Supplementary Materials to Establishing the reliability and validity of measures extracted from long-form recordings

Contents

Recalculate everything or not?

If RECALC is set to TRUE, then the ICC tables will be re-generated and the simulation re-ran. Notice that you can only do this if you have access to underlying data. If you are not one of the paper’s authors, please email us for access to reproduce this section. You do not need access to reproduce all the rest.

Read in all data

SM A: Simulation to better understand ICC

We were uncertain of how to interpret ICC’s numeric values. It is described as “proportion of variance explained”, but we do not know if it should be considered as a percentage (like R^2) or a correlation (like r).

We therefore simulated data controlling the underlying r between paired datapoints to see how ICC recovered that underlying r.

We will inform the simulation by the data we have as follows:

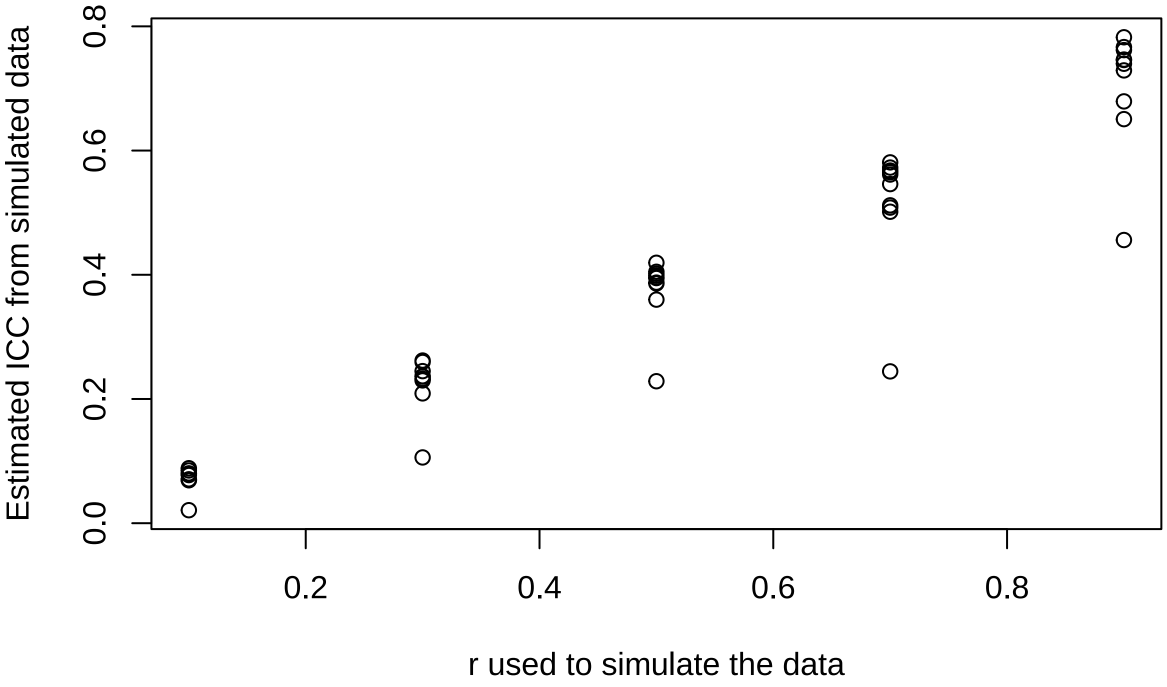
- we'll have the same N of corpora, and of children in each corpus
- we'll have the same metrics for each (i.e., CVC, AWC, etc) – and these metrics will have the same mean & SD for day 1 of recordings as in observed data

We'll depart from reality as follows:

- we will not consider the r across multiple days observed in the data, but instead generate data points to vary r from a small correlation (r=.1), a moderate one (r=.3), a larger one (r=.5). It is unlikely that test-retest correlations in infancy would be much greater than .5 (based on previous studies using test-retest in experimental tasks), but for completeness, we also use .7 and .9
- we will not consider child age, nor variable re-recording periods
- we will have a single pair of recordings (rather than variable N of re-recordings)

We use simstudy, a package created for such simulations, following the vignette https://cran.r-project.org/web/packages/simstudy/vignettes/correlated.html to create correlated data providing a correlation matrix

In the following plot, each point represents the ICC extracted from a mixed model applied to one metric, combining data from all corpora. It appears that ICC values reflect underlying r values, but underestimating r more the larger r is.



SM B: More information for benchmarking our results against previously reported reliability studies

