American Journal on Intellectual and Developmental Disabilities

# Semi-Automatic Assessment of Vocalization Quality for Children with and without Angelman Syndrome

--Manuscript Draft--

| | |
|---|---|
| **Manuscript Number:** | AJIDD-D-21-00085R2 |
| **Article Type:** | Research Report |
| **Keywords:** | vocal maturity, naturalistic recording, Angelman syndrome, vocalizations, babble |
| **Corresponding Author:** | Lisa R. Hamrick<br>Purdue University<br>West Lafayette, IN UNITED STATES |
| **First Author:** | Lisa R. Hamrick |
| **Order of Authors:** | Lisa R. Hamrick |
| | Amanda Seidl |
| | Bridgette L. Kelleher |
| **Manuscript Region of Origin:** | UNITED STATES |
| **Abstract:** | Automated methods for processing of daylong audio recordings are efficient and may be an effective way of assessing developmental stage for typically developing children; however, their utility for children with developmental disabilities may be limited by constraints of algorithms and the scope of variables produced. Here, we present a novel utterance-level processing (ULP) system that 1) extracts utterances from daylong recordings, 2) verifies automated speaker tags using human annotation, and 3) provides vocal maturity metrics unavailable through automated systems. Study 1 examines the reliability and validity of this system in low-risk controls (LRC); Study 2 extends the ULP to children with Angelman syndrome (AS). Results showed that ULP annotations demonstrated high coder agreement across groups. Further, ULP metrics aligned with language assessments for LRC but not AS, perhaps reflecting limitations of language assessments in AS. We argue that ULP increases accuracy, efficiency, and accessibility of detailed vocal analysis for syndromic populations. |

# Semi-Automatic Assessment of Vocalization Quality for Children with and without Angelman Syndrome

## Abstract

Automated methods for processing of daylong audio recordings are efficient and may be an effective way of assessing developmental stage for typically developing children; however, their utility for children with developmental disabilities may be limited by constraints of algorithms and the scope of variables produced. Here, we present a novel utterance-level processing (ULP) system that 1) extracts utterances from daylong recordings, 2) verifies automated speaker tags using human annotation, and 3) provides vocal maturity metrics unavailable through automated systems. Study 1 examines the reliability and validity of this system in low-risk controls (LRC); Study 2 extends the ULP to children with Angelman syndrome (AS). Results showed that ULP annotations demonstrated high coder agreement across groups. Further, ULP metrics aligned with language assessments for LRC but not AS, perhaps reflecting limitations of language assessments in AS. We argue that ULP increases accuracy, efficiency, and accessibility of detailed vocal analysis for syndromic populations.

**Key words**: vocal maturity, naturalistic recording, Angelman syndrome, vocalizations, babble

Atypical speech and language development are often concerns for families of children with severe intellectual and developmental disabilities (IDDs). However, most of our knowledge about infants' early vocalizations relies on prospective information gathered from children with typical development (Fisher, 2017) or retrospective reports of children later determined to be at-risk (Belardi et al., 2017; Patten et al., 2014). As such, little is known about early vocalization patterns, particularly vocal maturity, as they emerge in children with IDDs, including those resulting from neurogenetic syndromes. Wearable devices such as the Language ENvironment Analysis (LENA, Gilkerson & Richards, 2009) recorder offer a promising option for collecting and analyzing high-volume audio data on the vocal development of children with IDDs. However, the usefulness of LENA's automated output for children with IDDs warrants exploration, as algorithms are primarily based on samples with normative volubility and further, may not produce key information relevant to at-risk populations, such as utterance-level measures of canonical babbling (vocalizations containing a rapid transition between consonant and vowel), which are known to predict later language outcomes (e.g., Patten et al., 2014; Lang et al., 2019; Oller et al., 1999; Roche et al., 2018). Therefore, additional annotation is necessary to access details about metrics of vocal development that may be particularly useful in assessing vocal development in children with IDDs.

The present study introduces the utterance-level processing (ULP) system, which we developed to integrate automated LENA output with strategic, rapid human coding to generate key metrics of the quantity and quality of vocalizations that assist in mapping the early language trajectories of infants with IDDs. In Study 1, we establish the ULP procedure, including the reliability of annotations and validity of the ULP metrics, in a sample of low-risk controls (LRC) who are expected to demonstrate typical language trajectories. In Study 2, we extend the use of

the ULP to characterize early vocalizations in a separate sample of children with Angelman syndrome (AS), a rare neurogenetic syndrome (1 in 15,000 live births) associated with severe developmental delays, particularly in the domain of communication. This population is an ideal group for exploring this method of vocal maturity assessment, as there is a pressing need to establish rapid and valid communication assessments in children with AS to be used in clinical trials (Kolevzon et al., 2021). As such, our goal was to examine the reliability and validity of the ULP components in LRC, and to test this system in a pilot sample of children with AS.

## Study 1: Reliability and Validity of Utterance-Level Processing System in LRC

The present study introduces the ULP system, a semi-automated system that results in metrics of quantity and quality of a child's early vocalizations. A variety of metrics and methods can be used to capture these skills, many of which can prospectively inform risk for later atypical outcomes. One metric for capturing the *quantity* of child vocalizations is the child vocalization rate (CVR), or the rate at which a child produces speech-like syllables. CVR increases with age (Gilkerson et al., 2017; Paul et al., 2011) and is positively associated with later standardized language scores (Gilkerson et al., 2017; Wang et al., 2020). Children with disorders affecting speech (e.g., expressive language impairment, autism spectrum disorder) typically have a reduced CVR (Rescorla & Ratner, 1996; Warren et al., 2010). The *quality* of child vocalizations is also important and can be captured by the canonical babbling ratio (CBR), the proportion of canonical syllables (syllables that include a rapid consonant-vowel transition) to all syllables produced. An increase in CBR reflects more consistent use of advanced (i.e., higher quality) syllables. CBR increases with age in early development (Cychosz et al., 2019; Lee et al., 2018) and is positively associated with later language outcomes (Lohmander et al., 2017; McCune & Vihman, 2001; Nyman et al., 2021). Atypical language learners often demonstrate delayed canonical babbling

onset and reduced CBR (Lang et al., 2019; Nyman & Lohmander, 2018; Overby et al., 2019; Patten et al., 2014). Of note, CBR may be a particularly useful metric for pre-verbal children and those with severe delays, as it is not reliant on functional language (Hamrick et al., 2019).

Both CVR and CBR are directly derived from samples of child vocalizations, which can be captured in a variety of contexts (e.g., structured lab-based tasks, daily activities at home) using a variety of annotation techniques. Capturing child vocalizations within a naturalistic context may offer the most valid representation of the child's skills, as children often behave differently in laboratory settings and with unfamiliar examiners (Bornstein et al., 2000; Lewedag et al., 1994). Indeed, naturalistic observations provide more flexibility in terms of setting, communication partner, and context. However, naturalistic observations also produce expansive volumes of "big data" that, if not processed using automated methods, require significant time and resources to code and annotate, limiting their scalability and utility in applied research settings.

Efficient segmentation and annotation procedures are needed to circumvent these barriers. Wearable recorders with related data processing algorithms, such as the LENA system (Gilkerson & Richards, 2009), offer a potential solution by producing automated, standardized summary metrics from naturalistic data. The LENA system –arguably the most widely-used naturalistic annotation system– includes a small recorder worn in children's clothing that records the child's vocalizations and sounds in their immediate environment. Daylong LENA recordings can be analyzed through proprietary software that "tags" the sound type (e.g., Key Child, Female Adult, Silence) and time of each recorded utterance. Further algorithms integrate this information and produce summary scores. The efficiency with which the LENA system produces information about vocal behaviors, including data that can be used to calculate CVR, makes it an attractive method

for studying and monitoring early speech and language development, including in applied settings (Weil & Middleton, 2010).

However, there are some limitations to the LENA system. First, there is evidence that LENA's automated algorithms may systematically misidentify certain types of vocalizations, such as adult females using infant-directed speech, despite generally adequate levels of accuracy in identifying child vocalizations (VanDam & Silbert, 2016, Bulgarelli & Bergelson, 2019). Second, LENA provides limited insight into the *quality* of young child vocalizations, particularly related to child's use of canonical syllables. LENA's metric of canonical syllable usage (i.e., the Vocal Productivity Measure; Du et al., 2017) has not been well-validated in populations beyond the normative sample used to create the measure and is limited to utterances occurring in the context of back-and-forth interactions (additional limitations of LENA specific to special populations are discussed in Study 2). Furthermore, LENA does not provide utterance-level measures of canonical and non-canonical utterances, making manual calculation of frequently used and predictive quality metrics, such as CBR, unfeasible. Thus, although the convenience of LENA is promising and useful for assessing the quantity of child vocalizations in clinical settings, its utility in research on the quality of vocalizations in high-risk populations may be limited.

To address this gap, we developed the ULP system, which extracts a subset of child vocalizations from the LENA recording to be annotated by human coders and used to generate key metrics of early vocal development, including a precise measure of CVR and CBR. The ULP system capitalizes on LENA's initial automated speaker classifications and extends this process by using human coders to 1) clean utterances tagged as the Key Child and 2) categorize the vocal maturity of early vocalizations, ultimately producing validated and more nuanced metrics of early vocal development than LENA while minimizing the amount of time and resources required to

obtain this information. The end-product of the ULP is a set of supplemental variables that are important signposts of early vocal development (Gilkerson et al., 2017; Lohmander et al., 2017; McCune & Vihman, 2001; Yuanyuan Wang et al., 2020) and would be prohibitively time-consuming to hand-code without the support of semi-automated approaches such as the ULP system.

Study 1 addresses two primary research questions: 1) Do annotations produced by ULP coders demonstrate acceptable reliability? 2) Do ULP metrics (i.e., CVR and CBR) demonstrate convergent validity with standardized language assessments?

## Method

Additional materials, including raw data, R scripts for data construction and analysis, Python scripts for data extraction and database management, and supplementary materials referenced in this manuscript, can be found on this study's OSF page at https://osf.io/d948m/?view_only=bac60180bd7c43438051e780b8d45f55.

### Participants

Participants were 39 LRC children ($M_{age}$=11.64 months, range 4-25 months; 21 female) drawn from a series of internally and externally funded studies. Families were recruited via social media postings and word of mouth. Inclusion criteria included residing in the United States and English as the primary language spoken in the home. Demographic information is reported in Table 2. Participants were primarily White, non-Hispanic, and had mothers who had completed at 4-year degree or higher.

### Measures

**Naturalistic Language Recording.** Daylong recordings obtained from the LENA system were analyzed by LENA's proprietary automated cloud-based program, which segments the

recording audio, assigns a pre-determined speaker label to each segment (e.g., Key Child, Female Adult, Silence), and provides counts of the number of child utterances, adult words, and conversational turns occurring during the recording.

**Utterance-level processing (ULP)**. Detailed ULP procedures can be found in S1 at https://osf.io/rj2ts?view_only=bac60180bd7c43438051e780b8d45f55. The ULP system involves three components: extraction, annotation, and generation of output.

