

Supplemental Material S5. Data sheets.

HOW TO READ INTRAWIDE AND INTRARATERLONG DATA FILES

Project

Title: Reliability of diagnosing childhood apraxia of speech

Authors: Elizabeth Murray^{1,2}, Shelley Velleman³, Jonathan Preston⁴, Robert Heard¹, Akhila Shibu¹ and Patricia McCabe¹

¹ The University of Sydney, Sydney, Australia

² Remarkable Speech + Movement, Sydney, Australia

³ The University of Vermont, Burlington, Vermont, USA

⁴ Syracuse University, Syracuse, NY, USA.

Funding Statement:

This study was funded by a grant from Apraxia Kids (formerly CASANA) to E. Murray (PI). Primary data collection was also supported by grants from the National Institutes of Health R03DC012152 and R15DC016426 (J. Preston, PI).

Ethical approval

The research was approved by The University of Sydney Human Ethics Research Committee (approval number: 2019/270). Data collection for the samples and their use for this purpose was also approved by The University of Sydney Human Ethics Research Committee (approval number: 2019/270), Syracuse University IRB (approval numbers: 17-177, 14-117), and The University of Vermont IRB (approval number: 17-0661)

Data Files

1. INTERWIDE_labels.csv is a comma separated version (csv) file of recorded and calculated variables for the original examinations and diagnoses of 92 children, described below, used to estimate inter-rater reliability, and forming the initial data for intra-rater calculations.
2. INTERWIDE_nolabels.csv is a csv format of INTERWIDE.sav, showing the number data for Likert-style items.
3. INTERWIDENAMES.csv is a csv format list of all measures in INTERWIDE.sav together with a description of each measure.
4. INTRARATERLONG_labels.csv is a csv version of INTRARATERLONG.sav, showing the word labels for Likert-style items.
5. INTRARATERLONG_nolabels.csv is a csv version of INTRARATERLONG.sav, showing the number data for Likert-style items.
6. INTRARATERLONGNAMES.csv is a csv format list of all measures in INTRARATERLONG files together with a description of each measure. Initial ratings have "Time 1" in the label and re-ratings have "Time 2".

Reading the INTERWIDE Data Files

The data files are in 'wide' format, with one row of data per child diagnosed. Each child was rated by three clinicians. The set of clinicians differed for each child. Their code identities are in data columns 9.1, 9.2 and 9.3. All children and clinicians are identified by

their own, unique, arbitrary code number. Records linking human identities to arbitrary codes are confidential and will not be released.

Ratings the three clinicians gave their child are adjacent for each item rated. For example, Q1_1 asked "What features does the child have in their communication? - Inconsistency". The rating by clinician 1 is in data column Q1_1.1, Clinician 2 in Q1_1.2 and Clinician 3 in Q1_1.3. All variables with a separate response from each clinician end in ".1", ".2" or ".3".

Some data columns combine data from all three ratings for a child. These data columns do not end in ".1", ".2" or ".3".

Missing data appear in the .sav file as a full stop. In the csv files missing data are shown by "NA".

Reading the INTRARATERLONG files

The data are in "long" format. Each child was examined by three clinicians. The child has three rows of data, one row for each clinical examination. 92 children x 3 examinations = 276 rows of data. The 20 re-ratings (16 children) by a clinician repeating an initial examination have data in extra columns on the right side of the data sheet. The order of these columns matches the order of the original examination data columns.

Missing data appear in the .sav file as a full stop. In the csv files missing data are shown by "NA".

Method

Information below is extracted from the Method section of Reliability of diagnosing childhood apraxia of speech by Elizabeth Murray, Shelley Velleman, Jonathan Preston, Robert Heard, Akhila Shibu and Patricia McCabe.

The research design is a reliability study of expert raters who were asked to determine the presence of discriminative speech characteristics and diagnose a selection of children aged 2-17 years as having CAS only, CAS plus another disorder, or a disorder other than CAS.

Participants and Recruitment

Children

Due to the low prevalence of CAS, we used area sampling designed to identify potential participants who have CAS or disorders that are similar to CAS. Two samples of children were used, as the nature of the clinical tasks and speech samples differed by age. Thirty-six children aged 30-85 months with moderate-severe speech sound disorders were recruited and prospectively assessed in Sydney, Australia and Vermont, USA. Audio and video data were available for this group ("video + audio children"). A separate cohort of 56 children aged 84-208 months had completed data collection for other studies in Syracuse, NY, USA. Only audio data were available for this group ("audio only children").

