CHILD DEVELOPMENT

Child Development, March/April 2021, Volume 92, Number 2, Pages 672–681

Comparing Automatic and Manual Measures of Parent–Infant Conversational Turns: A Word of Caution

Naja Ferjan Ramírez, Daniel S. Hippe, and Patricia K. Kuhl
University of Washington

The Language ENvironment Analysis system (LENA) records children's language environment and provides an automatic estimate of adult–child conversational turn count (CTC). The present study compares LENA's CTC estimate to manually coded CTC on a sample of 70 English-speaking infants recorded longitudinally at 6, 10, 14, 18, and 24 months of age. At each age, LENA's CTC was significantly higher than manually coded CTC (all *ps* < .001, Cohen's *ds*: 0.9–2.05), with the largest discrepancies between the two methods observed at younger ages. The Limits of Agreement Analyses confirm wide disagreements between the two methods, highlighting potential problems with automatic measurement of parent–infant verbal interaction. These findings suggest that future studies should validate LENA's CTC estimates with manual coding.

Decades of research demonstrate significant and strong associations between children's early language environments and subsequent language and cognitive outcomes (Hart & Risley, 1995; Hoff, 2003; Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991; Rowe, 2012; Tamis-LeMonda, Bornstein, Kahana-Kalman, Baumwell, & Cyphers, 1998). In recent years, the cornerstone method employed across many studies that measure children's language input is the Language Environment ANalysis (LENA™), which facilitates day-long audio recordings in children's natural environments. The system provides a light-weight recorder that can store up to 16 hr of sound and can be snapped into a chest pocket of children's clothing to record everything that the child produces and hears. The LENA audio recording is analyzed by the LENA software, which uses proprietary algorithms to provide an estimate of three primary measures: the number of adult words heard by the child (Adult Word Count, AWC), the number of child's language-related vocalizations (Child Vocalization Count, CVC), and the number of adult–child back and forth exchanges (conversational turn count, CTC).

Over the last decade, LENA has been used across many languages to study the links between the language input and language outcomes, to characterize specific language environments, to study clinical populations, and to provide linguistic feedback to caregivers (see the list of over 100 publications at www.lena.org/research/). Of the three primary LENA measures, CTC has recently received the most attention, as it is interpreted as a proxy for high-quality "serve and return" caregiver–child interactions, child engagement, and adult responsiveness, and thus a key component of high-quality language environments (Christakis, Lowry, Goldberg, Violette, & Garrison, 2019; Gilkerson et al., 2017, 2018; Perry et al., 2018; Zimmerman et al., 2009; see also http://lena.org/conversational-turns). In the present study, we take a closer look at this measure.

Theoretical Perspectives on Turn-Taking

Language, social-emotional, and cognitive development in infancy are highly influenced by turn-

The authors thank Denise Padden and Neva Corrigan for valuable assistance throughout data collection. They also thank Julia Mizrahi, Bo Woo, Nicole Dunn, Spencer Sherwin, Amy Busch, Laura Purdy, Sarah Hillegass, A'Dawnah Pangelinan, Annique Nyman, Margaux Bouchegnies, Kyli McGillivrey, and Ella Deaton for assistance in participant recruitment, data coding, and manuscript preparation. The research described here was supported by the Overdeck Family Foundation, the University of Washington Institute for Learning & Brain Sciences Ready Mind Project, and the University of Washington's Language Acquisition and Multilingualism Endowment. The authors have no conflict of interest.

Correspondence concerning this article should be addressed to Naja Ferjan Ramírez, Department of Linguistics, University of Washington, Guggenheim Hall 415B, Box 352425, 3940 Benton Ln NE, Seattle, WA 98195-2425. email: naja@uw.edu.

0009-3920/2021/9202-0016
DOI: 10.1111/cdev.13495

taking between caregivers and children (Golinkoff, Hoff, Rowe, Tamis-LeMonda, & Hirsh-Pasek, 2019; Leech & Rowe, 2020; Meltzoff & Brooks, 2009; Tamis-LeMonda, Kuchirko, & Song, 2014), and some theories propose that this mechanism may have been essential for the evolution of language (Levinson, 2016). As such, it is important to consider which characteristics of turn-taking may be critical for its contribution to infants' language learning. Unlike overheard speech, or speech from an electronic source (Kuhl, Tsao, & Liu, 2003; Shneidman, Arroyo, Levine, & Goldin-Meadow, 2013; Weisleder & Fernald, 2013), turn-taking allows caregivers to provide contingent feedback that is constantly adjusted to their infant's linguistic needs, and infants, in turn, adjust their vocalizations in response, thereby creating a feedback loop (Braarud & Stormark, 2008; Goldstein & Schwade, 2008; Smith & Trainor, 2008; Warlaumont, Richards, Gilkerson, & Oller, 2014). Turn-taking provides opportunities for temporal contiguity and contingency, fluency, connectedness, and joint engagement between parents and children, which are critical in word learning and predict children's subsequent language skills (Bornstein, Tamis-LeMonda, & Haynes, 1999; Conboy, Brooks, Meltzoff, & Kuhl, 2015; Hirsh-Pasek et al., 2015; Tamis-LeMonda et al., 2014). Finally, recent brain studies propose that, through contiguity, contingency, connectedness, and social feedback, turn-taking shapes the social circuitry of the language-related brain areas (Merz, Maskus, Melvin, He, & Noble, 2019; Romeo, Leonard, et al., 2018; Romeo, Segaran, et al., 2018). Taken together, current theories of language acquisition propose that early language learning crucially relies on parent–child interaction, social processes, and the social neural circuitry (Hoff, 2006; Kuhl, 2007; Tamis-LeMonda et al., 2014). Turn-taking is hypothesized to play a key role in this process because it provides opportunity for continued social engagement, interaction, and feedback.

LENA's Automatic Assessment of Turn-Taking

Considering these theoretical perspectives, along with the ubiquitous use of LENA's automatic CTC as a measure of turn-taking, it is important to consider two questions: First, *how* does LENA estimate CTC (i.e., what does it actually measure)? Second, to what degree does LENA's CTC estimate agree with human understanding and human counts of turn-taking?

With regard to the first question, LENA's algorithms look for instances of alterations between adult and child speech in close temporal proximity. Importantly, the algorithms do *not* differentiate between child-directed and overheard speech, which means that an unknown proportion of CTCs are identified in error, such as when a parent is talking on the phone and the infant is babbling to herself nearby (i.e., accidental contiguity).

With regard to the second question, LENA's sensitivity for automated speaker segmentation ranges between 68% and 82% (Oller et al., 2010; Xu, Richards, & Gilkerson, 2014; Zimmerman et al., 2009) suggesting that human transcription and LENA agree fairly well when labeling speaker and sound types. However, as noted by a recent review study (Cristia, Bulgarelli, & Bergelson, 2020), the majority of LENA validation studies have not been peer-reviewed, or have failed to report key methodology and results, suggesting that further research is needed. This is especially true for CTC, which is a derived measure with at least two distinct sources of error: inaccurate speaker labels and accidental contiguity.

To our knowledge, only four published peer-reviewed studies have considered the relation between LENA's CTC estimates and human CT counts. These studies considered CTs in different languages, and in children of varying ages, only on occasion matching those on which the LENA algorithms were trained (English-speaking infants between 2 and 48 months of age; see Oller et al., 2010). One study reports a significant correlation after removing three outliers (Gilkerson et al., 2015), one reports no correlation unless five samples that contained lots of overlapping speech and crying are removed (Pae et al., 2016), one reports a moderate correlation (Busch, Sangen, Vanpoucke, & van Wieringen, 2018), and one reports a strong correlation (Ganek & Eriks-Brophy, 2018). Only the Busch et al. study assessed the *agreement*, as opposed to simple correlations, which can mask systematic biases (see Method). The inconsistency of these findings warrants further investigation.

The Present Study

The present study is exploratory in nature. It was not preregistered, and we did not have strong predictions at the time of data collection. Our curiosity about the relation between LENA's CTC estimates and human CT counts was sparked while we were conducting a parent coaching study (Ferjan Ramírez, Lytle, & Kuhl, 2020), which required careful listening to LENA audio snippets in order to identify intervention behaviors (among others, conversational turns). During this process, it became apparent that LENA's automatic identification of CTs was often inconsistent with what we considered turn-taking between caregivers and children. Particularly problematic were scenarios where infants babbled to themselves (a ubiquitous and typical behavior in the 6–24 month age range), and parents were nearby and talking to one another or on the phone, but not to the child. Because such examples were ubiquitous in our data, we decided to further pursue this question within our data set. The present study considers correlation and agreement between LENA's automatic CTC estimate and manually coded CTC, on a sample of 70 English-speaking infants recorded longitudinally at five time points between 6 and 24 months of age. A portion of each recording was manually coded to identify CTs, allowing for a side-by-side comparison with LENA's CTC estimates. Critically, cases where the child and the adult vocalized in close temporal proximity, but there was no vocal interaction between them, were not counted as CTs by the human coders. Because LENA includes such cases in its estimate of CTC, we hypothesize that the automatic CTC estimates will be significantly higher than the manual counts. Furthermore, based on the recent findings that the proportion of child-directed speech increases with age (Ferjan Ramírez, Lytle, Fish, & Kuhl, 2018), and that the proportion of overheard speech in infants' input declines with age (Bergelson et al., 2019), we also hypothesize that correlation and agreement between the two methods will increase with infant age.

Method

Participants and Data Collection

Seventy-nine families from the Seattle metro area were recruited through the University of Washington Subjects Pool. All participating infants were full-term (within ±14 days of due date), of normal birth weight (6–10 lbs.), and without any birth or postnatal complications. Of the 79 families, 70 (35 with girls) completed the recordings at all time points (6, 10, 14, 18, and 24 months) and are included in the present data set. Sixty-four were White, four were of mixed race, and two were of unknown race. Socioeconomic status was measured with the Hollingshead Index (Hollingshead, 1975) and ranged between 30 and 66 ($M = 49.6$, $SD = 10.2$). Experimental procedures were approved by the Institute Review Board of the University of Washington, and informed consent was obtained from parents. The study conforms to the U.S. Federal Policy for the Protection of Human Subjects. The data were collected between October 1, 2016 and August 5, 2018.

