**Comments from the Editor:**

I have read this revised manuscript along with two expert reviewers. Reviewer 1 provided several moderate to minor comments. For example, he noted that it is difficult for readers to access and contextualize the findings of the simulations after moving the simulation results to the supplemental materials. I agree and suggest that additional information (e.g., briefing and discussion) remain in the main text, while more detailed results may be moved to the appendix (not supplemental). Please check all his detailed comments in the end of this correspondence. I also think it is acceptable to slightly exceed the typical page limit.

**Response**: We have provided additional information related to the data analysis methodology for these simulations in the manuscript and moved a short description of the findings to the results section and a full description and visualization of the results to the Appendix.

The methods section of the manuscript has been updated as follows: “We also examined the stability of results when generating multiple synthetic datasets. Specifically, our approach involved generating 100 different synthetic datasets for each original dataset. A statistical model with the original dataset was fit, and the *p*-value and effect size were recorded. We then evaluated whether 95% of *p*-values and effect sizes from the synthetic datasets demonstrated a similar result as the original study, Measures of effect size and their interpretation for each study are provided in Table 3. Additional information related to these analyses, as well as results and figures are provided in Appendix A. Since Curtis et al. (2023) did not perform inferential statistical models, we directly compared each synthetic dataset to the original data with a zero-inflated beta multilevel model with the *gamlss* package (version 5.4.3) (Stasinopoulos & Rigby, 2007), which included a fixed effect of dataset type and a random intercept of participant. The *p*-value from both zero-inflated and beta portions of the model were evaluated and *p* < .05 was interpreted as no statistically significant difference between the synthetic and original dataset. Since Robinaugh et al. (2024) used a Bayesian analysis, we compared synthetic versus original model estimates. The pre-registered analysis plan and corresponding deviations are publicly available on the Open Science Framework (https://osf.io/vhgq2). (lines 249 – 264)

We have updated the results section (one example), as follows: “Findings from the 100 generated synthetic datasets indicate that 71% of datasets demonstrated the same inferential result (i.e., a statistically significant *p*-value) and 57% maintained their effect size category.” (lines 302 – 304)  
  
Below are my own comments and suggestions for improvement. Some of them appear critical, but are straightforward to address (if the authors agree), Thus I marked it as “minor revision”. The manuscript will be considered for publication when all these concerns are addressed.  
  
I think the overall concept of sharing and using synthetic data has value to the field. However, the biggest concern I have is that there are a few bold claims that may be misleading. Sharing and using synthetic data can be beneficial for educational purposes and particularly useful for reproducibility in analysis and methodological development. However, I do not believe that synthetic aggregate data are suitable for (new) hypothesis validation or for generating new research questions. Please adjust in the statement ending on line 126 (change-tracked version, same below) accordingly.  
**Response**: We have updated this statement to better reflect the utility of synthetic data for educational purposes and methods development. Specifically, we have updated the manuscript, as follows: “Importantly, synthetic data retains the statistical properties and relationships of the original data, enabling readers to evaluate key aspects of the study’s analysis workflow, such as data pre-processing and statistical modeling, as well as develop and extend methodologies.” (lines 109 – 112)

Additionally, the next sentence on line 126 may be misleading and should be rephrased appropriately to tone down the claim: "Synthetic data generation is widely used across medical research, industry, and government agencies, most notably by the United States Census Bureau (Jarmin et al., 2014)." I do not see “widely used” in their paper. Instead, Jarmin et al. (2014) actually emphasized concerns and challenges when discussing the “expansion” of synthetic data use: "Sophisticated users are rightly concerned that inferences drawn from synthetic data may not always be valid; a problem exacerbated in the case of small area applications." Using synthetic data requires users to combine multiple implicates to obtain valid estimates. While this may not be burdensome for sophisticated users, recent experience with multi-year estimates from the Census Bureau’s American Community Survey suggests that a significant portion of the user community may struggle to understand the limitations and to perform the additional computations required for statistical inference.

**Response**: The goal of this statement was to illustrate how synthetic data is currently utilized, and that its use is common in government agencies like the US Census Bureau and the Government of Canada. Admittedly, this Jarmin citation is somewhat outdated; therefore, we have provided a more recent citation from the Government of Canada’s Directive on Open Government, who have expertise on synthetic data and describe its use (particularly with hierarchical data). We hope to collaborate with them in the future to illustrate newer methods for synthesis that can accommodate more complex, hierarchical data sets. We have updated the manuscript, as follows: “Synthetic data generation has been used in government agencies, such as the United States Census Bureau (Jarmin et al., 2014) and the Government of Canada’s Directive on Open Government (Gauvin et al., 2021)to promote greater access to data and information.” (lines 112 – 115)  
  
My second concern is that it should be made clear at the beginning that the synthetic data here refer to synthesized analysis/aggregate data (as explained in the second section of the Introduction). I was initially confused between synthetic data and (raw) data augmentation until I reached the Methods section, where I found clarity.