***Extraction***. We developed Python scripts ([*masked*]) in collaboration with computer scientists to perform the functions needed for identifying and extracting utterances from the LENA recording to be annotated by ULP coders. Consistent with prior studies (Lee et al., 2018), the Python scripts identified 30 five-minute segments from the LENA recording containing speech produced by the Key Child (150 minutes total): 10 five-minute segments with the highest volubility (i.e., the highest number of child utterances as determined by LENA's automated output) and 20 additional five-minute segments randomly selected from remaining segments. As expected, more utterances were extracted per participant from high volubility segments ($M$=435, $SD$=103, range: 212-727) than from randomly selected utterances ($M$=218, $SD$=117, range: 63-500). Using LENA's labels, we extracted Key Child audio occurring during the selected 30 five-minute segments to be further tagged by ULP coders (Figure 1), which resulted in collections of 275-1,191 utterances per participant ($M$=653, $SD$=196).

***Annotation***. Each extracted utterance was presented via a user-friendly interface (Figure 2) and tagged by three randomly selected coders who were naïve to the child's demographic information. Coders tagged each utterance with an annotation (i.e., a vocal maturity label) that either indicated the highest vocal maturity level of syllables present in the utterance (e.g., if multiple syllable types were present in the utterance, the coders assigned a single annotation using

a hierarchy, where "Word" syllables are given highest priority, then "Canonical", then "Non-Canonical", then "Crying" or "Laughing"), or that the utterance contained only vegetative sounds (e.g., burps, coughs), significant overlapping noise/speech (e.g., another speaker, toy noises), or did not contain an utterance from the Key Child (e.g., was misclassified by LENA as the Key Child; the annotation for these utterances was "Don't Mark"). For speech utterances ("Word", "Canonical", "Non-Canonical"), coders marked the number of canonical, non-canonical, and word syllables. Given the developmental age of our sample, we focus primarily on canonical and non-canonical syllables; word syllables (1.16% of all LRC syllables) were captured as canonical or non-canonical. Crying (5.11% of all LRC utterances) and laughing (2.03% of all LRC utterances) were not analyzed for the purposes of this study. Utterances were excluded from analyses if all 3 coders disagreed on the vocal maturity annotation (n=889 utterances; 3.49% of all LRC utterances). Twelve coders (undergraduate students in Psychology or Speech Language, Hearing, and Sciences) participated and attended regular meetings during which an expert coder ([*masked*]) reviewed utterances with the coders and provided feedback to minimize drift. Detailed coder training procedures and materials can be found in S1 and S2 at https://osf.io/kt4am/?view_only=bac60180bd7c43438051e780b8d45f55.

***ULP output.*** For each participant, the ULP system generates output of the overall counts of utterances and syllables in each vocal maturity category. From this output, we calculated two subtypes of CVR (vocalization quantity) – rate of canonical (CVR-C) and non-canonical (CVR-N) syllables per minute – by dividing the number of canonical or non-canonical syllables, respectively, by 150 (the total length of the extracted segments). We also calculated CBR (vocalization quality) by dividing number of canonical syllables by total number of speech syllables (i.e., # canonical/(# canonical syllables + # non-canonical syllables).

**Standardized Language Assessments**. We used two well-established standardized measures of communication -the Vineland Adaptive Behavior Scales, 3rd Edition (VL-3; Sparrow et al., 2016) and the Communication and Symbolic Behavior Scales – Infant-Toddler Checklist (CSBS; Wetherby & Prizant, 2003) - to assess the concurrent validity of CVR and CBR. The VL-3 is a semi-structured parent interview that assesses child adaptive behaviors. The CSBS is a 24-item parent-report screening checklist used to identify children at risk for social communication delays. VL-3 and CSBS scores are reported in Table 2. We expected most children in the LRC group to demonstrate average performance on these measures; however, to maintain sample size and represent natural heterogeneity in language skills observed within low-risk populations, we did not exclude participants if they scored outside the average range. We used the Expressive and Receptive Language subscale raw scores of the VL-3 Communication domain and the Speech scale and Understanding subscale raw scores of the CSBS. Importantly, neither the VL-3 nor the CSBS are designed specifically to capture vocal quality. As such, we also examined two items from the VL-3 and four items from the CSBS expected to capture vocal quality (described in Table 5). Four participants were missing VL-3 and CSBS data.

**Procedure**

The [*masked*] IRB approved all study activities. Families were provided with a LENA recorder and vest and were asked to have their child wear the recorder for at least 12 hours. Families were provided a "scrub sheet" (S3 at https://osf.io/rgysb?view_only=bac60180bd7c43438051e780b8d45f55) in case they wanted to redact any data. Participants were included in the present study if they had completed a LENA recording for which there was ULP data at the time of analyses. Each participant's caregiver completed the VL-3 with a trained examiner in person or over the phone and the CSBS as part of

a set of online forms. The VL-3 and CSBS were completed within 1 month of the LENA recording on average (VL-3 range: 0-16 months; CSBS range: 0-7 months). We covaried the length of time that passed between the LENA recording and the VL-3 and CSBS in analyses to account for variation among participants.

**Analytic Plan**

**Coder reliability.** To test our first research question, we calculated percent agreement (a measure of absolute agreement among coders) and inter-rater reliability (a measure of consistency among coders that is adjusted for chance agreement) for ULP annotations, canonical syllable counts, and non-canonical syllable counts. We calculated percent agreement for annotations by dividing the number of utterances for which 2 or more coders assigned the same vocal maturity annotation by the total number of utterances. We calculated canonical and non-canonical syllable percent agreement by dividing the number of utterances for which 2 or more coders agreed on the number of canonical or non-canonical syllables contained in the utterance, respectively, by the total number of utterances. We used Gwet's AC1 (a reliability metric that calculates the probability the two randomly selected coders will agree and is more stable than Cohen's Kappa; Wongpakaran et al., 2013) to calculate reliability of categorical annotations and average two-way random effects ICCs to calculate reliability of continuous syllable counts. We assessed reliability using conventional thresholds (i.e., percent agreement > 80%; inter-rater reliability coefficients [AC1/ICC] ≥ .75; Cohen, 1960; Koo & Li, 2016).

**Convergent Validity.** To test our second research question, we conducted partial Spearman's correlations of the ULP metrics and scores from the standardized language assessments, covarying the length of time between when LENA and language assessments were collected. We corrected for multiple comparisons using the Holm-Bonferroni correction. We

hypothesized that higher CVR-C, CVR-N, and CBR would be positively associated with age and scores from standardized language assessments.

## Results

### Coder Reliability

Coders were highly reliable. Across all 25,456 utterances, 24,567 (97%) were assigned the same annotation by two or more coders. Inter-rater reliability for annotations was moderate (Gwet's AC1 = .702 [.698, .707], $p < .001$). Percent agreement for canonical and non-canonical syllable counts were similarly high (canonical: 97%, non-canonical: 84%; Table 3) and demonstrated excellent inter-rater reliability (canonical: ICC = .980 [.977, .982], $p < .001$; non-canonical: ICC = .941 [.939, .943], $p < .001$).

### Convergent Validity

**Canonical and Non-Canonical Vocalization Rate.** Rho, p-values, and adjusted p-values are reported in Table 6. CVR-C significantly increased with age ($\rho$=.76, $p$'<.001), whereas CVR-N was not associated with age (Figure 3). CVR-C was positively associated with VL-3 Expressive and Receptive scores (VL-3 Expressive: $\rho$=.66, $p$'<.001; VL-3 Receptive: $\rho$=.68, $p$'<.001), as well as the CSBS Speech score ($\rho$=.62, $p$'=.007) and the CSBS item capturing the ability to string two or more syllables together ($\rho$=.61, $p$'=.007). Notably, CVR-C also demonstrated medium associations (i.e., $\rho \geq .30$; Cohen, 1988) with most other raw subscale scores and item scores on the standardized language assessments; however, these associations did not survive corrections for multiple comparisons. CVR-N was not associated with any scores from the VL-3 or CSBS.

**Canonical Babbling Ratio.** Rho, p-values, and adjusted p-values are reported in Table 6. CBR significantly increased with age ($\rho$=.79, $p$'<.001; Figure 3) and was positively associated with VL-3 Receptive/Expressive scores, CSBS Speech scores, and CSBS items capturing whether

the child strings two-syllable sounds together and the number of consonants the child uses. CBR also demonstrated medium associations (i.e., $\rho \geq .30$; Cohen, 1988) with several other scales and items, however associations did not survive corrections for multiple comparisons.

**Summary**

Results support the use of the ULP system for generating reliable and valid estimates of CBR and CVR in LRC. Annotations produced by ULP coders demonstrated a high percentage of agreement across raters (>97%) and moderate to excellent inter-rater reliability. With respect to validity, both CVR-C and CBR demonstrated expected positive associations with several language assessments. Thus, it appears that these metrics are valid estimates of vocal maturity, and canonical syllable usage corresponds to overall levels of expressive and receptive language in this group. Further, as expected based on the literature (e.g., Cychosz et al., 2019), both CVR-C and CBR increased with age, demonstrating increased canonical syllable usage overall and increasing ratios of canonical syllables to overall syllables produced with age for LRC children. Unlike CBR and CVR-C, CVR-N was not associated with age or language assessments. Thus, in the LRC group, overall volubility of non-canonical syllables may be less informative than rate of canonical syllable usage and CBR. This may be partially due to the age of the sample, as most typically developing children at this age are increasing canonical syllable usage as a precursor to using meaningful words (Cychosz et al., 2019).

Given evidence that ULP annotations and metrics can be reliably and validly used to probe early language development in LRC, we next explored the utility of the ULP system in characterizing language in children with severe delays. Study 2 examined the use of ULP system in children with Angelman syndrome (AS). Individuals with AS exhibit severe language delays, with most acquiring only a handful of meaningful words in their lifetimes (Pearson et al., 2019).

Monitoring early language in AS is complicated by the lack of valid assessment tools that can capture small but meaningful changes in skills over time. We anticipated that ULP system would be useful for quantifying early language in AS and provide a novel method for mapping developmental skills and treatment response in this population.

**Study 2: Utterance-Level Processing of Vocalizations from Children with AS**

Understanding early vocal features of children with IDD is a high priority across a number of specific disorder communities (Berry-Kravis et al., 2013; Wheeler et al., 2017). Families of children with AS, for example, rate accurate measures of language and communication as a top unmet need and priority for clinical trials (Willgoss et al., 2021). Indeed, although many standardized language assessments exist, few can be used successfully with individuals with severe IDDs such as AS, particularly in early development (Soorya et al., 2018). This is because these measures were often normed for a broader range of skills, thus children who have mastered very few skills receive the lowest possible score, even as their skills progress over time (e.g., Summers, 2019). These "floor effects" may mask variability and prevent the detection of small but meaningful changes in skills over time or through interventions. Standardized tools can be further limited by items that assume that a child can express themselves verbally, or that require informants to make assumptions about the child's internal states (Grieco et al., 2019). Furthermore, efforts to administer clinical assessments to publisher standards can be complicated by co-occurring medical, motor, and behavioral challenges common to AS (Wheeler et al., 2017). Therefore, despite the importance of assessment tools for AS, standardized tools often fail to meet the needs of AS and other IDD communities.

Naturalistic assessments of development offer a promising alternative for monitoring skills in AS and other populations with severe IDD (Handen et al., 2018; Wandin et al., 2020). There are

several notable benefits that extend beyond those previously described for LRC. First, naturalistic methods can be useful for highlighting skills that are meaningful and relevant for specialized populations, complementing standardized assessments that often focus on skill deficits and providing a strengths-based approach desired by caregivers of children with AS and other IDDs (Kelleher et al., 2020). Furthermore, although naturalistic observations can offer more accurate samples of any child's day-to-day experiences relative to their performance in a lab, children with severe IDD may be even more sensitive to contextual changes and unfamiliar people, particularly if they have common co-occurring conditions such as anxiety or autism, making naturalistic data even more valuable. However, despite these benefits, *quantifying* naturalistic observations can be challenging; whereas standardized assessments can be scored quickly, naturalistic data requires substantial time and resources to clean, score, and code. These data demands are particularly challenging for researchers engaged in treatment trials for rare disorders, which require key scores be generated quickly and accurately.

As in LRC populations, wearable devices like the LENA system can partially address these challenges by generating rapid estimates of vocalization quantity. Indeed, LENA and other recording systems have been used extensively in research on language development through studies on a wide range of ages, languages, and developmental risks (Greenwood et al., 2018; Ye Wang et al., 2017) including individuals with severe IDD similar to AS (Rankine et al., 2017; Reisinger et al., 2019). However, it is important to consider some additional limitations of LENA in IDD populations. LENA developers report that algorithms rely on a combination of biological information (e.g., age, sex) and input characteristics (e.g., utterance pitch) to accurately tag the speaker of each utterance (Xu et al., 2008); thus, it is possible that misclassification is exacerbated in populations with differences in biological maturity and/or craniofacial abnormalities, such as in

AS and other genetically-based IDDs (Rankine et al., 2017). In addition, the standardized summary scores produced by LENA face similar pitfalls to other standardized language assessments when used with IDDs with regards to reduced variability in scores and floor effects. Furthermore, given that preliminary research has detected reduced CVR and CBR in AS compared to LRC (Grieco et al., 2018; Semenzin et al., 2021), LENA's limited output related to canonical syllable usage may be particularly detrimental for children with IDD whose developmental progress may not be detected on standardized tools (Thurm et al., 2020). Thus, despite some promising features, the appropriateness of LENA for use in populations with IDD remains in question.

In contrast, the CVR and CBR metrics generated from the ULP system offer several advantages for use in IDD populations: (1) CVR and CBR are ability-unbiased, meaning they can be collected from even the lowest-performing participants; (2) because these metrics are generated from naturalistic daylong recordings, they are more accessible and likely more ecologically valid than metrics gathered from laboratory settings; and (3) these metrics are likely more sensitive to small but meaningful change between children who demonstrate significant delays or within the same individuals over time. Thus, the ULP system may improve our assessment of vocal development for children with IDD.

In Study 2, we apply the ULP system to characterize early speech and language development in children with AS. We aim to address two research questions: 1) Do annotations produced by ULP coders demonstrate acceptable reliability in AS? 2) Do ULP metrics demonstrate concurrent associations with language assessments in AS?

## Method

Participants were 27 children with AS ($M_{age}$=37.98 months, range 11-62 months; 15 female). Demographic information is reported in Table 1. AS participants were primarily White,

non-Hispanic, and had mothers who had completed at 4-year degree or higher (Table 2). For a subset of analyses, we compare AS to the LRC group from Study 1. Given the smaller sample of children with AS (reflecting the low incidence of this neurogenetic syndrome), we did not attempt to match the AS and LRC groups on age or any developmental variables. The LRC group is significantly younger than the AS group, which we expected would better approximate the language level observed in older children with AS; however, children in the AS group still demonstrated lower language skills on the VL-3 and the CSBS than the younger LRC group (Table 2). AS families were recruited via social media postings and through syndrome foundations and registries. Inclusion criteria for the AS families included residing in the United States, English as the primary language spoken in the home, and providing confirmation of their child's medical diagnosis of AS (78% confirmed via genetic report). Measures and procedures are identical to those implemented in Study 1. Two AS participants were missing data from both the VL-3 and CSBS, and one AS participant was missing data from just the VL-3.

**Analytic Plan**

**Preliminary Analyses.** We first provide descriptive information about the utterances extracted by the ULP system for AS participants, including number of utterances extracted overall and from high volubility vs. randomly selected segments.

**Coder Reliability.** To test our first research question, we calculated percent agreement and inter-rater reliability using the same procedures as Study 1. We assessed reliability using conventional thresholds (i.e., percent agreement > 80%;inter-rater reliability coefficients [AC1/ICC] ≥ .75).

**Concurrent Associations with Language Assessments.** Given known challenges and lack of appropriate methods of assessing language in children with AS, it is difficult to assess the

validity of ULP metrics. While both concurrent validity measures used for the LRC group have been implemented with children with AS (Hamrick & Tonnsen, 2019; Peters et al., 2004), the range of scores in AS is more limited than LRC (see S4 at https://osf.io/6ck8z?view_only=bac60180bd7c43438051e780b8d45f55), and in some cases, we observed significant floor effects (e.g., all AS informants endorsed "0" for the CSBS assessing the use of two-word phrases, and the standard deviation of most scores is smaller in AS relative to LRC). With these limitations in mind, we addressed our second research question by conducting partial Spearman's correlations of the ULP metrics and standardized language assessment metrics, covarying the length of time between when LENA and language assessments were collected. We corrected for multiple comparisons using the Holm-Bonferroni correction across all statistical tests. Of note, we are not "validating" ULP metrics against these measures, but rather contextualizing how ULP metrics relate to the existing measures that are commonly used in clinical trials.

**Sensitivity to the Angelman Syndrome Phenotype.** Due to the difference in the chronological ages and language abilities of the AS and LRC groups, direct comparisons of ULP metrics between these groups are difficult to interpret. Nevertheless, given the extent of language delay characteristic of AS, we still expect – despite the large age difference – that the AS group will demonstrate delays in ULP metrics compared to the LRC group. In this way, comparing the AS and LRC groups can provide a sense of the ULP metrics' sensitivity to the AS phenotype. Thus, we conducted exploratory Mann-Whitney U-tests comparing the rate and percentage of each vocal maturity category and CBR between groups. We hypothesized that the LRC group would demonstrate higher rates and percentages of all categories of speech syllables and CBR relative to AS, despite being younger than the AS group on average. We also conducted Mann-Whitney U-

tests to test differences in non-consensus and "Don't Mark" utterances between the AS and LRC groups with the expectation that there would be no difference in the rate or proportion of non-codable or "Don't Mark" utterances for the AS and LRC groups.

## Results

### Preliminary Analyses

The number of extracted utterances per AS participant ranged from 178 to 932 (*M*=538, *SD*=224). Similar to LRC, more utterances were extracted per participant from high volubility segments (M=396, SD=152, range: 158-658) than from randomly selected utterances (M=143, SD=85, range: 7-371).

### Coder Reliability

Percent agreement and inter-rater reliability suggested that coders were able to reliably code AS utterances. Coders were highly reliable in assigning annotations to AS utterances. Of 14,534 utterances, 14,009 (96%) were assigned the same annotation by two or more coders (Table 3) and demonstrated good inter-rater reliability (Gwet's AC1 = .749 [.743, .755], $p < .001$). Canonical and non-canonical syllables counts were similarly high (canonical: 99%; non-canonical: 87%) and demonstrated excellent inter-rater reliability (canonical: ICC = .930 [.929, .932], $p < .001$; non-canonical: ICC = .953 [.951, .956], $p < .001$).[1]

### Concurrent Associations with Language Assessments

Test statistics, p-values, adjusted p-values, and Cohen's *d* effect sizes are reported in Table 6. Results revealed no significant associations of CVR-C, CVR-N or CBR with scores from the VL-3 or CSBS in the AS group. Further, neither CVR-C nor CVR-N increased with age in the AS

[1] Reliability estimates for two coders were explored by averaging estimates of each coding pair (i.e., Coders 1 and 2, Coders 2 and 3, and Coders 1 and 3). Gwet's AC1 and ICCs were comparable, while percent agreement was lower for two coders compared to three coders, particularly for non-canonical syllable counts (see S5 at https://osf.io/uh2bk?view_only=bac60180bd7c43438051e780b8d45f55).

group (Figure 3). CBR decreased with age in the AS group, though this association did not remain significant after correcting for multiple comparisons ($\rho$=-.42, $p$'=1.000; Figure 3).

**Sensitivity to the Angelman Syndrome Phenotype**

Test statistics, p-values, adjusted p-values, and Cohen's *d* effect sizes are reported in Table 5. Children with AS used 1.22 canonical syllables per 5-minute segment, with 5.82% of their overall syllable usage accounted for by canonical syllables, both of which, as expected, were statistically lower than those observed in the LRC group, (rate=7.75 syllables per segment, $d$=.99; proportion=22.29%, $d$=1.18). Also consistent with our hypotheses, children with AS demonstrated significantly lower CBR than the LRC group (AS: CBR=.06, LRC: CBR=.22, $d$=1.16). Contrary to our predictions, children with AS used a similar rate of non-canonical syllables (23.69 syllables per segment) to the LRC group (19.93 syllables per segment, $d$=.35), which resulted in a higher proportion of non-canonical syllables for AS (93.97%) than the LRC group (76.88%, $d$=1.17).

Mann-Whitney U-tests indicated nonsignificant differences in the rate or proportion of non-consensus utterances (rate: $d$=.26; proportion: $d$=.15), or rates and proportions of "Don't Mark" utterances, (rate: $d$=.13; proportion: $d$=.29) between the AS and LRC groups, with approximately 21% of utterances assigned to this category in the AS group compared to 18% in the LRC group (Table 5).

## Summary

Our preliminary data support the utility of the ULP system in AS. Annotations and syllable counts demonstrated high percent agreement and moderate to excellent inter-rater reliability. Thus, it appears that annotations produced by ULP coders provide sufficiently reliable data about early vocalizations of children with AS. ULP-derived metrics were not associated with standardized language assessments measures or age in AS, potentially reflecting the atypical development of

language in AS and the well-established limitations of existing standardized language assessments for this population (Pearson et al., 2019). Indeed, upon visual examination, ULP metrics demonstrated improved granularity among participants with the same scores on the VL-3 and CSBS (Figure 4-6 and S6 at https://osf.io/brv48?view_only=bac60180bd7c43438051e780b8d45f55), suggesting that ULP metrics may produce more nuanced information about vocal development than standardized language assessments can provide.

## Discussion

The present study established the initial "proof of concept" of a novel utterance-level processing system to efficiently phenotype language development among LRC children and children with AS. We observed strong coder reliability across groups, and we also found that ULP metrics aligned with validity metrics (age, language assessments) for the LRC group. Notably, ULP metrics did not demonstrate similar associations for children with AS. Here we discuss the implications of these data in terms of ULP system's utility in IDD and LRC populations.

### ULP annotations are reliable and ULP metrics are valid in LRC

Overall, data from the ULP system demonstrated good coder reliability and produced valid metrics of vocal maturity in the LRC group, suggesting a potential advancement in how vocalization data can be processed. Until now, obtaining CVR and CBR estimates has either (1) required trained coders to manually identify utterances in a speech sample, which requires significant time and resources, or (2) required reliance on "black box" algorithms, which in some cases have demonstrated subpar detection accuracy (Cristia et al., 2020; Marchman et al., 2020; Rankine et al., 2017). The ULP system offers an alternative that balances the automated efficiency of LENA with the gold standard annotation of human coders to produce reliable estimates of CVR

and CBR. Given that these vocal features are known to predict language outcomes (e.g., Patten et al., 2014; Lang et al., 2019; Oller et al., 1999; Roche et al., 2018), the ULP system provides an innovation in the field's approach to capturing early vocal features. Use of this procedure may facilitate more efficient examination of vocal maturity patterns over time, opening the door to answering big data questions that linger beyond our field's computational capacities. For example, we can employ the ULP system to efficiently examine how child language changes in response to other aspects of their language environment, such as rate or quality of adult input, or how language changes in response to intervention.

**Annotations obtained using the ULP system are reliable in Angelman syndrome**

The ULP system also generated promising results in AS and supported the use of LENA as a recording tool in AS, a population characterized by severe language delays. ULP coder agreement was good to excellent for AS, suggesting that coders were consistent in their ability to annotate vocal maturity and provide syllable counts of AS utterances. Furthermore, rates of usable utterances were comparable to those of LRC, suggesting that neither group stood out in terms of mislabeled or unusable utterances identified by LENA. These data provide a promising first step in determining whether LENA is a valid system for collecting naturalistic language samples in AS and other IDDs. Notably, our analyses do not directly address LENA's accuracy in AS, as the ULP system only focuses on child utterances, and the ULP "Don't Mark" category stringently excludes utterances that have significant overlapping speech or sound. Therefore, with these data, we cannot make any evaluations of LENA's accuracy in identifying other speaker types or LENA's potential misidentification of child speech as other speaker types. Furthermore, because ULP coders do not listen to utterances in the context of the full recording, it may be particularly difficult for them to identify when LENA has misclassified utterances from other children on the recording who are

similar in age and/or language level as the target child (though this misclassification appears to happen rarely; see Cristia et al., 2021). Nevertheless, and with these limitations in mind, our data suggest that some aspects of the LENA system (i.e., its use as a diarization tool) may be just as useful for children with IDD as it is for children with expected typical development.

**The ULP system may fill a gap in existing language assessment approaches**

The lack of valid tools for assessing communication in AS is a known barrier in the field (Wheeler et al., 2017). As such, the lack of association we observed between ULP outputs, standardized language assessments, and age may be unsurprising. It is promising, however, that the ULP system used naturalistic data to generate objective metrics of vocalization quantity and quality in AS, which provide a granularity that is often not afforded by standardized language assessments (Figures 4-6 and S6 on OSF).

In the present study, the associations between ULP metrics and age were non-significant in AS (ranging from $d$=.22-.42 in magnitude). While this finding may reflect a true lack of growth in vocal development over time in AS, our cross-sectional sample may also be insufficient to detect change over time across this relatively restricted age range given the variability in the AS phenotype and the slower expected rate of growth in vocal development. Thus, future longitudinal work – including in the context of language intervention – is needed to determine whether ULP metrics can detect acute changes in development over time and to provide further evidence of the improved sensitivity of ULP metrics.

There continues to be a need to determine how to best assess validity of new language measures in AS given the lack of "gold standard" metrics. It is possible that objective metrics such CVR and CBR could serve as valuable anchors for future work by evaluating whether estimates of verbal expression – such as caregiver-report tools – align with the day-to-day patterns of child

output. Given language and communication is a broad category of constructs, however, it is important to consider whether misalignment across existing tools (e.g., scores that do not align with "validation" metrics) provides meaningful information. For example, the present study used language assessments that generally focus on the *presence or absence* of certain speech skills. They do not, however, capture the *frequency* with which children use such skills, which is the focus of CVR and CBR. Thus, not only might the ULP metrics capture aspects of early speech skills that are not captured using standard language assessments, but it also may highlight a crucial area of individual differences among individuals with AS that previous standard language assessments are too coarse to detect.

Importantly, given that individuals with AS are on the extreme end of language impairment (as evidenced by the range of VL-3 and CSBS scores; Table 2), exploring the validity of ULP metrics for children with less severe delays will provide a more comprehensive understanding of the utility of the ULP system for high-risk populations. It is unclear whether the patterns observed in this study are reflective of the AS phenotype specifically or whether they reflect the ULP system's performance with atypical populations more generally.

**The ULP system offers a novel approach for characterizing vocal maturity in IDD**

Identifying a reliable method for assessing early vocalizations across ability levels – such as the ULP system – opens the door for cross-group comparisons that can help us understand longitudinal phenotypes. As expected, vocalization patterns of children with AS differed in important ways from the LRC group, despite the differences of age and language level between the two groups, suggesting the ULP system generates data that is sensitive to the AS phenotype. Most notably, children with AS used a much lower rate of speech syllables and lower proportion of canonical syllables which did not increase with age. These data converge with evidence that

individuals with AS typically do not advance past a language level of approximately 12-24 months (Andersen et al., 2001). Furthermore, we observed a trend toward decreasing CBR with age in AS that is distinct from what we see in our separate LRC sample, in which CBR increases across early development. If corroborated by future longitudinal work, these data suggest that vocal development in AS may look quite different from LRC, and the benchmarks we place on evaluating interventions and growth in AS may need to be aligned to these trajectories.

While most vocal metrics differed in AS relative to LRC, it is interesting to note that children with AS and in the LRC group did not differ in their rate of non-canonical syllables. This suggests that although children with AS may vocalize less overall and use less advanced utterances, the rate of non-canonical syllables they produce may be similar to low-risk children with similar developmental levels. Again, given that the AS and LRC groups were not matched on age or developmental skills, it is possible that the rate of non-canonical syllables in the AS group is still characteristic of a much younger language level. Now that we have provided preliminary data on the reliability of ULP annotations in a sample of children with AS, as well as the validity of ULP metrics in LRC children using standardized language assessments, our next steps will be to examine these patterns longitudinally across multiple high- and low-risk groups simultaneously, including groups matched on both age and developmental level. Mapping early developmental trajectories in this way can inform both the nature and course of typical and atypical language development, setting the stage for more effective, targeted intervention.

**Conclusion**

Although wearable vocal recording devices provide much promise for understanding how language develops, a number of practical constraints limit their utility. The present study introduced a novel post-processing system that generates estimates of vocal maturity that are not

available with automated methods. We found that the ULP system can be used to capture features of early vocal development related to volubility and canonical syllable usage. This method demonstrated good coder reliability for both LRC children as well as children with AS. Data from ULP also aligned with standardized measures of language in the LRC group. Together, these data suggest that the ULP system offers a method for systematically analyzing naturalistic samples to capture metrics of vocal development that are unavailable via automated methods or that are costly to process using human coders. We anticipate that the ULP system will advance the field by making the analysis of early vocal features more efficient and accessible, addressing a significant gap in available outcome measures for quantifying vocal maturity development in high-risk groups.

## References

Andersen, W. H., Rasmussen, R. K., & Strømme, P. (2001). Levels of cognitive and linguistic development in Angelman syndrome: a study of 20 children. *Logopedics Phoniatrics Vocology*, *26*(1), 2–9. https://doi.org/10.1080/14015430117324

Belardi, K., Watson, L. R., Faldowski, R. A., Hazlett, H., Crais, E., Baranek, G. T., McComish, C., Patten, E., & Oller, D. K. (2017). A Retrospective Video Analysis of Canonical Babbling and Volubility in Infants with Fragile X Syndrome at 9–12 Months of Age. *Journal of Autism and Developmental Disorders*, *47*(4), 1193–1206. https://doi.org/10.1007/s10803-017-3033-4

Berry-Kravis, E., Hessl, D., Abbeduto, L., Reiss, A. L., Beckel-Mitchener, A., & Urv, T. K. (2013). Outcome measures for clinical trials in fragile X syndrome. *Journal of Developmental and Behavioral Pediatrics*, *34*, 508–522. https://doi.org/10.1097/DBP.0b013e31829d1f20

Bornstein, M. H., Haynes, O. M., Painter, K. M., & Genevro, J. L. (2000). Child language with mother and with stranger at home and in the laboratory: A methodological study. *Journal of Child Language*, *27*(2), 407–420. https://doi.org/10.1017/S0305000900004165

Bulgarelli, F., & Bergelson, E. (2019). Look who's talking: A comparison of automated and human-generated speaker tags in naturalistic day-long recordings. *Behavior Research Methods*. https://doi.org/10.3758/s13428-019-01265-7

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. In L. Erlbaum (Ed.), *Statistical Power Analysis for the Behavioral Sciences* (Vol. 2nd, Issue 2). Lawrence Erlbaum Associates. https://doi.org/10.1234/12345678

Cristia, A., Bulgarellib, F., Bergelson, E., Bulgarelli, F., & Bergelson, E. (2020). Accuracy of the

language environment analysis system segmentation and metrics: A systematic review. *Journal of Speech, Language, and Hearing Research*, *63*(4), 1093–1105. https://doi.org/10.1044/2020_JSLHR-19-00017

Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., Bunce, J., & Bergelson, E. (2021). A thorough evaluation of the Language Environment Analysis (LENA) system. *Behavior Research Methods*, *53*(2), 467–486. https://doi.org/10.3758/s13428-020-01393-5

Cychosz, M., Cristia, A., Bergelson, E., Casillas, M., Baudet, G., Warlaumont, A. S., Scaff, C., Yankowitz, L., & Seidl, A. (2019). *Canonical babble development in a large-scale crosslinguistic corpus*. 1–48.

Du, S., Xu, D., Richards, J. A., Hannon, S. M., & Gilkerson, J. (2017). *The LENA ™ Vocal Productivity Measure (Technical Report No. LTR-11-1)*. https://www.lena.org/wp-content/uploads/pdf/technical-reports/LTR-11-1_Vocal Productivity.pdf

Fisher, E. L. (2017). A systematic review and meta-analysis of predictors of expressive-language outcomes among late talkers. *Journal of Speech, Language, and Hearing Research*, *60*(10), 2935–2948. https://doi.org/10.1044/2017_JSLHR-L-16-0310

Gilkerson, J., & Richards, J. A. (2009). Impact of adult talk, conversational turns, and TV during the critical 0-4 years of child development. *The Power of Talk*, 1–36.

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H. L., & Paul, T. D. (2017). Mapping the Early Language Environment Using All-Day Recordings and Automated Analysis. *American Journal of Speech-Language Pathology*, *26*(2), 248–265. https://doi.org/10.1044/2016_AJSLP-15-0169

Greenwood, C. R., Schnitz, A. G., Irvin, D., Tsai, S. F., & Carta, J. J. (2018). Automated

language environment analysis: A research synthesis. *American Journal of Speech-Language Pathology*, *27*(2), 853–867. https://doi.org/10.1044/2017_AJSLP-17-0033

Grieco, J. C., Bahr, R. H., Schoenberg, M. R., Conover, L., Mackie, L. N., & Weeber, E. J. (2018). Quantitative Measurement of Communication Ability in Children with Angelman Syndrome. *Journal of Applied Research in Intellectual Disabilities*, *31*(1), e49–e58. https://doi.org/10.1111/jar.12305

Grieco, J. C., Romero, B., Flood, E., Cabo, R., & Visootsak, J. (2019). A Conceptual Model of Angelman Syndrome and Review of Relevant Clinical Outcomes Assessments (COAs). *Patient*, *12*(1), 97–112. https://doi.org/10.1007/s40271-018-0323-7

Hamrick, L. R., Seidl, A., & Tonnsen, B. L. (2019). Acoustic properties of early vocalizations in infants with fragile X syndrome. *Autism Research*, *12*(11), 1663–1679. https://doi.org/10.1002/aur.2176

Hamrick, L. R., & Tonnsen, B. L. (2019). Validating and Applying the CSBS-ITC in Neurogenetic Syndromes. *American Journal on Intellectual and Developmental Disabilities*, *124*(3), 263–285. https://doi.org/10.1352/1944-7558-124.3.263

Handen, B. L., Mazefsky, C. A., Gabriels, R. L., Pedersen, K. A., Wallace, M., Siegel, M., Erickson, C., Gabriels, R. L., Kaplan, D., Morrow, E. M., Righi, G., Santangelo, S. L., Wink, L., Benevides, J., Beresford, C., Best, C., Bowen, K., Dechant, B., Dixon, J., … Tager-Flusberg, H. (2018). Risk Factors for Self-injurious Behavior in an Inpatient Psychiatric Sample of Children with Autism Spectrum Disorder: A Naturalistic Observation Study. *Journal of Autism and Developmental Disorders*, *48*(11), 3678–3688. https://doi.org/10.1007/s10803-017-3460-2

Kelleher, B., Halligan, T., Garwood, T., Howell, S., Martin-O'Dell, B., Swint, A., Shelton, L. A.,

& Shin, J. (2020). Brief Report: Assessment Experiences of Children with Neurogenetic Syndromes: Caregivers' Perceptions and Suggestions for Improvement. *Journal of Autism and Developmental Disorders*, *50*(4), 1443–1450. https://doi.org/10.1007/s10803-020-04363-0

Kolevzon, A., Ventola, P., Keary, C. J., Heimer, G., Neul, J. L., Adera, M., & Jaeger, J. (2021). Development of an adapted Clinical Global Impression scale for use in Angelman syndrome. *Journal of Neurodevelopmental Disorders*, *13*(1), 1–13. https://doi.org/10.1186/s11689-020-09349-8

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Lang, S., Bartl-Pokorny, K. D., Pokorny, F. B., Garrido, D., Mani, N., Fox-Boyer, A. V., Zhang, D., & Marschik, P. B. (2019). Canonical Babbling: A Marker for Earlier Identification of Late Detected Developmental Disorders? *Current Developmental Disorders Reports*, *6*(3), 111–118. https://doi.org/10.1007/s40474-019-00166-w

Lee, C. C., Jhang, Y., Relyea, G., Chen, L. mei, & Oller, D. K. (2018). Babbling development as seen in canonical babbling ratios: A naturalistic evaluation of all-day recordings. *Infant Behavior and Development*, *50*(September 2017), 140–153. https://doi.org/10.1016/j.infbeh.2017.12.002

Lewedag, V. L., Oller, D. K., & Lynch, M. P. (1994). Infants' vocalization patterns across home and laboratory environments. *First Language*, *14*(42–43), 49–65.

https://doi.org/10.1177/014272379401404204

Lohmander, A., Holm, K., Eriksson, S., & Lieberman, M. (2017). Observation method identifies that a lack of canonical babbling can indicate future speech and language problems. *Acta Paediatrica, International Journal of Paediatrics*, *106*(6), 935–943. https://doi.org/10.1111/apa.13816

Marchman, V. A., Weisleder, A., Hurtado, N., & Fernald, A. (2020). Accuracy of the Language Environment Analyses (LENATM) system for estimating child and adult speech in laboratory settings. *Journal of Child Language*, 1–16. https://doi.org/10.1017/S0305000920000380

McCune, L., & Vihman, M. M. (2001). Early Phonetic and Lexical Development. *Journal of Speech, Language, and Hearing Research*, *44*(3), 670–684. https://doi.org/10.1044/1092-4388(2001/054)

Nyman, A., & Lohmander, A. (2018). Babbling in children with neurodevelopmental disability and validity of a simplified way of measuring canonical babbling ratio. *Clinical Linguistics and Phonetics*, *32*(2), 114–127. https://doi.org/10.1080/02699206.2017.1320588

Nyman, A., Strömbergsson, S., & Lohmander, A. (2021). Canonical babbling ratio – Concurrent and predictive evaluation of the 0.15 criterion. *Journal of Communication Disorders*, *94*(November). https://doi.org/10.1016/j.jcomdis.2021.106164

Overby, M., Belardi, K., & Schreiber, J. (2019). A retrospective video analysis of canonical babbling and volubility in infants later diagnosed with childhood apraxia of speech. *Clinical Linguistics & Phonetics*, *00*(00), 1–18. https://doi.org/10.1080/02699206.2019.1683231

Patten, E., Belardi, K., Baranek, G. T., Watson, L. R., Labban, J. D., & Oller, D. K. (2014). Vocal Patterns in Infants with Autism Spectrum Disorder: Canonical Babbling Status and

Vocalization Frequency. *Journal of Autism and Developmental Disorders*, *44*(10), 2413–2428. https://doi.org/10.1007/s10803-014-2047-4

Paul, R., Fuerst, Y., Ramsay, G., Chawarska, K., & Klin, A. (2011). Out of the mouths of babes: Vocal production in infant siblings of children with ASD. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *52*(5), 588–598. https://doi.org/10.1111/j.1469-7610.2010.02332.x

Pearson, E., Wilde, L., Heald, M., Royston, R., & Oliver, C. (2019). Communication in Angelman syndrome: a scoping review. *Developmental Medicine and Child Neurology*, *61*(11), 1266–1274. https://doi.org/10.1111/dmcn.14257

Peters, S. U., Goddard-Finegold, J., Beaudet, A. L., Madduri, N., Turcich, M., & Bacino, C. A. (2004). Cognitive and adaptive behavior profiles of children with Angelman syndrome. *American Journal of Medical Genetics*, *128 A*(2), 110–113. https://doi.org/10.1002/ajmg.a.30065

Rankine, J., Li, E., Lurie, S., Rieger, H., Fourie, E., Siper, P. M., Wang, A. T., Buxbaum, J. D., & Kolevzon, A. (2017). Language ENvironment Analysis (LENA) in Phelan-McDermid Syndrome: Validity and Suggestions for Use in Minimally Verbal Children with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *47*(6), 1605–1617. https://doi.org/10.1007/s10803-017-3082-8

Reisinger, D., Shaffer, R., Pedapati, E., Dominick, K., & Erickson, C. (2019). A Pilot Quantitative Evaluation of Early Life Language Development in Fragile X Syndrome. *Brain Sciences*, *9*(2), 27. https://doi.org/10.3390/brainsci9020027

Rescorla, L., & Ratner, N. B. (1996). Phonetic profiles of toddlers with specific expressive language impairment (SLI-E). *Journal of Speech and Hearing Research*, *39*(1), 153–165.

https://doi.org/10.1044/jshr.3901.153

Semenzin, C., Hamrick, L. R., Seidl, A., Kelleher, B. L., & Cristia, A. (2021). Describing vocalizations in young children: A big data approach through citizen science annotation. *Journal of Speech, Language, and Hearing Research*, *64*(7), 2401–2416. https://doi.org/10.1044/2021_JSLHR-20-00661

Soorya, L., Leon, J., Trelles, M. P., & Thurm, A. (2018). Framework for assessing individuals with rare genetic disorders associated with profound intellectual and multiple disabilities (PIMD): the example of Phelan McDermid Syndrome. *Clinical Neuropsychologist*, *32*(7), 1226–1255. https://doi.org/10.1080/13854046.2017.1413211

Sparrow, S. S., Cicchetti, D. V., & Saulnier, C. A. (2016). *Vineland-3 Vineland Adaptive Behavior Scales - Third Edition*. NCS Pearson, Inc.

Summers, J. (2019). Using Behavioral Approaches to Assess Memory, Imitation and Motor Performance in Children with Angelman Syndrome: Results of a Pilot Study. *Developmental Neurorehabilitation*, *22*(8), 516–526. https://doi.org/10.1080/17518423.2019.1619857

Thurm, A., Kelleher, B., & Wheeler, A. (2020). Outcome Measures for Core Symptoms of Intellectual Disability: State of the Field. *American Journal on Intellectual and Developmental Disabilities*, *125*(6), 418–433. https://doi.org/10.1352/1944-7558-125.6.418

VanDam, M., & Silbert, N. H. (2016). Fidelity of automatic speech processing for adult and child talker classifications. *PLoS ONE*, *11*(8), 1–13. https://doi.org/10.1371/journal.pone.0160588

Wandin, H., Lindberg, P., & Sonnander, K. (2020). Development of a tool to assess visual attention in Rett syndrome: a pilot study. *AAC: Augmentative and Alternative*

*Communication*, *36*(2), 118–127. https://doi.org/10.1080/07434618.2020.1798507

Wang, Ye, Hartman, M., Aziz, N. A. A., Arora, S., Shi, L., & Tunison, E. (2017). A Systematic Review of the Use of LENA Technology. *American Annals of the Deaf*, *162*(3), 295–311. https://doi.org/10.1353/aad.2017.0028

Wang, Yuanyuan, Williams, R., Dilley, L., & Houston, D. M. (2020). A meta-analysis of the predictability of LENA™ automated measures for child language development. *Developmental Review*, *57*(May), 100921. https://doi.org/10.1016/j.dr.2020.100921

Warren, S. F., Gilkerson, J., Richards, J. A., Oller, D. K., Xu, D., Yapanel, U., & Gray, S. (2010). What automated vocal analysis reveals about the vocal production and language learning environment of young children with autism. *Journal of Autism and Developmental Disorders*, *40*, 555–569. https://doi.org/10.1007/s10803-009-0902-5

Weil, L. W., & Middleton, L. (2010). Use of the LENA Tool to Evaluate the Effectiveness of a Parent Intervention Program. *Perspectives on Language Learning and Education*, *17*(3), 108–111. https://doi.org/10.1044/lle17.3.108

Wetherby, A. M., & Prizant, G. (2003). *CSBS DP Manual. First Normed Edition*. Brookes Publishing.

Wheeler, A. C., Sacco, P., & Cabo, R. (2017). Unmet clinical needs and burden in Angelman syndrome: A review of the literature. *Orphanet Journal of Rare Diseases*, *12*(1), 1–17. https://doi.org/10.1186/s13023-017-0716-z

Willgoss, T., Cassater, D., Connor, S., Krishnan, M. L., Miller, M. T., Dias-Barbosa, C., Phillips, D., McCormack, J., Bird, L. M., Burdine, R. D., Claridge, S., & Bichell, T. J. (2021). Measuring What Matters to Individuals with Angelman Syndrome and Their Families: Development of a Patient-Centered Disease Concept Model. *Child Psychiatry and Human*

*Development*, *52*(4), 654–668. https://doi.org/10.1007/s10578-020-01051-z

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology*, *13*(1), 1–7. https://doi.org/10.1186/1471-2288-13-61

Xu, D., Yapanel, U., Gray, S., & Gilkerson, J. (2008). Signal processing for young child speech language development. *Wocci*. http://www.researchgate.net/profile/Umit_Yapanel/publication/242130394_Signal_Processing_for_Young_Child_Speech_Language_Development/links/0046353a4991aa7c7c000000.pdf

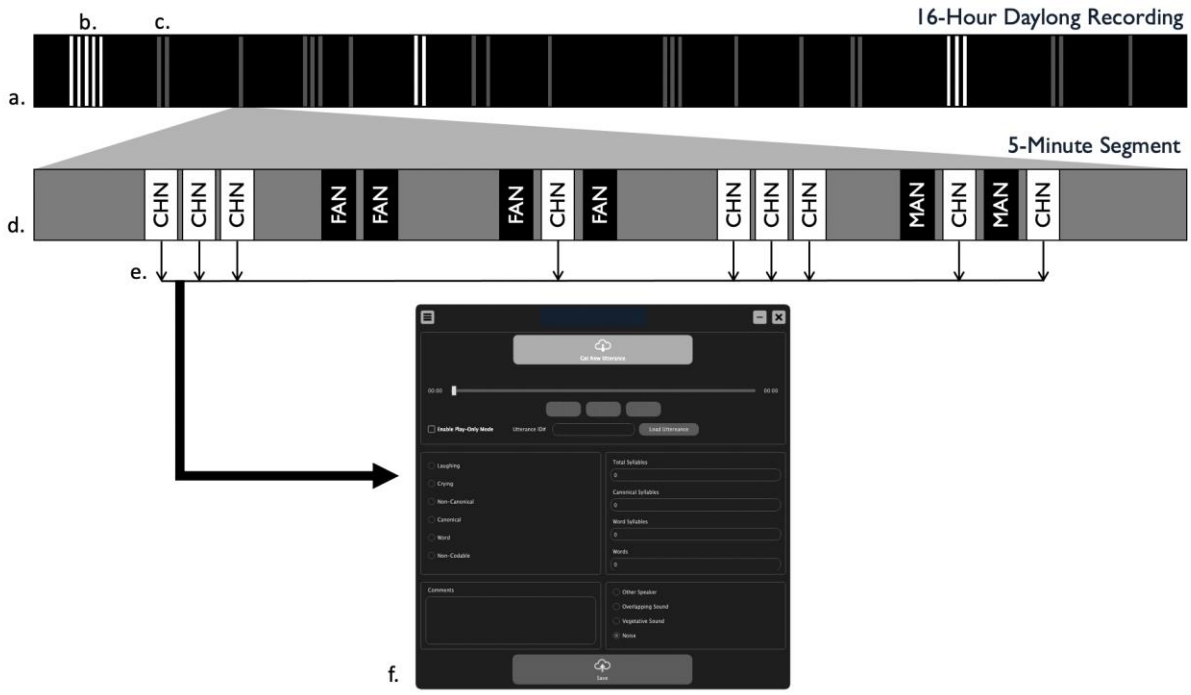


Figure 1. Graphic of the ULP utterance selection process. a) LENA divides a full daylong recording into 5-minute segments. b) The ULP scripts select 10 5-minute segments from the daylong recording that correspond to the periods of highest child volubility. c) The ULP scripts randomly select 20 5-minute segments from the remaining 5-minute segments of the daylong recording. d) An example of a single selected 5-minute segment, which contains utterances tagged by LENA as the Key Child (CHN) and utterances tagged by LENA as other speakers (e.g., FAN). e) Utterances tagged by LENA as the Key Child are selected for further annotation. f) Selected utterances are tagged by coders using the ULP coding interface.

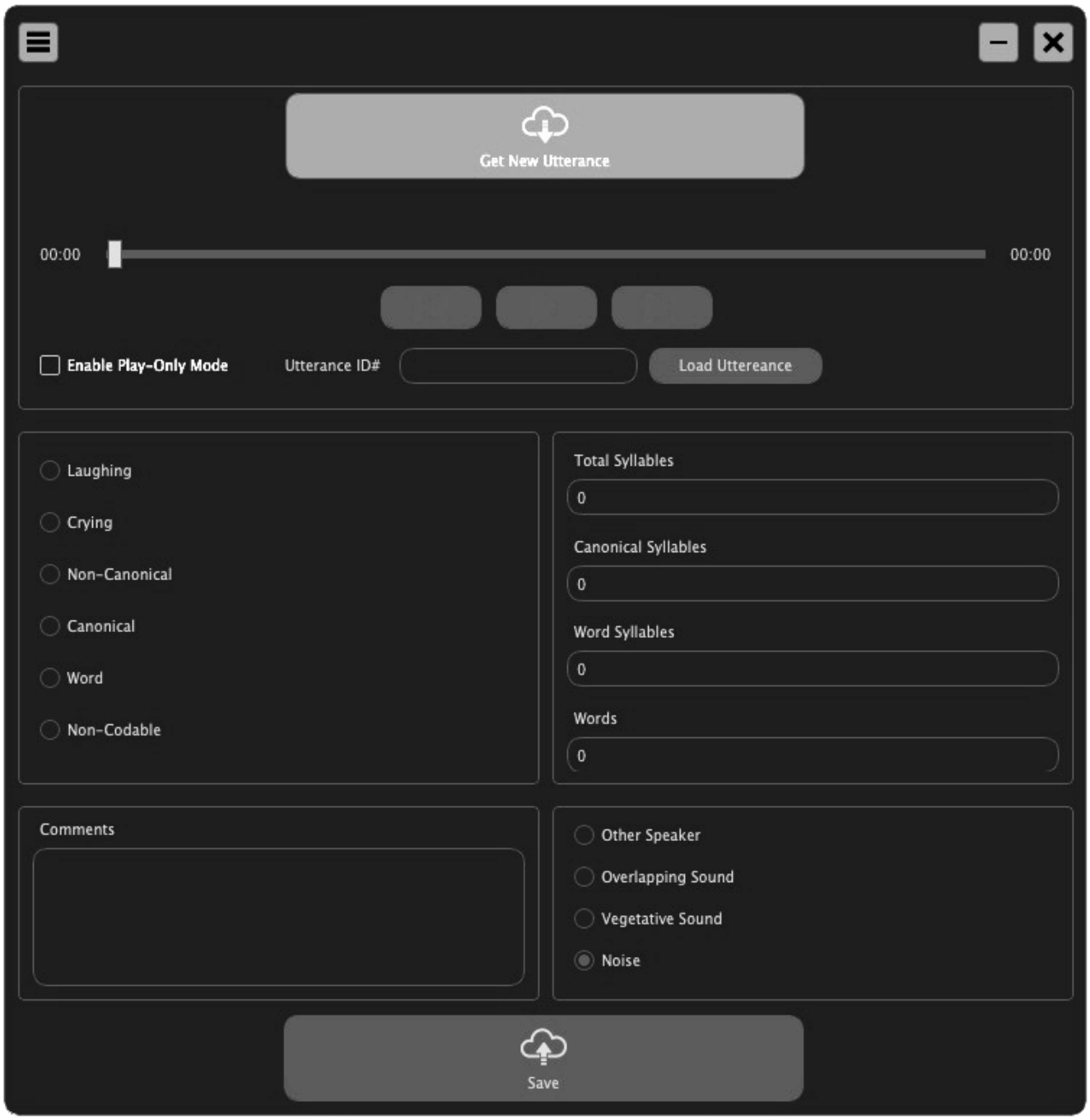


Figure 2. ULP Coding Interface.

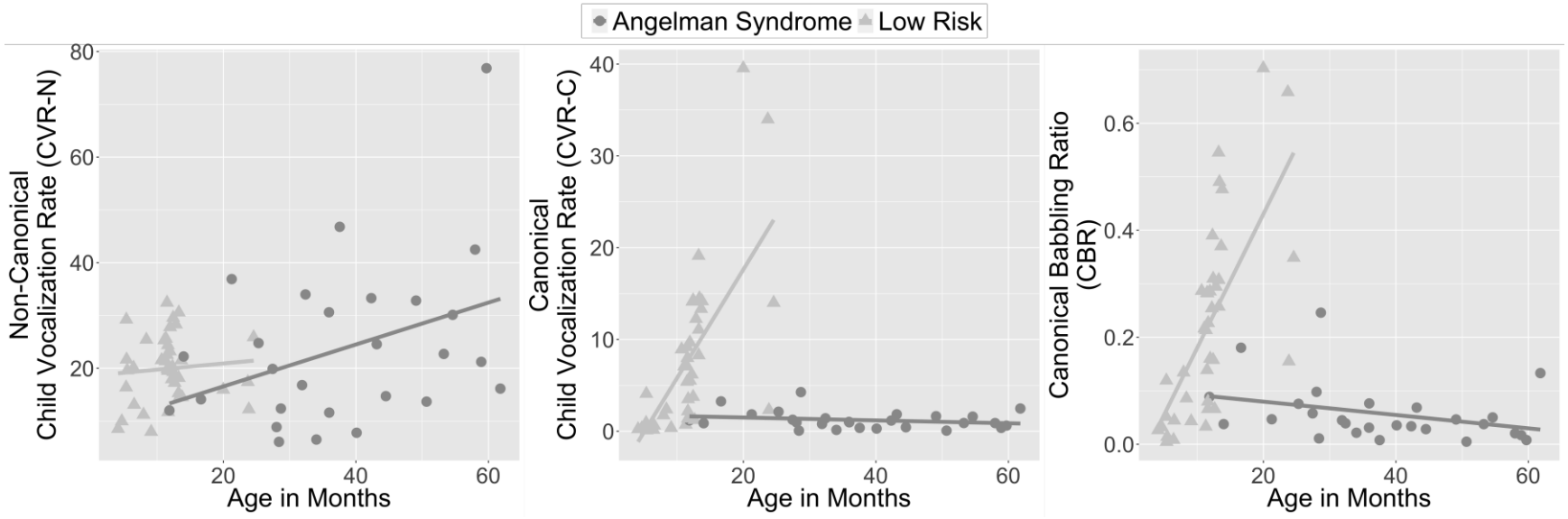


Figure 3. Associations of ULP Metrics with Age in Low-Risk Controls and Angelman Syndrome.

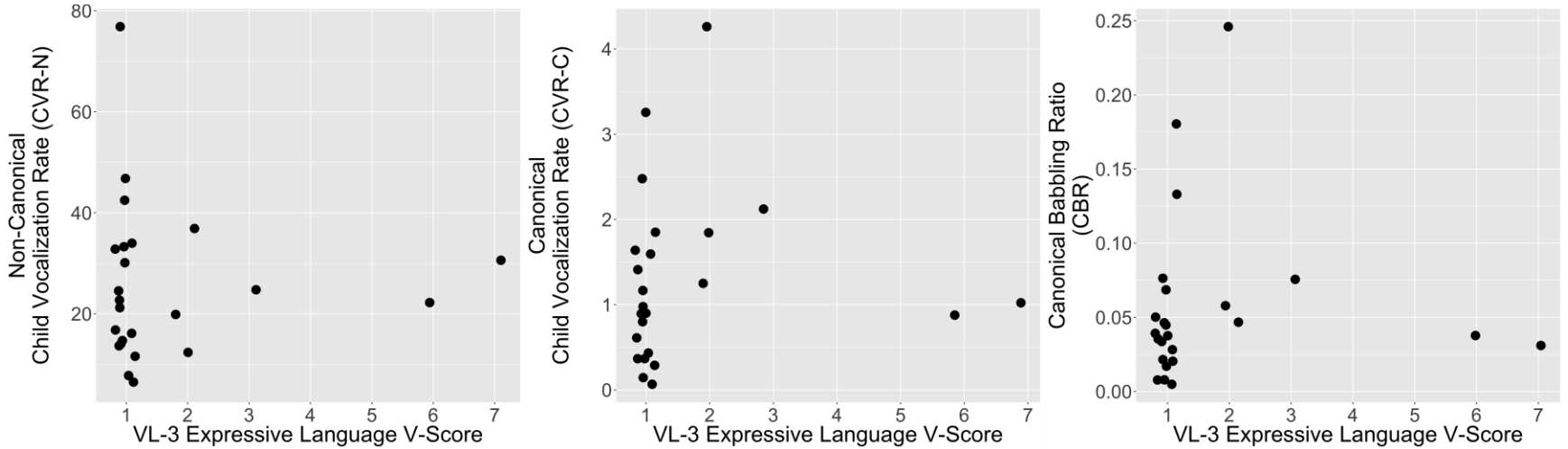


Figure 4. Distribution of ULP Metric Scores Within VL-3 Expressive V-Scores for AS Sample.
*Note*. ULP metrics demonstrate variability among participants who received the same VL-3 Expressive Language V-Score.

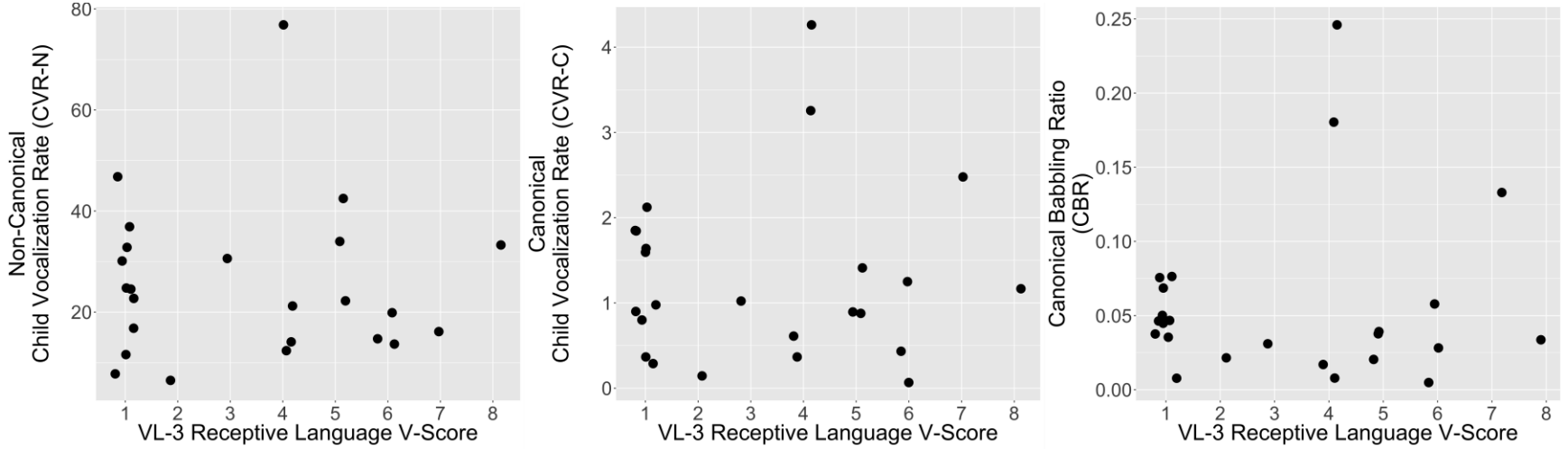


Figure 5. Distribution of ULP Metric Scores Within VL-3 Receptive V-Scores for AS Sample.
*Note*. ULP metrics demonstrate variability among participants who received the same VL-3 Receptive Language V-Score.

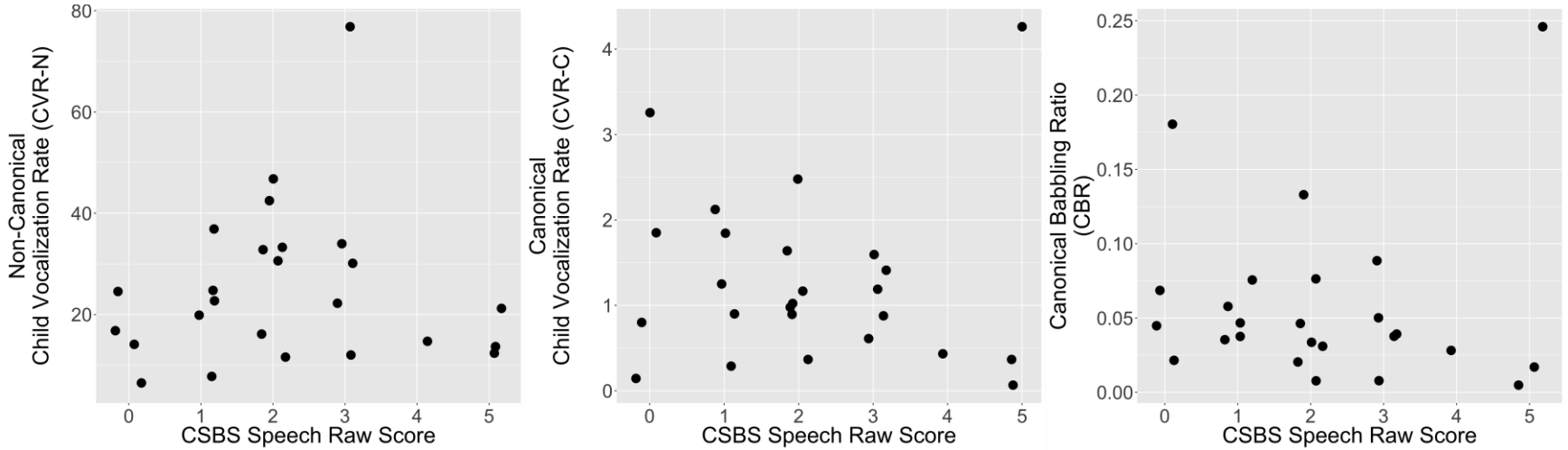


Figure 6. Distribution of ULP Metric Scores Within CSBS Speech Raw Scores for AS Sample.
*Note*. ULP metrics demonstrate variability among participants who received the same CSBS Speech Raw Score.

Table 1.

*Abbreviations*

| | |
|---|---|
| LENA | Language ENvironment Analysis |
| ULP | Utterance-level processing |
| LRC | Low-risk control |
| AS | Angelman syndrome |
| IDD | Intellectual and developmental disabilities |
| CVR | Child vocalization rate |
| CVR-N | Rate of non-canonical vocalizations |
| CVR-C | Rate of canonical vocalizations |
| CBR | Canonical babbling ratio |
| VL-3 | Vineland Adaptive Behavior Scales, 3rd Edition |
| CSBS | Communication and Symbolic Behavior Scales – Infant Toddler Checklist |

Table 2.

*Demographic Information*

| | LRC (*n*=39) | | AS (*n*=27) | |
|---|---|---|---|---|
| | ***M (SD)*** | **Range** | ***M (SD)*** | **Range** |
| Age in Months at Recording | 11.64 (4.87) | 4.11-24.57 | 37.98 (14.41) | 11.86-61.80 |
| Household Income | $73,514 ($51,127) | $23,000-$300,000 | $130,240 ($84,526) | $0-$300,000 |
| **Race** | ***n*** | **%** | ***n*** | **%** |
| White | 32 | 82% | 17 | 63% |
| Black | 1 | 3% | 0 | 0% |
| More than One Race | 5 | 12% | 3 | 11% |
| Not Reported | 1 | 3% | 7 | 26% |
| **Ethnicity** | ***n*** | **%** | ***n*** | **%** |
| Hispanic/Latino | 1 | 3% | 0 | 0% |
| Not Hispanic/Latino | 36 | 92% | 20 | 74% |
| Not Reported | 2 | 5% | 7 | 26% |
| **Maternal Education** | ***n*** | **%** | ***n*** | **%** |
| High School Degree | 1 | 3% | 1 | 4% |
| Some College | 6 | 15% | 5 | 19% |
| 2-Year Degree | 2 | 5% | 2 | 7% |
| 4-Year Degree | 11 | 28% | 8 | 30% |
| Professional Degree | 15 | 38% | 9 | 33% |
| Doctoral Degree | 3 | 8% | 2 | 7% |
| Not Reported | 1 | 3% | 0 | 0% |
| **Deletion Type** | ***n*** | **%** | ***n*** | **%** |
| Maternal Deletion | -- | -- | 21 | 78% |
| UBE3A Mutation | -- | -- | 1 | 4% |
| Paternal Uniparental Disomy | -- | -- | 1 | 4% |

| | | | | |
|---|---|---|---|---|
| Unknown | -- | -- | 4 | 15% |
| **Standardized Language Scores** | ***M (SD)*** | **Range** | ***M (SD)*** | **Range** |
| VL-3 Adaptive Behavior Composite[a] | 101.34 (12.21) | 82-128 | 49.25 (7.74) | 34-69 |
| VL-3 Receptive V-Score[b] | 14.29 (2.95) | 8-18 | 3.29 (2.31) | 1-8 |
| VL-3 Receptive Age Equivalent | 10.86 (6.21) | 0-30 | 9.33 (5.21) | 2-19 |
| VL-3 Expressive V-Score[b] | 15.34 (2.46) | 11-24 | 1.67 (1.58) | 1-7 |
| VL-3 Expressive Age Equivalent | 11.94 (8.46) | 2-52 | 7.42 (4.58) | 1-18 |
| CSBS Total Standard Score[a,d] | 94.63 (15.71) | 70-128 | 66.40 (3.13) | 65-72 |
| CSBS Speech Composite Score[c,d] | 9.30 (3.17) | 3-17 | 4.60 (3.05) | 3-10 |

*Note.* CSBS = Communication and Symbolic Behavior Scales; VL-3 = Vineland Adaptive Behavior Scales, 3rd Edition; M = mean, SD = standard deviation.[a]VL-3 Adaptive Behavior Composite and CSBS Total Standard Score have a mean of 100 and standard deviation of 15. [b]VL-3 V-Scores have a mean of 15 and a standard deviation of 3. [c]CSBS Speech Composite Scores have a mean of 10 and a standard deviation of 2.[d]CSBS normative scores are only available for 30 LRC participants (77%) and 5 AS participants (23%) who were within the CSBS normative age range (i.e., 6-24 months).

Table 3.

*ULP Utterance Agreement in LRC*

| | 0% Agreement[a] | | 66% Agreement[b] | | 100% Agreement[c] | |
|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % |
| Annotation | 889 | 3% | 9,570 | 38% | 14,997 | 59% |
| # Canonical Syllables | 780 | 3% | 5,782 | 23% | 18,894 | 74% |
| # Non-Canonical Syllables | 4,088 | 16% | 11,846 | 47% | 9,522 | 37% |

*Note*. Rows sum to the total number of LRC utterances coded (i.e., 25,456).

Table 4.

*ULP Utterance Agreement in AS*

| | 0% Agreement[a] | | 66% Agreement[b] | | 100% Agreement[c] | |
|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % |
| Annotation | 525 | 4% | 4,348 | 30% | 9,661 | 66% |
| # Canonical Syllables | 141 | 1% | 1,811 | 12% | 12,582 | 87% |
| # Non-Canonical Syllables | 1,885 | 13% | 6,646 | 46% | 6,003 | 41% |

*Note*. Rows sum to the total number of AS utterances coded (i.e., 14,534).

Table 5.

*ULP Metric Descriptives and Group Comparisons*

| | | Low-Risk Controls | | Angelman Syndrome | | Mann-Whitney U Test | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | M (SD) | Range | M (SD) | Range | U | p | p' | d |
| Canonical Babbling Ratio | | 0.22 (0.18) | 0.00-0.70 | 0.06 (0.05) | 0.00-0.25 | **207** | **<.001** | **<.001** | **1.16** |
| Speech Utterances | Rate | 15.63 (5.43) | 6.43-25.67 | 13.18 (6.41) | 3.90-27.27 | 392.5 | .082 | 1.000 | .43 |
| | %[a] | 71.56 (12.16) | 34.11-93.10 | 72.28 (12.59) | 44.10-97.15 | 528 | .990 | 1.000 | .06 |
| Word Syllables | Rate | 0.33 (0.79) | 0.00-3.94 | 0.05 (0.14) | 0.00-0.58 | *314.5* | *.002* | *.128* | *.46* |
| | %[b] | 0.84 (1.56) | 0.00-6.64 | 0.21 (0.68) | 0.00-3.25 | *311* | *.002* | *.112* | *.50* |
| Canonical Syllables | Rate | 7.75 (8.68) | 0.10-39.55 | 1.22 (0.98) | 0.07-4.26 | **231.5** | **<.001** | **.007** | **.99** |
| | %[b] | 22.29 (17.83) | 0.46-66.53 | 5.82 (5.63) | 0.49-25.63 | **206** | **<.001** | **<.001** | **1.18** |
| Non-Canonical Syllables | Rate | 19.93 (6.30) | 7.96 (32.43) | 23.69 (15.38) | 6.06-76.84 | 564 | .629 | 1.000 | .35 |
| | %[b] | 76.88 (18.53) | 26.83-99.54 | 93.97 (5.96) | 74.37-99.51 | **858** | **<.001** | **<.001** | **1.17** |
| Non-Speech Utterances | Rate | 1.62 (1.79) | 0.10-10.37 | 0.74 (0.74) | 0.13-2.87 | *233.5* | *.012* | *.682* | *.59* |
| | %[a] | 7.54 (9.07) | 0.45-51.75 | 4.03 (3.12) | 0.48-10.72 | 284 | .092 | 1.000 | .48 |
| Don't Mark Utterances | Rate | 3.81 (2.45) | 0.90-13.23 | 3.50 (2.33) | 0.43-8.90 | 471 | .473 | 1.000 | .13 |
| | %[a] | 17.84 (9.52) | 5.03-43.11 | 20.81 (11.72) | 1.54-52.31 | 609 | .285 | 1.000 | .29 |
| Non-Consensus Utterances | Rate | 0.76 (0.44) | 0.13-1.77 | 0.65 (0.44) | 0.03-1.73 | 438 | .251 | 1.000 | .26 |
| | %[a] | 3.35 (1.40) | 0.96-6.97 | 3.61 (2.10) | 0.56-8.05 | 517 | .907 | 1.000 | .15 |

*Note.* Bolded values are significant after correcting for multiple comparisons; italic values are significant but not after correcting for multiple comparisons; M = mean, SD = standard deviation, U = Mann-Whitney U test statistic, p = unadjusted p-value, p' = adjusted p-value, d = Cohen's d. [a]Speech, Non-Speech, Don't Mark, and Non-Consensus utterance percentages indicate the proportion of total utterances represented by each category (i.e., percentages of Speech, Non-Speech, Don't Mark, and Non-Consensus utterances should add up to approximately 100%). [b]Word, Canonical, and Non-Canonical syllable percentages indicate the proportion of total speech syllables represented by each category (i.e., percentage of Word, Canonical, and Non-Canonical syllables should add up to approximately 100%).

Table 6.

*Concurrent Associations of ULP Metrics with Parent-Reported Language Assessments*

| | Low-Risk Controls | | | | | | | | | | Angelman Syndrome | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CVR-C | | | CVR-N | | | CBR | | | | CVR-C | | | CVR-N | | | CBR | | |
| | M (SD) | ρ | p | p' | ρ | p | p' | ρ | p | p' | M (SD) | ρ | p | p' | ρ | p | p' | ρ | p | p' |
| Age at Recording | 11.64 (4.87) | **.76** | **<.001** | **<.001** | .06 | .716 | 1.000 | **.79** | **<.001** | **<.001** | 37.98 (14.41) | -.22 | .266 | 1.000 | .30 | .126 | 1.000 | -.42 | .029 | 1.000 |
| VL-3 | M (SD) | ρ | p | p' | ρ | p | p' | ρ | p | p' | M (SD) | ρ | p | p' | ρ | p | p' | ρ | p | p' |
| Receptive Language raw score | 21.37 (12.98) | **.68** | **<.001** | **<.001** | -.01 | .941 | 1.000 | **.68** | **<.001** | **<.001** | 18.12 (11.72) | .02 | .932 | 1.000 | .07 | .763 | 1.000 | -.16 | .454 | 1.000 |
| Expressive Language raw score | 18.80 (14.81) | **.66** | **<.001** | **<.001** | -.11 | .536 | 1.000 | **.70** | **<.001** | **<.001** | 12.00 (5.79) | .02 | .931 | 1.000 | .19 | .387 | 1.000 | -.21 | .325 | 1.000 |
| EL Item: (3 one-syllable speech sounds) | 1.79 (0.59) | *.51* | *.003* | *.162* | .13 | .481 | 1.000 | *.48* | *.005* | *.285* | 1.46 (0.88) | .21 | .332 | 1.000 | .40 | .057 | 1.000 | -.04 | .853 | 1.000 |
| EL Item: (Babbles in strings of sounds) | 1.71 (0.72) | *.53* | *.001* | *.092* | .01 | .950 | 1.000 | **.56** | **.001** | **.048** | 0.46 (0.83) | .14 | .534 | 1.000 | .05 | .811 | 1.000 | -.02 | .929 | 1.000 |
| CSBS | M (SD) | ρ | p | p' | ρ | p | p' | ρ | p | p' | M (SD) | ρ | p | p' | ρ | p | p' | ρ | p | p' |
| Speech scale raw score | 5.97 (3.18) | **.62** | **<.001** | **.007** | -.07 | .681 | 1.000 | **.66** | **<.001** | **<.001** | 2.12 (1.54) | -.17 | .426 | 1.000 | .06 | .792 | 1.000 | -.24 | .251 | 1.000 |
| Understanding subscale raw score | 2.94 (1.70) | *.54* | *.001* | *.068* | .05 | .777 | 1.000 | *.52* | *.002* | *.109* | 2.96 (1.67) | -.01 | .961 | 1.000 | .17 | .423 | 1.000 | -.18 | .387 | 1.000 |
| Sounds Item: (uses sounds to get attention) | 1.49 (0.70) | .21 | .242 | 1.000 | .14 | .425 | 1.000 | .22 | .216 | 1.000 | 1.08 (0.70) | -.10 | .648 | 1.000 | .16 | .459 | 1.000 | -.17 | .434 | 1.000 |
| Sounds Item: (strings sounds together) | 1.09 (0.92) | **.61** | **<.001** | **.007** | -.09 | .623 | 1.000 | **.63** | **<.001** | **.007** | 0.20 (0.50) | -.05 | .835 | 1.000 | *-.44* | *.031* | *1.000* | .07 | .738 | 1.000 |
| Sounds Item: (number of consonants) | 2.34 (0.94) | *.53* | *.001* | *.092* | -.24 | .176 | 1.000 | **.57** | **.001** | **.035** | 0.72 (0.79) | -.13 | .553 | 1.000 | .12 | .580 | 1.000 | -.23 | .276 | 1.000 |
| Words Item: (number of single words) | 0.91 (1.12) | *.52* | *.002* | *.109* | -.04 | .809 | 1.000 | *.54* | *.001* | *.074* | 0.12 (0.33) | *-.43* | *.034* | *1.000* | .10 | .658 | 1.000 | *-.52* | *.010* | *.570* |
| Words Item: (two words together) | 0.14 (0.43) | *.45* | *.008* | *.476* | .02 | .917 | 1.000 | *.43* | *.011* | *.594* | 0.00 (0.00) | -- | -- | -- | -- | -- | -- | -- | -- | -- |

*Note.* Bolded values are significant after correcting for multiple comparisons; italic values are significant but not after correcting for multiple comparisons. CSBS = Communication and Symbolic Behavior Scales; CBR = canonical babbling ratio; CVR-C = canonical child vocalization rate; CVR-N = non-canonical

child vocalization rate; VL-3 = Vineland Adaptive Behavior Scales, 3rd Edition; M = mean, SD = standard deviation, ρ = Spearman's rho, p = unadjusted p-value, p' = adjusted p-value, d = Cohen's d.