At each data collection time point, families received two LENA recorders and were instructed to use one recorder on each day of a typical weekend. The average duration of the recordings was 12 hr and 50 min (for more information, see Supporting Information).

Selection of Audio Segments for Analysis and Data Coding Protocol

Following previously described procedures (see Ferjan Ramírez et al., 2020; Ramírez-Esparza García-Sierra, & Kuhl, 2014, 2017a, 2017b), the daylong audio files were processed using the LENA Advanced Data Extractor Tool to automatically identity for manual analyses of CTs. Each participant's two daily recordings were segmented into 30-s intervals. For each of the two recording days, 50 intervals with the highest AWC that were at least 3 min apart were automatically selected, for a total of one hundred 30-s coding intervals per participant per age, yielding a total of 17,500 min of audio data for the study. Four research assistants listened to each audio interval and noted the total number of CTs per interval in a separate spreadsheet, using the previously described training and reliability assessment procedures (see Supporting Information for details). As with the LENA algorithm, CTs were counted in discrete pairs, and pauses of 5s or more constituted the end of a conversation. Intercoder reliability was 0.98.

Statistical Analyses

All statistical analyses were conducted on the total counts obtained from one hundred 30-s intervals per participant per age. Overall agreement between automatic and manual CTC estimates was summarized at each age using Pearson's correlation coefficient (r) and the intraclass correlation coefficient (ICC). Pearson's r indicates the scatter of values around the line of best fit and quantifies random error, but not the systematic biases that may exist between two different measurements. By contrast, ICC considers the absolute agreement between the two methods, and is sensitive to systematic shifts or biases. The ICC ranges from 0 (no agreement) to 1 (perfect agreement). The presence of a systematic shift in one of the measurements, all else being equal, would decrease the ICC but not affect Pearson's *r*.

Agreement was further analyzed using the techniques of Bland and Altman (Bland & Altman, 1986, 1999). Bland–Altman analysis helps characterize differences between two methods (henceforth referred to as "errors") in multiple ways. Errors are decomposed into systematic biases and random errors in either direction around the bias (limits of agreement [LoA]). The bias in the automatic CTC was estimated as the mean difference between automatic and manual CTC and was tested against 0 using the paired *t*-test. The LoA was estimated as mean difference ±2 × standard deviation of differences. The LoA is an interval which is expected to contain 95% of differences that might be observed between automatic and manual CTCs. Bland-Altman plots were generated to display the differences (automatic minus manual CTC) versus the average of the two. This provides a way to visualize how the magnitudes of the errors vary across the range of CTC values (i.e., whether errors are smaller when CTCs are smaller but bigger when counts are bigger, or whether the amount of error is similar regardless of underlying CTC values). A random scattering of points that is centered around 0 on the *y*-axis (appearing "flat" across the range of CTC values) would indicate a lack of apparent bias. Differences were analyzed on the original scale (absolute differences) and on the log-scale (relative or percent differences; Bland & Altman, 1996). Throughout, two-tailed statistical tests were used, with statistical significance defined as $p < .05$.

Results

Table 1 summarizes the means and standard deviations for all three automatic LENA measures, and for the manually measured CTCs, in one hundred 30-s segments per participant at each age. In agreement with findings from previous studies considering 12-hr recordings (Gilkerson et al., 2017), AWC was not significantly correlated with infants' age (Spearman's $r = .03$, $p = .44$), while automatic CVC ($r = .59$, $p < .001$) and automatic CTC ($r = .57$, $p < .001$) were. Manual CTC was also significantly correlated with infant's age ($r = .81$, $p < .001$).

A comparison of the automatically and manually measured CTs is shown in Figure 1. Paired sample *t*-tests conducted separately for each age show that the mean number of counts was significantly higher by the automatic method than the manual method at each age (mean difference in one hundred 30-s segments: 80–136 counts across the age groups, $p < .001$ for each, Cohen's *d*s between 0.9 and 2.05; see also Table 3).

Table 2 presents the overall agreement between the automatically and manually measured CTs, as measured by both Pearson's *r* and the ICC. Considering the Pearson's *r*, the two measures of CTC had low correlations at 6 and 10 months ($r = .28$ and .18, respectively) but had higher correlations at 14–24 months ($r = .54$–.75, $p < .001$). However, the ICC, which is sensitive to systematic biases, demonstrates a low absolute agreement at both 6 and 10 months (ICC = .03–.04, $p > .23$) and 14–24 months (ICC = .24–.36, $p > .073$).

Table 3 shows the absolute and percent difference between the two methods. At 6 months, the average CTC estimates were 143 ± 57 and 25 ± 15 for the automatic and manual methods, respectively, corresponding to a mean absolute difference of 118 (95% CI [105, 131]) and a mean percent difference of 745% (95% CI [548, 941]). At 24 months, the CTC estimates were substantially larger than at 6 months, with averages of 343 ± 150 and 207 ± 66, respectively. Compared to the differences at 6 months, the mean absolute difference was similar at 136 (95% CI [105, 166]), while the mean percent difference was smaller but still substantial at 72% (95% CI [55, 89]). Over the ages of

Table 1

Summary of Counts in one hundred 30-s Segments Per Participant at Each Age

	6 Months (*N* = 70)		10 Months (*N* = 70)		14 Months (*N* = 70)		18 Months (*N* = 70)		24 Months (*N* = 70)	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Automatic AWC	9,445	4,395	8,646	3,865	7,894	3,213	8,946	3,804	9,583	4,134
Automatic CVC	271	119	275	99	279	127	487	238	710	322
Automatic CTC	143	57	148	52	156	69	260	126	343	150
Manual CTC	25	15	47	26	77	49	131	70	207	66

Note. AWC = Adult Word Count; CVC = Child Vocalization Count; CTC = Conversational Turn Count.

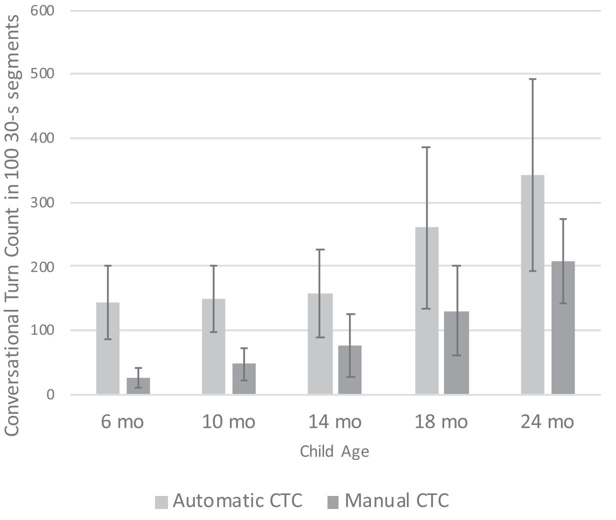

Figure 1. Mean number of conversational turns per child in one hundred 30-s segments per participant at 6, 10, 14, 18, and 24 months, as estimated automatically by Language ENvironment Analysis (lighter shade) and as counted manually by a human coder (darker shade). Error bars represent one standard deviation. CTC = conversational turn count.

Table 2
Correlation Between Automatic and Manual Conversational Turn Counts

Age	ICC	(95% CI)	*p*-value	*r*	(95% CI)	*p*-value
6 months	.03	(−.04, .12)	.24	.28	(.05, .48)	.018
10 months	.04	(−.04, .14)	.23	.18	(−.06, .39)	.14
14 months	.30	(−.10, .62)	.10	.60	(.43, .74)	< .001
18 months	.36	(−.10, .68)	.097	.75	(.63, .84)	< .001
24 months	.24	(−.07, .49)	.073	.54	(.34, .68)	< .001
All ages	.45	(−.09, .73)	.066	.76	(.71, .81)	< .001

Note. ICC = intraclass correlation coefficient; *r* = Pearson's correlation coefficient.

6–24 months, the mean absolute difference generally stayed within a narrow range (80–136) while the percentage difference decreased (745%–72%). This suggests that a relative fixed amount of extraneous counts is being recorded by LENA on average, no matter what the underlying true count is, small or large. However, as the manual CTC count naturally increases with age, this overestimation becomes less prominent as a percentage of the true CTC.

The LoA estimates in Table 3 provide ranges of likely differences between the two methods. At 6 months, the LoA was 114% to 2,611% while it was −24% to 231% at 24 months. These ranges are expected to include approximately 95% of the differences that could be observed between the methods if the experiment were repeated. The LoA estimates indicates that CTC will be nearly always overestimated by a substantial amount at 6 and 10 months (lower bounds are 114% and 38%, respectively). At older ages, it was possible for the automatic method to underestimate CTC, with LoA lower bounds of −10%, 4%, and −24%, respectively, while on average CTC was overestimated substantially. These patterns are also illustrated in Figure 2.

In a follow-up analysis, we asked whether the discrepancies between the two methods were affected by the household size. The number of siblings was grouped as 0 versus ≥1 (30 infants had ≥1 siblings) and the number of adults was grouped as <3 versus ≥3 (only 6 had >3 adults). We then used a multivariable model to assess differences in the ratio of automatic CTC: manual CTC by number of siblings, number of adults, and age in a multivariable model. CTC was overestimated more by LENA when there were ≥1 siblings in the household (Table 4; automatic CTC: manual CTC ratio; +31%; 95% CI [10, 54]; p = .002), adjusting for age and number of adults in the household. There was also a tendency, though not statistically significant, for greater overestimation of CTC when there were ≥3 adults in the household (Table 4; automatic CTC: manual CTC ratio; +19%; 95% CI [−4, 47]; p = .11), adjusting for age and siblings in the household.

