Supplemental Information

Table of Contents

Supplement 1: Q&A, Quick links, and Primary Metrics overview

Supplement 1 contains a **Q&A** with the most likely questions, **quick links** to the preregistration documents, and definitions of the **primary metrics** used for analyses.

1A Q&A

N.b.: This information is all detailed further in the manuscript or other supplements. We provide it here as a simple digest for reading ease.

- **What is adult talk?** Number of utterances in near and clear adult speech (from both male and female adults) estimated by the LENA algorithm. (SI3B)
- **What is child speech?** Number of child speech vocalizations (including babbling and more mature speech utterances) estimated by the LENA algorithm. (SI3B)
- **Is the dataset representative?** We systematically asked all researchers working with LENA data outside the USA and with atypical populations to share their data, and a majority agreed. Individual datasets are not nationally representative. (SI3A)
- **What are the SES backgrounds studied here?** We use maternal education as our SES proxy. It has a wide spread in the raw data, from children whose mothers have no formal schooling to those with advanced degrees. For the analyses, we use five levels: less than high school (roughly 0-10 years of education), completed high school (12 years), some college (roughly 13-15 years of education), completed college/university (16 years), or above (>16 years). (SI2B)
- **What are multilingual children?** Those reported to hear >1 language. (cf. Methods for full language list)
- **What is non-normative?** Children who were diagnosed with and/or who had a characteristic that is associated with high risk for language delays and deficits (SI2A).
- **What do we mean by two-step preregistration?** We decided how to split the data into exploration and confirmation subsets, and pre-registered that first. We then used the exploration set to examine relationships among variables, make decisions about random effects structure and interactions, and generate hypotheses to test in the (completely held-out) confirmation subset. (SI3B, SI3C)
- **What models did we fit?** We used a mixed model to predict **child speech**: output_model = lmer(CVCr_s ~ gender + ses.5f + overlapr_s + normative* AVCr_s* age_s + monoling* AVCr_s* age_s + (1 + overlapr_s + AVCr_s |corpus) + (1 | corpus:corp_chi), weights = rec_rat, data = bsl_conf_scaled), where "_s" indicates standardized, rec_rat is length of the ratio of each recording's length divided by maximum recording length. (SI3E). We fit a similar model predicting **adult talk**: input_model = lmer(AVCr_s ~ gender * ses.5f + overlapr_s + normative*age_s + monoling*age_s + (1 + overlapr_s |corpus) + (1 | corpus:corp_chi), weights = rec_rat, data = bsl_conf_scaled). See Tables 1 and 2 in the main manuscript.
- **Are your results due to the LENA algorithm and/or "mixing together" many corpora?** No, we replicated key conclusions using a different algorithm (SI3F) and in more homogeneous subsets (SI3G). See further robustness checks (SI3H) and evidence against alternative interpretations (SI4).
- **Why do you refer to individual datasets with fruit names in some places?** In order to preserve anonymity regarding some sensitive aspects of the data, individual datasets have a fruit name (e.g. kiwi, orange) in some of our historical or supplemental documents (we use #1-18 in the manuscript instead). Figure 1 shows where each dataset came from and the relevant citation.
- **Where can I find the pdf's/docx mentioned in this document?** In our Main OSF project page

1B Quick Links

Main OSF project page
Preregistration 1: Exploratory analysis
Preregistration 2: Confirmatory analysis

1C Primary Metrics Definition

Adult vocalization count rate (AVCr): This metric estimates the quantity of adult speech (from both male and female adults) produced near the child, measured in number of LENA delineated 'utterances' (i.e. vocalizations). Only LENA-classified "near and clear" speech (speech that is not faint or distant, and sufficiently noise-free) is included. The r in AVCr indicates that the raw value was transformed into a rate reflecting adult utterances/hour. We call this **adult talk** in non-technical portions of the paper.

Child linguistic vocalization rate (CVCr): This metric was calculated from LENA's child vocalization count (CVC) measure, which estimates the number of linguistically-relevant vocalizations (including early vocalizations, canonical babble and more mature speech utterances) produced by the child wearing the recorder. The r in CVCr indicates that the raw value was transformed into a rate reflecting child vocalizations/hour. We call this **child speech** in non-technical portions of the paper.

1D Primary Metrics Conceptualization

Based on their names, both adult vocalization count rate (AVCr) and child vocalization count rate (CVCr) may seem to capture only quantity. This is not the case, for two reasons. First, AVCr is based on speech recognition systems's search for speech; CVCr is computed based on an age- and sex-dependent models of children's linguistic vocalizations (rather than e.g. 'anything made by a vocal tract'). Second, conceptually, both measures are used here as proxies for language input and output, respectively, which intrinsically meld notions of quantity, complexity, and quality.

In particular, CVCr is used as a proxy for language as a whole in the child, similarly to how previous work has used parental checklists of children's vocabulary or measures derived from transcriptions of naturally-occuring production samples (e.g., Bornstein et al., 2014). Similarly, AVCr is used as a proxy of language input to the child in general, just as previous work uses short transcribed samples (sometimes with external observers or in the lab), or measures derived from such production samples. These proxies are well-supported by prior research, namely:

1) Findings showing that CVC is correlated concurrently and predictively with standardized measures of language (a meta-analysis in Wang et al., 2020);
2) Quantity and complexity of language are robustly and repeatedly found to be correlated (e.g., word tokens and types at r~.7, tokens and mean length of utterance [MLU] at r~.5; Choi et al., 2020; tokens and types at r~.9, tokens and MLU at r~.4; Rowe, 2008). That is, one can surely imagine a scenario in which 'cat, cat, cat, cat, cat, cat' and 'let's pet

> the cute cat' are the same "amount" of speech but differ in complexity. However, in practice, natural language samples yield a strong and persistent correlation between amount of speech and its complexity (e.g. tokens and type counts of words in this example) (Rowe, 2012; Malvern et al, 2004).

To provide further evidence that AVCr and CVCr are not merely proxies for amount of talk, here we engage in two specific analyses linking these metrics to other field-standard measures reflecting the diversity of words that children are hearing or producing.

Adult speech (ACVr) correlates with word types.
Within the subset of corpora for which our team has access to the audio, 300 minutes were subsampled and manually segmented and transcribed (>150 hours of researcher effort including training). We observed highly robust correlations between ACVr and the manually transcribed number of word types produced by adults (i.e. the lexical diversity of the speech). Number of distinct word types is a well-established language "quality" measure, which captures not just how many words are said but how many *different* words are said (as in the 'let's pet the cat' example above).

More specifically, we found that AVCr and the number of adult word types (or tokens for that matter) correlated with a Pearson's R>.8 (p<.0001); with zero values excluded, the correlation remains >.7 (p<.0001). While it would take ~500 people working full time for a year to do this type of transcription for our full sample, rendering it an impractical approach, this robust association between adult speech quantity and the quality of that speech as indexed by lexical diversity leaves us confident in this measure.

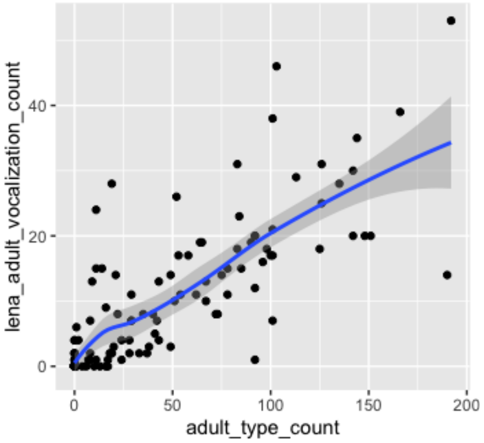


Figure S1D.1. Correlation between our LENA adult speech (AVCr) measure and manually transcribed adult word types. Each point depicts the LENA value (y-axis) and adult type count (x-axis) in a 2-minute annotation sample drawn from a daylong recording including in our analysis. Line and band depict local estimated line of best fit and standard error.

Child speech (CVCr) correlates with reported vocabulary.
CVCr has been found to be a good proxy for development *among young children,* as established, for instance, by Gilkerson and colleagues (2017, AJSLP; data kindly shared for this supplemental analysis). In a sample of nearly 200 U.S. children aged 0-3 years, they report that LENA's CVCr correlated with a gold-standard measure of child language development in this age range (including when partialling out the effect of age).

More specifically, this gold-standard measure was a vocabulary checklist designed to capture a qualitative dimension of language knowledge: which words the child produces from a list selected to adequately represent individual variation in language development among North American participants (the Macarthur Communicative Development Inventory (CDI); Fenson et al 1994). Our automated child speech measure correlates with this gold-standard at R=.7 (p<.0001), again highlighting that, in early childhood, child speech-like vocalizations yield a valid and robust measure of language knowledge (not just chattiness).

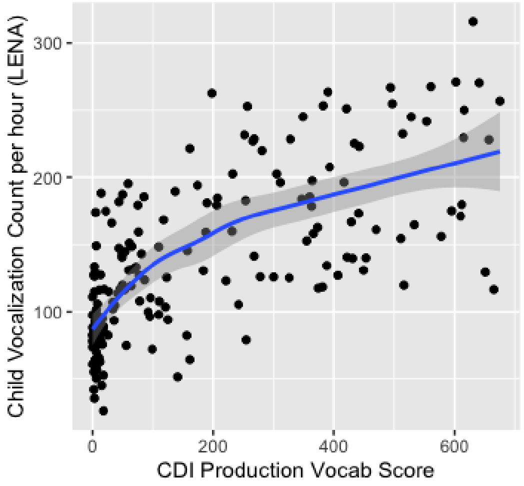


Figure S1D.2. Correlation between our LENA child speech (CVCr) measure and children's CDI Vocabulary Production scores. Each point depicts the LENA value (y-axis) and child vocabulary total (x-axis) from the CDI. Line and band depict local estimated line of best fit and standard error.

1E Primary Metrics Reliability

To confirm our measures' validity, we evaluated them against human annotations (and did the same for our alternative VTC algorithm; Section SI3F). As we describe in "Evaluation Against Human Annotation" in the main manuscript methods, we find similar reliability (median Human vs. Algorithm correlation = .74) to other cognitive measures used in early development, both clinically and in research settings (e.g., Dale, 1991; Feldman et al., 2005; Velikonja et al, 2017). For a full report on the details of our validation, see EL1000-Details_on_validation_check.pdf.

Table S1E.1. Association between manual (human) vs. automated counts by corpus. Numbers are correlation coefficient (Pearson's r), excepting ganek (Spearman's rho) and weisleder (% agreement). NA indicates no human annotation was available.

Setting	Language	Corpus	AVCr r	CVCr r
Urban	English	bergelson	.76	.79
		kalashnikova	NA	NA
		lucid	.75	.82
		rague	NA	.78
		vandam	.4	.51
		warlaumont	.75	.89
		winnipeg	.64	.69
		kidd	NA	NA
	English/Spanish	ramirez-esparza	.59	.68
	Dutch	alphen	.87	.77
	Finnish	elo	NA	NA
	French	lyon	.64	.71
	Spanish	weisleder	66.1%	64.4%
	Swedish	swedish	.78	NA
	Vietnamese	ganek	rho = .70	
Rural	Tsimane	tsimane	.74	.77
	Wolof	senegal	.48	.71
	Yélî	rossel	.77	.67
		median	**.74**	**.74**

To put it plainly: The idea that LENA's algorithm output is systematically more accurate in language/population A than language/population B is not supported by the literature (Cristia et al., 2020, 2021). Table S1E.1 shows different estimates across individual corpora, with each estimate resulting from a cluster of properties, including conceptual ones like population or language, but also methodological ones. Although it is impossible to attribute variance around point estimates to these disparate sources of difference, inspection of these results leads us to highlight methodological differences. For instance, accuracy diverged in the Bergelson versus Vandam corpora, despite both datasets' inclusion of English-learning American children of similar demographic background to each other and to the LENA training sample. This is likely due to the divergent human annotation approaches in these datasets: While the human annotation in the Bergelson corpus was based on a random sample within recordings by annotators specifically trained on precision in utterance segmentation, that in VanDam sampled sections with more speech within recordings, and annotators' priority was transcription rather than precise utterance segmentation. More generally, the human annotation in the Senegal, Vandam, Weisleder, and Ramirez-Esparza corpora were undertaken with different goals, which did not entail precise vocalization-level segmentation, and thus represent a lower bound on reliability.

1C-E References

Bornstein, M. H., Hahn, C. S., & Haynes, O. M. (2004). Specific and general language performance across early childhood: Stability and gender considerations. *First language, 24*(3), 267-304.

Choi, B., Nelson, C. A., Rowe, M. L., & Tager-Flusberg, H. (2020). Reciprocal influences between parent input and child language skills in dyads involving high- and low-risk infants for autism spectrum disorder. *Autism Research, 13*(7), 1168-1183.

Cristia, A., Bulgarelli, F., & Bergelson, E. (2020). Accuracy of the Language Environment Analysis (LENA^{TM}) system segmentation and metrics: A systematic review. Journal of Speech, Language, and Hearing Research.

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, 467–486.

Dale, P. S. (1991). The validity of a parent report measure of vocabulary and syntax at 24 months. *Journal of Speech, Language, and Hearing Research, 34*(3), 565-571.

Feldman, H. M., Dale, P. S., Campbell, T. F., Colborn, D. K., Kurs‐Lasky, M., Rockette, H. E., & Paradise, J. L. (2005). Concurrent and predictive validity of parent reports of child language at ages 2 and 3 years. *Child development, 76*(4), 856-868.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., & Stiles, J. (1994). Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development*, *59*(5), i–185. JSTOR.

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Oller, D. K., 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.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development* (pp. 16-30). New York: Palgrave Macmillan

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of child language, 35*(1), 185-205.

Rowe, M. L. (2012). A Longitudinal Investigation of the Role of Quantity and Quality of Child-Directed Speech in Vocabulary Development. *Child Development*, *83*(5), 1762–1774. https://doi.org/10.1111/j.1467-8624.2012.01805.x

Velikonja, T., Edbrooke, Childs, J., Calderon, A., Sleed, M., Brown, A., & Deighton, J. (2017). The psychometric properties of the Ages & Stages Questionnaires for ages 2-2.5: a systematic review. *Child: care, health and development*, *43*(1), 1-17.

Wang, Y., Williams, R., Dilley, L., & Houston, D. M. (2020). A meta-analysis of the predictability of LENA^{TM} automated measures for child language development. *Developmental Review*, *57*, 100921.

Back to Table of Contents

Supplement 2: Detailed justification for demographic variables

Supplement 2 provides detailed justification for demographic variables.

2A: Classification of non-normative development

Description of the procedure used to determine whether specific participant characteristics identified in the corpora qualified as “non-normative” for the purpose of this study.

2B: Socioeconomic Status (SES) and Maternal Education

Description of the how and why we decided to use maternal education as our SES proxy, and how the raw data from each corpus were transformed for analysis.

Back to Table of Contents

2A: Classification of non-normative development

Goal

To determine whether children in our dataset might be expected to demonstrate non-normative language abilities due to developmental, physical, environmental or genetic factors.

Procedure

For each corpus, data stewards provided an initial classification of each child as normative or non-normative and provided their operationalization of normativity. These classifications were then subject to a further review as described below.

For a factor to be classified as non-normative, it had to meet **one of the following criteria**:

A. A meta-analysis found a language difference effect of at least .50 (Cohen's *d* or Hedges' *g*) between those with and without the factor. “Language differences” were defined as skills related to the expressive language, receptive language, and/or verbal IQ.
B. The group means for individuals with versus without the factor differ by more than two standard deviations on at least one language measure in a meta-analysis OR two or more peer-reviewed papers with at least 20 infants per sample.
C. A meta-analysis OR at least one peer-reviewed paper finds that this factor doubles the likelihood of a language disorder diagnosis (i.e., odds ratio >= 2).

Procedures for conducting the literature reviews for each factor were as follows:

1. Identify a meta-analysis to fulfill Criterion A by searching PubMed and Google Scholar using keywords "language" AND infancy AND meta-analysis AND [keyword for the factor, e.g., premature, autis*, etc.]. Keywords were added to refine the search if the initial search provides too many results.
2. If the first 20 results did not produce a relevant meta-analysis, a small systematic review was conducted to fulfill Criterion B or Criterion C. This review included 10 articles total, with 3-5 articles identified as relevant based on authors’ existing knowledge and the remaining 5-7 identified using a literature search "language" AND infancy AND [keyword for the factor] on PubMed. Keywords were added to refine the search if the initial search provided too many results.
3. At all stages, search terms and deviations from the steps outlined above were recorded.

Across the corpora, there were initially 11 categories of non-normativity tagged. After our literature-based assessment, the ones in bold were retained. **(1) prematurity (< 37 week gestation), (2) diagnosed speech or language delay, (3) family history of speech language impairment and/or dyslexia (4) global developmental delay, (5) low birth weight (<2,500 g), (6) hearing loss,** chronic otitis media, or **having hearing aids or cochlear implants, (7) familial risk of autism spectrum disorder, (8) genetic syndrome**, (9) chronic condition leading to regular interactions with health care system (10) child of a multiple birth pregnancy, or (11) parent with psychiatric diagnosis of anxiety or depression. Note that there was variability in reporting across the corpora (e.g. not all corpora asked about familial risk for autism).

Decisions about how to classify children from (9) and (10) (i.e. multiple birth pregnancies and parent psychiatric diagnoses) were made based on discussion with data stewards and authors, as described below. **For each of the remaining factors, we conducted the review process described above to identify evidence indicating whether that factor was likely to increase a child's likelihood of having non-normative language.**

Results of Normativity Classification following Literature Review

Factors with a Non-Normative Final Classification

(1) Preterm birth
Decision: Non-normative
Children born prematurely (< 37 weeks) were classified as non-normative based on meta-analyses finding large language difference effects between children born prematurely and those born full term (Allotey et al., 2018; Barre, Morgan, Doyle, & Anderson, 2011; Mulder, Pitchford, Hagger, & Marlow, 2009; Van Noort-Van Der Spek, Franken, & Weisglas-Kuperus, 2011; Zimmerman, 2018). Notably, many of the studies from these meta-analyses included children born very prematurely (< 32 weeks) but not those born prematurely only (33-37 weeks). This distinction was not available for all corpora included in our study; thus, we are unable to separate out this effect in our sample.

(2) Diagnosed speech or language delay
Decision: Non-normative
Children with diagnosed speech or language delays were classified as non-normative based on meta-analyses indicating that children with developmental language or speech disorders show poorer performance compared to controls on standardized or experimental measures of language abilities or language-related skills (Graf Estes, Evans, & Else-Quest, 2007; Kan & Windsor, 2010; Krok & Leonard, 2015; Snowling & Melby-Lervag, 2016).

(3) Familial risk of speech language impairment and/or dyslexia
Decision: Non-normative
Children (3 years and older) who are at family risk for developmental dyslexia exhibit significantly lower performance in assessments of oral language abilities and phonological awareness (Snowling & Melby-Lervag, 2016).

(4) Global developmental delay (GDD)
Decision: Non-normative
Several corpora indicated children with cognitive or motor delays, which we classified under a broader category of global developmental delay. No meta-analyses were identified that directly compared the language abilities of children with GDD to those of typically developing children. Children with GDD were found to have language abilities that were not significantly different from those of children with SLI (Kim, Jeon, Park, Chung, & Song, 2014; Shevell, Majnemer, Platt, Webster, & Birnbaum, 2005). Given evidence from our review of literature on language outcomes of children with diagnosed speech or language delays that led us to classify children

with SLI as non-normative, we infer that children with GDD would be expected to demonstrate similar delays in language abilities that would justify classifying them as non-normative.

(5) Low birth weight (<2,500 grams, when specified)
Decision: Non-normative
Children with low birth weight were classified as non-normative based on meta-analyses indicating that children with low birth weight tend to have lower language skills than children with normal birth weight (Barre, Morgan, Doyle, & Anderson, 2011; Zimmerman, 2018).

(6) Hearing loss, hearing aids or cochlear implants
Decision: Non-normative
Children with hearing loss, hearing aids or cochlear implants were classified as non-normative based on meta-analyses indicating that children with cochlear implants and hearing loss tend to have lower spoken language skills compared to children without hearing loss (Lund, 2015).

Note: In contrast, children with ear infections/otitis media were classified as normative since the available meta-analyses and empirical studies did not meet our criteria for non-normativity (Harsten, Nettelbladt, Schalén, Kalm, & Prellner, 1993; Roberts, Rosenfeld, & Zeisel, 2004).

(7) Familial risk of autism spectrum disorder (ASD)
Decision: Non-normative
Children with an older sibling with ASD were classified as non-normative based on meta-analyses demonstrating that having an older sibling with ASD increased the likelihood of being diagnosed with a non-ASD language delay (Garrido, Petrova, Watson, Garcia‐Retamero, & Carballo, 2017; Marrus et al., 2018).

(8) Other relevant genetic syndromes
Decision: Non-normative
Children with other relevant genetic syndromes were classified as non-normative based on evidence that the presence of a genetic syndrome increases the likelihood of having a speech delay (Cafferkey, Ahn, Flinter, & Ogilvie, 2014).

Factors with a Normative Final Classification

(6) Chronic Otitis Media
Decision: Normative.
Children with ear infections/otitis media were classified as normative since the available meta-analyses and empirical studies did not meet our criteria for non-normativity (Harsten, Nettelbladt, Schalén, Kalm, & Prellner, 1993; Roberts, Rosenfeld, & Zeisel, 2004).

Note: In contrast, children with hearing loss, hearing aids, or cochlear implants were classified as non-normative (see above).

(9) Multiples

Decision: Normative
There were 14 sets of twins included across corpora. Only one child from each set of twins was included in the dataset and was classified as normative or non-normative based on any additional normativity tags. (*Note*: non-twin siblings are also not included in the dataset).

(10) Parent psychiatric diagnoses
Decision: Normative
Children who had a parent diagnosed with anxiety or depression (n = 19) were originally labelled as non-normative in one corpus, but as this criteria was only noted in one corpus, and none of these children were tagged with any other potential source of non-normativity, these were retagged as normative for consistency.

(11) Chronic condition leading to regular interactions with the healthcare system
Decision: Normative
No meta-analyses were identified that reported broadly on the effect of frequent contact with the healthcare system or having a chronic illness. Two meta-analyses focused on language outcomes of individuals with specific chronic illnesses suggested that even if there is an effect of frequent exposure to the healthcare system or the presence of chronic illness on language outcomes, the effect is likely smaller than .50, especially in children younger than 6 years (Karsdorp, Everaerd, Kindt, & Mulder, 2006; Schatz, Finke, Kellett, & Kramer, 2002). Thus, children when chronic conditions and/or regular contact with the healthcare system were classified as normative.

N.B. Further details regarding the procedure used to determine the classification of non-normativity can be found in the document EL1000-Non_Normativity_review.docx.

Back to Table of Contents

2A References and Language Metric(s) Used

Allotey, J., Zamora, J., Cheong-See, F., Kalidindi, M., Arroyo-Manzano, D., Asztalos, E., … Thangaratinam, S. (2018). Cognitive, motor, behavioural and academic performances of children born preterm: a meta-analysis and systematic review involving 64 061 children. *BJOG: An International Journal of Obstetrics & Gynaecology*, *125*(1), 16–25. https://doi.org/10.1111/1471-0528.14832.
Language Metric(s) Relevant to Decision: **Verbal IQ** (as measured by the Wechsler Intelligence Scales for Children, Stanford Binet Scale, Kaufman Assessment Battery for Children or McCarthy Scales of Children's Abilities)

Barre, N., Morgan, A., Doyle, L. W., & Anderson, P. J. (2011). Language abilities in children who were very preterm and/or very low birth weight: A meta-analysis. *The Journal of Pediatrics, 158*(5), 766–774.e1. https://doi.org/10.1016/j.jpeds.2010.10.032.
Language Metric(s) Relevant to Decision: **expressive language** (tasks that measure the participant's overall verbal expression and are unable to be further classified as semantics or grammar; e.g., indices from PLS or CELF); **receptive language** (tasks that measure the participant's overall understanding of verbal expression, at the level of the word or sentence and are unable to be further classified as semantics or grammar; e.g., indices from PLS or CELF); **receptive-semantics** (tasks that measure the level of understanding of the meaning of verbal information; e.g., PPVT-R in which the participant is asked to point to the picture of the named object).

Cafferkey, M., Ahn, J. W., Flinter, F., & Ogilvie, C. (2014). Phenotypic features in patients with 15q11. 2 (BP1‐BP2) deletion: Further delineation of an emerging syndrome. *American Journal of Medical Genetics Part A, 164*(8), 1916-1922. https://doi.org/10.1002/ajmg.a.36554.
Language Metric(s) Relevant to Decision: **speech and language milestones delayed**; **speech delay** (diagnostic criteria not specified)

Garrido, D., Petrova, D., Watson, L. R., Garcia‐Retamero, R., & Carballo, G. (2017). Language and motor skills in siblings of children with autism spectrum disorder: A meta‐analytic review. *Autism Research, 10*(11), 1737-1750. https://doi.org/10.1002/aur.1829.
Language Metric(s) Relevant to Decision: **expressive and receptive language** (as measured by the MSEL, RDLS, CELF-P, MCDI, or VABS).

Graf Estes, K., Evans, J. L., & Else-Quest, N. M. (2007). Differences in the Nonword Repetition Performance of Children With and Without Specific Language Impairment: A Meta-Analysis. *Journal of Speech, Language, and Hearing Research*, *50*(1), 177–195. https://doi.org/10.1044/1092-4388(2007/015)
Language Metric(s) Relevant to Decision: **non-word repetition** (as measured by the CNRep, NRT, Montgomery nonword set, and three- to four-syllable set).

Harsten, G., Nettelbladt, U., Schalén, L., Kalm, O., & Prellner, K. (1993). Language development in children with recurrent acute otitis media during the first three years of life. Follow-up study from birth to seven years of age. The Journal of Laryngology & Otology, 107(5), 407-412. https://doi.org/10.1017/S0022215100123291
Language Metric(s) Relevant to Decision: **phonology** (as measured by a phoneme test), **grammar** (as measured by the Ringsted material and sequential and thematic picture tasks), **interaction** (as measured by sequential and thematic picture tasks) and **auditory discrimination**

Kan, P. F., & Windsor, J. (2010). Word Learning in Children With Primary Language Impairment: A Meta-Analysis. *Journal of Speech, Language, and Hearing Research*, *53*, 739–756. https://doi.org/10.1097/aud.0000000000000560
Language Metric(s) Relevant to Decision: **novel word learning** (as measured by production tasks [producing target words], comprehension tasks [identifying items/pictures], and/or recognition tasks [forced-choice comprehension]; specific measures not listed).

Karsdorp, P. A., Everaerd, W., Kindt, M., & Mulder, B. J. (2006). Psychological and cognitive functioning in children and adolescents with congenital heart disease: a meta-analysis. *Journal of Pediatric Psychology, 32*(5), 527-541. https://doi.org/10.1093/jpepsy/jsl047.
Language Metric(s) Relevant to Decision: **Verbal IQ** (as measured by British Ability Scale, Bayley Scales of Infant Development, Differential Ability Scale, Hamburger Wechsler Intelligence Test for Adults, Hamburger Wechsler Intelligence Test for Children, Hamburger Wechsler for Children in

Pre-school Age, Kaufman Assessment Battery for Children, Leiter International Scale, McCarthy Scales of Children's Abilities, Stanford Binet Scale, Wechsler Intelligence Test for Children, or Wechsler Preschool and Primary Scale of Intelligence)

Kim, S. W., Jeon, H. R., Park, E. J., Chung, H. J., & Song, J. E. (2014). The differences in clinical aspect between Specific Language impairment and Global Developmental delay. *Annals of Rehabilitation Medicine, 38*, 752-758. https://doi.org/10.5535/arm.2014.38.6.752.
Language Metric(s) Relevant to Decision: **expressive and receptive language** (as measured by the Sequenced Language Scale for Infants or Preschool Receptive-Expressive Language Scale)

Krok, W. C., & Leonard, L. B. (2015). Past tense production in children with and without specific language impairment across Germanic languages: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *58*, 1326–1340. https://doi.org/10.1044/2015_JSLHR-L-14-0348.
Language Metric(s) Relevant to Decision: **past tense production** (using elicitation probes, conversational language samples, a combination of probes and conversation, and narrative tasks to elicit past tense production from the participants)

Kwok, E. Y., Brown, H. M., Smyth, R. E., & Cardy, J. O. (2015). Meta-analysis of receptive and expressive language skills in autism spectrum disorder. *Research in Autism Spectrum Disorders, 9*, 202-222. https://doi.org/10.1016/j.rasd.2014.10.008.
Language Metric(s) Relevant to Decision: **receptive and expressive language** (as measured by Peabody Picture Vocabulary Test, Expressive Vocabulary Test, Preschool Language Scale, Reynell Developmental Language Scales,MacArthur-Bates Communicative Development Inventory, Schlichting Test for Dutch language production, Mullen Scales of Early Learning; Sequenced Inventory of Communication Development, Vineland Adaptive Behavior Scales, Clinical Evaluation of Language Fundamentals, British Picture Vocabulary Scale, Test for Reception of Grammar, Renfrew Word Finding Vocabulary Test, and Action Picture Test Grammar Scale).

