Dear Dr. Stepp,

Thank you very much for these encouraging reviews. We are resubmitting this paper with modifications addressing all of the reviewers' and editorial points. Saliently, we:

- reorganized the Introduction and added structure
- more clearly motivated the two samples of participants included
- added mentions of limitations in the Discussion
- made all other suggested edits

A full point-by-point reply can be found below with replies in blue.

We additionally improved the reproducibility and legibility of the code, and thoroughly proofed the manuscript.

Looking forward to hearing from you,

A. Cristia, on behalf of all authors

####################################

Ref.: Ms. No. JSLHR-20-00661

Describing vocalizations in young children: A big data approach through citizen science annotation

Journal of Speech, Language, and Hearing Research

Dear Dr. Cristia,

We are interested in publishing your submission in the Journal of Speech, Language, and Hearing Research. The manuscript does, however, require some further revision. We ask that you revise with respect to the changes suggested.

Please ensure you display the changes to your revised manuscript by using either the highlighter function in MS Word, track changes, or by using bold, underlined, or colored text. If you upload a tracked changes version of the main manuscript text file, please also include a clean copy of the main text file with all the changes accepted. This will greatly help peer reviewers evaluate your revised submission.

We also request that you write a point-by-point response to the reviewer comments. You should first include the reviewer comment and then follow with your response (which should start "RESPONSE:").

Your revision is due by Feb 05, 2021.

To submit a revision, go to https://www.editorialmanager.com/jslhr/ and log in as an Author. You will see a menu item call Submission Needing Revision. You will find your submission record there.

Kind regards,

Cara E. Stepp

Editor

Journal of Speech, Language, and Hearing Research

Comments from the Editor:

Thank you for sending your work to JSLHR. I have received two thorough and informed reviews. Both reviewers find merit in your submission and I agree with their assessment that this work is appropriate for JSLHR. They both have identified several minor edits that will improve the clarity and impact of the work. I won't repeat them here, but I agree with their assessments and look forward to a revision of this work.

Comments from Reviewers:

Reviewer #1:

1. Overall Strengths

Reviewer 1: Well-conducted research, of likely interest to a broad spectrum of JSLHR readers.

> Thank you for highlighting the broad relevance of this manuscript.

2. Importance

Reviewer 1: This work is important as it articulates and evaluates a novel approach to measuring/evaluating child vocalizations.

> Thank you for highlighting the novelty of these results.

3. Justification/Rationale

Reviewer 1: The topic is of interest to readers of JSLHR both from the clinical perspective and researchers who study typical child language development.

While the topic is well-motivated, the authors can do a better job of articulating their objectives and how that related to the existing literature.

> We clarified objectives and improved flow in the Introduction by addressing your detailed comments below.

4. Methods/Approach

Reviewer 1: Overall, the study is of very high quality. The primary methodological limitation relates to the small sample and limitations on generalizability. But as a first study of this type, those limitations are expected. The authors hold to very high Open Science best practices for reproducibility.

> Thank you for highlighting the open science angle of our work, and the high quality of the study.

5. Results/Findings

Reviewer 1: With some minor exceptions as described below, the results are very clearly laid out.

6. Discussion/Conclusions

Reviewer 1: The Discussion is very clear in articulating the findings and strengths/weaknesses of the approach. Some additional work to identify limitations of their results (as opposed to their approach) would strengthen the paper.

> We added limitations of our results in the Discussion of this revised version, by pointing out:

- **remaining methodological problems with the LENA pipeline (p. 32);**
- **the relatively small sample sizes of the two groups we studied (p. 35);**
- **possible restrictions on generalization from this work (and other similar work) (pp. 32-33).**

Could the manuscript benefit from the addition of supplemental material?

Reviewer 1: None:

Is additional information regarding the research methodology needed to replicate the study?

Reviewer 1: No:

Reviewer 1 Comments to Author: This study used a novel citizen science approach to categorize the vocalizations of typically-developing infants and young children with Angelman syndrome. Derived metrics of proportions of linguistic and canonical vocalizations were examined. Comparisons were made of the citizen science data to data collected from trained research assistants in the laboratory. Overall the citizen science data was similar to that of the trained RAs, showing promise of this technique for future studies. Some limitations were found with lower precision/recall in some cases, in particular confusion between non-canonical and crying, and between laughing and non-canonical/junk categories.

I enjoyed reading this manuscript. It is overall well-written and provides a number of insights that are beneficial for the advancement of the field. I have some relatively minor comments for the authors' consideration, that are primarily about clarity of presentation.

My one more major comment is that as a reader **I felt the Introduction could use a little more work to clearly identify the specific goals of the project up front and have a more coherent flow of topics.** There are a number of competing topics: comparing the vocalizations of children with Angelman syndrome with typically developing children, assessing the feasibility of a citizen science approach to labeling vocalizations, creating (and assessing) specific derived metrics, etc. and the Introduction seems to jump around these topics a little too much. Some specific points that may help illustrate:

> We have broken up the first paragraph of the paper so that the overview of the Introduction structure stands out more (p. 3), and added more subsection headers to help readers notice the structure (throughout the Introduction).

* On page 4, the authors state that they will focus on the two descriptors (Linguistic and Canonical Proportion), but then the subsequent sections launch into methodological concerns with longform audio, automated processing and classification algorithms. (And it's not clear that we ever truly return to this topic in a robust way.) I found the discussion of Table 1 particularly challenging to follow, in where the data presented originate, how it should be interpreted, and what its relevance is for the current study.

> We have:

- **removed the mention of the two descriptors at the beginning of this subsection;**
- **reorganized the contents of this subsection, for instance by moving Table 1 until after the newer algorithms are introduced;**
- **added a brief point about methodological concerns in the Discussion (to reply to your parenthetical comment, and to address your request for highlighting of limitations in our results, p. 32-3 in both tracked changes & clean versions);**
- **added information regarding Table 1 when introduced (pp. 5-6); we also refer to it in the Discussion (p. 33), so hopefully it's clearer how it is informative.**

* Similarly, on page 6, the authors introduce derived metrics, but then shift to a discussion of LENA's vocalization counts. They note that it is a "promising" metric, but since a specific goal hasn't been clearly identified, it's not obvious what it is promising for.

> We have (on p. 7):

- **clarified that LENA's vocalization counts *are* a derived metric: "a derived metric because it is not simply a description of sections of the audio, but instead it integrates over the whole recording length";**
- **added "of individuals' vocalization development" "after a promising metric".**

* On page 7 at the bottom, the authors argue that there are two "outstanding challenges" for the use of daylong recordings to describe child vocalizations. Their first, the cost of human annotation, seems clearly discussed in their proposed solution of crowdsourcing, but the second, "how useful descriptors are extracted", seems a little more vague.

> We have moved this paragraph into the following subsection, and rewritten the second challenge so it is clearer how our proposed method helps with it as well (p. 9): "The second relates to how descriptors of vocal development are generated, i.e., how

annotations of individual audio sections are integrated over all data for a given child to derive child-level vocal development metrics.".

* The justification for the particular sample annotated needs clarification. The use of the term "low risk control" for the typically developing children suggests that there is an intended comparison to be made across these groups, perhaps to assess the feasibility of this approach for clinical assessment purposes, but the authors are (I believe rightly) very careful about the comparative conclusions they draw from the two samples, and no analyses are made to assess the ability of this approach to discriminate the two samples. On page 11 they describe their sample as based on the need for "generalizability", but it's far from clear how generalizable the results of this study would be beyond these two specific (and small) samples. The claim that a "wide age range" was selected also needs to be tempered as the typically-developing children were restricted to the ages 4-18 months. To be clear, I don't think this is a deal-breaker for the study, as this is simply a preliminary study of feasibility, but the selected samples do need to be more carefully and clearly motivated, and the limitations on generalizability described explicitly.

> We have:

- **more clearly motivated the two samples at the end of the Introduction (p. 13-14);**
- **reworded the mention of "wide age range" (p. 13): "We therefore included a group of low-risk control infants, who are close to the population most commonly sampled in studies using LENA® (see meta-analyses in Wang et al., 2020; Cristia et al., 2020); as well as children who had been diagnosed with Angelman syndrome and whose ages spanned a wide age range";**
- **noted in what ways these results cannot be widely generalized in the Discussion (p. 32-33).**

* Some additional information about Angelman syndrome would be helpful to understand the findings. Were the different patterns between the canonical and linguistic proportions predicted by the already-known characteristics of this syndrome?

> We have said this more clearly at the end of the Introduction (p. 13).

Other comments

* I think it is worth explicitly noting in the descriptive results (though it is clear in the graph and noted later in the discussion) that the children with Angelman syndrome scored much lower than the typically-developing infants on the canonical measure but higher on the linguistic measure.

> We have added this (p. 21): "We draw the readers' attention to the fact that the children diagnosed with Angelman syndrome had overall higher levels of Linguistic Proportion than the typically-developing infants, but the opposite was true for Canonical Proportion.".

* I did not follow the argument on page 21 beginning "Above, we conclude that derived metrics seem more promising…". It wasn't clear to me where this would have been articulated, nor what the relationship is between derived/raw segment measures and clips vs. full segments - these seem like orthogonal considerations.

> We have made the following changes to make sure that the logic is clear; specifically we have:

- **made sure we use the words *clips, segments, vocalizations, and chunks*, consistently throughout the manuscript;**
- **added a figure (Figure 2) that portrays the matching across these words/concepts;**
- **added wording reminding readers of the difference between derived versus first-pass metrics in several places, including in p. 24: "Above, we concluded that derived metrics integrating information across audio clips (Linguistic and Canonical Proportion, which can be derived from segment- or chunk-level labels) seem more promising than segment-level data (where individual segments are classified into Crying, Laughing, etc.)".**

* In the following paragraph, it would be helpful to more directly compare these clip-level findings with segment-based ones - the Discussion does a good job of describing these similarities in more detail, but it would be helpful to have a one-to-one lining up of the findings to better compare them without flipping between pages of results. Perhaps a table where the findings can be directly aligned would help.