We also considered, for each participant at each age, the number of 30-s segments in which no adult was directing speech to the infant, but LENA tagged a nonzero value for CTs (for details about coding of child-directed speech, see Supporting Information). In such segments, LENA's CTC constitutes "accidental contiguity," as we know from manual analyses that no adult is directing speech to the infant wearing the recorder. While the number of such segments was high at all ages (range 23–15 out of 100 segments, see Table S1), it decreased with infants' age (r = −.26, p < .001).

Discussion

The present study considered correlation and agreement between LENA's automatic CTC estimate and manual transcription of parent–infant turn-taking on a sample of 70 English-speaking families recorded at 6, 10, 14, 18, and 24 months. Confirming our hypothesis, the results show the LENA's CTC estimate was significantly above the manual

Table 3
Agreement Between Automatic and Manual Conversational Turn Counts

	Technique		Absolute difference				Percent difference			
Age	Automatic	Manual	*M*	(95% CI)	*p*-value[a]	LoA	*M*	(95% CI)	*p*-value[a]	LoA
6 months	143 ± 57	25 ± 15	118	(105, 131)	< .001	(8, 228)	745%	(548, 941)	< .001	(114, 2,611)
10 months	148 ± 52	47 ± 26	101	(88, 114)	< .001	(−7, 208)	336%	(236, 437)	< .001	(38, 1,578)
14 months	156 ± 69	77 ± 49	80	(66, 93)	< .001	(−31, 190)	211%	(125, 298)	< .001	(−10, 937)
18 months	260 ± 126	131 ± 70	129	(108, 150)	< .001	(−45, 303)	153%	(95, 211)	< .001	(4, 621)
24 months	343 ± 150	207 ± 66	136	(105, 166)	< .001	(−119, 390)	72%	(55, 89)	< .001	(−24, 231)
All ages	210 ± 126	97 ± 82	113	(101, 124)	< .001	(−53, 279)	303%	(250, 367)	< .001	(−7, 1,786)

Note. LoA = limits of agreement.
[a] Test comparing the mean difference between techniques to zero.

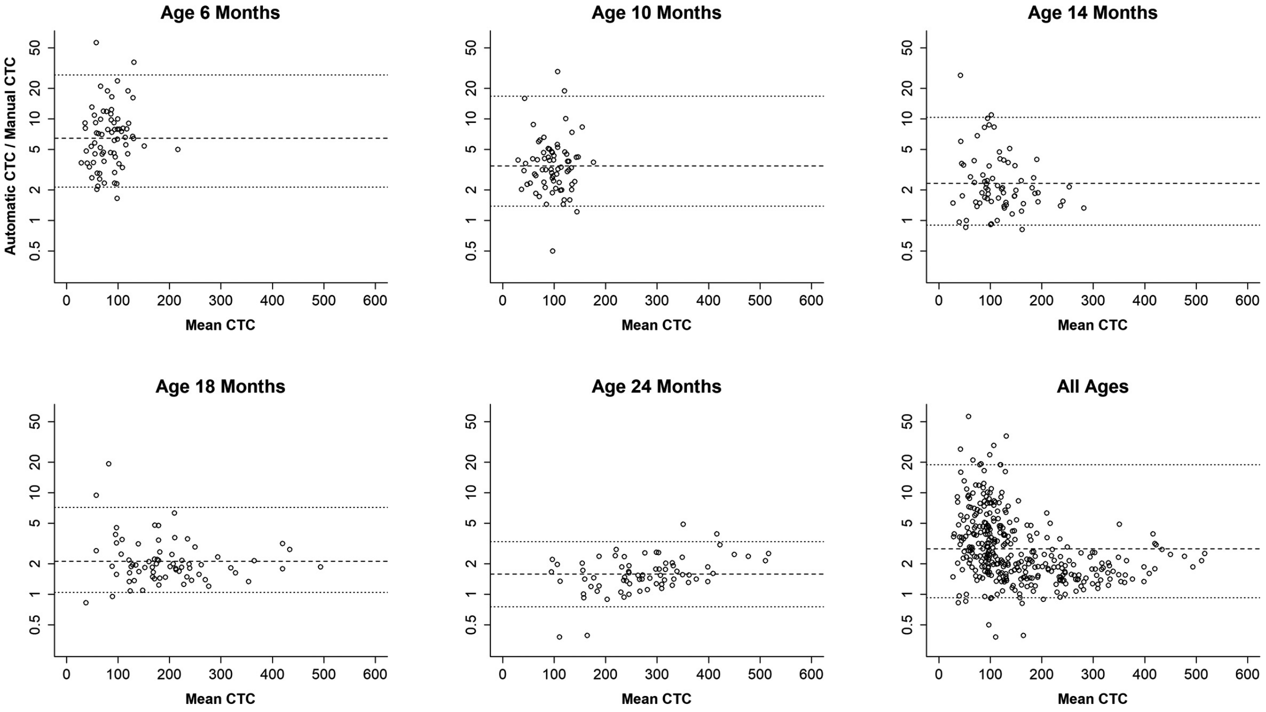

Figure 2. Bland-Altman plots comparing the conversational turn counts (CTC) estimated by Language ENvironment Analysis and as counted manually by a human coder. The mean difference of the automatic and manual turn counts was calculated after log-transforming the counts to reduce right skewness—that is, log(automatic CTC) − log(manual CTC)—and then exponentiated to present differences as ratios, which are more interpretable than differences in log(CTC) values (Bland & Altman, 1996). The dashed line indicates the mean ratio of automatic to manual turn counts. The dotted lines indicate the limits of agreement.