First, we looked for measures of language development used with observational data that can be employed with children aged 0-3 years, and which are available at least in English. All of the instruments we found rely on reports from caregivers, who are basing their judgments on their cumulative experience with the child (e.g., the Child Observation Record Advantage, Schweinhart, McNair, & Larner, 1993; the Desired Results Developmental Profile, REF; the MacArthur-Bates Communicative Development Inventory, REF). Readers are likely most familiar with the MB-CDI, Fenson et al., 1994 report a correlation of r=.95 in their sample of North American, monolingual infants. We did not find a systematic review or meta-analysis providing more such estimates. However, @frank2021 analyzed data archived in a CDI repository, concentrating on American English and Norwegian, where longitudinal data was available. They found that for both datasets, correlations within 2-4 months were above r=.8, with correlations at 16 months of distance (i.e., a form filled in at 8 and 24 months) at their lowest, r=.5. These correlations are very high when considering that the CDI tends to have ceiling and floor effects at these extreme ages. Another report looked at correlations when parents completed two versions of the form in two media (e.g., short form in paper and long form online, or vice versa) within a month. Here, the correlation was r=.8 for comprehension and r=.6 for production. It is worth bearing in mind that test-retest reliability in parental report measures does not depend only on the consistency in the individual infants' ranking for a given behavior, but also on the consistency of the adult in reporting it. Moreover, they are based on cumulative experience, rather than a one-shot observation, as in the case of long-form recordings. Therefore, they do not constitute an appropriate comparison point, and by and large we can be quite certain that they will yield higher reliability than metrics based on the children themselves. For example, a meta-analysis of infant laboratory tasks, including test-retest data for sound discrimination, word recognition, and prosodic processing, found that the meta-analytic weighted average was not different from zero, suggesting that performance in these short laboratory tasks may not be captured in a stable way across testing days. Thus, parental report (or short lab studies) may not be the most appropriate comparisons for our own study.

Second, we did a bibliographic search for systematic reviews of test-retest reliability of standardized instruments to measure language development up to age three years that are available at least in English. Although reliability in these ones may also partially reflect consistency in the adults' reports, they are at least based on a one-shot observation of how the child behaves, rather than the adult's cumulative experience with the child. Note that some promising tests are currently in development or have only recently been released, including the NIH Baby Toolbox (REF) and GSED (REF), and thus we cannot include them in the present summary.

The Ages and Stages Questionnaire (ASQ) is a screening tool to measure infants and children's development based on observation of age-specific behaviors: For instance, at 6 months, in the domain of communication, one of the items asks whether the child smiles. The ASQ's reliability has been the object of a systematic review of independent evidence (Velikonja et al., 2017). Across 10 articles with data from children in the USA, Canada, China, Brazil, India, Chile, the Netherlands, Korea, and Turkey, only three articles reported test-retest correlations, two in the USA (r=.84-1) and one in Turkey (r=.67). However, the meta-analysis authors judged these three studies to be "poor" in quality. Moreover, the ASQ is a questionnaire that an observer fills in, but it can also be administered as a parental questionnaire, reducing its comparability with our purely observational method. For the other tests, reliability is mainly available from reports by the companies commercializing the test. The Goldman Fristoe Articulation Test – 3rd edition (GFTA-3) (REF) focuses on the ability to articulate certain sounds through picture-based elicitation and can be used from two to five years of age. Available in English and Spanish for a USA context, it has a reported reliability of r=.92 – although we do not know whether it is this high for two-year-olds specifically. The Preschool Language Scales – 5th edition can be used to measure both comprehension and production from birth to 7 years of age. According to Pearson's report, its test-retest reliability is .86 to .95, depending on the age bracket (0;0-2;11, 3;0-4;11, and 5;0-7;11, 0-7 years). The test has also been adapted to other languages, with good reported test-retest reliability (r=.93; Turkish, Sahli & Belgin, 2017). One issue we see with both of these reports is that children tested varied greatly in age, and the correlation seems to have been calculated based on the raw (rather than the normed) score. As a result, children's ranking may have been stable over test and retest mainly because they varied greatly in age. The Expressive Vocabulary Test – 2nd Edition (EVT-2) is a picture-based elicitation method that can be used with children from 2.5 years of age. The company Springer reports a test-retest reliability of r=.95 by age (REF).

Many other standardized tests exist for children over three years of age. Given that most children included in the present study were under three, these other tests do not constitute an ideal comparison point. Nonetheless, in the spirit of informativeness, it is useful to consider that a systematic review and meta-analysis has been done looking at all psychometric properties (internal consistency, reliability, measurement error, content and structural validity, convergent and discriminant validity) for standardized assessments targeted at children aged 4-12 years (REF). Out of 76 assessments found in the literature, only 15 could be evaluated for their psychometric properties, and 14 reported on reliability based on independent evidence (i.e., the researchers tested and retested children, rather than relying on the company's report of reliability). Among these, correlations for test-retest reliability averaged r=.67, with a range from .35 to .76. The authors concluded that psychometric quality was limited for all assessments, but based on the available evidence, PLS-5 (whose test-retest reliability was r=.69) was among those recommended for use.

Third, and perhaps most relevant, we looked for references that evaluated the psychometric properties of measures extracted from wearable data. We found no previous work attempting to do so on the basis of completely ecological, unconstrained data like ours. The closest references we could find reported on reliability and/or validity of measurements from wearable data collected in constrained situations, such as having 4.5 year old children wear interior sensors and asking them to complete four tests of balance (e.g., standing with their eyes closed; Liu et al., 2022). It is likely that consistency and test-retest reliability are higher in such cases than in data like ours, making it hard to compare. Nonetheless, to give an idea, a recent meta-analysis of wearable inertial sensors in healthy adults found correlations between these instruments and gold standards above r= .88 for one set of measures (based on means) but much lower for another (based on variability, max weighted mean effect r = .58). Regarding test-retest reliability, the meta-analysts report ICCs above .6 for all measures for which they could find multiple studies reporting them. However, those authors point out that the majority of the included studies were classified as low quality, according to a standardized quality assessment for that work.