Lund, E. (2015). Vocabulary knowledge of children with cochlear implants: A meta-analysis. *Journal of deaf studies and deaf education*, *21*(2), 107-121. https://doi.org/10.1093/deafed/env060.
Language Metric(s) Relevant to Decision: **receptive and expressive language** (as measured by the British Picture Vocabulary Scale, Expressive One-Word Picture Vocabulary Test, Expressive Vocabulary Test, Lexical Phonological Test, Peabody Picture Vocabulary Test, Third Edition, Peabody Picture Vocabulary Test, Fourth Edition, and Receptive One Word Picture Vocabulary Test)

Marrus, N., Hall, L. P., Paterson, S. J., Elison, J. T., Wolff, J. J., Swanson, M. R., ... & Zwaigenbaum, L. (2018). Language delay aggregates in toddler siblings of children with autism spectrum disorder. *Journal of Neurodevelopmental Disorders, 10*(1), 29. doi: https://doi.org/10.1186/s11689-018-9247-8.
Language Metric(s) Relevant to Decision: **receptive and expressive language** (as measured by the Mullen Scales of Early Learning)

Mulder, H., Pitchford, N. J., Hagger, M. S., & Marlow, N. (2009). Development of executive function and attention in preterm children: a systematic review. *Developmental Neuropsychology, 34*(4), 393–421. https://doi.org/10.1080/87565640902964524
Language Metric(s) Relevant to Decision: **verbal fluency**, specifically **semantic fluency** (as measured by semantic or category fluency tasks, where participants have to name as many items as possible in a given category (e.g., toys) within a time limit; specific measures not listed) and **phonetic fluency** (as measured by phonemic fluency tasks, where participants have to name as many words as possible starting with a certain letter within a time limit; specific measures not listed)

Ozonoff, S., Young, G. S., Carter, A., Messinger, D., Yirmiya, N., Zwaigenbaum, L., ... & Hutman, T. (2011). Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study. *Pediatrics, 128*(3), e488. https://doi.org/10.1542/peds.2010-2825.
Language Metric(s) Relevant to Decision: N/A

Riou, E. M., Ghosh, S., Francoeur, E., & Shevell, M. I. (2009). Global developmental delay and its relationship to cognitive skills. *Developmental medicine and child neurology, 51*, 600-606. https://doi.org/10.1111/j.1469-8749.2008.03197.x.
Language Metric(s) Relevant to Decision: **Verbal IQ** (as measured by the Wechsler Preschool and Primary Scale of Intelligence, third edition).

Roberts, J. E., Rosenfeld, R. M., & Zeisel, S. A. (2004). Otitis media and speech and language: a meta-analysis of prospective studies. *Pediatrics*, *113*(3), e238-e248. https://doi.org/10.1542/peds.113.3.e238.
Language Metric(s) Relevant to Decision: **receptive and expressive language** (as measured by the Preschool Language Scale, Sequenced Inventory of Communication Development, Reynell Development Language Scales).

Rovers, M. M., Black, N., Browning, G. G., Maw, R., Zielhuis, G. A., & Haggard, M. P. (2005). Grommets in otitis media with effusion: an individual patient data meta-analysis. *Archives of disease in childhood, 90*(5), 480-485.
Language Metric(s) Relevant to Decision: **language z-scores** (as measured by Reynell language scores).

Schatz, J., Finke, R. L., Kellett, J. M., & Kramer, J. H. (2002). Cognitive functioning in children with sickle cell disease: a meta-analysis. *Journal of Pediatric Psychology, 27*(8), 739-748. https://doi.org/10.1093/jpepsy/27.8.739.
Language Metric(s) Relevant to Decision: **verbal or language functions** (specific measures not listed).

Shevell, M., Majnemer, A., Platt, R. W., Webster, R., & Birnbaum, R. (2005). Developmental and functional outcomes in children with global developmental delay or developmental language impairment. Developmental medicine & child neurology, 47(10), 678-683.
Language Metric(s) Relevant to Decision: **communication** (as measured by the Batelle Developmental Inventory and the Vineland Adaptive Behavior Scale).

Snowling, M. J., & Melby-Lervag, M. (2016). Oral language deficits in familial dyslexia: a meta-analysis and review. *Psychological Bulletin*. https://doi.org/10.1037/bul0000037.
Language Metric(s) Relevant to Decision: **articulatory accuracy**, **vocabulary knowledge**, **phonological memory, phoneme awareness, rhyme awareness, and rapid naming** (specific measures not listed).

Van Noort-Van Der Spek, I. L., Franken, M.-C., & Weisglas-Kuperus, N. (2011). Language functions in preterm children: A systematic review and meta-analysis. *Pediatric Research, 70*(S5), 357. https://doi.org/10.1038/pr.2011.582.
Language Metric(s) Relevant to Decision: **complex total language, complex receptive and complex expressive language** (as measured by the Clinical Evaluation of Language Fundamentals)

Zimmerman, E. (2018). Do Infants Born Very Premature and Who Have Very Low Birth Weight Catch Up With Their Full Term Peers in Their Language Abilities by Early School Age? *Journal of Speech, Language, and Hearing Research, 61*(1), 53–65. https://doi.org/10.1044/2017_JSLHR-L-16-0150.
Language Metric(s) Relevant to Decision: **total language, receptive language, expressive language** (as measured by the Clinical Evaluation of Language Fundamentals, Preschool Language Scales, Test di Vocabolario Figurato, Batteria per la Valutazione del Linguaggio: Naming and Lexical Comprehension, Boston Naming Test, language sampling, Test of Language Development–Primary: Speaking Quotient and Listening Quotient, Woodcock Johnson Test of Achievement: Picture Vocabulary and Understanding Directions, The Token Test for Children, Marburg Language Comprehension Test for Children: Language Comprehension, Peabody Picture Vocabulary Test–Revised, Developmental NEuroPSYchological Assessment–Second Edition: Comprehension of Directions)

Back to Table of Contents

2B: Socioeconomic Status (SES) and Maternal Education

Goal

To determine the best classification system for SES/Maternal Education given both the literature and the available data within our dataset, harmonizing across different classifications used in the different corpora.

Procedure

SES can be conceptualized and measured in different ways. Typical components of SES include level of education, household income, and parents' occupations. Maternal education was the most common SES metric used across all of the original corpora, and is an SES proxy used widely in the literature. Having decided to use maternal education, we reviewed the possible different subcategorizations based on the available information across corpora, and data balance concerns.

Results

We first recategorised each dataset's original SES delineation into a 5-level classification system based on maternal education attainment for the SES variable:

1) BHD: **b**elow **h**igh school **d**egree
2) HD: **h**igh school **d**egree or equivalent
3) SC: **s**ome **c**ollege, vocational or associate degree level training
4) CD: university/**c**ollege **d**egree (e.g. B.A., B.S.)
5) AD: **a**dvanced **d**egree (master's, PhD, JD, MD, etc.)

Rationale for using maternal education as a proxy for SES

Maternal education may be one of the most important components of SES for child development (APA Task Force on SES, 2007). In U.S. samples, maternal education has been found to predict child language skills (Hoff, Burridge, Ribot, & Giguiere, 2018), and differences in maternal education levels have been associated with differences in parental behaviors related to language, such as quantity and quality of caregiver speech (Rowe, 2012). In some U.S. samples, maternal education was found to be a stronger predictor of caregiver speech than income (Huttenlocher et al., 2007). Indeed, following this same logic, LENA's statistical models and algorithms that are used to process audio files were trained using human transcriptions of audio from U.S. families of different SES backgrounds defined based on maternal education attainment (Gilkerson, Coulter, & Richards, 2008)

Another consideration in our choice of SES indicator was consistency across corpora. We wanted to use the same indicator for all children because SES indicators are not interchangeable: Different measures index different components of SES, and each component may contribute to variation in parenting and child development in different ways (Bornstein et al., 2003; Diemer et al., 2013; Rowe, 2018).

Maternal education information was available for all 18 original corpora. Of the 18 corpora, 15 reported maternal educational attainment as the primary indicator of SES[1], 2 reported maternal education along with other indicators of income and education (such as paternal education, family income), and 1 included maternal education information as part of the Hollingshead Index (a composite score based on marital status, employment status, educational attainment, and occupational prestige), along with raw maternal education information.

Rationale for standardizing all corpora to 5 education levels

One consideration was the need to balance the number of groups and the number of cases in each group. Before we had explored the distribution of all predictors we had initially considered that 5 levels would allow for reasonable spread and granularity of measurement to represent differences in educational attainment. The 5-level scale we used is consistent with (in some cases, even more fine-grained than) the granularity of measurement used in previous studies of child language and maternal education (e.g., Hoff et al., 2018; Huttenlocher et al., 2007; 2010; Rowe, 2008, 2012).

Most corpora included in our analyses originally reported maternal education using a classification system based on educational attainment levels similar to the 5 levels used here (Table S2B.1). Five corpora used a system with levels equivalent to the 5 levels above (i.e., less than high school, high school or equivalent, some college, etc.), and five corpora used a classification system with additional levels (e.g., less than elementary school, PhD coursework completed) that were reduced to five (for example, collapsing masters degree and PhD). The other eight corpora used systems that were mapped to the five levels after discussing with the data stewards. Note that in some cases the mapping was imperfect: Because educational systems are not the same across different countries, sometimes it was necessary to adapt the original categories to fit the 5 levels. We relied extensively on the data steward's expertise and guidance.

Table S2B.1 shows a description of the different classification systems used across different corpora and how these were standardized into the 5 levels we used in this study

[1] In one corpus reporting maternal and paternal education attainment, parents were identified as caregiver 1 and caregiver 2 and it was not possible to identify mother and father separately. For most families in the sample, the education level of the two caregivers was equal. When this was not the case (n=7) we took the mean across parents.

Table S2B.1. Original and standardized SES classification schema used in the corpora included in the dataset. Thick line indicates the 3-way split, which collapses levels 1, 2, and 3. (BHD = below high school degree, HD = high school degree, SC = some college, CD = college degree, AD = advanced degree, GED = general education diploma, bacc = Baccalaureate diploma)

			Standardization				
Scheme	# corpora	Possible SES levels in original dataset classification	1 - BHD	2 - HD	3 - SC	4 - CD	5 - AD
2 levels	1	1) vocational college or bachelor's degree; 2) master’s or doctoral degree				1	2
3 levels	2	1) BHD; 2) HS or community college; 3) university	1		2	3	
		1) none; 2) preschool or koranic school; 3) some secondary school or above	1, 2	3			
4 levels	3	1) bacc+3; 2) bacc+4; 3) bacc+5; 4) PhD			1	2	3, 4
		1) year 9; 2) year 12; 3) college/higher vocational; 4) university	1	2	3	4	
		1) BHD; 2) GED; 3) HS; 4) SC	1	2, 3	4		
5 levels	5	1) BHD; 2) HS; 3) SC; 4) CD; 5) AD	1	2	3	4	5
7 levels	4	1) elementary school; 2) middle school; 3) HS; 4) >HS; 5) associates/technical degrees; 6) 4 year university degree; 7) graduate degree	1, 2	3	4, 5	6	7
		1) <year 10; 2) year 10 completed; 3) year 12 completed; 4) trade or college certificate or diploma; 5) university degree; 6) master’s degree; 7) doctorate	1, 2	3	4	5	6, 7
		1) year 10; 2) year 12; 3) trade certificate; 4) diploma; 5) bachelor; 6) master’s; 7) PhD	1, 2	3	4	5	6, 7
		1) BHD; 2) HS; 3) SC; 4) 2-year degree; 5) 4-year degree; 6) professional degree; 7) doctorate	1	2	3, 4	5	6, 7
12 levels	1	1) elementary school; 2) jr. high school; 3) GED; 4) HS; 5) 1 or more years of technical/vocational school; 6) completed technical/vocational school; 7) completed 1 or more years of university/college; 8) bachelor’s degree; 9) completed 1 or more years of graduate school; 10) master’s degree; 11) course work completed for PhD but no dissertation, law degree without bar, medical degree without internship completed; 12) PhD, law degree with bar, med degree with internship completed	1, 2	3, 4	5:7	8	9:12
Years mat. ed.	2	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12	0:6	7:9	10:12		
		0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10	0:7	8:10			

SES Levels for Final Analysis

After our initial data examination, we pre-registered an analysis with a 3-way split in SES (collapsing levels 1-3) due to an uneven distribution of data across these levels (with most children having a mother with SES level 1 and 4; see Fig. 4 and Table 3, main text). Based on reviewer feedback we opted to reinstate all 5 levels in the final analysis.

Alternatives to our method

There are many ways to measure SES in child development research (Diemer et al., 2013; Hoff, Laursen, & Bridges, 2012). Other possibilities include using multiple indicators instead of a single indicator (maternal education and income, for example) and using a composite index. Two well-known composite indices include the Hollingshead Four-Factor Index of Social Status, which is based on education and occupation, and the Socioeconomic Index of Occupations, based on occupational prestige (Bornstein et al., 2003). An important drawback to these two indices is their outdated system for classification of occupation categories and social prestige (Diemer et al., 2013).

Another alternative is to transform educational attainment levels into years of education by assigning a fixed number of years to each category (e.g., Huttenlocher et al., 2010; Rowe, 2012: below high school = 10 years, high school degree = 12 years, some college or associate degree = 14, college degree = 16, advanced degree = 18). The drawback of this strategy is that the values might not reflect true years of education (e.g., “some college” is anywhere from 1 to 3 years; “advanced degree” can be anywhere from 1 to 6 years) so we’d be necessarily adding noise in this interpolation/extrapolation. In the interest of completeness, we did this analysis, and our conclusions do not change (see EL1000-9_other-checks.pdf, Check #1.4).

Yet another alternative is to dichotomize education level based on typical academic achievement in the relevant country. We also did this, and again, it did not change our conclusions (see 3H; EL1000-9_other-checks.pdf, Check #1.5).

Finally, it’s worth noting that while we felt maternal education was the right solution for the reasons above, other approaches are surely possible, and we hope will be taken up in future research. The corpora in our final dataset represent 12 countries (Australia, Bolivia, Canada, Finland, France, Netherlands, Papua New Guinea, Senegal, Sweden, United Kingdom, United States, and Vietnam); including geographic or region-level variables, such as economic inequality (e.g., Gini coefficient), access to education, or urban vs. rural residence might prove fruitful (APA Task Force on SES, 2007).

2B References

American Psychological Association, Task Force on Socioeconomic Status. (2007). *Report of the APA Task Force on Socioeconomic Status.* Washington, DC: American Psychological Association.

Bornstein, M. H., Hahn, C.-S., Suwalsky, J. T. D., & Haynes, O. M. (2003). Socioeconomic status, parenting, and child development: The Hollingshead Four-Factor Index of Social Status and The Socioeconomic Index of Occupations. In M. H. Bornstein & R. H. Bradley (Eds.), *Monographs in parenting series. Socioeconomic status, parenting, and child development* (pp. 29-82). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Diemer, M. A., Mistry, R. S., Wadsworth, M. E., López, I., & Reimers, F. (2013). Best practices in conceptualizing and measuring social class in psychological research. *Analyses of Social Issues and Public Policy (ASAP), 13*(1), 77-113.

Gilkerson, J., Coulter, K. K., & Richards, J. A. (2008). Transcriptional analyses of the LENA natural language corpus (LENA Technical Report LTR-06-2). Available https://www.lena.org/technology/#tech-reports

Hoff, E., Burridge, A., Ribot, K. M., & Giguere, D. (2018). Language specificity in the relation of maternal education to bilingual children's vocabulary growth. *Developmental Psychology, 54*(6), 1011-1019.

Hoff, E., Laursen, B., & Bridges, K. (2012). Measurement and model building in studying the influence of socioeconomic status on child development. In M. Lewis & L. Mayes (Eds.), *A developmental environmental measurement handbook*, pp. 590-606. Cambridge, England: Cambridge University Press.

Huttenlocher, J., Vasilyeva, M., Waterfall, H. R., Vevea, J. L., & Hedges, L. V. (2007). The varieties of speech to young children. *Developmental Psychology, 43*(5), 1062-1083.

Huttenlocher, J., Waterfall, H. R., Vasilyeva, M., Vevea, J. L., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology, 61*(4)*,* 343-365.

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language, 31*(1), 185-205.

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development, 83*(5), 1762-1774.

Back to Table of Contents

Supplement 3: Analysis pipeline

Supplement 3 provides a historical account of our decision making and further explanation of our full pipeline.

3A: Constituting the corpus

Information on how data were obtained, preprocessed, preregistered (preregistration 1) and split into exploration and confirmation subsets.

3B: Choosing measures based on the exploration subset

How predictor and outcome variables were chosen using the exploration subset.

3C: Selection of the model based on the exploration subset

How the model structure was chosen.

3D: Deriving hypotheses to test in the confirmation subset

How final pre-registered hypotheses (preregistration 2) were decided upon.

3E: From exploration to confirmation, to final analyses

Information on decisions for the confirmatory analyses, differences between results for the exploration versus confirmation subsets, and changes between registered analyses and final analyses included in the manuscript.

3F: Convergent evidence for our main results using a different speech analysis algorithm

Replication of our main analysis predicting child vocalizations using a different algorithm, which provides convergent results.

3G: Convergent evidence from exploratory analysis subsetting to North American infants

Replication of our main analysis predicting child vocalizations in the subset of data collected in North America, which provides convergent results.

3H: Additional robustness checks

Evidence that our results hold with different implementations of SES, when splitting infants into older and younger groups, and when considering potential confounds.

Back to Table of Contents

3A: Constituting the corpus

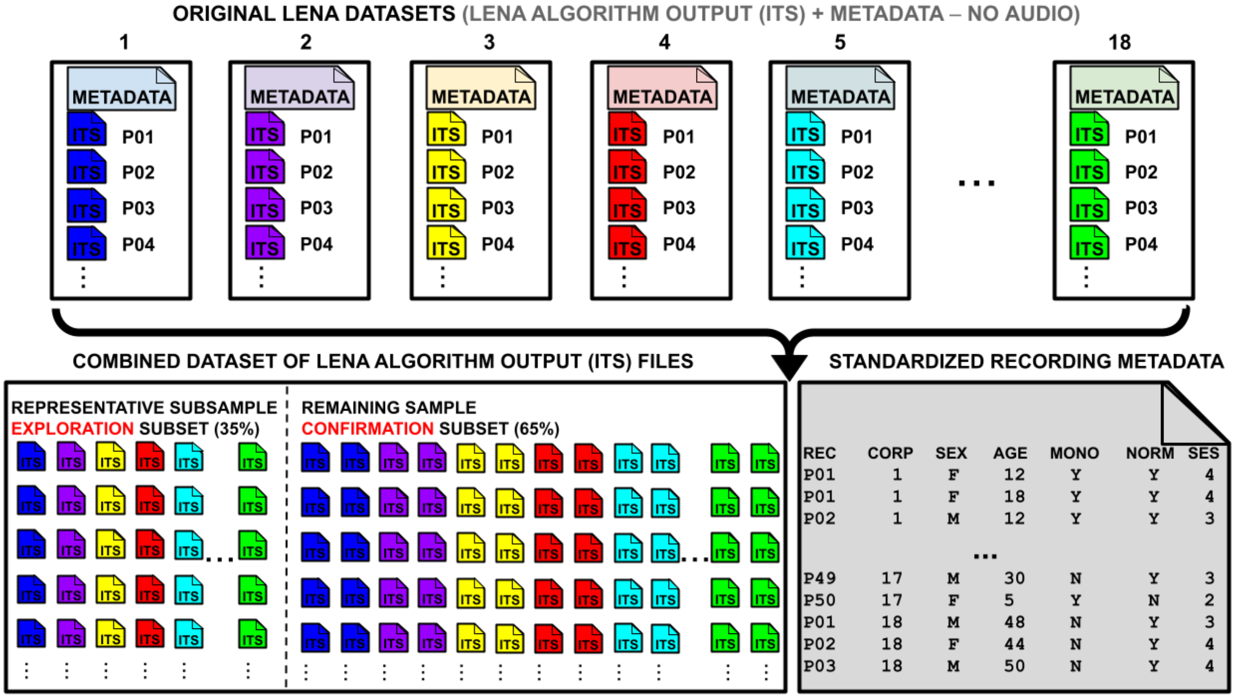
Obtaining the data

To counter a prevalent bias for normative American data, we reached out to all researchers working with non-normative or non-American LENA™ data, based on a systematic review of the literature (Ganek & Eriks-Brophy, 2018), supplemented by a list of recent conference presenters not represented in that review. Out of the 33 researchers discovered with eligible data, 6 did not respond and 9 declined to share their data, with the remaining 18 contributing data. They contributed LENA .its files (i.e., not the audio recordings, but rather the output of LENA analyses) as well as metadata.

Each data steward provided (or made available through HomeBank, VanDam et al., 2016) their raw .its files (the LENA™ 'interpreted time segment' files output by its proprietary algorithm), along with metadata indicating the child's age (months), gender (binary), monolingualism (binary), socioeconomic status (converted to 5 levels), normative development (binary), and language(s) being learned (not analyzed). If there were multiple children from the same family, we included the child with more recordings, or if equal, picked randomly.

Preprocessing the data

The data was pre-processed as described in Figure S3A.1.


ORIGINAL LENA DATASETS (LENA ALGORITHM OUTPUT (ITS) + METADATA – NO AUDIO)
1
2
3
4
5
18
METADATA
ITS
P01
P02
P03
P04
...
COMBINED DATASET OF LENA ALGORITHM OUTPUT (ITS) FILES
REPRESENTATIVE SUBSAMPLE
EXPLORATION SUBSET (35%)
REMAINING SAMPLE
CONFIRMATION SUBSET (65%)
STANDARDIZED RECORDING METADATA

REC	CORP	SEX	AGE	MONO	NORM	SES
P01	1	F	12	Y	Y	4
P01	1	F	18	Y	Y	4
P02	1	M	12	Y	Y	3
		...				
P49	17	M	30	N	Y	3
P50	17	F	5	Y	N	2
P01	18	M	48	N	Y	3
P02	18	F	44	N	Y	4
P03	18	M	50	N	Y	4

Figure S3A.1. Schematic display of the dataset assembly and subsampling process used for analysis. The final combined dataset (lower box) was assembled from 18 independent LENA datasets (upper boxes); each original data steward contributed a collection of automated speech annotation files

(LENA's "its" files) and their corresponding participant and recording metadata. The full dataset was then divided into two subsets: one for exploratory analysis and one for confirmatory analysis.

Creating the exploration and confirmation subsets

Following Anderson and Magruder (2017) we split our data into an exploration subset (containing 35% of the data) and a confirmation set (containing the remaining 65%). This split was chosen in order to follow a hybrid approach: some of our hypotheses came from the literature and the theory, and others from the data. Thus, this is not a pure case, in which a different split would have been preferable (e.g. 15/85%, to maximize power for confirmation for hypotheses coming from other work; or 50/50%, to allow more detailed exploration of this data set). For more details about the split procedure, see EL1000-1-prereg_before_split.pdf document; for characteristics of the two sets, see the EL1000-2-split_202003.pdf.

3A References

Anderson M, Magruder J (2017) Split-Sample Strategies for Avoiding False Discoveries (National Bureau of Economic Research, Cambridge, MA).

Ganek H, Eriks-Brophy A (2018) Language ENvironment analysis (LENA) system investigation of day long recordings in children: A literature review. Journal of Communication Disorders 72:77–85.

VanDam M, et al. (2016) HomeBank: An online repository of daylong child-centered audio recordings. Seminars in Speech and Language 37(2). doi:10.1055/s-0036-1580745.

Back to Table of Contents

3B: Choosing measures based on the exploration subset

Goal

To select among the possible language input and output metrics, based on analysis of their psychometric properties within the exploration subset.

Procedure

We used LENA's interpreted-time segment (.its) files, and thus our dataset provided the potential for a large number of possible input (Table S3B.1) and output (Table S3B.2) metrics. We therefore examined these candidates to select the single best measure for input and output.

Table S3B.1. Definitions of all explored input metrics.

	Calculated from raw .its			LENA primary measure
	MAN	FAN	FAN+MAN	FAN+MAN
Number of events per hour	MAN_AVCr	FAN_AVCr	AVCr	NA
Total duration per hour	MAN_durr	FAN_durr	FAN_MAN_durr	NA
Avg. event duration	MAN_avgdur	FAN_avgdur	FAN_MAN_avgdur	NA
Total word count per hour	MAN_AWCr	FAN_AWCr	FAN_MAN_AWCr	AWCr
Avg. word count per event	MAN_avgAWC	FAN_avgAWC	FAN_MAN_avgAWC	NA

Table S3B.2. Definitions of all explored output metrics.

	Calculated from raw .its				LENA primary measure
	cry	linguistic	vfx	all	linguistic
Number of events per hour	CHN_cry_nr	CHN_ling_nr	CHN_vfx_nr	CHN_nr	CVCr
Total duration per hour	CHN_cry_durr	CHN_ling_durr	CHN_vfx_durr	CHN_durr	NA
Avg. event duration	CHN_cry_avgdur	CHN_ling_avgdur	CHN_vfx_avgdur	CHN_avgdur	NA
Proportions based on N of events	CHN_cry_n_prop	CHN_ling_n_prop	CHN_vfx_n_prop	NA	NA
Proportions based on total dur.	CHN_cry_dur_prop	CHN_ling_dur_prop	CHN_vfx_dur_prop	NA	NA

We relied on the following desiderata:

- output and input metric(s) should be easy to explain to readers and reviewers, and should be 'mutually reasonable'. That is, we'd rather not have a proportion based on N for output and a straight quantity for input. We thought child vocalization count (CVC) would be easy for our readers to understand as a child speech measure; Ns would be easier than total duration, and these than average duration. Proportions are potentially interesting because they control for quantity in theory, which we'd ideally separate from quality.
- output and input metrics should be stable in a re-recording setting. We are doing individual differences research, and therefore want measures that have a high test-retest reliability. We look at this by assessing the percentage of variance attributable to the child/corpus via the random effect structure (and thus to the stability of each measure) in a subset of the data with multiple within-child recordings carried out close together in time (Intraclass Correlation Coefficient, ICC).
- output and input metrics' stability should not be affected by which corpora are included. We check this by re-calculating ICC dropping one corpus at a time ("leave on corpus out" validation).
- our output metric should be highly correlated with age, in a reasonable direction. In this dataset, our best proxy for development is age. Therefore, we want something that correlates with linguistic development ergo age.

Full technical details on the LENA algorithm and variables can be found via LENAs technical reports: https://www.lena.org/technology/#tech-reports.

Results

Full documentation of this step is available in EL1000-3-reliabChecks.pdf.

The input metric chosen was Adult vocalization count rate: i.e. the average number of vocalizations (i.e. LENA "utterances") produced by male or female adults per hour (AVCr).

Rationale: (a) AVCr has one of the highest ICC proportion variance explained by corpus and child ID in recordings separated by <2 months (minimally 50%); and (b) it does not vary much when individual corpora are held out from this analysis ("leave one corpus out" validation, see EL1000-3-reliabChecks.pdf, 4.A.2, figure on p. 30).

The output metric chosen was Number of child vocalizations per hour (CVCr).
Note that this is a metric of *linguistic* vocalizations. I.e. the "cry" and "vegetative" categories that LENA provides for the key child only (not other talkers), for which we provide descriptive analysis in the main manuscript, are not counted in CVCr.

Rationale: (a) CVCr has one of the highest ICC proportion variance explained by corpus and child ID in recordings separated by <2 months (minimally 80%); (b) it does not vary much when individual corpora are held out from this analysis ("leave one corpus out" validation, see EL1000-3-reliabChecks.pdf, 3.B.4, figure on p. 20). Additionally, (c) it has one of the highest betas for age in mixed models that control for corpus and child ID as random structure (~.06; see EL1000-3-reliabChecks.pdf, 3.A.2, p. 9).

Back to Table of Contents

3C: Selection of the model based on the exploration subset

Goal

To select the optimal analysis model structure to be used in confirmation analyses based on the chosen metrics (see 3B) and desiderata detailed in this section.

Procedure and results

Full documentation of this step is available in EL1000-4-modelTest.pdf.

Prior to inspecting any models (in the exploration or confirmation set), we decided on the following desiderata:

- Interpretability: We favor models that we can interpret as cognitive scientists.
- We are mindful of power when it comes to fixed effects and interactions: We have only kept fixed effects and interactions that are central, to avoid overfitting.
- We play it safe with variables that are not central to our analyses, but we would not want confounding things: We keep corpus in the random structure, and overlap quantity (see Figure S3C.1) among the fixed effects, because we want to make sure we do not have them as confounds. For overlap in particular, since overlap across talkers (or with background noise) will result in lower adult and child vocalization counts, not including overlap would artificially inflate the correlation between our measure of adult talk (AVCr) and child speech (CVCr).