**Response**: We have updated the introduction to specify that synthetic data can be applied to any data type (raw, intermediate, or analysis data). This section now reads as, “This approach can be applied to a wide variety of data types *(e.g., raw, intermediate, or analysis data)* and variables (e.g., qualitative demographic information or quantitative outcome measures).” (lines 104-108) We have also further specified that our aim is to test the feasibility and utility of using synthetic data to generate analysis data: “To address this gap, the present study aimed to explore the feasibility and preliminary utility of *generating synthetic analysis data* in CSD.” (lines 159-160)

Below are some more specific suggested edits.  
  
Title: make it more clear with the appropriate scope and expectation of this study. One example may be to add words “in Analysis” in the end. Based on the manuscript and author responses, the current study focuses on synthesizing analysis/aggregate data to promote transparency reproducibility in (data) analysis, as also stated in the text. Since the data are not raw, it is important to clarify that this method does not support direct reproduction of scientific findings. This distinction should be reflected in the title.

**Response**: We have revised the title to better reflect the scope of the study and our focus on analysis data, as follows: “Using Synthetic Data in Communication Sciences and Disorders to Promote Computational Reproducibility and Transparency”.

Also, there is no definition or clear explanation of “transparency” at the beginning. I assume it refers to transparency in data analysis (e.g., statistical analysis) in this study. Please clarify this explicitly in the introduction.

**Response**: We have revised the introductory paragraph to define research transparency. It now reads as: “Transparency and openness are fundamental tenets of science, *with transparency referring to the ability of others to clearly understand how scientific conclusions were reached, including the data, methods, and analyses used (Aguinis et al., 2018). A key aspect of transparency is computational reproducibility,* which refers to the ability to recreate a study’s results using the original data. Nowadays, the vast majority of scientific studies use some degree of computation in processing data, conducting descriptive or inferential statistics, and visualizing results. When these computations are reproducible, the transparency and confidence in findings are enhanced. Achieving computational reproducibility, however, requires authors to share their data. Both the National Institutes of Health and the National Science Foundation mandate data sharing and management plans to ensure that scientific data supporting a study is shared upon publication and aligns with FAIR (Findability, Accessibility, Interoperability, and Reuse) principles of digital assets (Watson et al., 2023; Wilkinson et al., 2016).” (lines 58-69)

Figure 5 from the original version has been removed in this revision. I agree with Reviewer 1 that while the decision tree in the former Figure 5 does not constitute a framework, it is nonetheless useful. In my opinion, the decision tree clearly illustrates a strategy for sharing different types of data, and it helps clarify that “synthetic data” in this manuscript refers to synthetic analysis data—which is critical. I suggest reinstating Figure 5 but do not refer to it as a framework. It could be placed at the end of the first subsection of the Introduction (page 6, tracked changes version), where data sharing is discussed.

**Response**: We have added this figure to the manuscript.  
  
Table 3 has a symbol display issue (some symbols are shown as rectangles), although they display correctly in the separate Word document. This appears to be a PDF generation issue—please check this carefully.

**Response**: This symbol will be corrected during the proofing stage with the ASHA journal.

The example on page 11 involving ID, age, and weight is very confusing. Participant ID is simply a random code and should be randomly assigned (as also noted by Reviewer 1) and does not need to be synthesized from “its observed distribution”. The goal in the example appears to be to create a “fake” dataset. In addition, I think age and weight should be completely independent of ID.

**Response**: This paragraph has been revised for accuracy and clarity. It now reads as, “For example, consider a dataset containing four variables: participant ID, sex (categorical), age (continuous), and weight (continuous). The process would begin by synthesizing sex by estimating its distribution based on the original data and sampling synthetic values from that distribution. Age would then be synthesized conditionally based on the synthetic sex values, using a model that captures the relationship between sex and age in the original data. Finally, weight would be synthesized conditionally on both sex and age, again based on relationships estimated from the observed data. Participant IDs, which serve only as identifiers and do not contain meaningful information, would be randomly assigned after data synthesis is complete.” (lines 198-206)  
  
I am not sure why the discussion on raw vs. intermediate data was removed. I found it interesting and suggest considering reintroducing that discussion.