SM C: Code to reproduce Table 2

SM D: Code to reproduce Fig. 2

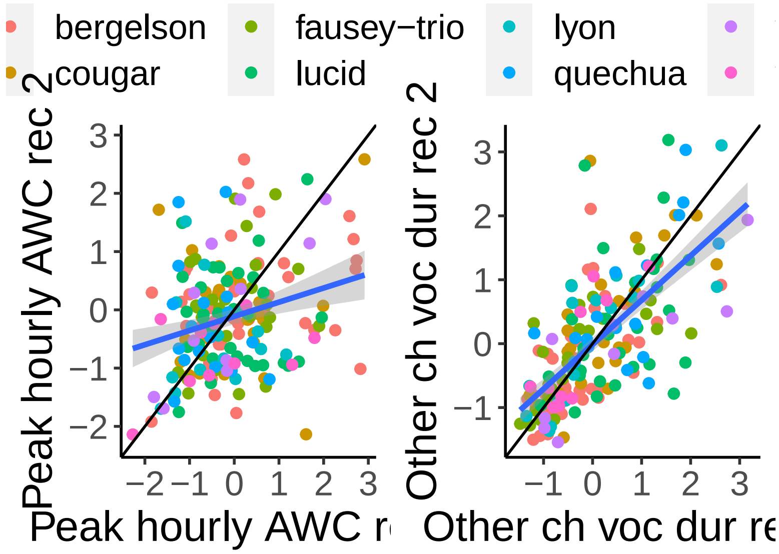


Figure 1: (A) scatterplot for one variable with relatively low ICCs versus (B) one with relatively higher ICCs (see Tables 1-2 for details)

SM E: Code to reproduce text at the beginning of the "Setting the stage" section

```
##           aclew             lena
## AWC      ".55 [.47,.63]" ".52 [.4,.64]"
## CVC      ".8 [.76,.84]"  ".7 [.64,.76]"
## CTC      ".76 [.72,.8]"  ".69 [.61,.77]"
## Chi vocs ".69 [.63,.75]" ".6 [.52,.68]"
```

Out of our 8 corpora and 191 children, 148 children (belonging to 7 corpora) could be included in this analysis, as some children did not have recordings less than two months apart (in particular, no child from the Warlaumont corpus did).

SM F: Exploration: Is lower Child ICC than correlations due to the fact that we are controlling for age?

May it be that correlation coefficients are higher than Child ICC because we control for age in the latter but not the former? To assess this, we take a metric that shows a large difference across the correlation and the Child ICC analysis, ACLEW's CVC per hour. We then refit our mixed model, but this time we do not declare age nor corpus, so that the situation is more similar to the simple correlations. When we do this, Child ICC goes up from 0.42 to 0.67.

This is still lower than the correlation observed for this same variable, 0.8. We believe this is due to another difference between the two analyses, namely that here we are including all recordings, whereas the other analysis is focused on recordings less than 2 months away.

SM G: Code to reproduce Figure 3

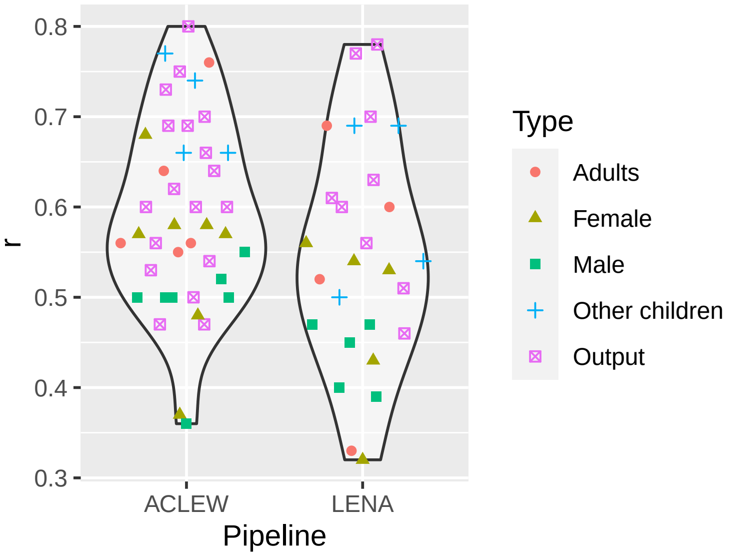


Figure 2: Distribution of correlation coefficients.

SM H: Code to reproduce text under Figure 3

To see whether correlations in this analysis differed by talker types and pipelines, we fit a linear model with the formula $lm(cor\ type * pipeline)$, where type indicates whether the measure pertained to the key child, (female/male) adults, other children; and pipeline LENA or ACLEW. We found an adjusted R-squared of 31%, suggesting this model did not explain a great deal of variance in correlation coefficients. A Type 3 ANOVA on this model revealed a significant effect of pipeline (F = 3.55, p = 0.06), due to higher correlations for ACLEW (M = 0.59, SD = 0.1) than for LENA metrics (m = M = 0.55, SD = 0.12). See below for fuller results.