We sought 100 children aged 2-18 years with CAS or non-CAS SSD but achieved a sample of 92. This was due to prospective data collection being ceased due to ongoing COVID restrictions affecting face-to-face clinical work. Approximately 40% of the

children in each group had CAS, according to those who tested them for research purposes; however, the current study does not rely on prior diagnoses.

“Audio + video Children” aged 30-85 months: The 36 Australian-English and American-English speaking children aged 2;0 to 7;11 years with any moderate-severe speech sound disorder including CAS were included. All participants met the following inclusion criteria:

- a. Aged between 2;0 to 7;11 years at the time of data collection.
- b. Had caregivers that were concerned that their child had significant SSD as measured by a score of ≤ 4 on the Intelligibility in Context Scale (McLeod et al, 2015; McLeod et al, 2017)
- c. Could say at least 5 words.
- d. Had normal or corrected to normal hearing and vision.
- e. Have English as a primary language and have at least one parent who has English as a primary language.

Children were not excluded for having syndromes, intellectual disability, or other neurodevelopmental disorders as CAS can co-occur with these conditions (ASHA, 2007; Peter et al., 2013).

“Audio-only Children” aged 84-208 months: This was a retrospectively collected sample of 56 American English-speaking children aged 8-17 years with speech sound disorders, including CAS as well as residual speech errors. If a speech sound disorder was present, co-occurring speech disorders (e.g., mild dysarthria, stuttering) were not exclusionary. Only audio recordings of assessment sessions were available for rating this group.

The inclusion criteria were:

- (a) diagnosis of CAS and/or residual speech errors,
- (b) no structural deficits (e.g., cleft palate),
- (c) nonverbal intelligence, receptive language, and oral-facial exam within normal range,
- (d) normal (or corrected to normal) hearing and vision,
- (e) no other developmental diagnosis (e.g., autism), and
- (f) American English as first and primary language.

Expert Speech-Language Pathologist Raters

Raters were recruited via passive recruitment using flyers on Twitter, Facebook, and email distribution via Apraxia Kids and the International Association of Communication Sciences and Disorders (IALP). Further raters across countries not well represented by the initial recruitment round were emailed directly via an independent research assistant and asked, using a Qualtrics form, if they wanted to participate. We sought at least eight raters to meet our study aims.

Raters were asked to complete a Qualtrics questionnaire to ensure they met the following inclusion criteria. To be eligible, raters had to self-report that they:

- (a) were a practicing SLP who had worked with clinical populations of children with CAS and other SSDs for a minimum of 5 years,
- (b) had some international recognition for expertise in CAS (e.g., >1 peer-reviewed journal article and/or multiple national/ international workshop presentations) ascertained by assessing the SLP’s resume or sighting evidence of the work)
- (d) had normal (corrected to normal) hearing and vision,
- (e) use any dialect of English as their primary language,

(f) had a capacity to rate a minimum of 10 samples (9 original and 1 repeat sample completed for intra-rater reliability).

Sample Preparation

All children's speech samples were prepared by an independent research assistant for rating. Preparation included (a) isolating tasks into separate samples for viewing (e.g., diadochokinesis, inconsistency tasks); (b) selecting 4 minutes of connected speech for "video + audio children" in which the child was most talkative or minutes 2-5 of the sample for "audio only children" and (c) editing of pauses and interruptions within tasks (e.g., pauses, off-topic talk, and non-useful comments). Audio files were prepared in PRAAT and were sampled at 22.1 kHz and video files were prepared in Adobe Premier in mp4 format.

Sample Content

Recordings of assessment tasks totaling 20 minutes per child were compiled. The assessment constructs were similar across "video + audio children" and "audio only children" samples; however, the specific tasks were chosen according to the age and severity of the participants so they were accessible and diagnostically helpful.