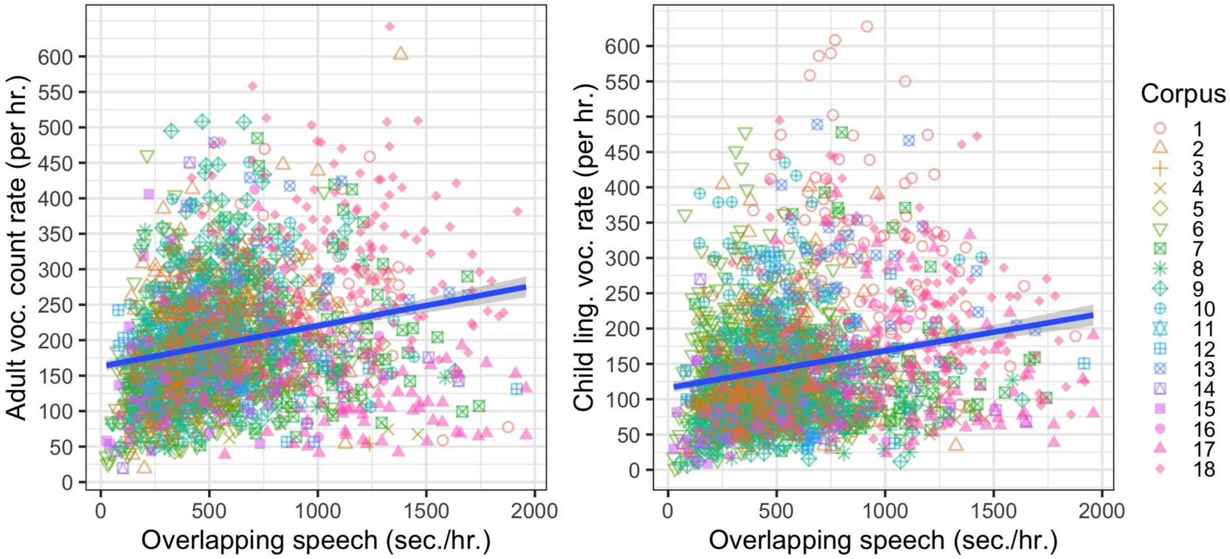


Figure S3C.1. Relationship between the amount of overlap and our input and output metrics. Overlap was retained in the model because it is conceptually necessary to capture potential confounds, and it is significantly related to our key variables.

The random effect structure is always (1|corpus) + (1|corpus:child). Variables considered were: age, gender, SES (in various groupings) and NAE (North American English), as well as monolingual/multilingual status and normativity (i.e. typically developing or not).

To reduce parameters/increase power, the following decisions were made:

- NAE (i.e. whether the language spoken in the household was North American English) was removed because there was no theoretical motivation to examine this variable
- normative*monolingual*ses and normative*monolingual were removed as interactions due to a lack of sufficient observations to test these effects.

Based on model performance in the exploration set:

- The random structure initially selected was the most appropriate (e.g., no random slopes by corpus)
- Age-squared did not improve performance and was therefore not added
- AVC-squared did improve performance but in the interest of ease of interpretation of the model output, it was not included in the final models
- Three-way interactions between age, SES, and AVCr were not theoretically interpretable, and were therefore not included

Back to Table of Contents

3D: Deriving hypotheses to test in the Confirmation Set

Goal

To explain how the final set of hypotheses we tested in our confirmation set were derived.

Procedure

We used a hybrid approach to hypothesis-generation, meaning that we combined predictions from previous work together with tests in the exploration subset. In more detail, we carried out literature searches to identify theoretical positions and previous results to make predictions about each factor and interaction, removing interactions for which there was no theoretical or empirical motivation. Among those that remained, we tested them in the exploration subset, and pared down the model to the most complex model upheld by the exploration data. We also investigated how to ensure model convergence, and checked normality of residuals.

Results

Documentation of this step is available in EL1000-5-prereg_zafter_explo.pdf.

Since our confirmation subset included all of the children with non-normative development and all of the multilingual children, our hypotheses for these groups could not be derived from the exploration subset results. While prior literature readily justified including monolingual status and normative status as predictors in our child speech model, there was more uncertainty regarding how these factors may be connected with language input (i.e. adult talk). We thus conducted literature reviews to guide predictions for our confirmatory subset analysis of adult talk.

Monolingual Status. We found that very few studies reported observed measures of bilingual infants' overall language exposure. The existing results suggest that multilingual caregivers' language input is a significant predictor of multilingual infants' language output when the two measures are collapsed across languages, likely similarly to monolinguals' input.
See EL1000-Multilingual_Language_Environment.docx for more details.

Normative Status. Our review aimed to determine whether there is evidence in the literature that children in non-normative groups are more likely to hear less adult talk or receive less benefit from exposure to adult talk than those in normative groups. Overall, our review returned relatively few studies addressing such links. Furthermore, many of the studies that were identified presented conflicting findings, even within a particular non-normative group. Given the vast heterogeneity in the present study, there does not appear to be sufficient evidence for making a directional prediction of the amount of adult speech or the effect of adult speech on child vocalization in non-normative groups compared to the normative group.
See EL1000-Non_Normative_Input_review.docx for more details.

Back to Table of Contents

3E: From exploration to confirmation, to final analyses

Goal

To explain our shift from a frequentist to a Bayesian framework, and a return to a frequentist approach.

Procedure

The exploratory analyses established the measures (3B) and models (3C), and that allowed us to pre-register and test hypotheses (3D) in the held-out confirmation subset. Since some of the hypotheses were predictions of the null, we decided to change the analytic framework from frequentist mixed models to Bayesian mixed models, since the latter can quantify support for the null (rather than just fail to disconfirm it). However, comments from reviewers and additional methodological papers documenting the instability of Bayesian-based analyses (since this is an approach in very active development currently) led us to return to a frequentist approach, where we are certain that our code and approach will remain stable in years to come.

<u>The final reported model</u>[2] <u>on which main results are based is</u>:

```
output_model = lmer(CVCr_s ~ gender + ses.5 + overlapr_s  +
           normative* AVCr_s* age_s + monoling* AVCr_s* age_s +
           (1 + overlapr_s + AVCr_s |corpus) + (1 | corpus:corp_chi), weights = rec_rat, data =
bsl_conf_scaled
  )
```

<u>The final reported model predicting adults' talk was</u>:

```
input_model = lmer(AVCr_s ~    gender * ses.5 + overlapr_s +
         normative*age_s +  monoling*age_s +
         (1 + overlapr_s |corpus) + (1 | corpus:corp_chi),
         weights = rec_rat, data = bsl_conf_scaled,
   )
```

Model Detail Notes:

- variables ending in _s are centered and scaled
- Baseline levels for categorical variables are: gender: female; ses.5: some university (3); normative, and monolingual

[2] The final <u>pre-registered</u> model predicting infants' linguistic vocalizations was: **output_model** = brm(CVCr_s | weights(rec_rat) ~ gender + ses.3f + overlapr_s + normative* AVCr_s* age_s + monoling* AVCr_s* age_s + (1 |corpus) + (1 | corpus:corp_chi), data = bsl_conf_scaled, iter=niter, warmup=nwarmup, chains=4,cores=2, seed=12). The final pre-registered model predicting adults' talk was: **input_model** = brm(AVCr_s | weights(rec_rat) ~   gender * ses.3f + overlapr_s + normative*age_s + monoling*age_s + (1 |corpus) + (1 | corpus:corp_chi), data = bsl_conf_scaled,   iter=niter, warmup=nwarmup, chains=4,cores=2, seed=12). For both Bayesian models, priors were defined as prior("student_t(3,0,1)") for scaled variables, prior("student_t(5,0,5)" for unscaled ones. The number of iterations niter=4000, nwarmup=500. Using our pre-registered analyses, we found that 11 out of our 12 pre-registered hypotheses were confirmed. The one divergence is that a gender*SES interaction (preregistered hypothesis 1) obtains in the exploration but not the confirmation subset. This may be because it was a false positive in the exploration subset, or because the sample composition changed by design (i.e. by the inclusion of non-monolingual and non-normative children in the confirmation sample only). Since this is not central to our key analyses in the paper, however, we have not taken this matter further.

Following reviewers' feedback, we added an analysis to check whether SES effects emerged when AVCr_s was not included as a predictor:

```
output_model_noAVCr = lmer(CVCr_s ~ gender + ses.5 + overlapr_s  +
            normative* age_s + monoling* age_s +
            (1 + overlapr_s |corpus) + (1 | corpus:corp_chi), weights = rec_rat, data =
bsl_conf_scaled
  )
```

Data processing and analysis were conducted using R (Version 4.1.2; R Core Team, 2020) and the R-packages lme4 (Version 1.1.27.1; Bates et al., 2015)], papaja (Version 0.1.0.9997; Aust & Barth, 2020), and tidyverse (Version 1.3.1; Wickham et al., 2019).

Results

Documentation of this step is available in EL1000-6-z_confirmation_nonBayesian.pdf, which also includes model diagnostics.

Results based on the reported models are in the main text (Tables 1-2). The added model predicting output without including AVCr as a fixed effect revealed a main effect of normative (ß = -.23, SE .09); age (ß = -.64, SE .03); normative*age (ß = -.25, SE .05). The ß for SES were between -.07 and .04, suggesting a very small (non-significant) effect (see EL1000-6-z_confirmation_nonBayesian.pdf, p. 6).

3E References

Aust F, Barth M (2020) papaja: Create APA manuscripts with R Markdown Available at: https://github.com/crsh/papaja.

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67(1):1–48.

R Core Team (2020) R: A language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, Austria) Available at: https://www.R-project.org/.

Wickham H, et al. (2019) Welcome to the tidyverse. Journal of Open Source Software 4(43):1686.

Back to Table of Contents

3F: Convergent evidence for our main results using a different speech analysis algorithm

Goal

The key results we highlight in the paper and that we need to replicate with VTC are:

1) The effect of language input on language output is roughly the same size as normative status, and roughly half the size of the effect of age on language output
2) SES does not significantly predict language output

To what extent might our key results be limited to data from the LENA algorithm? This section sought convergent external evidence. Such an analysis was possible for the subset of our data for which raw audio-recordings were available, and therefore could be reanalyzed using a wholly different speech analysis algorithm. We used an open source alternative called Voice Type Classifier (VTC, Lavechin et al., 2020). VTC was developed using the combination of various corpora in several languages (including Minn, French, Ju|'hoan, Tsimane', English, and several others, in rough order of quantity of data). These corpora were internally split as a function of individual infants into independent training, development, and test sets. As reported on Lavechin et al. (2020), F-score performance on the test set from this multilingual corpus was 77.3% for key child, 82.4% for female adult, and 42.2% for male adult (which is a rarer class overall). Comparison of performance with the state-of-the-art LENA algorithm in a completely held-out, English-only, test set revealed similar levels of performance for VTC and LENA on these three classes.

As we did for LENA, we established the association between automated and human counts, this time using the VTC algorithm.

Table S1F.1. Association between manual (human) vs. automated counts by corpus. Numbers are correlation coefficient (Pearson's r), excepting ganek (Spearman's rho) and weisleder (% agreement). NA indicates no human annotation was available.

			VTC	
Setting	Language	Corpus	AVCr	CVCr
Urban	English	bergelson	0.79	0.81
		kalashnikova	NA	NA
		lucid	0.76	0.74
		rague	NA	NA
		vandam	0.30	0.51

		warlaumont	0.73	0.78
		winnipeg	0.74	0.78
		kidd	NA	NA
	English/Spanish	ramirez-esparza	NA	NA
	Dutch	alphen	NA	NA
	Finnish	elo	NA	NA
	French	lyon	NA	NA
	Spanish	weisleder	NA	NA
	Swedish	swedish	NA	NA
	Vietnamese	ganek	NA	NA
Rural	Yélî	rossel	0.93	0.40
	Wolof	senegal	NA	NA
	Tsimane	tsimane	0.74	0.75

Procedure

We were able to reanalyze nearly every recording in 11 (out of the 18) corpora, except a small handful of recordings for which the LENA data was available but not the audio-recording itself. Given that less data was available, we did not subset these data using the same exploration/confirmation split, but instead included all recordings to maximize power. There were 1,510 recordings included, from 417 children, with the demographic split shown on Figure S3F.1.

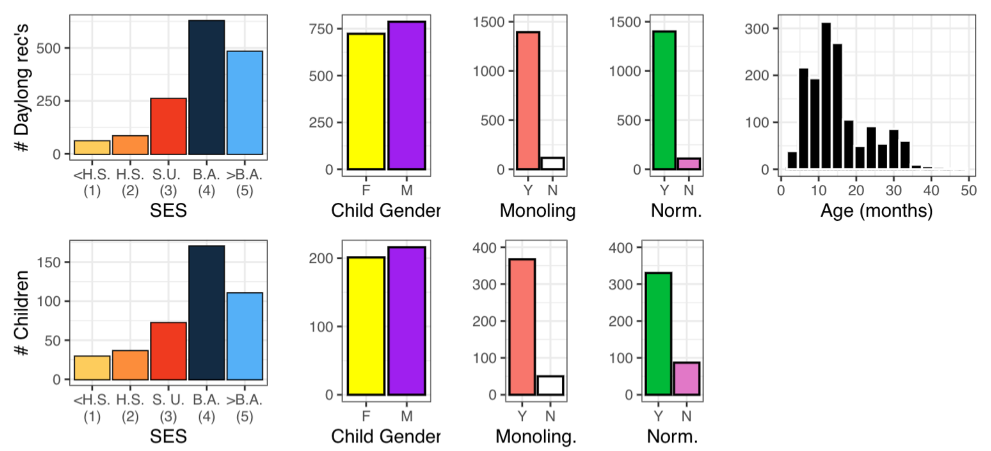


Figure S3F.1. Demographic split based on recordings (top) and children (bottom).

We note that while we have ~500 fewer recordings than in the main analysis (which relies on the 2016 recordings in the confirmation set), we nevertheless have similar proportions of children/recordings as a function of gender, monolingualism and normativity. We note however a trend for fewer children older than 36 months in this reanalysis compared to the main analyses (i.e., older children are relatively under-represented here when compared to analyses in the main manuscript); and fewer recordings of low SES families (i.e., low SES families are relatively under-represented here when compared to analyses in the main manuscript).

Our main analysis uses LENA data as both the language input (AVCr) and language output (CVCr). In this reanalysis, we fit three models predicting language output, by drawing either the predictor, the outcome, or both from our new VTC automated annotations (rather than the LENA automated annotations.) Given that VTC allows overlap between talkers, we did not include the overlap control variable.

The models predicting infants' linguistic vocalizations (i.e. child speech) were the following:

```
output_model_ChildLENA_AdultVTC = lmer(CVCr_s  ~
   gender + ses.5 + overlapr_s  +
     normative* vc_adu_ph_s* age_s +
     monoling* vc_adu_ph_s* age_s +
  (1     |corpus) + (1 | corpus:corp_chi),
  weights = rec_rat,
  data = bsl_conf_scaled
  )

output_model_ChildVTC_AdultVTC = lmer(vc_chi_ph_s  ~
   gender + ses.5 + overlapr_s  +
     normative* vc_adu_ph_s* age_s +
     monoling* vc_adu_ph_s* age_s +
  (1     |corpus) + (1 | corpus:corp_chi),
  weights = rec_rat,
  data = bsl_conf_scaled
  )
output_model_ChildVTC_AdultLENA = lmer(vc_chi_ph_s  ~
   gender + ses.5 + overlapr_s  +
     normative* AVCr_s* age_s +
     monoling* AVCr_s* age_s +
  (1     |corpus) + (1 | corpus:corp_chi),
  weights = rec_rat,
  data = bsl_conf_scaled
  )
```

Variables are as in all preceding sections and the main manuscript, and additionally:

- vc_chi_ph_s is the VTC (new algorithm) equivalent to CVCr_s
- vc_adu_ph_s is the VTC (new algorithm) equivalent to AVCr_s

Results

Full results are available in EL1000-7_alternative-algo.pdf.