	Sum Sq	Df	F value	Pr(>F)
Type	0.28	4	7.92	0.00
data_set	0.03	1	3.55	0.06
Type:data_set	0.03	4	0.73	0.57
Residuals	0.52	58	NA	NA

SM I: Code to reproduce text at the beginning of the "Overall reliability" section

Out of the 68 fitted models, 64 could be fit with the full model, yielding a measure of Corpus ICC. Of these, 97% had Corpus ICCs smaller than .2, consistent with the idea that LENA and ACLEW metrics are robust to corpus differences. For the 4 for which the full model was singular, we fit the data with the No Corpus model, and none was singular then, allowing us to have Child ICC for all 68 metrics.

Figure 3 shows the distribution of Child ICC across all 69 metrics, separately for each pipeline. The majority of measures had Child ICCs between .3 and .5. 7 measures had Child ICCs higher or equal to .5. Surprisingly, the top 6 metrics in terms of Child ICC corresponded to the "other child" category, known to have the worst accuracy according to previous analyses (Cristia et al., 2020). In an analysis fully reported in the SM, we find some evidence that this may be due to the presence versus absence of siblings. The next metric with the highest Child ICC corresponded to an output measure, namely the total vocalization duration per hour extracted from ACLEW annotations (voc_dur_chi_ph, aclew), with a Child ICC of 0.5. Among adult metrics, the average vocalization duration for female vocalizations for ACLEW (avg_voc_dur_fem, aclew) and the ACLEW equivalent of CTC had the highest Child ICC (0.45 and 0.46, respectively).

SM J: Exploration: Are high Child ICCs for "other child" measures due to number or presence of siblings?

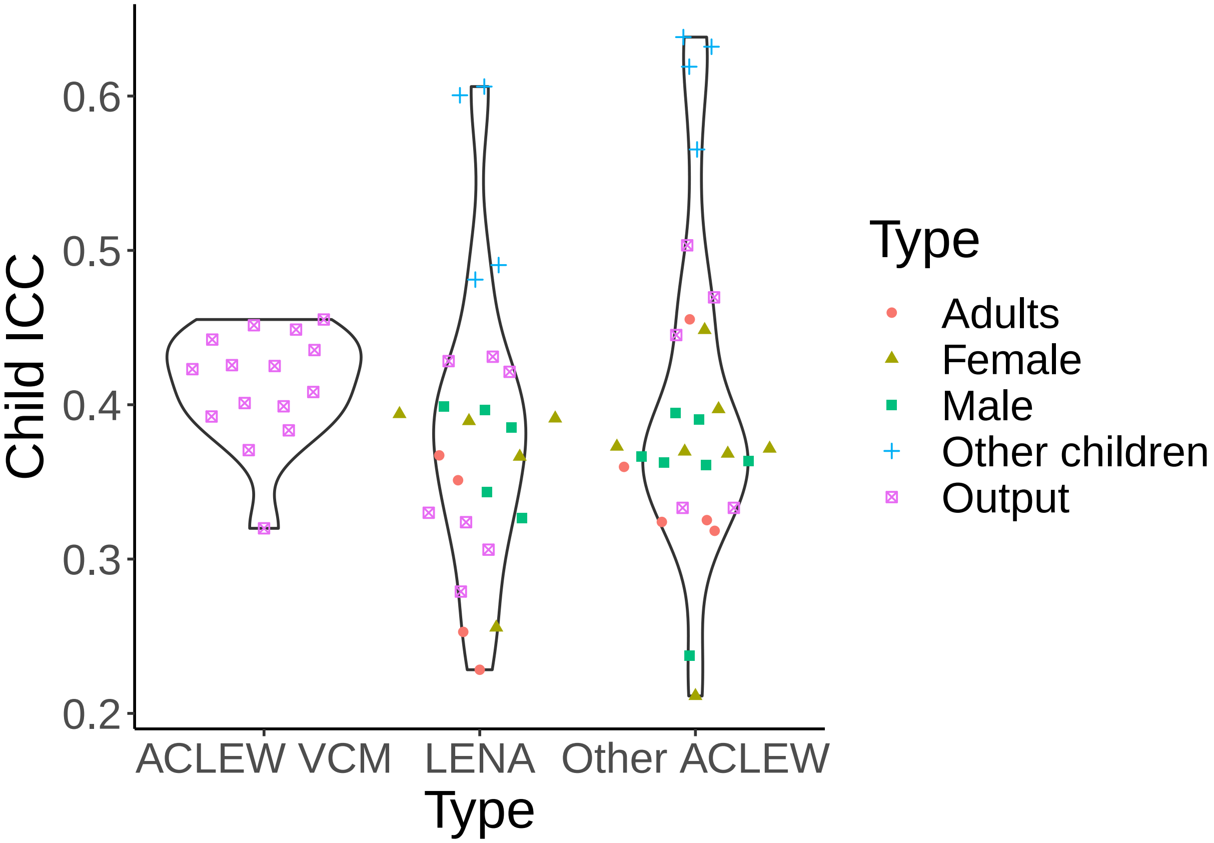
We reasoned the high Child ICC for metrics related to other children may be because children in our corpora vary in terms of the number of siblings they have, that siblings' presence may be stable across recordings, and that a greater number of siblings would lead to more other child vocalizations. As a result, any measure based on other child vocalizations would result in stable relative ranking of children due to the number of siblings present. To test this hypothesis, we selected the metric with the highest Child ICC, namely ACLEW's total vocalization duration by other children. We fit the full model again to predict this metric, but this time, in addition to controlling for age, we included sibling number as a fixed effect $lmer(metric\ age + sibling_{n}umber + (1|corpus/child))$, so that individual variation that was actually due to sibling number was captured by that fixed effect instead of the random effect for child. We had sibling number data for 874 recordings from 120 in 5 corpora (bergelson, lucid, quechua, warlaumont, winnipeg). The number of siblings varied from 0 to 7, with a mean of 0.9 and a median of 1. Results indicated the full model was singular, so we fitted a No Corpus model to be able to extract a Child ICC. As a sanity check, we verified that the number of siblings predicted the outcome, total vocalization duration by other children – and found that it did: ß = 0.2, t = 4.73, p < .001. This effect is relatively small: It means that per additional sibling, there is a .2 standard deviation increase in this variable. Turning now to how much variance is allocated to the random factor of Child, there was no difference in Child ICC in our original analysis (0.64) versus this re-analysis including the number of siblings (0.64).

Perhaps it is not so much the sheer number of siblings that explains variance, but the sheer presence versus absence. After all, we can imagine that the effect of the number of siblings is not monotonic. We therefore repeated the analysis above but rather than adding the actual number of siblings, we had a binary variable that was "present" if the child had any siblings, and "absent" otherwise.

As in the sibling number analysis, the full model was singular, so we fitted a No Corpus model to be able to extract a Child ICC. We again verified that sibling presence predicted the outcome, total vocalization duration by other children – and found that it did: ß = 0.97, t = 9.91, p < .001. This effect is, as expected, sizable: It means that there is nearly one whole standard deviation increase in this variable when there are any siblings. In addition to being a better predictor, in this model, the amount of variance allocated to individual children as measured by Child ICC was considerably higher in our original analysis (0.64) than in this re-analysis including sibling presence (0.52).

SM K: Exploration: are "bad" output measures those coming from VCM?

Among ACLEW measures, a fair number of them come from VCM, a module that classifies child vocalizations in terms of vocal maturity types into cry, canonical, and non-canonical categories. In unpublished analyses, we have found that VCM labels are inaccurate when compared to human labels of the same vocalizations, relatively to other metrics. In this analysis, we checked whether VCM-derived measures had lower Child ICC than other ACLEW measures. As shown in the next Figure, this was not the case: Some output measures from the ACLEW pipeline have lower Child ICC than VCM ones.



SM L: Code to reproduce Figure 4

SM M: Code to reproduce text below Figure 4

Next, we explored how similar Child ICCs were across different talker types and pipelines. We fit a linear model with the formula $lm(icc\_child\_id\ type * pipeline)$, where type indicates whether the measure pertained to the key child, (female/male) adults, other children; and pipeline LENA or ACLEW. We found an adjusted R-squared of 61%, suggesting much of the variance across Child ICCs was explained by these factors. A Type 3 ANOVA on this model revealed type was a signficant predictor (F(4) = 26.7, p<.001), as was pipeline (F(1) = 4.4, p = 0.04); the interaction between type and pipeline was not significant. The main effect of type emerged because output metrics tended to have higher Child ICC (M = .4, SD = .06) than those associated to adults in general (M = .33, SD = .07), females (M = .36, SD = .06), and males (M = .36, SD = .04); whereas those associated with other children had even higher Child ICCs (M = .58, SD = .06). The main

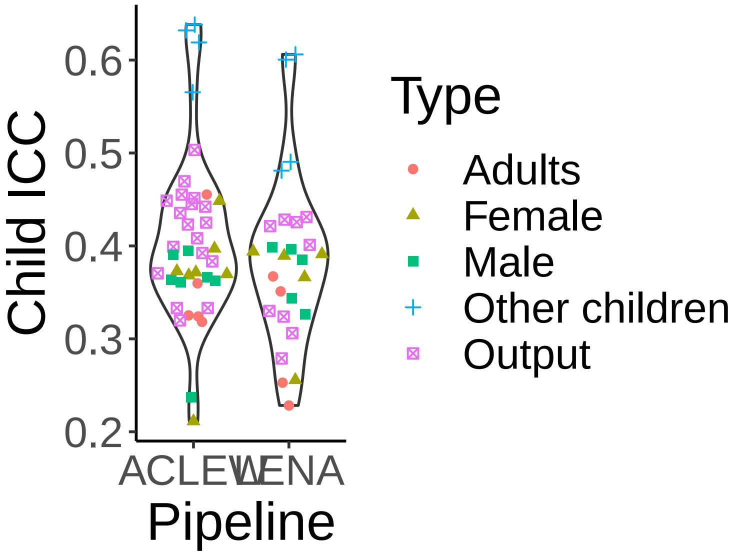


Figure 3: Distribution of ICC attributed to corpus (a) and children (b), when combining data from all corpora.

Table 1: Most commonly used metrics.

metric	LENA ICC	ACLEW ICC
wc_adu_ph	0.37	0.33
lena_CVC_ph	0.43	0.42
lena_CTC_ph	0.35	0.46
voc_fem_ph	0.39	0.37
voc_chi_ph	0.42	0.45

effect of pipeline arose because of slightly higher Child ICCs for the ACLEW metrics (M = .41, SD = .09) than for LENA metrics (M = .38, SD = .09).

