JSLHR

Review Article

Accuracy of the Language Environment Analysis System Segmentation and Metrics: A Systematic Review

Alejandrina Cristia,[a] Federica Bulgarelli,[b] and Elika Bergelson[b]

Purpose: The Language Environment Analysis (LENA) system provides automated measures facilitating clinical and nonclinical research and interventions on language development, but there are only a few, scattered independent reports of these measures' validity. The objectives of the current systematic review were to (a) discover studies comparing LENA output with manual annotation, namely, accuracy of talker labels, as well as involving adult word counts (AWCs), conversational turn counts (CTCs), and child vocalization counts (CVCs); (b) describe them qualitatively; (c) quantitatively integrate them to assess central tendencies; and (d) quantitatively integrate them to assess potential moderators.
Method: Searches on Google Scholar, PubMed, Scopus, and PsycInfo were combined with expert knowledge, and interarticle citations resulting in 238 records screened and 73 records whose full text was inspected. To be included, studies must target children under the age of 18 years and report on accuracy of LENA labels (e.g., precision and/or recall) and/or AWC, CTC, or CVC (correlations and/or error metrics).
Results: A total of 33 studies, in 28 articles, were discovered. A qualitative review revealed most validation studies had not been peer reviewed as such and failed to report key methodology and results. Quantitative integration of the results was possible for a broad definition of recall and precision (M = 59% and 68%, respectively; N = 12–13), for AWC (mean r = .79, N = 13), CVC (mean r = .77, N = 5), and CTC (mean r = .36, N = 6). Publication bias and moderators could not be assessed meta-analytically.
Conclusion: Further research and improved reporting are needed in studies evaluating LENA segmentation and quantification accuracy, with work investigating CTC being particularly urgent.
Supplemental Material: https://osf.io/4nhms/

Over the past decade, there has been a sea change in how early childhood specialists think about and analyze the language environment that infants experience and contribute to. This has been driven in no small part by improving technology, which allows for longer, less obtrusive recordings and automated analyses of recording contents. The Language Environment Analysis (LENA; Greenwood et al., 2011) system is perhaps the most prominent "off the shelf" system, providing researchers, clinicians, and early educators a snapshot of young children's input and speech productions with virtually no effort. The system combines a small wearable device that records for up to 16 hr, with automated speech analyses of these long recordings, determining who is talking, how much, and when. These kinds of numbers, in turn, can be used to assess where a given child's vocalizations fall relative to age-matched peers and to guide interventions seeking to assess and increase children's "language nutrition" (e.g., Oller et al., 2010; Suskind et al., 2013). Two key advantages of the LENA system are that it provides automated output without requiring any manual annotation and that, by measuring an entire day, it provides more ecologically valid data than can be gathered when infants come into a lab or clinic, or in brief home recording contexts (e.g., Bergelson et al., 2018; Tamis-LeMonda et al., 2017).

Thus, the LENA system has been an important addition to clinicians', psychologists', and linguists' toolkits, as evidenced by a growing literature in both the basic science and intervention realms (Adams et al., 2018; Sosa, 2016; Suskind et al., 2016, 2013; Wood et al., 2016). A number of independent evaluations of the LENA system have also become public since validation of several LENA metrics was published as part of LENA Foundation reports (Xu et al., 2009).

[a]Laboratoire de Sciences Cognitives et Psycholinguistique, Département d'Études Cognitives, ENS, EHESS, CNRS, PSL University, Paris, France
[b]Psychology & Neuroscience, Duke University, Durham, NC
Correspondence to Alejandrina Cristia: alecristia@gmail.com

Editor-in-Chief: Bharath Chandrasekaran
Editor: Chao-Yang Lee
Received June 18, 2019
Revision received October 1, 2019
Accepted December 30, 2019
https://doi.org/10.1044/2020_JSLHR-19-00017

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

We aimed to summarize work comparing the LENA automatic output against human annotation of the same data, for both basic and derived LENA annotations. Regarding basic annotations, the system divides the audio into a few classes to decide who speaks when. The speech classes are as follows: Female Adult (FAN), Male Adult (MAN), Key Child (CHN), and Other Child (CXN). The nonspeech classes are as follows: Silence, TV and Electronic Noise, Undefined Noise, and Overlap (among any of the above). Additionally, LENA can label a section as "far" versions of the other tags (e.g., FA Far) when the model finds the section is equally likely to be FAN or Silence. Adopting a standard approach, we aimed to calculate classification accuracy as recall and precision. Details on how this was operationalized are below, but in short, recall indicates how accurate the system is in detecting an event (e.g., a given speech class) when the event occurs, and precision indicates how often the system is right when it says there was an event.

As for derived metrics, the LENA system also provides quantitative information about how much talking is happening, that is, adult word count (AWC), child vocalization count (CVC), and conversational turn count (CTC). We sought to quantitatively integrate estimates of the accuracy of these metrics, which by virtue of being numeric (rather than categorical such as speech class) can be evaluated using correlation coefficients (i.e., how much variance is shared across estimates provided by the LENA system and by human coders) and relative error rates (RERs; i.e., how far off the LENA system's estimates and the human coders' estimates are).

Additionally, we considered two sets of potential moderators. LENA's algorithms were originally trained on data from a sample of 1- to 42-month-old children learning North American English (NAE), balanced in gender and stratified by maternal education. Specifically, the training used 135 hr of data from 309 separate recordings from as many infants, sampling a 30-min section from each recording (Gilkerson et al., 2008).[1] We intended to look at accuracy as a function of several participant- and method-related moderators to assess the generalizability of LENA output beyond the kind of data in the original training set. For participant-related moderators, we considered native language, matching status (i.e., whether the participants matched LENA's training sample or not), and mean age, and age range. We expected that studies that did not match LENA's training sample (e.g., because it was a different language or an older mean age) would have lower reliability. Relatedly, wider age ranges may lead to higher reliability because the participants may be more diverse, and thus observations for each of them may be more different from each other.

For methods-related moderators, we considered how clips were selected and presented to human annotators and the total duration of annotations cumulated across participants in each study (further details in the Method section). These methodological factors were relevant because they affect the extent to which we can view validation results as biased by the same algorithm they intend to evaluate and how those results may generalize to daylong data at large. For example, if human annotators know that the LENA system classified a stretch of audio as "FAN," it may take more effort for them to notice that this stretch of audio also contains some silence or noise or overlaps with a different talker. In contrast, human annotators who did not have access to LENA's labels cannot be influenced by them. As for total duration, we reasoned this may be a measure of data quantity and could potentially point to evidence of biased reporting, if there were any. In many meta-analyses, an anticorrelation between data quantity and effect size (i.e., smaller studies have better results than bigger studies) is consistent with this kind of selective reporting. Following the same rationale, we reasoned that an anticorrelation of reliability and amount of data may be consistent with selective reporting.

Method

All decision steps are documented in the supplementary materials (Cristia, Bulgarelli, & Bergelson, 2019), which also provide access to code necessary to reproduce all analyses below. The eligibility criteria for inclusion in the qualitative analyses of the systematic review were as follows: (a) The LENA device was worn by children aged 18 years or less, with no limits placed on native language, population sample, recording setting, or recording duration (which led to the inclusion of two studies at least partially based on very short recordings); (b) the LENA audio was coded by a human annotator; and (c) accuracy was estimated via agreement for categorical decisions (e.g., speaker labels) and correlations and/or error rates for derived estimates (AWC, CVC, CTC). As a result of these deliberately widely inclusive criteria, studies were only excluded if there was no accuracy discussed.

To be included in the quantitative analyses, an article is needed to report quantitatively on one or more of the following metrics: precision and/or recall for categorical decisions; Pearson correlations on AWC, CVC, and CTC; RERs for AWC, CVC, and CTC; or means for LENA and human AWC, CVC, and CTC (so that we could derive RERs).

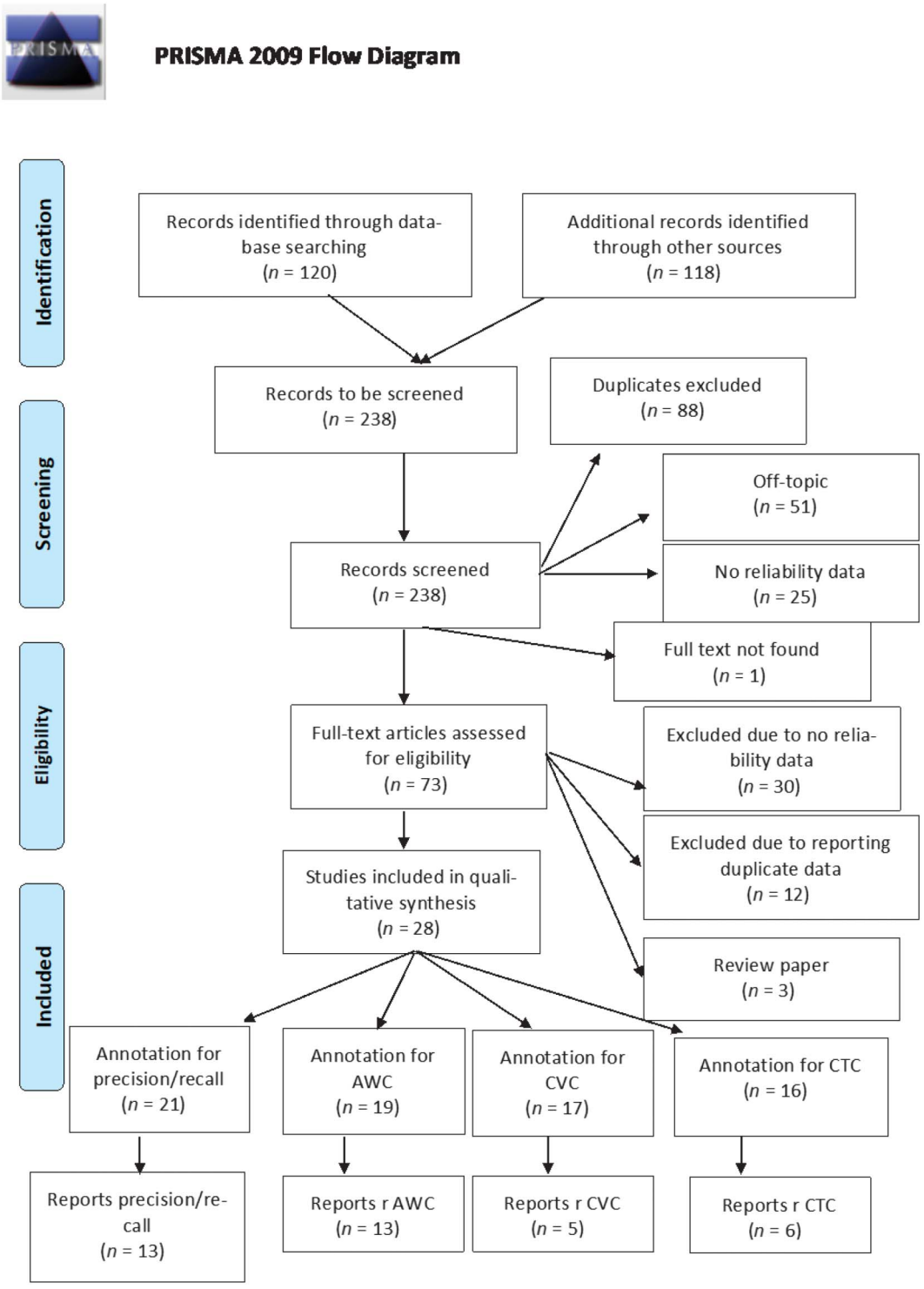
The information sources used to compose the initial list included suggestions by experts (authors of this work as well as Melanie Soderstrom); three Google Scholar searches (example of a full key word set: "LENA, speech reliability, male-adult, female-adult, child"); and a Google alert complemented with searches in Scopus, PubMed, and PsycInfo as well as articles cited as containing an evaluation by any of the full articles inspected. We stopped including

[1]Although other documents from the LENA Foundation state this age range is 2–48 months (e.g., Richards et al., 2008), we use the range in the document that specifically reports on the transcriptions of the training and test sets used for training acoustic models for segmentation (Gilkerson et al., 2008).

studies on September 10, 2019. The PRISMA flowchart is in Figure 1.

Data were coded independently by the first and second authors, with disagreements resolved by discussion. The full list of variables coded is available in the supplementary materials (Cristia, Bulgarelli, & Bergelson, 2019). The most relevant to the present analyses are discussed below.

Figure 1. PRISMA flowchart. AWC = adult word count; CTC = conversational turn count; CVC = child vocalization count.



We coded number of children included and their demographic characteristics. These demographics included mean age, age range, native language, and whether they belong to a special population (e.g., typically developing, at risk for autism or other developmental disorders, bilingual, twin, or a mix[2]). Based on this description, we classified a sample as matching the LENA training sample (if children were typically developing, 42 months old or less, learners of NAE, with no specific bias for a given socioeconomic status [SES] group) or not.

In addition, we coded the size of the samples presented to the annotators. Specifically, we coded whether the annotator was presented with continuous samples (i.e., a long section of audio) or with segments (i.e., a short stretch that LENA had classified as a specific talker type); when given segments, annotators sometimes were able to inspect a broader context and, at times, were able to adjust the LENA-generated boundaries.

We also noted the type of data selection. We coded whether the recordings presented to annotators were chosen based on high volubility (CTC, CVC, or AWC), whether they were based on the algorithm in some other way (using information from the LENA segmentation algorithm), or whether it was random, chosen independently of the LENA segmentation algorithm.[3]

Additionally, total quantity of the audio data annotated was coded, and this was done at two levels: sample duration in minutes (how long were continuous segments presented to annotators) and total duration (total duration in minutes, collapsing across children and samples).

Some of our analyses below rely on error rates, which are estimates of how close LENA counts are to human counts. When RER was reported ($N = 8$), we noted how it was implemented. In eight further cases, RER was not reported but it could be calculated from information provided in the text. In the latter case, we estimated RER as the system estimate minus the human estimate divided by the latter and multiplied by 100 to have a percentage. Notice that this number is positive when the LENA overestimates (reports a higher number than the human) and negative when it underestimates.

Risk assessment at the level of articles was done by assessing whether the authors acknowledged the LENA Foundation as their affiliation or as a funding source. If any of the authors had such an affiliation or acknowledged the Foundation, then the article as a whole was tagged as being at risk. Regarding other risks at the individual level (such as outcome selection or other forms of bias), we systematically coded methodological characteristics that may affect validation results as noted above; for example, the way in which data are selected may involve the LENA algorithm (in which case, generalization to the whole data set may be compromised). We present these methodological variables in context in the Results section. Risk assessment for the whole body of literature was not possible.

Some of the articles (e.g., Bergelson et al., 2018; Bulgarelli & Bergelson, 2019) provided links to their data, allowing us to calculate relevant validation metrics to include here using scripts that can be found in our supplementary materials (Cristia, Bulgarelli, & Bergelson, 2019). Finally, as we will see below, some articles did not provide any metrics that could be integrated quantitatively (Bredin-Oja et al., 2018), but they otherwise match our inclusion criteria and thus are included in qualitative analyses.