Regarding our key conclusions:

1) **The effect of language input on language output is roughly the same size as normative status, and roughly half the size of the effect of age on language output**

This was based on the standardized betas, as reported in this section of the manuscript's text:

> In contrast, young children’s speech production correlated with the amount of adult talk they heard (ß=0.27, SE=0.04), and this correlation strengthened with age (ß=0.13, SE=0.02). This effect is a substantial one. Taking the effects of age and normativity as convenient (but unrelated) gauges for what counts as a considerable effect, we see that the effect of adult talk is about a third of that for age and similar to that for normativity (adult talk: 0.27; interaction adult talk by age: 0.13; age: 0.65; non-normative development: -0.22; interaction non-normative by age: -0.22; all effects are expressed in standard deviations).

Table S3F.2 provides the standardized betas for comparison across the 4 models.

Table S3F.2. Standardized betas predicting child speech from adult talk, age, and non-normative in the analyses reported on in the main text, as well as 3 re-analysis using a software that is not LENA for adult talk (AdultVTC), child speech (ChildVTC), or both. The top row is based on LENA data, reported on Table 1 in the main manuscript, and the three based on recalculated data, replacing CVCr and/or AVCr with its VTC equivalent. Table numbers refer to EL1000-7_alternative-algo.pdf. N_{recs} stands for number of data points (recordings); non-norm for non-normative. * indicates significant after within-model FDR correction.

	Adult talk	**Age**	**Non-norm**	**Age x adult**	**Age x non-norm**
Manuscript (LENA only)	0.26*	0.647*	-0.22*	0.125*	-0.217*
ChildLENA~AdultVTC (table 1, SM)	0.242*	0.488*	-0.207	0.155*	-0.385*
ChildVTC~AdultVTC (table 2, SM)	0.162*	0.426*	-0.122	0.067*	-0.285*
ChildVTC~AdultLENA (table 3, SM)	0.08*	0.453*	-0.127	0.044	-0.266*

This first conclusion remains unchallenged: we find convergent evidence from the VTC algorithm that adult talk has an effect comparable to that of normativity, and about a third of the massive age effect.

To help visualize the adult talk effect, we provide here the equivalent of the main manuscript's Figure 2a, using the VTC data (AdultVTC and ChildVTC, with a median rather than tertile split due to the smaller data quantity):

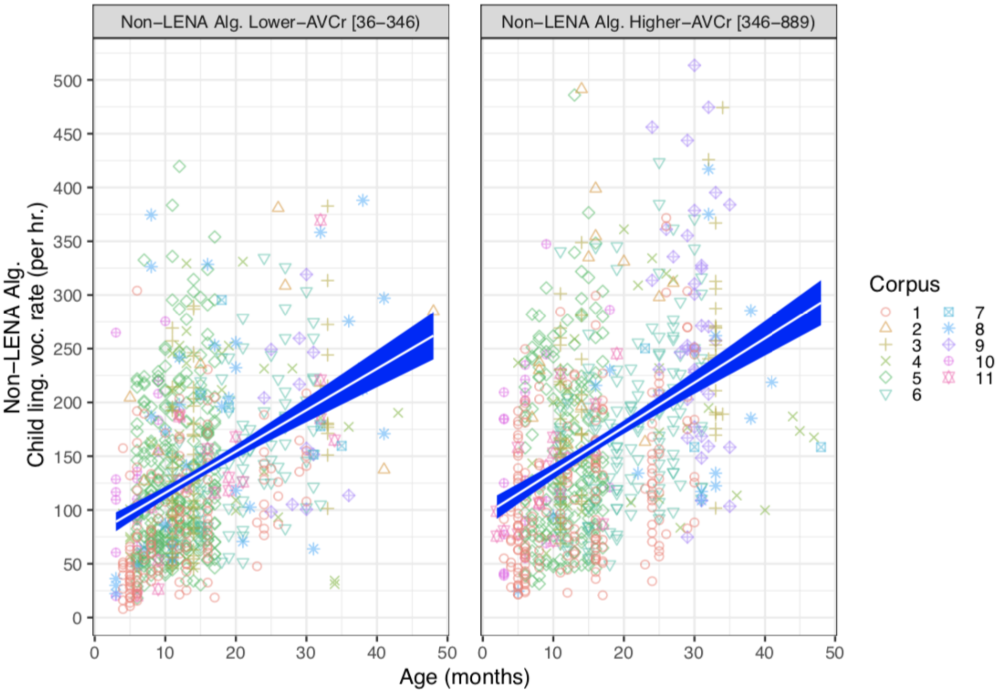


Figure S3F.2: Age by VTC-based CVCr, split by median VTC-based AVCr. Points show each daylong recording in the 11 corpora available for re-analysis by VTC; lines show linear regression with 95% Confidence Intervals (CI). Color-shape combinations show each unique corpus, numbered to preserve anonymity.

2) **SES** does not significantly predict **language output**

None of the 4 betas for SES were significant in any of the three analyses (ChildLENA~AdultVTC in Table S3F.3; ChildVTC~AdultVTC in Table S3F.4; and ChildVTC~AdultLENA in Table S3F.5). The betas tended to be small, and ranged between -.19 and .24 across levels and analyses, consistent with a small effect size (of less than .3 standard deviations).

Table S3F.3. Predicting CVCr extracted using LENA from AVCr extracted using VTC.

	Estimate	**Std. Error**	**q-value**	**stars**
(Intercept)	-0.018	0.133	0.991	
genderM	0.015	0.055	0.936	
ses.51	-0.191	0.17	0.522	
ses.52	-0.05	0.131	0.936	
ses.54	-0.003	0.085	0.991	
ses.55	0.063	0.087	0.707	
overlapr_s	-0.074	0.023	0.004	*
normativeN	-0.207	0.097	0.101	
vc_adu_ph_s	0.242	0.021	<.001	*
age_s	0.488	0.024	<.001	*
monolingN	-0.001	0.104	0.991	
normativeN:vc_adu_ph_s	0.023	0.066	0.936	
normativeN:age_s	-0.385	0.07	<.001	*
vc_adu_ph_s:age_s	0.155	0.023	<.001	*
vc_adu_ph_s:monolingN	0.049	0.063	0.707	
age_s:monolingN	-0.069	0.073	0.619	
normativeN:vc_adu_ph_s:age_s	-0.081	0.057	0.392	
vc_adu_ph_s:age_s:monolingN	0.083	0.063	0.417	

Table S3F.4. Predicting CVCr extracted using VTC from AVCr extracted using VTC.

	Estimate	Std. Error	q-value	stars
(Intercept)	0.358	0.162	0.115	
genderM	-0.116	0.072	0.24	
ses.51	0.219	0.21	0.421	
ses.52	0.094	0.164	0.62	
ses.54	0.1	0.11	0.437	
ses.55	-0.136	0.113	0.415	
overlapr_s	0.327	0.021	<.001	*
normativeN	-0.122	0.119	0.421	
vc_adu_ph_s	0.162	0.019	<.001	*
age_s	0.426	0.022	<.001	*
monolingN	0.249	0.132	0.152	
normativeN:vc_adu_ph_s	0.036	0.067	0.62	
normativeN:age_s	-0.285	0.08	0.002	*
vc_adu_ph_s:age_s	0.067	0.02	0.004	*
vc_adu_ph_s:monolingN	-0.071	0.065	0.421	
age_s:monolingN	0.099	0.069	0.302	
normativeN:vc_adu_ph_s:age_s	0.028	0.059	0.632	
vc_adu_ph_s:age_s:monolingN	-0.062	0.063	0.421	

Table S3F.5. Predicting CVCr extracted using VTC from AVCr extracted using LENA.

	Estimate	**Std. Error**	**q-value**	**stars**
(Intercept)	0.387	0.165	0.105	
genderM	-0.101	0.071	0.256	
ses.51	0.244	0.211	0.372	
ses.52	0.12	0.162	0.551	
ses.54	0.059	0.11	0.612	
ses.55	-0.169	0.113	0.256	
overlapr_s	0.371	0.021	<.001	*
normativeN	-0.127	0.118	0.393	
AVCr_s	0.08	0.02	<.001	*
age_s	0.453	0.022	<.001	*
monolingN	0.248	0.128	0.132	
normativeN:AVCr_s	0.07	0.072	0.426	
normativeN:age_s	-0.266	0.08	0.004	*
AVCr_s:age_s	0.044	0.021	0.105	
AVCr_s:monolingN	-0.042	0.082	0.612	
age_s:monolingN	0.089	0.063	0.256	
normativeN:AVCr_s:age_s	0.119	0.063	0.132	
AVCr_s:age_s:monolingN	-0.052	0.086	0.612	

3F References

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., & Cristia, A. (2020). An open-source voice type classifier for child-centered daylong recordings. Interspeech; also available from arXiv:2005.12656.

Back to Table of Contents

3G: Convergent evidence from analysis subsetting to North American infants

Goal

This section explored whether the two SES conclusions we find overall:
1) SES does not significantly predict language output
2) SES does not significantly predict language input

also hold in the subset of our data most similar to previous research that does report SES effects, namely North American infants.

Procedure

We created a subset which included only North American children and ran our child speech and adult talk models on this subset. We also provide a figure depicting child speech by SES in the North American vs. non-North American data.

The model predicting infants' linguistic vocalizations (i.e. child speech) within the North American infants was:

```
output_model =lmer(CVCr_s  ~
  gender + ses.5 + overlapr_s  +   normative* AVCr_s* age_s +   monoling* AVCr_s* age_s +
 (1     |corpus) + (1 | corpus:corp_chi),
 weights = rec_rat,
 data = nam
 )
```

The model predicting adult talk was:

```
input_model = lmer(AVCr_s  ~
  gender * ses.5 + overlapr_s +   normative*age_s +  monoling*age_s   +
 (1  + overlapr_s  |corpus)  + (1 | corpus:corp_chi) ,
 weights = rec_rat,
 data = nam
 )
```

Results

Full documentation of this step is available in EL1000-8_homogeneous-datasets.pdf. Since we are looking only at North American data, our data is limited to 642 daylong recordings from 206 infants in 7 corpora.

Results for children's linguistic vocalizations (i.e. child speech, CVCr) are shown in Figure S3G.1 and Table S3G.1. We replicate our overall (whole sample) conclusions in terms of AVCr and age being significant predictors, whereas gender and SES are not. The significant AVCr*age interaction is also replicated. The main effect of normativity is not replicated, but the age*normative interaction is larger than in the overall analysis.

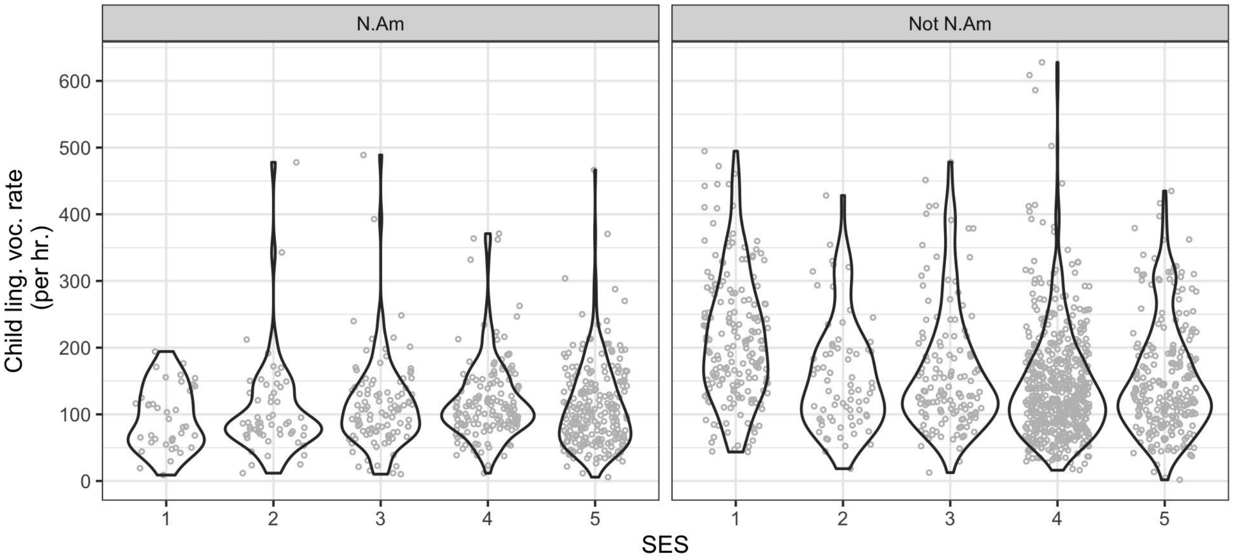


Figure S3G.1. CVCr by SES in North American (left) and Non-North American (right) data. Points = individual recordings, slightly jittered horizontally to avoid overlap.

Table S3G.1. Predicting CVCr in only the North American subset of the data

	Estimate	Std. Error	q	
(Intercept)	0.066	0.181	0.885	
genderM	-0.01	0.069	0.885	
ses.51	-0.179	0.175	0.62	
ses.52	0.02	0.136	0.885	
ses.54	0.02	0.106	0.885	
ses.55	-0.084	0.103	0.755	
overlapr_s	-0.019	0.031	0.885	
normativeN	-0.039	0.133	0.885	
AVCr_s	0.351	0.041	<.001	*
age_s	0.42	0.042	<.001	*
monolingN	-0.15	0.13	0.613	
normativeN:AVCr_s	0.057	0.109	0.885	
normativeN:age_s	-0.554	0.092	<.001	*
AVCr_s:age_s	0.187	0.034	<.001	*
AVCr_s:monolingN	0.021	0.084	0.885	
age_s:monolingN	0.081	0.074	0.613	

normativeN:AVCr_s:age_s	-0.177	0.09	0.154	
AVCr_s:age_s:monolingN	0.186	0.09	0.143	

To test the possibility that AVCr accounted for variance that would otherwise be accounted for by SES, we repeated the analysis but excluded the AVCr predictor. SES did not become a significant predictor in this analysis either (see Table S3G.2).

Table S3G.2. Predicting CVCr in only the North American subset of the data, excluding the AVCr predictor, to check whether SES gains importance here.