SM N: Code to reproduce Table 4

SM O: Code to reproduce text at the beginning of the "Reliability across age groups" section

Out of 408 fitted models (68 metrics times 6 age bins), 4 were singular when including a random intercept per child, and therefore they could not be included in these analyses at all. In addition, 126 were singular when including a random intercept per corpus. The remaining 278 could be analyzed with the full model.

SM P: Code to reproduce Figure 5

SM Q: Code to reproduce text below Figure 5

As we did in the previous section for corpus, we checked whether Child ICC differed by talker types and pipelines across age bins by fitting a linear model with the formula $lm(Child_ICC\ type * pipeline * age_bin)$. We found an adjusted R-squared of 41%, suggesting this model explained about a third of the variance in Child ICC. A Type 3 ANOVA on this model revealed type was a signficant predictor (F(4) = 26.7, p<.001),

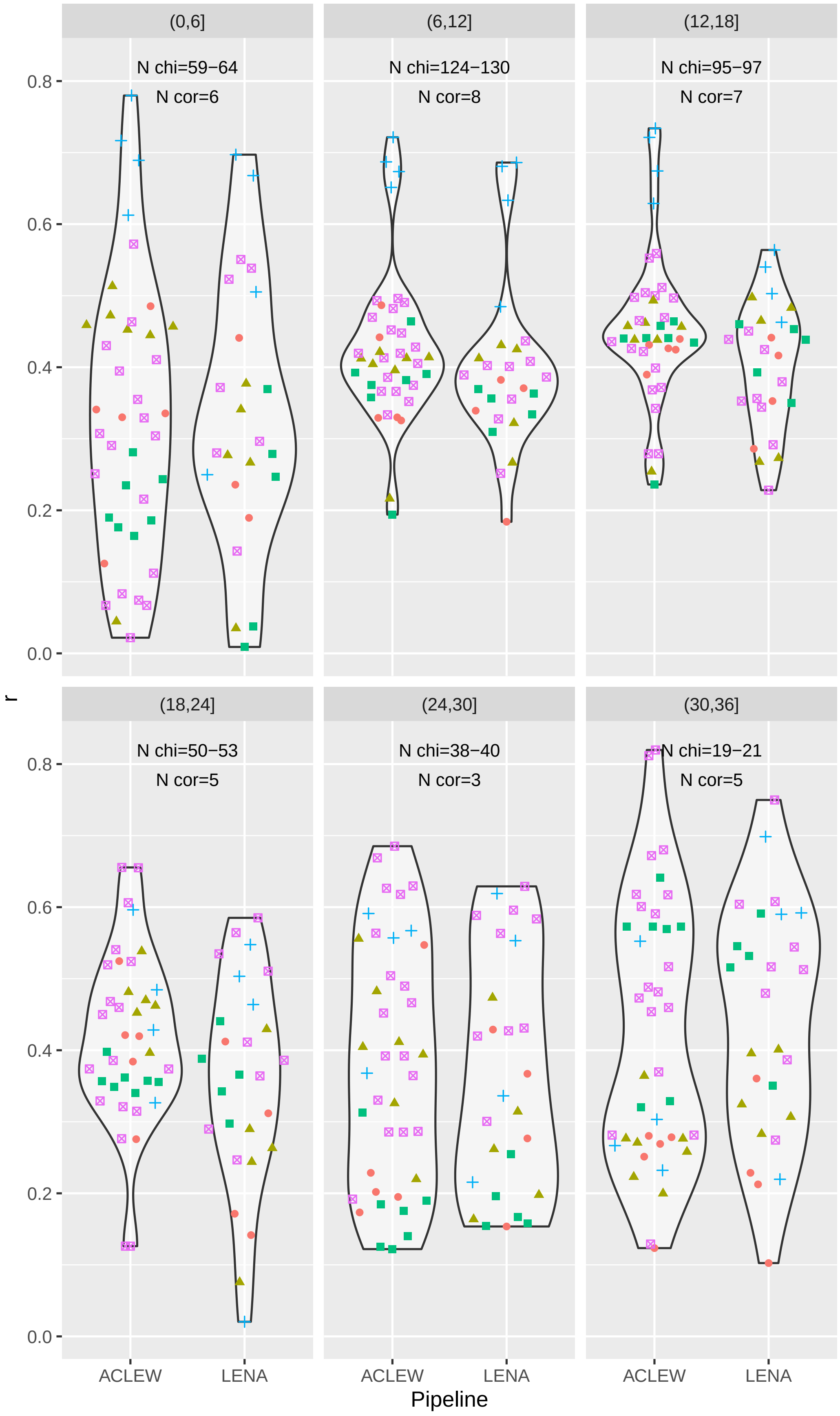


Figure 4: Distribution of ICC attributed to corpus (a) and children (b), when binning children's age.

whereas as was pipeline (F(1) = 4.4, p = 0.04); the interaction between type and pipeline was not significant. See below for more information.