Results

Qualitative Integration

We found a total of 28 articles reporting 33 (not mutually independent) validation studies (see Tables 1–3). Since only five studies involved one or more individuals with a LENA affiliation, we do not further discuss potential bias due to a conflict of interest. A total of 12 studies from 10 articles have been published in peer-reviewed journals as validation studies and thus have been (thoroughly) documented and evaluated as such (Bredin-Oja et al., 2018; Bulgarelli & Bergelson, 2019; Busch et al., 2018; Canault et al., 2016; Ganek & Eriks-Brophy, 2018; Gilkerson et al., 2015; Jones et al., 2019; Oetting et al., 2009; Orena et al., 2019; VanDam & Silbert, 2016). The remaining 21 studies from 18 articles are reported in preprints (Berends, 2015; Cristia, Lavechin, et al., 2019; Lehet et al., 2019); in theses (Elo, 2016); in conference proceedings, posters, or talks (McCauley et al., 2011; Schwarz et al., 2017; Soderstrom & Franz, 2016; van Alphen et al., 2017); in white papers (Xu et al., 2009); as secondary or preliminary methodological information in the service of a separate research question (Bergelson et al., 2018; Burgess et al., 2013; Caskey et al., 2014; D'Apice et al., 2019; Ko et al., 2016; Merz et al., 2019; Pae et al., 2016); or in an appendix or a supplementary

[2]Studies labeled as "mix" were the following: van Alphen et al. (2017) included children at a familial risk of dyslexia and controls; Bredin-Oja et al. (2018) had two children with autism diagnoses, two with Down syndrome, one with a chromosomal deletion, and one with a developmental delay; Lehet et al. (2019) included children with a range of hearing statuses, including children with normal hearing, children with hearing aids, and children with cochlear implants; Merz et al. (2019) included participants with a range of SES and maternal education statuses; Weisleder and Fernald (2013) included children from low-SES homes who were also bilingual; and Xu et al. (2009) included children with a range of SESs.

[3]Articles that were coded as random for data selection are as follows: In Bredin-Oja et al. (2018), all child vocalizations with clear vocal fold vibration that were not vegetative sounds were transcribed; in Bulgarelli and Bergelson (2019), all utterances containing concrete nouns were transcribed; Cristia, Lavechin, et al. (2019) used random or periodic sampling of 1–2 min; in Elo (2016), the recordings were transcribed in full; in Jones et al. (2019), a human transcriber started at the beginning of the recording until they encountered a child utterance, at which point they coded 5 full minutes, then skipped 10 min, and repeated this process; McCauley et al. (2011) selected three random 5-min segments; Oetting et al. (2009) used the full recording; Schwarz et al. (2017) selected at random; Soderstrom and Franz (2016) transcribed 15 min starting 1 hr into the recording; and Xu et al. (2009) selected two participants with different SESs.

Table 1. Articles and studies included in the systematic review: precision and recall.

ID	Article	Recall	Included in recall	Precision	Included in precision
1	van Alphen 2017				
2	Berends 2015				
3	Bergelson 2018			84	FA, MA
4	Bredin-Oja 2018+				
5	Bulgarelli 2019+	62	Adult, Child, CXN, FA, MA	63	Adult, Child, CHN, CXN, FA, MA
6	Burgess 2013				
7	Busch 2018+				
8	Canault 2016+				
9	Caskey 2014				
10	Cristia 2019	28	Adult, Child, CHN, CXN, FA, MA	41	Adult, Child, CHN, CXN, FA, MA
11	D'Apice 2019				
12.1	Elo 2016	86	CHN, CXN, FA, MA	81	CHN, CXN, FA, MA
12.2	Elo 2016	86	CHN, CXN, FA, MA	90	CHN, CXN, FA, MA
13	Ganek 2018+				
14	Gilkerson 2015*+	80	Adult, Child	46	Adult, Child, CHN, CXN, FA, MA
15.1	Jones 2019+	56	CHN		
15.2	Jones 2019+	11	CHN		
15.3	Jones 2019+	46	CHN	61	Adult, CHN
16	Ko 2016			84	CHN, FA
17	Lehet 2019	62	Adult, Child, FA, MA	64	Adult, Child, FA, MA
18	McCauley 2011	64	Adult, Child, CHN, CXN		
19	Merz 2019				
20	Oetting 2009+				
21	Orena 2019+				
22	Pae 2016*				
23	Schwarz 2017				
24	Seidl 2018			72	CHN, FA
25	Soderstrom 2016	45	Adult, Child, CHN, CXN, FA, MA	62	Adult, Child, CHN, CXN, FA, MA
26	VanDam 2016+			71	Adult, Child, FA, MA
27	Weisleder 2013				
28.1	Xu 2009*	79	Adult, Child	69	Adult, Child
28.2	Xu 2009*				
28.3	Xu 2009*				

Note. Not all studies contribute data for quantitative integration. Regardless of how many authors a publication has, articles are identified by the first author and year of publication. Articles with an asterisk (*) denote Language Environment Analysis affiliation; those with a plus sign (+) were published as evaluations in peer-reviewed journals. Recall and precision are provided in percentage points. "Included in recall" lists the sources of recall values included in that average recall; "included in precision", the same for precision: Adult = adult categories collapsed, Child = child categories collapsed, CHN = target child, CXN = other children, FA = female adult, and MA = male adult. All numeric predictors have been rounded for this display.

material (Seidl et al., 2018; Weisleder & Fernald, 2013). We include this "gray literature" for thoroughness but encourage any interested readers to filter the data and rerun the analyses with our openly provided code and data as they see fit.

Studies tended to have small sample sizes, with a median N of 11.50 children (range: 1–107, M = 22, total = 689). A majority (n = 18) focused on NAE; for the other 15, children were learning U.K. English, U.S. Spanish, Dutch, Finnish, Swedish, France French, Canada French, Korean, Mandarin and Shanghai Chinese, Tsimane', or Vietnamese. Very few (n = 8) reported on children who matched the LENA's training data sample. The other 25 came from populations that differed from the training sample, including children who were diagnosed with or at risk for autism spectrum disorder, of markedly low SES, preterm, twins, particularly high in language skills, bilingual, at risk for developmental delays, or a mixture of these groups. Children's age varied across studies (M = 35 months, Mdn = 26 months, range: 0–192 months). Within studies, children's age ranges varied between 0 and 144 months (M = 26 months, Mdn = 20 months).

Not all studies reported their process regarding data selection and human annotations. For those that did, 13 studies used a random clip selection algorithm or looked at a whole recording; six extracted sections based on high CVC, CTC, and/or AWC; and a further 10 were otherwise algorithm dependent. While 14 of the studies presented full clips (e.g., 5 continuous minutes) or whole recordings to annotators, nine presented sections that had been segmented by the LENA algorithm. In either case, LENA annotations may or may not have been visible to (and thus bias) the human annotators, and annotators may have been able to resegment them. Unfortunately, not all studies described this aspect of their annotation procedure clearly. The total cumulated duration of annotated data in minutes varied massively across studies (range: 25–4,200, M = 809, Mdn = 600).

We now turn to a qualitative overview of validation results. One class of reports pertains to the level of agreement

Table 2. Articles and studies included in the systematic review: adult word count (AWC), child vocalization count (CVC), and conversational turn count (CTC).