"Video + audio children" had a 2-hour play-based assessment (in 1-2 sessions) including use of the Language Neutral Assessment of Motor Speech in Young Children (LAMS) (Velleman et al., in press), which is appropriate for minimally verbal children and those with a range of motor disabilities, and also the Dynamic Assessment of Motor Speech Skills (DEMSS, Strand et al., 2013), which is for young children with greater speech severity. Both include cues to encourage verbal responses and therefore better understanding of a child's motor potential.

"Audio only children" completed tasks of greater complexity, such as rapid automatized naming under time pressure, to ensure that characteristics would be identified despite the likelihood of previous treatment. The selected tasks ensured raters could assess the constructs and characteristics of CAS and other SSDs on our rating form (see Table 1).

Development and piloting of the rating form

Ratings were completed using an online rating form in Qualtrics with raters accessing the audio or video files for each child via a secure Dropbox. The rating process is schematized in Figure 1. The broad process asked raters to (1) complete some background information on the sample from the Excel spreadsheet (rater's ID and dialect of English; child's age, gender, and dialect of English), (2) rate characteristics on a four-point scale and, if present, rate specific sub-characteristics, (3) rate diagnosis (CAS, CAS+ other disorders, or other disorders) (4) if CAS was present, rate CAS severity, (5) if other disorders were present, specify them and (6) rate their confidence in making a diagnosis. The rating characteristics included discriminative, theoretical and/or frequently used CAS characteristics as well as characteristics of other SSDs.

A pilot was conducted to ensure the Qualtrics online rating form's utility and to establish initial reliability. Pilot ratings were completed by the four SLPs who met the study's inclusion criteria (authors 1, 2, 3, and 6). Pilot 1 was completed on 2 samples of children aged 10-12 years who met the inclusion criteria from previous research (Preston et al., 2014) not later used in this research. Percentage exact agreement was 88% and Kappa was

$K = 0.7$ meeting the $K > 0.7$ criterion (Ogonowski et al., 2004). Overall, features showed agreement, but sub-features did not, resulting in changes to definitions. This pilot resulted in six changes to the online rating form (see Table 2).

A second pilot was completed with the same raters, using samples from four children aged 6-18 years who met the inclusion criteria. These samples were included in the inter-rater reliability numbers as no further changes were made to the form. Percent agreement was 81.25% and Kappa was $K = 0.41$, showing moderate agreement but not meeting the $K \geq 0.7$ criteria. The revised form was thought to be easier to use by the pilot raters. The final rating form is available as Supplemental Material A.

Training

Raters were trained to use the online Qualtrics rating form using a Zoom video and were provided with a training package including: (a) listening instructions for consistency across raters (see Supplemental Appendix B), (b) a list of rating definitions for consistent use of the rating form (see Supplemental Appendix C) and (c) word lists for all the structured tasks the children completed, for reference. To ensure understanding and compliance with the procedure, raters completed training on two test samples – one video of a “video + audio child” and one audio of an “audio only child” from previous research (Murray et al., 2015; Preston et al., 2014). These samples were not included in the research rating samples. They were not given feedback on their ratings as the goal was to train the raters in the procedures, not to calibrate them to a different rater's impressions. They were not given feedback on their ratings as the goal was to train the raters in the procedures, not to calibrate them to a different rater's impressions.

Rating procedures

Raters were sent rating packs in Excel spreadsheets with ten ratings per pack. Nine samples were unique, and one was a de-identified repeat of a sample already rated in the pack of ten to assess intra-rater reliability. Raters received links to a secure Dropbox folder for each child's de-identified and prepared samples. Within the folders, the first task (connected speech sample) and end task (sequences and diadochokinesis) remained consistent across children, but the middle 3 tasks were presented in randomized order for rating. Raters watched video recordings (“video + audio children”) or listened to audio recordings (“audio only children”). Each child's speech was rated by three blind raters (not involved in the original assessment) randomly assigned to the sample in sets of 10 ratings. To ensure the raters were reviewing child samples that were consistent with their expertise, raters were asked to identify the ages of children they were comfortable rating and they were provided with child samples within that range. To avoid raters relying on non-speech information, they were not provided with information about previous diagnoses or case history. All participants were invited to complete a second set of ten ratings so the minimum ratings a participant could complete was 10 and the maximum was 30 (excluding training). Listeners re-rated one sample in every set of ten ratings to determine intra-rater reliability.