	Estimate	**Std. Error**	**q**	**stars**
(Intercept)	0.088	0.235	0.967	
genderM	0.003	0.08	0.967	
ses.51	-0.065	0.205	0.967	
ses.52	0.097	0.158	0.967	
ses.54	0.007	0.12	0.967	
ses.55	-0.009	0.117	0.967	
overlapr_s	0.006	0.042	0.967	
normativeN	-0.175	0.152	0.967	
age_s	0.464	0.045	<.001	*
monolingN	-0.119	0.151	0.967	
normativeN:age_s	-0.619	0.102	<.001	*
age_s:monolingN	0.04	0.081	0.967	

Regarding adult talk, results show that there are no differences in AVCr as a function of any of our predictors (see Table S3G.3). Thus, our conclusions in the main paper also hold true of the North American subset.

Table S3G.3. Predicting AVCr in only the North American subset of the data

	Estimate	Std. Error	q	stars
(Intercept)	-0.279	0.225	0.459	
genderM	0.575	0.276	0.154	
ses.51	0.62	0.343	0.235	
ses.52	-0.096	0.317	0.991	
ses.54	0.342	0.234	0.388	
ses.55	0.282	0.234	0.459	
overlapr_s	0.168	0.063	0.154	
normativeN	0.05	0.216	0.991	
age_s	-0.023	0.068	0.991	
monolingN	0.005	0.215	0.991	
genderM:ses.51	-1.376	0.506	0.093	
genderM:ses.52	-0.373	0.444	0.642	
genderM:ses.54	-0.894	0.351	0.093	
genderM:ses.55	-0.304	0.341	0.642	
normativeN:age_s	0.002	0.152	0.991	
age_s:monolingN	0.013	0.124	0.991	

Additional analyses in EL1000-8_homogeneous-datasets.pdf also show that these conclusions hold when considering all English speaking corpora, and all urban corpora, rather than just North American children (along with the analogous graph of SES and child vocalizations in these subsets).

Back to Table of Contents

3H: Additional robustness checks

Goal

To summarize additional analyses performed to check whether our key conclusions were stable to alternative analytic decisions and/or alternative interpretations. Namely, we established:
1: Key results hold for alternative operationalizations of SES
2: Key results hold when adding corpus-level slopes for SES
3: Key results hold for children under and over 18 months of age
4: Key results hold when considering causal paths
5: Key results hold when considering other children's speech
6: Key results hold when using contrast sum coding for non-continuous variables

More detailed information on each check can be found in EL1000-9_other-checks.pdf.

Procedure

We performed several additional analyses on subsets of the data, explored directed acyclic graphs, and plotted the raw data in several ways.

Results

1: We checked whether the way SES was operationalized affected our conclusions by changing which level was declared as baseline, declaring SES as ordinal, declaring SES as a continuous measure based on the 5 levels, based on years of education, or through a binary classification based on country-level characteristics. The conclusion from all of these analyses is that, as in the main manuscript, increasing SES does not predict higher child vocalization rate or higher adult vocalization rate. See EL1000-9_other-checks.pdf, Check #1 for more information.

2: We checked whether including SES and adult talk as random slopes per corpus affected our conclusions. This would allow the size of the SES effect and adult talk to vary across corpora, which would be necessary if e.g., SES matters more in one corpus than another (due to differences across countries, cultures, or even the range of SES found in a given corpus). Including these led to singularity issues (because random slopes for SES within corpus were highly correlated with intercepts for some SES levels). But crucially, fixed effects for these models rendered the same patterns of significance as in the analyses we present in the main manuscript. Thus, including random slopes for SES within corpus or not does not affect our conclusions. See EL1000-9_other-checks.pdf, Check #2 for more information.

3: We checked whether our key conclusions held for recordings carried out when children were below and above 18 months, since vocal production changes quite a bit over the first three years of life. 18 months was selected given that this is (a) the mean age of children in our recordings, (b) when the commonly used MCDI vocabulary checklist (Fenson et al, 1994) switches from the ‘infant’ to ‘toddler’ version and, (c) an age where various screeners (e.g. Survey of Well-Being of Young Children, SWYC, Perrin et al., 2016) query milestones like beginning 2-word combinations. By and large, estimates fitted to data under and over 18 months were widely overlapping with each other and our full-group analyses. The small discrepancies we found were reasonable given that there was less data for each sub-analysis (for instance, for normativity, only a tiny proportion of children had been diagnosed before 18 months). See EL1000-9_other-checks.pdf, Check #3 for more information.

We also provide a by-corpus visualization of our age and normativity result here, splitting the continuous age variable at 18 months:

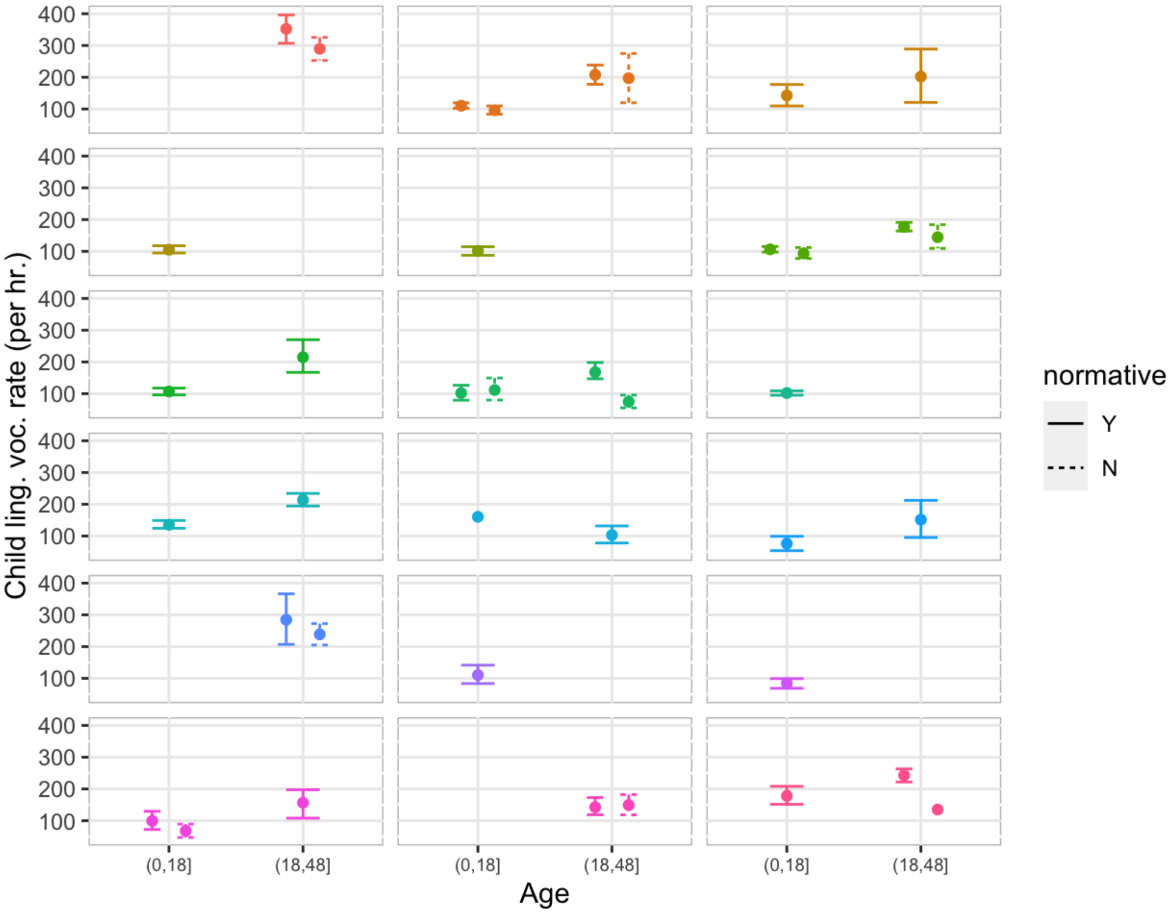


Figure S3H.1. Child speech as a function of age and normative development within individual corpora. Age is split at 18 months (the mean age in the dataset). Older children and children with normative development produce more speech. Black lines = 99% bootstrapped CIs of sample means, color = corpus.

4: Using causal path analyses, we established that our key conclusions were not affected by the absence of control for a confound (due to a fork or a pipe, whereby we observe a spurious correlation due to failing to control for a confound), or by the inappropriate inclusion of a variable (due to a collider, a variable that "creates" a spurious correlation which appears only when controlling for that variable). See EL1000-9_other-checks.pdf, Check #4 for more information.

5: We fit a new regression model predicting children's talk from the same predictors in the main analysis and adding predictors of input quantity from other children's speech. We found that our estimates were virtually unchanged by these additions, lending no support to the hypothesis that the sheer prevalence of other children's speech could cause adults' input to vary in importance. See EL1000-9_other-checks.pdf, Check #5 for more information.

6: We chose to use treatment coding for gender, normativity, and SES, because we find it easier to interpret main effects and interactions in this case. Others may prefer contrast sum coding. We re-fit our outcome and input models using the latter, to find the exact same patterns of significance/non-significance as in the analyses presented in the main paper. Estimates do change, but not in a way that threatens our interpretations. See EL1000-9_other-checks.pdf, Check #6 for more information.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., & Stiles, J. (1994). Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development*, *59*(5), i–185. JSTOR.

Perrin, E. C., Sheldrick, R. C., Visco, Z., & Mattern, K. (2016). The survey of well-being of young children (SWYC) user's manual. *Boston, MA: Floating Hospital for Children at Tufts Medical Center*.

Back to Table of Contents

Supplement 4: Evidence against alternative interpretations

Goal

To clarify the scope of our conclusion that *"Children who heard less talk from adults produced less speech."*

Results

We'd like to explicitly underscore that we do not make a causal claim, but rather a correlational one (see the "Adult Talk and Child Speech" section of the manuscript for a discussion of non-causal interpretations, including familial covariance).

That said, we'd like to highlight that it is **implausible that the correlation we find stems from recordings on special days.** While there will be enormous variance in documented activities (e.g., a day filled with TV in an urban recording vs. a day of gardening and family visits in a rural recording), two aspects of the data strengthen our belief that the results are an appropriate reflection of typical cross-population variation in everyday activity contexts: First, a universal feature of data collection across these corpora is that families make the recording on a "typical" day and, second, the target children, who are all young, need regular changing, feeding, and napping in all settings.

These routines certainly relate to speech prevalence in different ways: napping will have less speech than changing. By using day-long recordings, we smooth across these regularly occurring events. In fact, in other work (Bergelson et al, 2018) we've done a direct comparison of different aspects of the (manually annotated) nouns children hear both in a daylong audio-recording and an hour-long video-recording taken close together in time. In brief, while we find more talk in general in the shorter video-recordings, we find very consistent and robust correlations across almost all dimensions (quantity, talker, utterance type, etc.) confirming that even when one captures a context that's higher talk volume, the amount of talk is informative about the general amount of talk that child hears in a day. Converging evidence comes from other labs and comparisons (e.g. Tamis-Lemonda et al., 2017; Gilkerson et al., 2017). Indeed, using a large, dense sample of daylong recordings, Gilkerson et al. (2017) find high test-retest variability on LENA recordings collected 0-16 weeks apart for measures including Adult Word Count and Child Vocalization Count (from which our AVCr and CVCr are derived).

It is also **implausible that the correlation we find stems from some families being in more noisy environments than others**. Firstly, our random effects are structured to statistically account for family (and corpus) variation (see Methods, main manuscript). Moreover, all datasets were gathered while the child wore the recording device, on which there are no settings or knobs that can be adjusted. This ensures that the voices' amplitudes in the recordings cannot vary due to systematic differences in how far the recording device is from the child and the adult, as may happen in other datasets with an external recorder that may be placed closer or further away from the dyad. Samples vary in how much background noise there is, which is why we control for this using a variable that measures the amount of detected overlap, which would be higher for noisier environments (see SI3C). This concern is also addressed by the use of VTC as our second algorithm, which does not discard talker overlap, unlike the LENA algorithm (see Methods, main manuscript).

4 References

Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., & Tor, S. (2019). Day by day, hour by hour: Naturalistic language input to infants. *Developmental science*, *22*(1), e12715.

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., ... & 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.

Tamis‐LeMonda, C. S., Kuchirko, Y., Luo, R., Escobar, K., & Bornstein, M. H. (2017). Power in methods: Language to infants in structured and naturalistic contexts. *Developmental science*, *20*(6), e12456.

Back to Table of Contents

Supplement 5: Other Publications Based on Constituent Datasets

Corpus name	Corpus reference	Full reference (APA)	Brief summary
alphen	van Alphen et al 2020a		Used for a pilot study. Outcomes are not published in a scientific journal
	van Alphen et al 2020b	Blom, E., Fikkert, P., Scheper, A., van Witteloostuijn, M, & van Alphen, P. (accepted).The language environment at home of children with (a suspicion of) a Developmental Language Disorder and relations with standardized language measures. Journal of Speech, Language, and Hearing Research	The study compares the home language environments of Dutch children with (a suspicion of) Developmental Language Disorder (DLD) with that of children with typical development (TD). Toddlers with (a suspicion of) DLD vocalize at home less than their TD peers. They also hear fewer adult words and experience fewer conversational turns.
bergelson	Bergelson 2017		For space sake, see this list
elo	Elo 2016		No work has yet been published based on these data
ganek	Ganek & Eriks-Brophy 2019	Ganek, H. V., & Eriks-Brophy, A. (2018). A concise protocol for the validation of Language ENvironment Analysis (LENA) conversational turn counts in Vietnamese. Communication Disorders Quarterly, 39(2), 371-380.	This paper describes a validation effort based on a subset of recordings used here, together with human annotation, to assess accuracy of CTC.
		Ganek, H., Smyth, R., Nixon, S., & Eriks-Brophy, A. (2018). Using the Language ENvironment Analysis (LENA) system to investigate cultural differences in conversational turn count. Journal of Speech, Language, and Hearing Research, 61(9), 2246-2258.	A combination of Vietnamese and American data was employed to compare CTC across cultures and hearing statuses.

		Ganek, H., Nixon, S., Smyth, R., & Eriks-Brophy, A. (2019). A cross-cultural mixed methods investigation of language socialization practices. The Journal of Deaf Studies and Deaf Education, 24(2), 128-141.	Differences in CTC across cultures is interpreted based on language socialization practices.
kalashnikova	Brookman et al. 2020	Brookman, R., Kalashnikova, M., Conti, J., Xu Rattanasone, N., Grant, K. A., Demuth, K., & Burnham, D. (2020). Depression and anxiety in the postnatal period: An examination of infants' home language environment, vocalizations, and expressive language abilities. Child development, 91(6), e1211-e1230.	This paper reports significant differences in the early language environment experienced by infants whose mothers were and were not affected by postnatal depression and anxiety. LENA recordings were collected when the infants were 6 and 12 months, and measures of infant vocalizations, adult vocalizations, and conversational turns were extracted.
kidd	Kidd et al. 2018	Donnelly, S., & Kidd, E. (2021). The longitudinal relationship between conversational turn-taking and vocabulary growth in early language development. Child Development, 92(2), 609-625.	Growth curve analyses revealed a bidirectional relationship between conversational turns (from LENA recordings) and vocabulary growth (from parental reports) in English-learning children, controlling for the amount of words in children's environments.
lucid	Rowland et al. 2017		No work has yet been published based on these data
lyon	Canault et al. 2017	Canault, M., Le Normand, M. T., Foudil, S., Loundon, N., & Thai-Van, H. (2016). Reliability of the language environment analysis system (LENA™) in European French. Behavior research methods, 48, 1109-1124.	This paper assess the accuracy of LENA's algorithm for French, comparing AWC across LENA annotations and human transcriptions of 1h per day and per child, for a total of 54 recordings, which are a superset of those used in our paper.