	Sum Sq	Df	F value	Pr(>F)
Type	1.66	4	30.31	0.00
data_set	0.06	1	4.51	0.03
age_bin	0.53	5	7.73	0.00
Type:data_set	0.10	4	1.92	0.11
Type:age_bin	1.83	20	6.69	0.00
data_set:age_bin	0.10	5	1.53	0.18
Type:data_set:age_bin	0.40	20	1.45	0.10
Residuals	4.71	344	NA	NA

SM R: Code to reproduce Figure 6

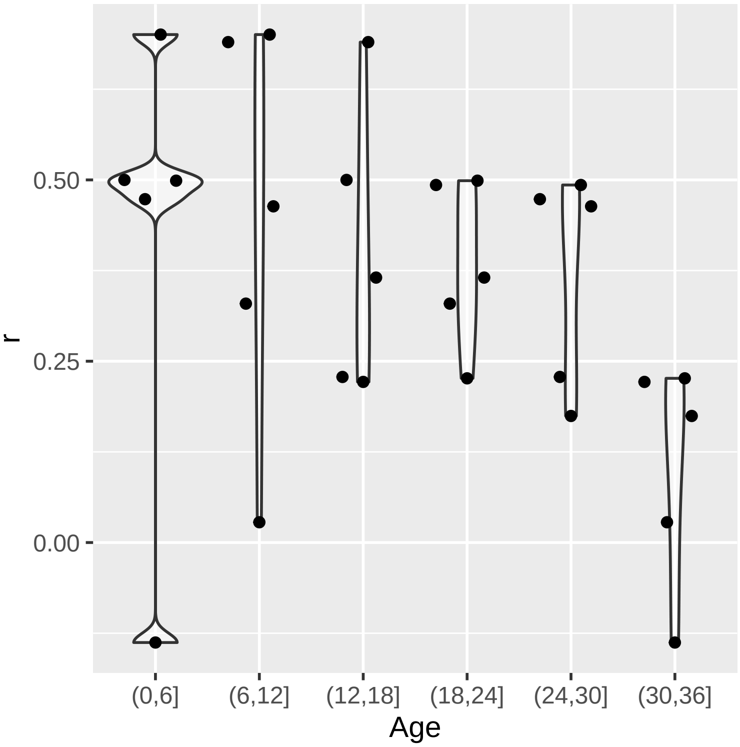


Figure 5: Correlations in Child ICC across age bins. Each point indicates the correlation in Child ICC for the age bin named in the x-axis with every other age bin.

SM S: Code to reproduce text at the beginning of the “Reliability within corpus” section

Figure 7 addresses this question, showing the distribution of ICC across our 68 metrics in each of the 8 included corpora. Out of 544 fitted models (68 metrics times 8 corpora), 21 were singular when including a random intercept per child, and therefore they could not be included in these analyses at all. (Including a random intercept per corpus is not relevant here, since only data from one corpus is included in each model fit.)

SM T: Code to reproduce Figure 7

SM U: Code to reproduce text below Figure 7

The fact that we cannot infer reliability from one corpus based on another one was confirmed statistically: We checked whether Child ICC differed by talker types and pipelines across corpora by fitting a linear model with the formula $lm(Child_ICC\ type * pipeline * corpus)$, where type indicates whether the measure pertained to the key child, (female/male) adults, other children; pipeline LENA or ACLEW; and corpus the corpus ID. We found an adjusted R-squared of 46%, suggesting this model explained nearly half of the variance in Child ICC. A Type 3 ANOVA on this model revealed several significant effects and interactions, including a three-way interaction of type, pipeline, and corpus (F(28) = 2.1, p<.001); a two-way interaction of type and corpus (F(7) = 3.6, p<.001); and a main effect of corpus (F(7) = 15.5, p<.001). See below for more information.

	Sum Sq	Df	F value	Pr(>F)
Type	0.59	4	8.93	0.00
data_set	0.10	1	6.02	0.01
corpus	1.81	7	15.52	0.00
Type:data_set	0.05	4	0.74	0.56
Type:corpus	4.76	28	10.22	0.00
data_set:corpus	0.42	7	3.61	0.00
Type:data_set:corpus	0.97	28	2.09	0.00
Residuals	7.36	443	NA	NA