**Response**: We have included the following discussion on raw versus intermediate data in the discussion, as follows: “It is important to highlight the many benefits of sharing either raw or intermediate data. Sharing these types of data enhances research transparency by enabling readers to re-analyze raw data for different purposes or reproduce the calculations behind analysis data. Different operational definitions or analysis steps are often a barrier to inclusion in a meta-analysis; therefore, sharing this type of data ensures that secondary analyses can be performed with alternate methodologies or operational definitions as the field progresses. In this sense, sharing raw or intermediate data facilitates the generation of new knowledge and accelerates scientific discovery. Despite its many benefits, there are instances where sharing raw or intermediate data may not be feasible. For example, researchers may not have obtained consent from participants for data sharing, or the institutional review board may impose project-specific guidelines that restrict sharing this type of data. In these instances, synthetic data fills an important gap by enhancing the transparency of analysis and methods workflows.”(lines 474-485)

**Comments from Reviewers:**

The reviewers were asked to consider and address the items below in your evaluation of this manuscript:    
 1. Overall Strengths    
 2. Importance    
 3. Justification/Rationale    
 4. Methods/Approach    
 5. Results/Findings    
 6. Discussion/Conclusions    
     
Please use the comments below to guide your revisions.

**Reviewer #1:**

Reviewer 1: The authors have made meaningful changes to the manuscript, and many of my comments have been thoroughly addressed. Descriptions of the utility and nuances around synthetic data are well done and thoughtful. The strategy of splitting up the manuscript into two distinct works that will each have the room to thoroughly address their aims is reasonable. A few remaining comments:  
  
 (1) While I understand the reasoning to move the simulation results to the supplemental materials, it unfortunately makes it difficult for the reader to access and contextualize these findings. Perhaps the authors could consider a few steps to incorporate this information into the manuscript by (1) briefly discussing simulation results for each study in the results (i.e., did the simulation match the results of the single study the authors have reported, or was there anything unexpected) and (2) moving these results to an appendix rather than a supplement to make them more accessible (3) including a brief summary of the simulation results in the discussion.  
  
I recognize that page length is often a limitation for research notes. If necessary, I think it is worth going slightly beyond the typical length of a research note for this addition. Otherwise, it seems to be that the impact of the quality simulation work conducted by the authors will be limited. If the authors wish to take more substantial steps to include the full simulation results in the body of the manuscript as well – I think that would also be acceptable, but it is not necessary.

**Response**: We have included the stability findings in the results section, as well as a full description and visualization in the appendix (see above response to the editor).  
  
(2) The inclusion of a SCED study is of clear benefit. However, I notice that it is not included in the supplemental material simulations. Second, while the synthpop package seems to have had difficulty reproducing the SCED data, in a Bayesian framework such as used by Robinaugh at colleagues, it is typically straightforward to create synthetic data that fits within the constraints of the study by sampling the posterior predictive distributions one or many times. This is one benefit of a fully Bayesian analytical approach – that it is relatively trivial to produce one or many synthetic datasets that meet the constraints of the original analytical approach (and perhaps this is one solution for the limitations of synthpop). There is a drawback is that additional work might be required to ensure the data is sufficiently anonymized if the posterior predictions are too similar to the real data. The authors might want to note this in their discussion where they mention difficulty creating synthetic datasets for hierarchical data.

**Response**: We have included the SCED study in the stability analysis, and updated the discussion to include the potential benefits of a fully Bayesian analytical approach, as follows: “Additionally, another potential method for handling more complex, hierarchical data involves creating synthetic datasets by sampling posterior predictive distributions through a fully Bayesian analytical approach; however, this method does not guarantee anonymization and should be used with caution when sharing confidential and highly sensitive data.” (lines 458 – 465)  
  
 Could the manuscript benefit from the addition of supplemental material?  
 Reviewer 1: No:  
  
 Is additional information regarding the research methodology needed to replicate the study?  
 Reviewer 1: No:

**Reviewer #2:**

Reviewer 2: I have functioned as Reviewer 2 for the previous version of the manuscript. The comments and suggestions from Reviewer 1 and me were highly overlapping. The authors have revised the manuscript accordingly and added sufficient caveats to nuance the claims and limit the scope of the paper. I thus consider the revision successful and recommend the paper for publication.

**Response**: Thank you.  
  
 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: Yes: