size of the mispronunciation sensitivity effect. Such a scenario could explain the publication bias suggested by the asymmetry for correct pronunciations in the funnel plot shown in Figure 2 (Simmons, Nelson, & Simonsohn, 2011). This could lead to an increase in significant results and even alter the developmental trajectory of mispronunciation sensitivity.”

p. 27, 659-664 could be clearer about what the difference is between  
what is referred to here as "Difference Score" and "Pre- vs. Post".  
From this page it sounds like the difference is basically the  
distinction between reporting a paired t-test, and a one-sample t-test  
of difference scores. This is very confusing, because these tests are  
really the same thing. I think what the authors meant to say is that  
in some studies, "pre" window fixations are aggregated over something  
(subjects, items, ...) and "post" window fixations are aggregated  
over the same things, and then a test compares them.  
  
If that is the case, then it is easier to understand why this analysis  
method does not show mispronunciation effects (Figure 11): it fails to  
aggregate over the right things, and measurement error blunts the effect.

*We appreciate the reviewer pointing out our error and we now explicitly state that the Difference Score and Pre- vs. Post reflect the same type of analysis in both the Data analysis section of the Methods as well as in the Dependent variable section of the Exploratory Analyses. We have also combined these two dependent variables in our Exploratory Analysis of Dependent variable on page 28:*

“As described in the Methods section, there was considerable variation across papers in whether the pre-naming phase was used as a baseline measurement (Difference Score or Pre- vs. Post) or whether the post-naming PTL was compared with a chance value of 50% (Post). Considering analyses of the dependent variables Difference Score or Pre- vs. Post produce the same result, we combined these two dependent variables into one condition, Baseline Corrected. To our knowledge, there is no theory or evidence that explicitly drives choice of dependent variable in preferential looking studies, which may explain the wide variation in dependent variable reported in the papers included in this meta-analysis. We next explored whether the type of dependent variable calculated influenced the estimated size of sensitivity to mispronunciations.”

*As a result, the models that we computed to explore the effect of dependent variable have now changed (see Dependent variable in the Exploratory Analysis Section, page 28). The comparison in mispronunciation sensitivity between the dependent variable Post and the now combined Baseline Corrected dependent variable is significant, with a higher estimate for the dependent variable Post. The moderator test including age, however, no longer finds a significant interaction between dependent variable, condition, and age. We have therefore removed the interpretation that dependent variable choice impacts the developmental trajectory of the mispronunciation sensitivity effect from the section Data Analysis Choices (bottom of page 33, top of 34).*

p. 29 I did not care for the editorializing occurring in the paragraph  
about vocabulary sizes. It is possible for researchers to not want to  
measure vocabulary sizes because it is not their research question.  
It is not evidence that researchers quail at the fearful prospect of  
maybe not finding a correlation.

*We have now changed this sentence on page 29 to read:*

“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.”

*We have also changed the wording regarding this point on page 19:*

“This decrease in mispronunciation sensitivity studies measuring and reporting vocabulary size correlations is surprising, considering its theoretical interest.”

p. 32, 791 I didn't follow why it would be that (e.g.) noticing "dog"  
having been pronounced with the wrong vowel would be harder if the  
distractor picture were a duck than if it were a truck. The notion of  
task demands interfering with performance is raised (based on Switch  
procedure studies) but I don't think distractor choice influences task  
difficulty in the relevant way. If anything I would expect that  
making the onset consonants the same would make a change in the vowel  
more detrimental to word recognition, for Cohort model sorts of  
reasons.

*If the target word “dog” is pronounced with the wrong vowel “dag”, but the distractor image is a truck, the child only needs to pay attention to the first phoneme, a consonant in order to disambiguate between the target and distractor images. If, however, the distractor image is that of a duck, then the two images can not be discriminated on the basis of the first phoneme alone. We have made this section in the Introduction on page 7 more clear based on the reviewer’s earlier suggestion:*

“For example, when examining sensitivity to a vowel mispronunciation of the target word "ball", the image of a ball would be paired with a distractor image that shares onset overlap, such as "bed", as opposed to a distractor image that does not share onset overlap, such as “truck”. This ensures that infants cannot use the onset of the word to differentiate between the target and distractor images (Mani & Plunkett, 2007).”

p. 33, 810 What is described as a prediction of TRACE would be a  
prediction of pretty much any theory or model of word recognition,  
wouldn't it?

*Although we are unaware of any model of word recognition that explicitly makes predictions about the speed of target looks in response to correct and mispronounced words, we agree that likely no model would dispute this claim. We have therefore rewritten this sentence on page 33:*

“Looks to the target in response to mispronunciations may be slower than that of correct pronunciations in infants (as predicted for example by TRACE; Mayor & Plunkett, 2014)…”

p. 49 etc It is good practice to choose colors that vary in lightness,  
so that they can be differentiated in black and white. Given that  
there are almost no constraints on which 3 colors you choose, why not  
use colors that work in grayscale?

*We had chosen these colors to be perceivable to those who are color blind, but we appreciate the reviewer’s point and have changed the color of our figures to grayscale.*

Proofing comments  
  
p. 3. "examines on" -> "examines"

*This has been changed.*

p. 3 line 34, more accurate to say "looking" than "looks".

*This has been changed.*

p. 3 line 42 if it were up to me, I wouldn't throw in the ManyBabies  
shoutout, which felt gratuitous.

*This has been removed*

p. 3 line 49 "infant's" -> "infants'"

*This has been changed.*

p. 8 line 170. The citation involves "Group"?

*The name in the last author position should be “The PRISMA Group”. This has now been*

*updated.*

p. 19 l 440 "studies ARE typically"

*This has been changed.*

p. 20 l 495 numer -> number

*This has been changed.*

p. 27, 672: "We next explored whether the type of dependent variable  
calculated influenced the estimated size of..."  
  
saying "influenced" presupposes a causal interpretation of a  
correlation between Type and Size.

*We have now replaced “influenced” with “was related to”.*

p. 29, 702. 6 to 30 months? I don't see 6 month olds on the plots,  
so this was puzzling

*Effect sizes from 6-month-olds come from the study of Bergelson & Swingley (2017) and can be seen for example in Figure 3a. The correct and mispronunciation conditions for these infants were between-subjects measurement, however, and were therefore not included in our graphical representation of mispronunciation sensitivity (e.g. Figure 3a) which depicts the within-subject group and within-study difference in effect sizes for correctly pronounced and mispronounced words. To make this more explicit, we have added the following to each figure that includes a graphical representation of mispronunciation sensitivity:*

“Effect sizes for mispronunciation sensitivity within subject group and study (correct - mispronunciations) …”

p. 29, 710. "carefully" cannot apply to a "lack of effects", as  
effects do not have intention or volition

*We have removed “carefully” from this sentence.*

p. 32, 777-8 "sensitivity", "have," and "trajectory" should agree in  
number (all plural or all singular)

*This has been changed.*

p. 32, 789 recieve -> receive

*This has been changed.*

p. 34, 833 "field standard" is an odd collocation

*This has been corrected to “field standards”.*