		Loukatou, G. R., Le Normand, M. T., & Cristià, A. (2019). Is it easier to segment words from infant-than adult-directed speech? Modeling evidence from an ecological French corpus. In CogSci (pp. 2186-2192).	Using the transcriptions from Canault et al., 2016, augmented by distinguishing child- from adult-directed speech, this paper uses computational modeling to study how easy it is to segment words from those two registers.
		Lavechin, M., de Seyssel, M., Métais, M., Metze, F., Mohamed, A., Bredin*, H., Dupoux*, E. & Cristia*, A. (2022). Early phonetic learning from ecological audio: Domain-general versus domain-specific mechanisms. https://psyarxiv.com/5tmgy	Using some of the audio, this paper uses computational modeling to study the kinds of mechanisms learners may bring to the task in order to learn language-specific phonological representations.
rossel/png2019	Cristia & Casillas 2020		No work has yet been published based on these data
rague	Hamrick et al. 2019	Hamrick, L.R., Seidl, A., & Kelleher, B.L. [In press]. Semi-Automatic Assessment of Vocalization Quality for Children with and without Angelman Syndrome. American Journal on Intellectual and Developmental Disabilities.	This study uses a subset of the rague corpus to present a post-processing system used to refine and expand on information about vocal features from LENA recordings (specifically, canonical and noncanonical syllable rate and canonical babbling ratio). The study demonstrates the feasibility of this post-processing system for children at low risk for developmental delays as well as children with Angelman syndrome. We found that vocal features demonstrated expected associations with other language measures in the low risk group, but not in the Angelman syndrome group, perhaps highlighting the limitations of standard measures in assessing vocal development in children with Angelman syndrome.
rague	Hamrick et al. 2019	Semenzin, C., Hamrick, L., Seidl, A., Kelleher, B., & Cristia , A. (2021). Describing vocalizations in young children: A big data approach through citizen science annotation. Journal of Speech, Language, and Hearing Research, 64, 2401-2416. 10.31219/osf.io/z6exv	This study uses a subset of the rague corpus to explore the extent to which citizen science coding of canonical babbling aligns with gold-standard laboratory coding for children at low risk for developmental delays and children with Angelman syndrome. Overall this study showed that citizen science coding aligned well with gold-standard coding for children both with typical language development as well as children with significant language delays.

rague	Hamrick et al. 2019	Kelleher, B. L., Halligan, T., Witthuhn, N., Neo, W. S., Hamrick, L., & Abbeduto, L. (2020). Bringing the laboratory home: PANDABox telehealth-based assessment of neurodevelopment risk in children. Frontiers in Psychology, 11(1634), 1-14. https://doi.org/10.3389/fpsyg.2020.01634	This study uses a subset of the rague corpus and explores the rate of child and adult speech occurring during a daylong LENA recording as well as LENA recordings collected during various standardized laboratory tasks for children with Down syndrome and their families. This study explored LENA as a component of a battery of methods being adapted for telehealth assessment of early development for children with neurogenetic syndromes. We found that rates of child and adult vocalizations demonstrated face validity across various tasks (e.g., more vocalizations were detected during tasks with interaction vs. independent play).
ramirez-esparza	Ramirez et al. 2014	Ramírez-Esparza, N., García-Sierra A., & Kuhl, K. P. (2014). Look who's talking: Speech style and social context in language input to infants is linked to concurrent and future speech development. Developmental Science, 17, 880-891. DOI: 10.1111/desc.12172	This study shows that parentese speech between parents and their infants is related to later word production. Overall this study shows that the quality of speech is what matters for later language development.
ramirez-esparza	Garcia-Sierra et al. 2016	García-Sierra, A., Ramírez-Esparza, N., & Kuhl, K. P. (2016). Relationships between quantity of language input and brain responses in bilingual and monolingual infants. International Journal of Psychophysiology, 110, 1-17.	This study shows that the quantity of speech as assessed by the LENA recorder relates to language development as assessed with ERPs.
ramirez-esparza	Ramirez et al. 2017a	Ramírez-Esparza, N., García-Sierra, A., & Kuhl, K. P. (2017). The impact of early social interaction on later language development in Spanish-English bilingual infants. Child Development, 88, 1216–1234.	This study shows the quality of speech used by parents in English and/or Spanish relates to later language development. This study also demostrates that cultural behaviors are intimately associated to the way bilingual children learn each of their languages.
ramirez-esparza	Ramirez et al. 2017b	Ramírez-Esparza, N., García-Sierra A., & Kuhl, K. P. (2017) Look who's talking NOW! Social interactions and language development across time. Frontiers in Psychology,	This study shows that the way parents speak to their children changes from when they are 1 year of age to 33 months of age. Overall parents decrease the proportion of time they use parentese speech and use more standard speech. Furthemore, parentese speech used during infancy predicts word production even at 33 months of age.

		Developmental Psychology, 8, DOI: 10.3389/fpsyg.2017.01008	
senegal		Weber, A., Fernald, A., & Diop, Y. (2017). When cultural norms discourage talking to babies: Effectiveness of a parenting program in rural Senegal. Child Development, 88(5), 1513-1526.	This study assessed the effectiveness of a parenting program designed to encourage verbal engagement between caregivers and infants in Wolof-speaking villages in rural Senegal. Results showed effects of the intervention on mothers' behavior from videos and in children's parent-reported vocabulary.
swedish	Marklund et al. 2019, Schwarz et al. 2019ab	Marklund, E., Schwarz, I. C., & Lacerda, F. (2019). Amount of speech exposure predicts vowel perception in four-to eight-month-olds. Developmental Cognitive Neuroscience, 36, 100622.	The study shows a correlation between the amount of speech input and the linguistic the component of the MMR discrimination response in 4- to 8-month-old Swedish learning infants.
		Schwarz, I. C., Botros, N., Lord, A., Marcusson, A., Tidelius, H., & Marklund, E. (2017). The LENATM system applied to Swedish: Reliability of the Adult Word Count estimate. In Interspeech 2017, Stockholm, Sweden, 20-24 August, 2017 (pp. 2088-2092).	An evaluation of the LENA AWC measure for Swedish based on four 30-month-old children shows sufficient correlation with manual word count.
tsimane	Scaff et al. 2020	Scaff, C., Casillas, M., Stieglitz, J., & Cristia, A. (2022). Characterization of children's verbal input in a forager-farmer population using long-form audio recordings and diverse input definitions. https://psyarxiv.com/mt6nz/	Based on human annotations of a superset of recordings (including those done with LENA but also USB and Olympus devices), this paper describes child-directed and total input for 27 Tsimane' young children. They find estimates change markedly across a permissive versus restrictive definitions of language input, whereas composition (proportion of child-directed versus overhearable; proportion from male adults, female adults, other children) were quite stable.
vandam/cougar	VanDam 2018		For space sake, see this list
warlaumont	Warlaumont et al. 2016		For space sake, see this list
weisleder	Weisleider & Mendelsohn 2019	Cychosz, M., Villanueva, A., & Weisleder, A. (2021). Efficient estimation of children's language exposure in two bilingual	The authors employed a general sampling with replacement technique to efficiently estimate two key elements of children's early language environments: (a) proportion of child-directed speech (CDS) and (b) dual language exposure. Approximately 49 min from each

		communities. Journal of speech, language, and hearing research, 64(10), 3843-3866.	recording or just 7% of the overall recording was required to reach a stable proportion of CDS and bilingual exposure.
winnipeg	McDivitt & Soderstrom 2016	Soderstrom, M., Grauer, E., Dufault, B., & McDivitt, K. (2018). Influences of number of adults and adult: child ratios on the quantity of adult language input across childcare settings. First Language, 38(6), 563-581.	This study examined the relationship between the number of adults (and adult:child rations) on AWC across three childcare settings. Different relationships were found across the settings. Note that only 2 participants from this study overlap with the current paper.
		Soderstrom, M., & Wittebolle, K. (2013). When do caregivers talk? The influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments. PloS one, 8(11), e80646.	This study examined the effect of activity context and time of day on AWC and child vocalizations across home and daycare settings. Similarities and differences were found in the relationships across the childcare contexts. Note that only 2 participants from this study overlap with the current paper.
		Cristia, A., Dupoux, E., Ratner, N. B., & Soderstrom, M. (2019). Segmentability differences between child-directed and adult-directed speech: A systematic test with an ecologically valid corpus. Open Mind, 3, 13-22.	This study used the corpus to examine computational approaches to discriminating IDS and ADS. Note that only 2 participants from this study overlap with the current paper.
		Ko, E. S., Seidl, A., Cristia, A., Reimchen, M., & Soderstrom, M. (2016). Entrainment of prosody in the interaction of mothers with their young children. Journal of child language, 43(2), 284-309.	This study examined prosodic entrainment effects in mother-infant dyads. Note that only 2 participants from this study overlap with the current paper.
		Bunce, J., Soderstrom, M., Bergelson, E., Rosemberg, C., Stein, A., Migdalek, M., & Casillas, M. (under revision). A cross-cultural examination of young children's everyday language experiences.	The Winnipeg corpus was part of a larger cross-cultural meta-corpus (using hand annotation only) to examine quantitative effects of various factors (community, speaker, etc.) on IDS and ADS
			For other references using this corpus see here.

Supplement 6: Glossary

Below we provide a glossary of variables and technical terms, as they are used in this SI and main text.

Adult Talk: Number of utterances in near and clear adult speech (from both male and female adults), estimated by the LENA™ algorithm.

Adult Vocalization Count rate (AVCr): This is the more technical name for 'adult talk'. Number of utterances in near and clear adult speech (from both male and female adults) estimated by the LENA™ algorithm. The count for the entire day is divided by the recording length, resulting in this rate of utterances per hour.

Child Speech: Number of child speech vocalizations (including babbling and more mature speech utterances) estimated by the LENA™ algorithm

Child Vocalization Count rate (CVCr): This is the more technical name for 'child speech'. Number of child speech vocalizations (including babbling and more mature speech utterances) estimated by the LENA™ algorithm. The count for the entire day is divided by the recording length, resulting in this rate of utterances per hour. This count excludes two other classes of vocalizations that the LENA™ system tags, 'cry' and 'vegetative'.

Confirmation subset: the subset of data reserved for confirmatory analysis. Here, after initial data-splitting into confirmation/exploration subsets, the confirmation subset was left untouched until we preregistered our planned confirmatory analysis.

Corpus: collection of LENA™ algorithm output and metadata for a set of daylong audio-recordings.

Cry: This metric was calculated from LENA™'s Cry measure which estimates the number of fixed signals produced by the child as emotional expressions. Here it is defined as the proportion of vocalizations that are classified as Cry.

Data Steward: each corpus has been collected by a research group; we asked each research group to name one individual who is our primary contact person, and who we call here data steward.

Discovery subset (a.k.a. **Exploration subset**): the subset of data reserved for exploratory analysis. Here, after initial data-splitting into confirmation/exploration subsets, the exploration subset was used to examine variables, plan models, and establish hypotheses to be tested in a confirmatory way in the confirmation subset.

Exploration subset (a.k.a. **Discovery subset**): the subset of data reserved for exploratory analysis. Here, after initial data-splitting into confirmation/exploration subsets, the exploration subset was used to examine variables, plan models, and establish hypotheses to be tested in a confirmatory way in the confirmation subset.

Homebank: a shared repository for daylong recordings and their metadata. Cf. https://homebank.talkbank.org

Its: LENA™'s proprietary automated speech algorithm generates .its files. ITS stands for interpreted time segment. All of our analyses are derived from these its files and concomitant metadata provided by data stewards

Language input: what a child receives from their environment, linguistically. While this could include speech or sign language input, here it largely refers to talk to and around the child from adults (i.e. *adult talk*).

Language output: what a child produces, linguistically. While this could include spoken or signed language produced by the child, here it largely refers to the child's early speech productions, ranging from babbling to sentences (i.e. *child speech*).

LENA™: A system including a small audio-recorder worn by a child, and the automated output (its file) produced by running the derived daylong audio-recording through a speech analysis algorithm. LENA stands for *l*anguage *en*vironment *a*nalysis.

Linguistic vocalization: a vocal behavior with language/speech content. For children, this is contrasted with two other types of vocalizations: vegetative and cry. (The LENA™ system does not track vegetative or cry vocalizations from talkers other than the child wearing the recorder).

Monolingual (monoling): someone who speaks (or is learning) one language.

Multilingual: someone who speaks (or is learning) >1 language.

Normative Development: development following a typical trajectory with no clinical diagnoses or high risk for language delays and deficits

Non-Normative Development: development following an atypical trajectory, based on a clinical diagnosis or high risk for language delays and deficits.

Open Science Foundation (OSF): an online platform where project details and preregistrations can be logged and documented. Cf. https://osf.io/

Overlap(r): The amount of audio that the LENA™ tags as overlap, divided by recording length to generate a rate (overlapr).The LENA™ algorithm separately categorizes speech sounds that *overlap* either with other speech/vocalizations, or with non-speech noise. Notably, within such overlap regions, the LENA™ algorithm does not report child vocalizations or adult utterances (from which our hourly child speech (CVCr) and adult talk (AVCr) values are derived). Overlap serves as a control variable to account for this in our models. Without this control, AVC(r) and CVC(r) counts would be artificially deflated.

Rec_rat: length of the *rat*io of each *rec*ording's length divided by maximum recording length

_s: a suffix we add to variables in our code that have been scaled (i.e. normalized).

Socioeconomic status (SES): We use maternal education as a proxy for SES based on both prior literature and the metadata available for our dataset. See 2b.

vc_adu_ph_s is the VTC equivalent to AVCr_s, the Adult Vocalization Count rate (AVCr) scaled.

vc_chi_ph_s is the VTC equivalent to CVCr_s, the Child Vocalization Count rate (CVCr) scaled. Note that VTC does not distinguish between speech and crying vocalizations and counts both together.

Vegetative: This is a metric derived from LENA's Vegetative (VEG) measure which estimates the quantity of respiratory or digestive sounds produced vocally. Here it is defined as the proportion of vocalizations that are classified as vegetative.

VTC: VTC is short for voice type classifier, a LENA open-source alternative software we used for the analyses.

Back to Table of Contents