> We have moved Table 2 up here.

* In the bottom paragraph on the same page, I found the wording "As for correlations between Zooniverse and laboratory-derived metrics, we observe very similar levels of correlation" - what is this similar to?

> We have added "(e.g., the correlation between the Linguistic Proportion as estimated using Zooniverse coding at the chunk level and the Linguistic Proportion via laboratory coding at the segment level)" (p. 25). We also made this clearer when discussing sampling as smaller number of segments (p. 26-27): "We repeated this process 50 times, to assess the extent to which the association between laboratory- and Zooniverse-derived metrics varied in each random sample (e.g., the correlation between Linguistic Proportion as estimated using Zooniverse coding on the one hand, and Linguistic Proportion as estimated using laboratory coding on the other)".

* On page 24, since (if I'm not mistaken) the age-related effects were not significant, it would be better to describe the Angelman syndrome group as being relatively stable with age, despite the (non-significant) trend to decrease with age.

> We corrected this: "In contrast, the two groups seem to differ in their trajectory for Canonical Proportion, with rapid increases up to 20 months in the low-risk group, and relatively stable levels of Canonical Proportion among the children in the Angelman syndrome group."

Reviewer #2:

1. Overall Strengths

Reviewer 2: This manuscript describes the feasibility of using crowdsourcing to code/annotate vocalizations collected via daylong audio recordings of the home language environment. The paper is well written and very informative. This work will be of broad interest to the field.

> Thank you for pointing out the relevance of this work.

2. Importance

Reviewer 2: Research using wearable recording devices (e.g., LENA) yields an abundance of data that is time consuming to transcribe and process. In addition, relying on algorithms to classify vocalizations and segment speech is not as precise or accurate as trained experimenters. This paper offers a promising methodological solution.

> Thank you for pointing out the promise of the methods proposed and evaluated here.

3. Justification/Rationale

Reviewer 2: The authors do a nice job demonstrating how this work builds on previous research. However, the authors do not discuss their justification for including a disordered population (i.e., Angelman syndrome) in addition to a typically developing population. Has other work included this population? Why was this population selected? Are there hypotheses about the validity between groups or is this more exploratory? I think the paper could be strengthened by addressing the dearth of LENA data with atypical populations, the additional challenges associated with processing daylong recordings with atypical populations, and the general need for this type of research.

> We have:

- **more clearly motivated the two samples at the end of the Introduction (p. 13-14);**

- **added a mention of the relative dearth of LENA data with atypical populations in the Introduction (p. 13-14), as well as in the Discussion (p. 35).**

4. Methods/Approach

Reviewer 2: Rigorous, clear. I appreciate the description of privacy concerns and measures. Do the authors intend to make their script available on open science platforms?

> Thank you for highlighting the attention to sensitivity of these data. Yes, all scripts are hosted on github. Those used to process the audio, push up clips for annotations, and retrieve annotations are available from github (https://github.com/psilonpneuma/Zooniverse). In addition, our manuscript was compiled using script, and it can be found (as well as the data and pre- and post-processing scripts) in https://osf.io/57yha/. Both of these links can be found in the manuscript.

5. Results/Findings

Reviewer 2: Clear and straightforward.

6. Discussion/Conclusions

Reviewer 2: I appreciate the discussion of applicability of derived metrics for studying group difference.

Could the manuscript benefit from the addition of supplemental material?

Reviewer 2: No:

Is additional information regarding the research methodology needed to replicate the study?

Reviewer 2: No:

IMPORTANT COMMENTS FROM THE EDITORIAL OFFICE.

In addition to completing the content revisions requested above please ensure you also attend to the following formatting issues:

* Please use a space around math operators reported in the text, e.g. (a + b + c) or < $500 or A = B. For P values reported in the manuscript, please use a lower case, italics font with no leading zero before the decimal point and a space before and after the math operator.

> We have:

- **added spaces around math operators**
- **used Italics for p values**
- **removed leading zeros for p-values and correlations**

* Please remove all shading from your tables. Bold, italic, or underlined text is allowed.

> We have removed shading and styling.

* Please do not embed figures in the manuscript text file. Each figure must be submitted as a separate graphics file. Acceptable file types include JPEG (.jpg, .jpeg), PNG (.png), TIF (.tif, .tiff), or PDF (.pdf) file format. Be sure to include figure callouts in the manuscript text (where the figure would be placed) and place the figure legends after the References listing of the manuscript. Note: All figures must have a minimum resolution of 200 dpi (preferably 300 dpi or greater). The figure titles and legends (captions) should be placed in the manuscript text file immediately following the References listing and not in the figures themselves.

> We have:

- **submitted figures separately;**
- **checked all figures are called out in the ms;**
- **placed figure legends after the references;**
- **checked that the figures are 300 dpi.**