CTC at each age. The mean absolute difference between the two methods was fairly consistent between 6 and 24 months, while the relative difference decreased with infants' age. The LoA analyses show that the automatic CTC will be nearly always overestimated, especially at younger ages (6 and 10 months), at which the automatic and manual CTC had low correlations and poor absolute agreement. At older ages (14, 18, and 24 months), the correlations were stronger, but the absolute agreement remained low and did not reach significance. Together, these data demonstrate wide discrepancies between the two measurements of CTC within the current data set, suggesting that the automatic and manual CTC measures are not identical and cannot be replaced by one another. Similar conclusions were previously drawn by Busch and colleagues (Busch et al., 2018), who studied a much smaller sample (six children, 240 min of transcribed data) and reported that LENA consistently counted

Table 4
Multivariable Analysis of Automatic CTC:Manual CTC Ratio by Household Size and Age (Age Adjustment Not Shown in Table)

	Relative difference (automatic/manual CTC)		
Variable	%Δ	(95% CI)	*p*-value
Adults in household ≥ 3	18.7	(−4.1, 47.0)	.11
Siblings in household ≥ 1	30.5	(10.4, 54.3)	.002

Note. %Δ = percent difference in mean automatic CTC:manual CTC ratio between groups.CTC = Conversational Turn Count.

fewer CTs than human coders. While these results might seem contradictory to our findings, it should be noted that the children in the Busch et al. study were older (2–5 years), and spoke a different language (Dutch), than the language spoken in recordings on which the LENA's algorithms were trained. Similar to our conclusions, however, Busch and colleagues argue that their data cast doubt on the comparability of LENA's CTC estimate across families, time, and environments.

The LENA algorithms are proprietary. As such, we are unable to fully describe the origins of the observed discrepancies. Nevertheless, our data suggest a few potential sources of bias. First, we found evidence suggesting inaccuracy in speaker labeling: CTC was overestimated more when there were siblings in the household, and there was a tendency for greater overestimation when there were more than two adults in the household. Second, we found evidence of substantial accidental contiguity for all ages in the sample, although its frequency declined with age. Third, it should be noted that human coding is not entirely error-free, although the procedures used here have been validated across a number of published studies, and intercoder reliability was high. Nevertheless, it may be that LENA is more permissive in selecting infants' language-like vocalizations compared to our criteria. Fourth, it is important to acknowledge that, in the present data set, we consider samples with high adult-speech activity. The present study was part of a larger project focusing on parent–child interaction, the goal of which was to select segments where adult–child verbal exchanges would be frequent. Although this is a speculation, it is possible that high volubility segments are also the ones where accidental contiguity is particularly high, or where LENA's performance on speaker identification is poorer than average.

For researchers who are inspired by the potential for further CTC validation, we make the following recommendations:

1. Systematically validate CTC for different ages between 2 and 48 months. Infants' vocalizations, their social behaviors, and responsiveness change drastically during this time window, and it is fully expected that adult–child speech in close temporal proximity will be a better proxy for turn-taking for some ages compared to others. Our data suggest that younger ages may be particularly challenging.
2. Systematically assess different environments, including high and low adult talk, high and low child talk, one versus multiple adults and children, high and low speech overlap, presence and absence of electronic media, and so forth. The present data set shows that at least some of these variables affect the accuracy of automatic CTC estimates.
3. Include correlation and agreement analyses. Unlike the standard correlation statistic, statistics for absolute agreement such as the ICC are sensitive to systematic biases, which we observed in the present data set. In this case, Pearson's *r* overstates the accuracy of the automatic CTC estimates because it does not reflect the systematic overestimation relative to the manual CTC estimates.
4. Prioritize validation of past studies that drew conclusions based on automatic assessments of CTC alone in very young infants, in contexts of multiple children or adults (such as larger families or schools), or in samples where participants were outside of the age on which the LENA algorithms were trained.

Our goal is by no means to dismiss LENA on the basis of the results presented here; by contrast, we find it to be an indispensable tool for collection of naturalistic, day-long recordings. Nevertheless, given LENA's ubiquitous use across various languages and ages, with typically developing and clinical populations, for the purposes of research and intervention in homes and schools, it is important that LENA's users understand its measurement limitations. Theories of language acquisition propose that turn-taking plays a key role because it provides opportunity for continued *social engagement*, *feedback*, and *interaction*. Our findings demonstrate that, at least in some contexts, there is a wide gap between LENA's estimate of adult–child speech in close temporal proximity and human measures of turn-taking in the audio recordings. Furthermore, the width of this gap is affected by multiple variables, such as the child's age and number of people present around the child. The present findings suggest that the current methods of automatically assessing caregiver–infant interaction are limited. Until systematic reliability estimates of turn-taking across different contexts are available, researchers should validate their conclusions and theoretical proposals by manual analyses.