SM V: Code to reproduce Figure 8

Save information about packages used

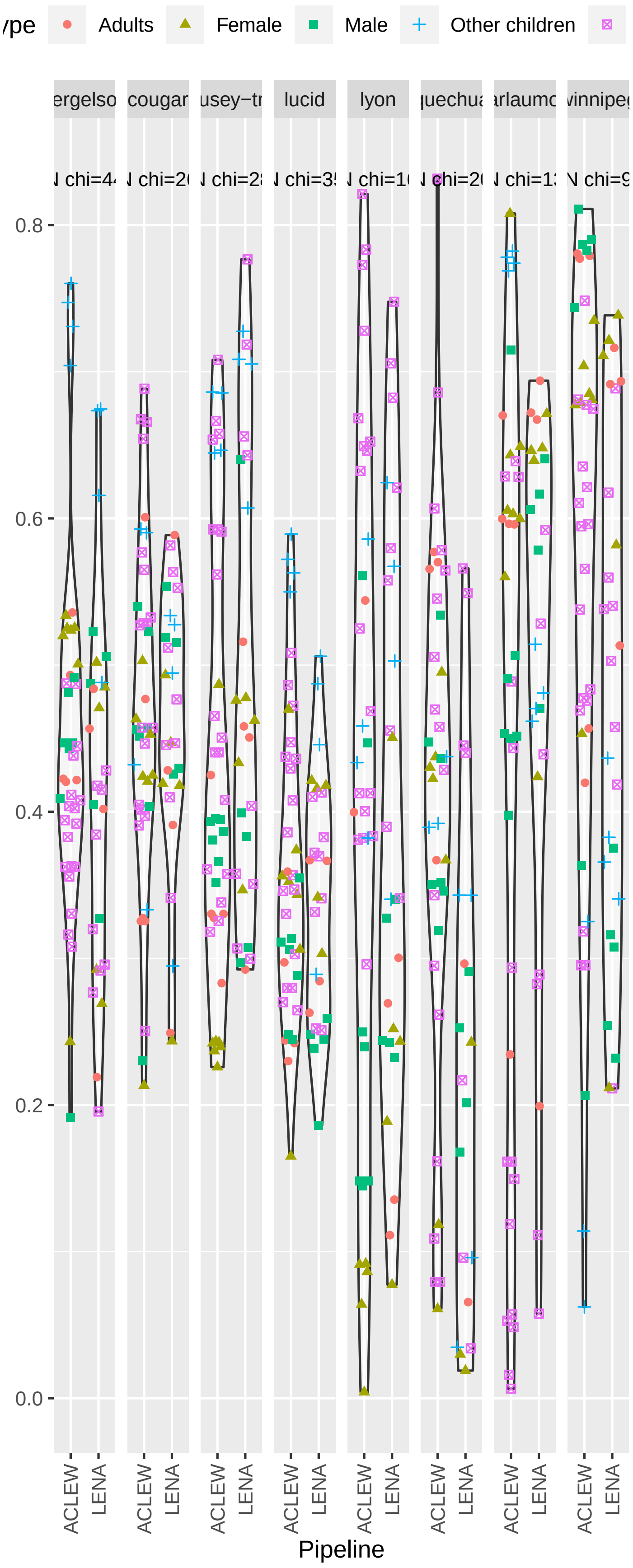


Figure 6: Child ICC by metric type and pipeline, when considering each corpus separately.

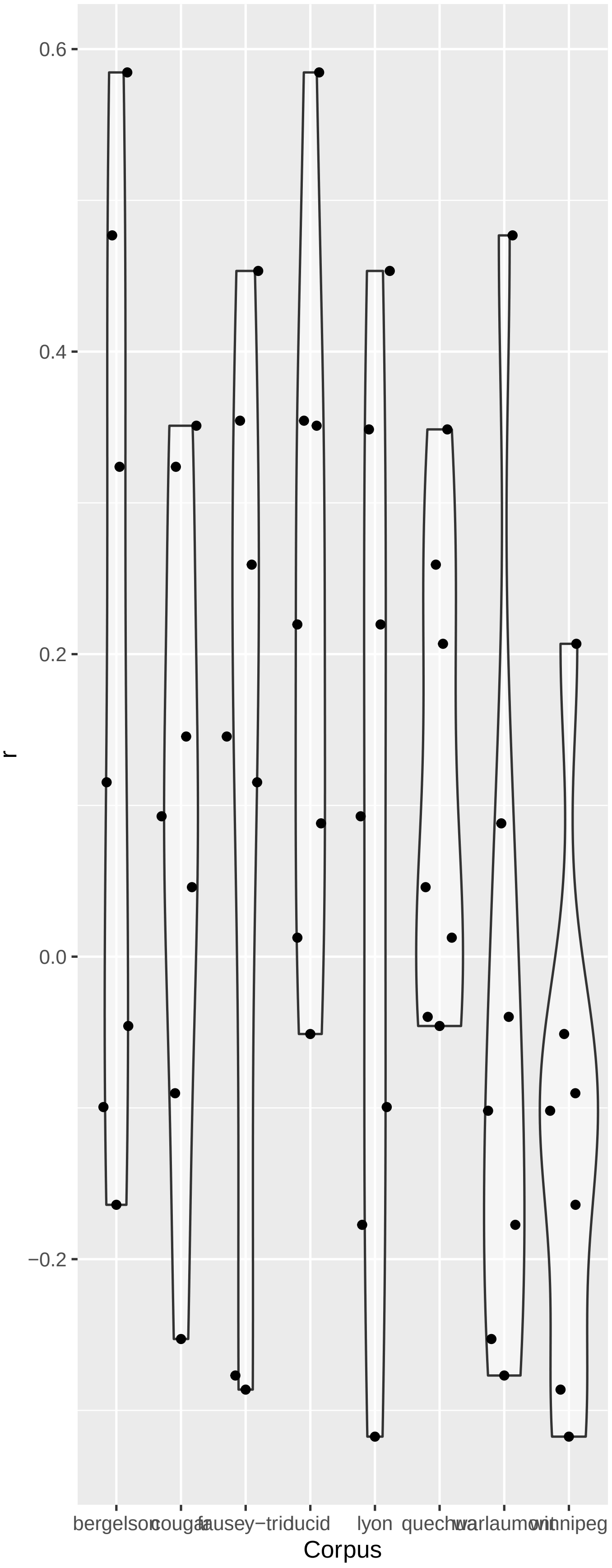


Figure 7: Correlations in Child ICC across corpora. Each point indicates the correlation in Child ICC for the corpus named in the x-axis with every other corpus.