Can automated vocal analyses over child-centered audio recordings be used to predict speech-language development?

Carissa M. Ott (carissaott@g.ucla.edu)
Department of Psychology, UCLA, 1202 W. Johnson Street
Los Angeles, CA, USA 90095

Margaret Cychosz (mcychosz4@ucla.edu)
Department of Linguistics, UCLA, 335 Portola Plaza
Los Angeles, CA, USA 90095

Abstract

Understanding how children's spontaneous language behavior relates to standardized metrics of language development remains a crucial challenge in developmental science, particularly given the time and resources required for many traditional lab-based assessments. This study investigates whether automated analysis of naturalistic, child-centered audio recordings can index the developmental trajectory of speech-language abilities. Using a longitudinal design following N=130 preschoolers, we employed deep learning methods to compute CANONICAL PROPORTION—a theoretically-motivated metric that reflects both speech motor control development and phonological representation building—from naturalistic, child-centered audio recordings at age 3 years. Canonical proportion measures significantly predicted multiple dimensions of speech-language development longitudinally, formally assessed in the lab one year later at age 4. The strongest relationships were found for consonant articulation skill and vocabulary size, suggesting that early speech production patterns may moderately index numerous later facets of language development. These findings outline a potential relationship between children's spontaneous, everyday language behavior and more traditional language development metrics, while demonstrating the potential for automated measures to expand and diversify research in developmental science.

Keywords: speech; language development; naturalistic observation; automated assessment; child development

Introduction

Children's speech capacities change rapidly in the first 5 years of life (Oller, 2000). In preschoolhood, the stages of both speech perception and production development are traditionally assessed via tightly controlled behavioral experiments such as looking while listening eye-tracking paradigms for perceptual development (Fernald, Zangl, Portillo, & Marchman, 2008) or targeted, elicited phoneme production (Edwards, Beckman, & Munson, 2004). While findings from this type of experimentation have increased our understanding of children's speech-language development, there are several drawbacks. First, the paradigms are almost exclusively conducted in formal university lab environments, skewing the samples and resulting in biases along several dimensions (e.g. linguistic, socioeconomic, cultural). For example, only 1.5% of the world's languages have ever been represented in mainstream language acquisition journals (Kidd & Garcia, 2022), a fact that limits the generalizability of results and proposed theories of language and cognitive development (Blasi, Henrich, Adamou, Kemmerer, & Majid, 2022). Second, in-lab assessments are extremely time-consuming and resource-intensive, for both children and researchers, leading to much work in developmental cognitive science lacking sufficient statistical power (Nielsen, Haun, Kärtner, & Legare, 2017). Finally, in-lab assessments, especially for children, may not reflect a child's true, underlying speech-language capacities if children are uncooperative, fatigued, or nervous during testing (Conti-Ramsden & Durkin, 2012). What is needed is a method to model children's developing speech-language capabilities that is portable, easy-to-administer, and ideally can be collected from children as they naturally go about their day.

Here we propose one such method, and we give an overview of the idea that we can automatically estimate stages of a child's speech-language development from long-form, child-centered audio recordings of children's everyday lives (Sy, Havard, Lavechin, Dupoux, & Cristia, 2023). Experimentalists and clinicians alike have traditionally assessed children's developing speech-language capacities using painstaking manual annotation and experimentation that require hours of labor and months of training. The ability to forgo some or part of these tasks would thus be a significant advance in developmental science, potentially allowing researchers to diversify and increase their participant samples.

Quantifying child speech maturity in preschoolhood

A key metric used to assess speech development in early childhood is CANONICAL PROPORTION (CP), which tracks the proportion of more complex, "canonical" vocalizations consisting of both consonants and vowels in quick transition (e.g. "ba") to all of child's speech-like vocalizations, including simpler, "non-canonical" vocalizations that consist of single vowels ("aa") or consonants ("mmm") (Oller, 2000). CP is a crucial measure of a child's developing speech capacities because it reflects the increasing complexity of vocalizations as children develop, providing insights into how children are reaching developmental milestones in speech and language. Critically, for our purposes, *CP continues to increase well past the pre-lexical period*. Recent work shows that CP continues to increase up through at least 6 years (Hitczenko et al., 2023), demonstrating the potential for CP to indicate how speech continues to develop and mature long after the transition from babbling to word use.Moreover, canonical babbling ratio (CBR), while widely used in infancy research, is similarly time-intensive to annotate and lacks strong predictive validity beyond early developmental stages, making auto-

mated CP estimation a contribution along both methodological and theoretical dimensions.

Relating canonical proportion to additional speech-language measures

While previous work has established that CP increases with age, and thus serves as a developmental indicator of *speech* maturity in children up to 6 years (Hitczenko et al., 2023; Long et al., 2024), there is limited understanding of how CP relates to other areas of child speech-language development. There are, however, clear reasons to anticipate that CP may predict some areas of children's developing speech and language. For one thing, there are well-known, positive relationships between 3- and 4-year-olds' speech articulation skills (e.g. number of consonants correctly produced as assessed in the lab during targeted phoneme elicitation tasks) and numerous areas of language development, including children's receptive and expressive vocabulary sizes (Sosa & Stoel-Gammon, 2012), phonological awareness (Foy & Mann, 2003), and phonological working memory (the ability to temporarily store and recall of sound-based information) (Adams & Gathercole, 1996). Rudimentary versions of these relationships between speech production and nascent language begin to develop as early as the first year of life: for example, infant speech volubility at 6 months of age positively predicts vocabulary development and language complexity abilities at one year (Lee, Jhang, Relyea, Chen, & Oller, 2018).

Evidence of connections between preschoolers' speech production maturity and phonological working memory have been documented using nonword repetition (NWR) tasks, where children are asked to repeat novel, phonotactically-probable words after a model speaker (e.g. "sudras"). Children's ability to faithfully repeat nonwords is strongly related, although not perfectly predicted by, their speech articulation skills (Gathercole, 2006). NWR requires children to process and store phonological information, which is directly relevant to articulating sounds accurately (Adams & Gathercole, 1996). Similarly, since CP tracks the transition to more mature speech, it may also relate to a child's ability to handle and integrate increasingly complex phonological information. This suggests a potential relationship between CP and NWR ability, supporting the idea that these measures might be linked indicators of broader language development, specifically in terms of phonological processing and speech articulation.

Current Study

In this study we ask the following overarching research question: How does automated computation of CP over daylong, child-centered audio recordings predict different areas of speech-language development in a large, longitudinal sample of 3- and 4-year-olds (N=130)? Specifically, using a pretrained child speech maturity classifier (details in Methods), we compute each child's CP over a longform audio recording made in the child's home at age 3 years. We then relate the measure of CP to different standardized, in-lab assessments of speech-language development that the *same* children completed 1 year later at age 4: (1) targeted consonant articulation skill, (2) receptive vocabulary, (3) phonological working memory (Edwards et al., 2004; Gathercole, 2006), (4) lexical-phonological access, and (5) phonological awareness. While previous research in each of these domains of speech and language suggests a positive relationship (e.g. relationships between child speech maturity and vocabulary size (Sosa & Stoel-Gammon, 2012)), we make the following more targeted hypotheses: since CP is foremost a speech production metric, that likewise indexes phonological representation building, we anticipate the strongest relationships between CP and those skills requiring integration of these cognitive systems (consonant articulation skill), followed by meta-phonological skills (phonological working memory and awareness), and finally a weaker, though still positive relationship between CP and skills related to the lexicon (lexical-phonological access, vocabulary size).

Methods

Participants in this study were part of a longitudinal study of preschool children (66 boys; 64 girls; 0.78% Asian, 14.73% African American, 82.17% white, 2.33% other; 97.67% non-Hispanic, 2.33% Hispanic). All children were typically-developing, monolingual speakers of American English and completed a daylong audio recording using the Language ENvironment Analysis (LENA) system (Xu, Yapanel, & Gray, 2009) at approximately 3 years of age ('Timepoint 1'; see Table 1 for exact ages), and then returned one year later at approximately age 4 ('Timepoint 2') to complete a number of controlled experimental and standardized speech-language measures. All participants passed a standard hearing screening at Timepoints 1 and 2.

Table 1: Participant information at time of daylong recording and descriptive statistics for speech utterances and clips extracted from daylong recordings.

	M	SD	Range
Demographics			
Age at LENA recording (mos)	33.75	4.30	28 - 49
Age at speech-language assessments (mos)	45.08	3.55	39 - 55
Maternal education (see text)	5.88	1.36	1 - 7
Num. of residents in household	4.23	1.09	2 - 8
Duration of LENA recording (hours)	15.77	1.13	6.65 - 16
Speech Utterances/Clips			
# of utterances/recording	3060.53	1447.75	121 - 7428
# of clips/recording	7407.43	3489.91	491 - 17478
Duration of processed clips (ms)	492.55	99.39	200 - 690

The maternal education variable was quantified on a scale from 1 to 7, with the categories representing increasing lev-

els of educational attainment: 1 = Less Than High School, 2 = GED, 3 = High School Diploma, 4 = Some College (<2 years), 5 = Some College (>2 years), 6 = College Degree, and 7 = Graduate Degree.

Data Collection

The child wore a small, lightweight LENA recorder (2"x3"; 2 oz.) in a specialized vest throughout the day (appx. 16 hours), excluding water activities. Devices and instructions were provided by mail or at the research lab, and recordings were made on typical non-school days to capture naturalistic language environments and ensure a relatively consistent recording environment between households. See Table 1 for additional household-level details.

Computing canonical proportion from recordings

Pre-processing child-centered audio recordings See Figure 1 for data processing flow. To pre-process the recordings, we follow methods laid out in Hitczenko et al. (2023) and Cychosz et al. (2021): The LENA system includes a proprietary speaker diarization algorithm that, among other features, segments raw audio into speaker categories (e.g. target child, adult female). We used this algorithm to identify all speech utterances from the target child in each recording (M=3060.53 target child utterances/recording (SD=1447.75) range=121-7428). To prepare the child speech utterances for automated classification, we then chopped each utterance into smaller audio clips of approximately 500-ms (M=7407.43 clips/child (SD=3489.91) range=80-17478). The 500-ms duration was chosen because the pre-trained classifier that we employed (see next section) was trained over child vocalization clips of this length. In all, our pre-processing pipeline generated 2,037,042 total clips from the 130 children for automated annotation.

Using a pre-trained child speech maturity classifier to compute canonical proportion We employed a previously-trained, deep learning child speech maturity classification algorithm, which classifies child vocalizations into five categories: "canonical", "non-canonical", "crying", "laughing", and "other/junk" (e.g. animal sounds, no sound). The model processes audio clips through a convolutional neural network (CNN) feature extractor. Features are then passed through a transformer architecture (12 layers, 768 hidden dimensions, 3072 inner dimensions, 8 attention heads) that was pre-trained on hundreds of hours of unlabeled LibriSpeech data (Baevski, Zhou, Mohamed, & Auli, 2020) and then fine-tuned on labeled task-specific data. Specifically, the model was fine-tuned on 46,674 audio clips of child vocalizations, approximately 500-ms in length, that were extracted from naturalistic, longform child-centered audio data. Children in the training data were aged 2 months-6 years and were acquiring a variety of languages, including English. None of the children in the current study were represented in the models' training data. See Table 2 for model performance statistics.

The model achieved classification accuracy comparable to human annotators (Cohen's K=0.451) and was robust across rural and urban child rearing settings. This agreement level reflects the task's difficulty, as hand-labeling canonical syllables has only moderate reliability due to noise, overlap, and subjective boundary judgments. Additional detail on model architecture, benchmarks, and train/dev/test data structures available in Zhang, Suresh, Hitczenko, Cristia, and Cychosz (under review).

Table 2: Performance statistics for child speech maturity classifier. UAR = Unweighted Average Recall. AUC = Area Under the Curve.

	UAR	AUC
Overall model performance	71.0	NA
Category		
Canonical	78.47	0.94
Non-Canonical	63.58	0.85
Crying	75.96	0.95
Laughing	69.70	0.95
Junk	67.42	0.92

Using this model, we classified each 500-ms clip into 1 of the 5 categories: (0) Canonical, (1) Non-Canonical, (2) Crying, (3) Laughing, or (4) Junk. Finally, for each child, we computed the CP by taking the number of clips classified as 'canonical' and dividing this by the sum of all speech-like clips (canonical+non-canonical).

Metrics of speech-language development assessed

To comprehensively evaluate the relationship between CP and preschoolers' future speech-language development, we assessed a battery of various speech-language measures when the children were approximately 4 years of age (see Table 3 for scores/results). When available, we employ standard scores, normalized for child gender and age. Not all children completed every assessment; while 130 children participated in the study, some children completed only certain assessments, leading to fewer than 130 children completing any one particular test. The number and percent of children who completed each specific assessment are provided in Table 3. Unless otherwise noted below, children's responses on the assessments were recorded in real time and final scoring was conducted following the test.

Consonant articulation skill was assessed with the Sounds-in-Words portion of the Goldman-Fristoe Test of Articulation, 2nd edition (GFTA-2) (Goldman & Fristoe, 2000); see Usha and Alex (2023) and Macrae (2017) for detail. Children's productions were audio-recorded for offline scoring following standardized instructions available in the test's instruction manual.

Receptive Vocabulary was assessed using the Peabody Picture Vocabulary Test, 4th edition (PPVT-4) (Dunn & Dunn, 2007). Children were presented with a series of images and asked to select the picture that best matched a spoken word. Responses were scored in real time and scores computed following test assessment, per test manual instructions.

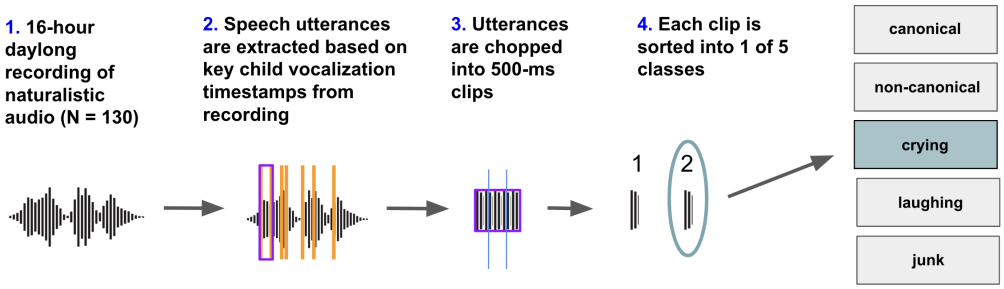

Figure 1: Pre-processing pipeline for audio recordings.

Table 3: Descriptive statistics for controlled, in-lab speech-language assessment completed at Timepoint 2. GFTA = Goldman Fristoe Test of Articulation-2. PPVT = Peabody Picture Vocabulary Test-4. CTOPP = Comprehensive Test of Phonological Processing-2. SS = Standard Score. ES = Elision Scaled. RWR = Real Word Repetition. NWR = Nonword Repetition.

	M	SD	Range	N (%) Children Completed
GFTA-2 (SS)	92.98	12.74	61.00 - 119.00	118 (90.77)
PPVT-4 (SS)	119.29	16.68	75.00 - 152.00	118 (90.77)
CTOPP-2 (ES)	10.76	2.23	6.00 - 16.00	102 (78.46)
RWR Accuracy	0.96	0.03	0.76 - 0.99	67 (51.54)
NWR Accuracy	0.95	0.05	0.79 - 1.00	56 (43.08)

Phonological awareness skills were evaluated using the elision subtest of the Comprehensive Test of Phonological Processing, Second Edition (CTOPP-2) (Wagner, 2013), which assesses the ability to delete specific sounds or parts of words; see Wagner (2013) for detail.

Lexical-phonological access was assessed through a picture-prompted verbal repetition task: real word repetition (RWR). The RWR task assesses how well children have established stable, accessible lexical-phonological representations of familiar words—a key developmental achievement reflecting the integration of lexical forms with practiced phonological/motor routines. Stimuli consisted of 23 familiar words selected from the "Toddler Says" portion of the MacArthur-Bates Communicative Development Inventory (Fenson et al., 2007). See Munson, Logerquist, Kim, Martell, and Edwards (2021) for further detail on RWR task administration and scoring.

Phonological working memory was assessed through a picture-prompted, verbal production task—nonword repetition (NWR)—which evaluated a child's ability to repeat novel, phonotactically-probable words, and measured the child's capacity to temporarily store and manipulate phonological information without confounds related to lexical access/frequency (Gathercole, 2006). Children repeated 23 bi-

Table 4: Descriptive statistics for the number of clips of each class used in computing each child's CP (clips/child).

	M	SD	Range
Canonical	2897.10	1693.76	5 - 6414
Non-Canonical	11760.21	969.75	13 - 4355
Crying	587.16	417.90	16 - 1964
Laughing	231.05	158.72	10 - 956
Junk	786.98	601.16	30 - 5033

syllabic nonwords, (e.g. "kaemig" [kæmɪg]). Responses were audio recorded; see Erskine, Munson, and Edwards (2020) for detail on administration and scoring.

Results

Computing canonical proportion at 3 years of age

We first computed CP from each child's recording at age 3 years. We found wide variability between children in CP (M=0.59 (SD=0.12) 0.20-0.76; here 0.59 would indicate that 59% of a child's speech-like vocalizations contained at least one consonant-vowel combination), even among a single cohort of three-year-olds. This result suggests that CP is detecting variability in children's early speech production. As expected, we also found that CP increased with age (r=0.28, p<.001, even within the 28-49 month age range in the sample.

Predicting standardized speech-language outcomes at age 4 from canonical proportion at age 3

To examine how CP at age 3 predicts speech-language outcomes at age 4, we conducted a series of separate multiple linear regression analyses to predict each speech-language outcome. Modeling progressed as follows: we first fit a baseline model that included only key demographic variables known to impact child language development: child gender (dummy coded as 0/1), child age at recording (coded continuously, in months, centered and scaled), and maternal education (coded continuously, on the 1-7 scale, centered and scaled). Next, we fit an expanded model, which added CP (centered and z-score normalized) at age 3 as an additional predictor of the speech-language outcome. This scaling allows us to interpret model outputs in terms of SDs, so we can assess how many SDs of CP relate to a certain change on the formal speech-language outcomes/assessments. Log likelihood ratio tests

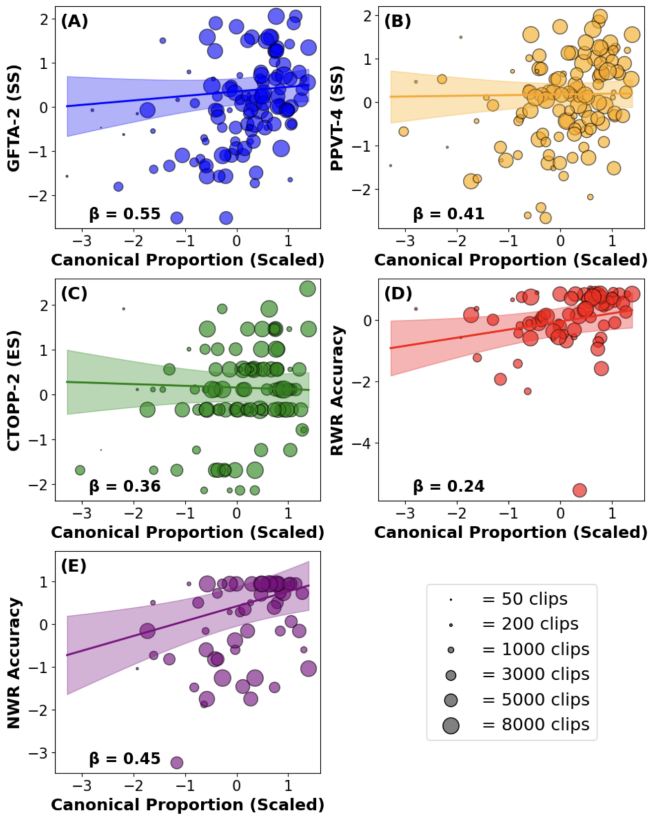

Figure 2: Scatter plots of child outcomes versus canonical proportion based on weighted and scaled model fits (to facilitate effect size comparison between assessments), with regression lines and 95% confidence intervals (shaded area) fitted from respective expanded models: (A) GFTA-2 Standard Scores, (B) PPVT-4 Standard Scores, (C) CTOPP-2 Elision, Scaled Scores, (D) RWR Accuracy, (E) NWR Accuracy. Each point represents an individual child. Point size corresponds to number of audio clips used to derive canonical proportion measure.

were conducted to determine best model fits. Model results and summary statistics are summarized in Table 5.

Consonant Articulation The baseline model, explained 9% (R^2=0.09) of the variance in articulation scores. The expanded model significantly improved model fit compared to the baseline (χ^2=13.61(1), p<.001), justifying the addition of CP. CP significantly predicted GFTA-2 scores 1 year later (β=4.28, SE=1.15, p<.001): for every SD increase in CP, the model predicts an additional 4.28 points on the GFTA-2. Given that a +/-10-point differential on the GFTA-2 standard score scale corresponds to a +/-1 SD, +/-4.28 represents a substantial effect—almost half a SD in articulation score.

Receptive Vocabulary The addition of CP to the model significantly improved model fit for receptive vocabulary (χ^2=8.29(1), p=0.004): CP showed a slightly weaker relationship with PPVT-4 scores than with speech articulation measures, though the relationship was still significant (β=4.01, SE=1.40, p=0.01, R^2=0.26), with a +/- 4.01 point increase in PPVT-4 scores for every 1 SD increase in CP, once again demonstrating an increase of nearly half an SD of the child's receptive vocabulary score for each SD increase in CP.

Phonological Awareness In the expanded model, CP showed a moderate relationship with CTOPP-2 scores (β=0.52, SE = 0.22, p = 0.02), explaining 11.5% of the variance (R^2=0.12) (loglikelihood ratio test justifying addition of CP to the baseline model: χ^2=5.61, df=(1), p=0.02). The relationship between CP and CTOPP-2 scores is thus substantially weaker than the effects observed for speech articulation and vocabulary (Figure 2c).

Lexical-Phonological Access CP did not show a reliable relationship with lexical-phonological access (β=0.01, SE=0.004, p=0.15; no improvement to baseline model fit (χ^2=2.23, df=(1), p=0.14), suggesting that CP did not increase the explanatory power of the model beyond demographic variables. It is also important to note the limited range in accuracy of RWR between children (76-99%).

Phonological Working Memory CP significantly predicted (β=0.02, SE=0.01, p=0.01) NWR scores and significantly improved model fit (χ^2=13.61, df=(1), p<.001).

Ensuring robustness to CP computation To ensure that our results were robust to differences in the number of clips used to compute each child's CP, we conducted an additional analysis where we weighted our models (full weighting process explained in project's Github repository[1]). Summarized results are presented in Table 5; in brief, we replicated all results: CP significantly predicts GFTA-2, PPVT-4, CTOPP-2, and NWR Accuracy. Relationships between CP and the outcome measures in the weighted models were just as, if not more, robust as the unweighted models.

Discussion

This study demonstrates that automated analysis of canonical proportion (CP) from naturalistic audio recordings at age 3 predicts various aspects of speech-language development at age 4. This research represents an important first step in using automated measures of child speech maturity (CP) to assess broader aspects of language development. The strongest relationships were observed for consonant articulation skills, where each SD increase in CP corresponded to approximately +.5 SD increase on the standardized measure. Contrary to our prediction, the relationship with receptive vocabulary was nearly as strong, though the relationship with articulation suggests that CP may more directly index the development of speech-motor control systems than lexical-semantic knowledge.

These findings have significant theoretical and methodological implications for the study of speech-language development. The more robust relationship between CP and later articulation skills suggests that the complexity of early vocalizations serves as a foundation for later speech sound development *even in the preschool years*. Such a finding aligns with theoretical frameworks typically conducted in infancy that propose that babbling creates a bridge between early vo-

[1] https://github.com/rissaott/spog-automated-vocal-analyses-cp

Table 5: Summary of statistical model results and log likelihood ratio tests. Baseline models refer to demographic-only models. Expanded models include CP as an additional predictor. Unscaled reported unscaled outcome measures (so coefficients can be interpreted in terms of SDs on the formal assessment) while Scaled permit comparison of effect sizes between assessments. Weighted models account for the number of canonical and non-canonical clips used to compute CP for each child, with higher weighting put on those with more clips. LRT = log likelihood ratio test. * = $P \leq 0.05, ** = P \leq 0.01, *** = P \leq 0.001$.

Variable	Model	β	SE	p-value	R^2	Log-Likelihood	χ^2	p-value (LRT)
GFTA-2 (SS)	Baseline (Unscaled)	-	-	-	0.09	-457.56	-	-
	Expanded (Unscaled)	4.28	1.15	<.001***	0.19	-450.76	13.61	<.001***
	Expanded (Scaled)	0.55	0.12	<.001***	0.21	-171.97	19.69	<.001***
	Weighted	7.01	1.55	<.001***	0.21	-469.16	19.69	<.001***
PPVT-4 (SS)	Baseline (Unscaled)	-	-	-	0.26	-476.99	-	-
	Expanded (Unscaled)	4.01	1.40	0.01**	0.31	-472.84	8.29	0.004**
	Expanded (Scaled)	0.41	0.10	<.001***	0.28	-155.25	15.31	<.001***
	Weighted	6.83	1.73	<.001***	0.28	-484.02	15.31	<.001***
CTOPP-2 (ES)	Baseline (Unscaled)	-	-	-	0.06	-220.55	-	-
	Expanded (Unscaled)	0.52	0.22	0.02*	0.12	-217.75	5.61	0.02*
	Expanded (Scaled)	0.36	0.12	0.003**	0.11	-145.68	9.15	0.002**
	Weighted	0.80	0.26	0.003**	0.12	-226.29	9.15	0.002**
RWR Accuracy	Baseline (Unscaled)	-	-	-	0.11	134.27	-	-
	Expanded (Unscaled)	0.01	0.00	0.15	0.14	135.38	2.23	0.14
	Expanded (Scaled)	0.24	0.15	0.11	0.16	-99.01	2.79	0.09
	Weighted	0.01	0.01	0.11	0.16	126.57	2.79	0.09
NWR Accuracy	Baseline (Unscaled)	-	-	-	0.07	89.95	-	-
	Expanded (Unscaled)	0.02	0.01	0.01**	0.20	94.01	8.13	0.01**
	Expanded (Scaled)	0.45	0.14	0.002**	0.30	-76.62	10.45	0.001**
	Weighted	0.02	0.01	0.002**	0.30	90.68	10.45	0.001**

cal exploration and mature speech production (Vihman, DePaolis, & Keren-Portnoy, 2014). Second, the moderate relationships found between CP and meta-phonological skills (phonological awareness and working memory) suggest that early speech production abilities may scaffold the development of phonological processing capabilities, supporting theories based on lab-based behavioral evidence that propose shared mechanisms between speech production and phonological processing systems in early development (DePaolis, Vihman, & Nakai, 2013).

Methodological Implications With further evaluation and scaling, this approach may enable more efficient and accessible methods to incorporate larger, more diverse samples into studies of cognitive development as these methods do not require an in-person visit and are potentially scalable across diverse cultural settings. By incorporating children from more diverse backgrounds, developmental scientists can stress test their theories over more representative samples.

Future Directions and Conclusion

We note that there was wide variability in the number of clips used to compute each child's CP (18-10,769). We fit additional, weighted models as one method to ensure that our results were robust to differences along this dimension, and determined that the relationships between CP and speech-language measures were not contingent upon the clip sample size. This result is not surprising given that systematic evaluations of measurement stability of a related speech metric in infancy, canonical babbling ratio, have found that N=100 samples are sufficient (Molemans, Van Den Berg, Van Severen, & Gillis, 2012). Nevertheless, CP is different from canonical babbling ratio along several dimensions (age of child, duration of typical clips) and these data were likewise noisier than many of the data reported in Molemans et al. Thus, going forward, we will evaluate the robustness of CP measures by clip sample size as well as within children by computing CP by, for example, hour of the day to compute a confidence interval around each child's measure. Analyses such as these will be important to clarify the stability and eventual practical use of automated vocal analyses. It is also important to note that the sample in this study was relatively homogeneous, with most participants being white, from mid to high SES backgrounds. Future work should aim to address this by including samples from a broader range of demographic backgrounds to ensure that CP can be reliably used as a speech-language measure across diverse groups.

Another important line of future research could be to understand how CP relates to speech-language measures in infancy, though standardized measures of phonological development are harder to derive at these early ages. Still, we believe that an important next step may be to, evaluate the relationship between CP and infant vocabulary, as reported in the Macarthur-Bates parent checklists. Such an analysis would also lend itself to cross-linguistic investigations of the relevance of CP to predict other areas of speech and language since CP is, in theory, relatively language-neutral (the model was trained on a diverse set of typologically-diverse languages) and there exist large databases of standardized reports of infants' developing vocabulary in many languages (Frank, Braginsky, Yurovsky, & Marchman, 2021) that do not exist for other outcomes such as consonant articulation skill.

Acknowledgments

The authors are grateful to Jan Edwards, Ben Munson, and Mary Beckman for generously sharing their data to be reused for this project, originally funded by National Institute on Deafness and Other Communication Disorders grant R01DC02932. Additional thanks to Theo Zhang for assistance in applying the model to this dataset.

References

Adams, A.-M., & Gathercole, S. E. (1996). Phonological Working Memory and Spoken Language Development in Young Children. *Quarterly Journal of Experimental Psychology: Section A*, *49*(1), 216–233. doi: 10.1080/027249896392874

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Proceedings of the 34th International Conference on NeurIPS Systems* (pp. 12 449–12 460).

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, S1364661322002364. doi: 10.1016/j.tics.2022.09.015

Conti-Ramsden, G., & Durkin, K. (2012). Language Development and Assessment in the Preschool Period. *Neuropsychology Review*, *22*(4), 384–401. doi: 10.1007/s11065-012-9208-z

Cychosz, M., Cristia, A., Bergelson, E., Casillas, M., Baudet, G., Warlaumont, A. S., … Seidl, A. (2021). Vocal development in a large-scale crosslinguistic corpus. *Developmental Science*, *24*(5), e13090. doi: 10.1111/desc.13090

DePaolis, R. A., Vihman, M. M., & Nakai, S. (2013). The influence of babbling patterns on the processing of speech. *Infant Behavior and Development*, *36*(4), 642–649. doi: 10.1016/j.infbeh.2013.06.007

Dunn, L. M., & Dunn, D. M. (2007). *PPVT-4: Peabody picture vocabulary test*. Pearson Assessments.

Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in novel word repetition. *Journal of Speech Language and Hearing Research*, *57*, 421–436.

Erskine, M. E., Munson, B., & Edwards, J. R. (2020, March). Relationship between early phonological processing and later phonological awareness: Evidence from nonword repetition. *Applied Psycholinguistics*, *41*(2), 319–346. doi: 10.1017/S0142716419000547

Fenson, L., Marchman, V., Thal, D. J., Dale, P., Reznick, J., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories User's Guide and Technical Manual* (2nd Edition ed.). San Diego, CA: Singular.

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In I. A. Sekerina, E. M. Fernández, & H. Clahsen (Eds.), *Language Acquisition and Language Disorders* (Vol. 44, pp. 97–135). Amsterdam: John Benjamins Publishing Company. doi: 10.1075/lald.44.06fer

Foy, J. G., & Mann, V. (2003). Home literacy environment and phonological awareness in preschool children: Differential effects for rhyme and phoneme awareness. *Applied Psycholinguistics*, *24*(01), 59–88. doi: 10.1017/S0142716403000043

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and Consistency in Early Language Learning: The Wordbank Project*. MIT Press.

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics*, *27*(4), 513–543. doi: 10.1017/S0142716406060383

Goldman, R., & Fristoe, M. (2000). *Goldman-Fristoe Test of Articulation-Second Edition (GFTA-2)* (2nd ed.). Circle Pines, MN: American Guidance Service.

Hitczenko, K., Bergelson, E., Casillas, M., Colleran, H., Cychosz, M., & Cristia, A. (2023). The development of canonical proportion continues past toddlerhood. In *Proceedings of the International Congress of the Phonetic Sciences.* Prague, CZ.

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First Language*, *42*(6), 703–735. doi: 10.1177/01427237211066405

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

Long, H. L., Ramsay, G., Bene, E. R., Su, P. L., Yoo, H., Klaiman, C., … Oller, D. K. (2024). Canonical babbling trajectories across the first year of life in autism and typical development. *Autism*, 13623613241253908. doi: 10.1177/13623613241253908

Macrae, T. (2017). Stimulus Characteristics of Single-Word Tests of Children's Speech Sound Production. *Language, Speech, and Hearing Services in Schools*, *48*(4), 219–233. doi: $10.1044/2017_{LSHSS}-16-0050$

Molemans, I., Van Den Berg, R., Van Severen, L., & Gillis, S. (2012). How to measure the onset of babbling reliably? *Journal of Child Language*, *39*(3), 523–552. doi: 10.1017/S0305000911000171

Munson, B., Logerquist, M. K., Kim, H., Martell, A., & Edwards, J. (2021). Does Early Phonetic Differentiation Predict Later Phonetic Development? Evidence From a Longitudinal Study of // Development in Preschool Children. *Journal of Speech, Language, and Hearing Research*, *64*(7), 2417–2437. doi: $10.1044/2021_{JSLHR}-20-00555$

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31–38. doi: 10.1016/j.jecp.2017.04.017

Oller, D. K. (2000). *The emergence of the speech capacity*. Mahwah, NJ: Lawrence Erlbaum Associates.

Sosa, A. V., & Stoel-Gammon, C. (2012). Lexical and Phonological Effects in Early Word Production. *Journal of Speech Language and Hearing Research*, *55*(2), 596. doi: 10.1044/1092-4388(2011/10-0113)

Sy, Y., Havard, W. N., Lavechin, M., Dupoux, E., & Cristia, A. (2023). Measuring Language Development From Child-centered Recordings. In *INTERSPEECH 2023* (pp. 4618–4622). ISCA. doi: 10.21437/Interspeech.2023-1569

Usha, G. P., & Alex, J. S. R. (2023). Speech assessment tool methods for speech impaired children: A systematic literature review on the state-of-the-art in Speech impairment analysis. *Multimedia Tools and Applications*, *82*(22), 35021–35058. doi: 10.1007/s11042-023-14913-0

Vihman, M. M., DePaolis, R. A., & Keren-Portnoy, T. (2014). The role of production in infant word learning. *Language Learning*, *64*(s2), 121–140. doi: 10.1111/lang.12058

Wagner, R. K. (2013). *CTOPP-2: Comprehensive Test of Phonological Processing* (Second ed.). Austin, TX: Pro-Ed.

Xu, D., Yapanel, U., & Gray, S. (2009). *Reliability of the LENA Language Environment Analysis System in young children's natural home environment* (Technical Report lTR-05-2). Boulder, CO: LENA Research Foundation.

Zhang, T., Suresh, M., Hitczenko, K., Cristia, A., & Cychosz, M. (under review). Employing self-supervised learning models for child speech maturity classification..