References

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019). What do North American babies hear? A large-scale cross-corpus analysis. *Developmental Science*, *22*. https://doi.org/10.1111/desc.1272

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, *327*, 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Bland, J. M., & Altman, D. G. (1996). The use of transformation when comparing two means. *British Medical Journal*, *312*, 1153. https://doi.org/10.1136/bmj.312.7039.1153

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*, 135–160. https://doi.org/10.1191/096228099673819272

Bornstein, M. H., Tamis-LeMonda, C. S., & Haynes, O. M. (1999). First words in the second year: Continuity, stability, and models of concurrent and predictive correspondence in vocabulary and verbal responsiveness across age and context. *Infant Behavior & Development*, *22*(1), 65–85. https://doi.org/10.1016/S0163-6383(99)80006-X

Braarud, H. C., & Stormark, K. M. (2008). Prosodic modification and vocal adjustments in mothers' speech during face-to-face interaction with their two- to four-month-old infants: A double video study. *Social Development*, *17*, 1074–1084. https://doi.org/10.1111/j.1467-9507.2007.00455.x

Busch, T., Sangen, A., Vanpoucke, F., & van Wieringen, A. (2018). Correlation and agreement between Language ENvironment Analysis (LENA™) and manual transcription for Dutch natural language recordings. *Behavior Research Methods*, *50*, 1921–1932. https://doi.org/10.3758/s13428-017-0960-0

Christakis, D. A., Lowry, S. J., Goldberg, G., Violette, H., & Garrison, M. M. (2019). Assessment of a parent-child interaction intervention for language development in children. *JAMA Network Open*, *2*, e195738. https://doi.org/10.1001/jamanetworkopen.2019.5738

Conboy, B. T., Brooks, R., Meltzoff, A. N., & Kuhl, P. K. (2015). Social interaction infants' learning of second-language phonetics: An exploration of brain-behavior relations. *Developmental Neuropsychology*, *40*, 216–229. https://doi.org/10.1080/87565641.2015.1014487

Cristia, A., Bulgarelli, F., & Bergelson, E. (2020). Accuracy of the language environment analysis system segmentation and metrics: A systematic review. *Journal of Speech, Language, and Hearing Research*, *6*, 1093–1105. https://doi.org/10.1044/2020_JSLHR-19-00017

Ferjan Ramírez, N., Lytle, S., Fish, M., & Kuhl, P. K. (2018). Parent coaching at 6 and 10 months improves language outcomes at 14 months: A randomized controlled trial. *Developmental Science*, *22*. https://doi.org/10.1111/desc.12762

Ferjan Ramírez, N., Lytle, S., & Kuhl, P. K. (2020). Parent coaching increases conversational turns and advances infant language development. *Proceedings of the National Academy of Sciences of the United States of America*, *117*, 3484–3491. https://doi.org/10.1073/pnas.1921653117

Ganek, H. V., & Eriks-Brophy, A. (2018). A concise protocol for the validation of Language-Environment Analulysis (LENA) Conversational turn counts in Vietnamese. *Communication Disorders Quarterly*, *39*, 371–380. https://doi.org/10.1177/1525740117705094

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Oller, D. K., & Hansen, J. H. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, *26*, 248–265. https://doi.org/10.1044/2016_AJSLP-15-016

Gilkerson, J., Richards, J. A., Warren, S. F., Oller, K., Russo, R., & Vohr, B. (2018). Language experience in the second year of life and language outcomes in late-childhood. *Pediatrics*, *142*, e20174276. https://doi.org/10.1542/peds.2017-4276

Gilkerson, J., Zhang, Y., Xu, D., Richards, J. A., Xu, X., Jiang, F., . . . Topping, K. (2015). Evaluating language environment analysis system performance for Chinese: A pilot study in Shanghai. *Journal of Speech Language and Hearing Research*, *58*, 445–452. https://doi.org/10.1044/2015_JSLHR-L-14-0014

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*, 515–523. https://doi.org/10.1111/j.1467-9280.2008.02117.x

Golinkoff, R. M., Hoff, E., Rowe, M. L., Tamis-LeMonda, C. S., & Hirsh-Pasek, K. (2019). Language matters: Denying the existence of the 30-million-word gap has serious consequences. *Child Development*, *90*, 985–992. https://doi.org/10.1111/cdev.13128

Hart, B., & Risley, T. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: P.H. Brookes.

Hirsh-Pasek, K., Adamson, L., Bakeman, R., Golinkoff, R. M., Pace, A., Yust, P., & Suma, K. (2015). The contribution of early communication to low-income children's language success. *Psychological Science*, *26*, 1071–1083. https://doi.org/10.1177/0956797615581493

Hoff, E. (2003). The Specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, *74*, 1368–1378. https://doi.org/10.1111/1467-8624.00612

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review*, *26*, 55–88. https://doi.org/10.1016/j.dr.2005.11.00

Hollingshead, A.B. (1975). *Four factor index of social status*, New Haven, CT: Yale University Department of Sociology.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, *27*, 236. https://doi.org/10.1037/0012-1649.27.2.236

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science*, *10*, 110–120. https://doi.org/10.1073/pnas.1532872100

Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 9096–9101. https://doi.org/10.1073/pnas.1532872100

Leech, K. A., & Rowe, M. L. (2020). An intervention to increase conversational turns between parents and young children. *Journal of Child Language*. https://doi.org/10.1017/S0305000920000252

Levinson, S. C. (2016). Turn-taking in human communication—Origins and implications for language processing. *Trends in Cognitive Sciences*, *20*(1), 6–14. https://doi.org/10.1016/j.tics.2015.10.010

Meltzoff, A. N., & Brooks, R. (2009). Social cognition and language: The role of gaze following in early word learning. In J. Colombo, P. McCardle, & L. Freund (Eds.), *Infant pathways to language: Methods, models, and research disorders* (pp. 169–194). New York, NY: Psychology Press.

Merz, E. C., Maskus, E. A., Melvin, S., He, X., & Noble, K. G. (2019). Socioeconomic disparities in language input are associated with children's language-related brain structure and reading skills. *Child Development*, *91*, 846–860. https://doi.org/10.1111/cdev.13239

Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., . . . Warren, S. F. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 13354–13359. https://doi.org/10.1073/pnas.1003882107

Pae, S., Yoon, H., Seol, A., Gilkerson, J., Richards, J. A., Ma, L., & Topping, K. (2016). Effects of feedback on parent–child language with infants and toddlers in Korea. *First Language*, *36*, 549–569. https://doi.org/10.1177/0142723716649273

Perry, L. K., Prince, E. B., Valtierra, A. M., Rivero-Fernandez, C., Ullery, M. A., Katz, L. F., . . . Messinger, D. S. (2018). A year in words: The dynamics and consequences of language experiences in an intervention classroom. *PLoS One*, *13*. https://doi.org/10.1371/journal.pone.0199893

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2014). Look who's talking: Speech style and social context in language input to infants are linked to concurrent and future speech development. *Developmental Science*, *17*, 880–891. https://doi.org/10.1111/desc.12172

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2017a). The impact of early social interactions on later language development in Spanish-English bilingual infants. *Child Development*, *88*, 1216–1234. https://doi.org/10.1111/cdev.12648

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2017b). Look who's talking NOW! Parentese speech, social context, and language development across time. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01008

Romeo, R. R., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Rowe, M. L., & Gabrieli, J. D. E. (2018). Beyond the 30-million-word gap: Children's conversational exposure is associated with language-related brain function. *Psychological Science*, *29*, 700–710. https://doi.org/10.1177/0956797617742725

Romeo, R. R., Segaran, J., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., . . . Gabrieli, J. D. E. (2018). Language exposure relates to structural neural connectivity in childhood. *Journal of Neuroscience*, *38*, 7870–7877. https://doi.org/10.1523/JNEUROSCI.0484-18.2018

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development*, *83*, 1762–1774. https://doi.org/10.1111/j.1467-8624.2012.01805.x

Shneidman, L. A., Arroyo, M. E., Levine, S. C., & Goldin-Meadow, S. (2013). What counts as effective input for word learning? *Journal of Child Language*, *40*, 672–686. https://doi.org/10.1017/S0305000912000141

Smith, N. A., & Trainor, L. J. (2008). Infant-directed speech is modulated by infant feedback. *Infancy*, *13*, 410–420. https://doi.org/10.1080/15250000802188719

Tamis-LeMonda, C. S., Bornstein, M., Kahana-Kalman, R., Baumwell, L., & Cyphers, L. (1998). Predicting variation in the timing of language milestones in the second year: An events history approach. *Journal of Child Language*, *25*, 675–700. https://doi.org/10.1017/s0305000998003572.

Tamis-LeMonda, C. S., Kuchirko, Y., & Song, L. (2014). Why is infant language learning facilitated by parental responsiveness? *Current Directions in Psychological Science*, *23*, 121–126. https://doi.org/10.1177/0963721414522813

Warlaumont, A. S., Richards, J. A., Gilkerson, J., & Oller, D. K. (2014). A social feedback loop for speech development and its reduction in autism. *Psychological Science*, *25*, 1314–1324. https://doi.org/10.1177/0956797614531023

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*, 2143–2152. https://doi.org/10.1177/0956797613488145

Xu, D., Richards, J. A., & Gilkerson, J. (2014). Automated analysis of child phonetic production using naturalistic recordings. *Journal of Speech, Language, and Hearing Research*, *57*, 1638–1650. https://doi.org/10.1044/2014_JSLHR-S-13-0037

Zimmerman, F. J., Gilkerson, J., Richards, J. A., Christakis, D. A., Xu, D., Gray, S., & Yapanel, U. (2009). Teaching by listening: The importance of adult–child conversations to language development. *Pediatrics*, *124*, 342–349. https://doi.org/10.1542/peds.2008-2267

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's website:

Appendix S1. Supplemental Materials