ID	Article	AWC_R	AWC_RER	CVC_R	CVC_RER	CTC_R	CTC_RER
1	van Alphen 2017		18				
2	Berends 2015						–90
3	Bergelson 2018						
4	Bredin-Oja 2018+						
5	Bulgarelli 2019+						
6	Burgess 2013		27				
7	Busch 2018+	.87	–20	.77	2	.52	–64
8	Canault 2016+	.64	–33	.71	–66		
9	Caskey 2014	.93					
10	Cristia 2019	.75	55	.8	–20	.59	–67
11	D'Apice 2019	.79	12				
12.1	Elo 2016		49		–10		
12.2	Elo 2016		67		–6		
13	Ganek 2018+					.7	43
14	Gilkerson 2015*+	.73	9			.22	7
15.1	Jones 2019+						
15.2	Jones 2019+						
15.3	Jones 2019+						
16	Ko 2016						
17	Lehet 2019						
18	McCauley 2011			.81	–45		
19	Merz 2019			.74			
20	Oetting 2009+	.85	18			.14	
21	Orena 2019+	.77	20				
22	Pae 2016*+	.72				–.03	
23	Schwarz 2017	.67					
24	Seidl 2018						
25	Soderstrom 2016	.82					
26	VanDam 2016+						
27	Weisleder 2013	.8					
28.1	Xu 2009*	.92	–2				
28.2	Xu 2009*		–0.4				
28.3	Xu 2009*		–27				

Note. Not all studies contribute data for quantitative integration. Regardless of how many authors a publication has, articles are identified by the first author and year of publication. Articles with an asterisk (*) denote Language Environment Analysis (LENA) affiliation; those with a plus sign (+) were published as evaluations in peer-reviewed journals. All numeric predictors have been rounded for this display. R = correlations between human and LENA counts; RER = relative error rate (in percentage points)

on the labels ascribed to a stretch of audio signal. Two measures are most often reported for this, recall and precision. Recall quantifies accuracy using human tags as the denominator, that is, (true positives) / (true positives + false negatives); put otherwise, it answers the question "Out of all the vocalizations by talker X, how many got the label X?". Precision quantifies accuracy using the system as the denominator, that is, (true positives) / (true positives + false positives); put otherwise, it answers the question "Out of all the vocalizations labeled as being talker X by the system, how many were really X according to the human annotator?". Ideally, such accuracy metrics are reported for each class separately, together with the frequency of that class: CHN, CXN, FAN, MAN, and none of the above.

However, few of the 33 studies reported even partial recall or precision metrics and often not at the speech class level of granularity: At most, 10 of 33 report precision/recall on a subset of labels (e.g., CHN vs. CXN, but not FAN and MAN). Instead or in addition, some (8/33) report recall and/or precision for broader classifications (e.g., adult vs. child, collapsing between FAN and MAN on the one hand and between CXN and CHN on the other). Most often than not, nonspeech and "far" categories are not discussed at all, and thus, it is unclear how any confusions with Silence, Overlap, and so forth were handled. A similarly sparse picture appears when we inspect the prevalence of AWC (13 correlations and 14 RERs reported), CVC (five correlations and six RERs reported), and CTC (six correlations and five RERs reported).

We wondered whether this may paint a direr picture than needed because perhaps the authors did not code their data in a way that would allow them to estimate LENA accuracy for talkers, or one or more of the other metrics. This case is hard to make for talker identification, which is fundamental to any other validation task: If the authors intend to validate CVC, then they need to decide when the key child is speaking and when he or she is not, thus incidentally producing data for this particular talker category; if they want to validate AWC, they do the same for the adult categories; and if they seek to validate CTC, they should

Table 3. Articles and studies included in the systematic review: demographic and methodological characteristics.

ID	Article	Language	Total	Sel	Size	Participant	Age	Range	*N*	Match
1	van Alphen 2017	Dutch				Mix	37.9		42	No
2	Berends 2015	Dutch		High		Drisk	36	30–42	14	No
3	Bergelson 2018	NAE		algo_driven	segm	Typ	11.5	5–20	61	Yes
4	Bredin-Oja 2018+	NAE	138	Random	cont	Mix	36	28–46	6	No
5	Bulgarelli 2019+	NAE	1,932	Random	segm	Typ	6.5	6–7	44	Yes
6	Burgess 2013	NAE	465	High	cont	ASD	51	35–67	10	No
7	Busch 2018+	Dutch	240	algo_driven	cont	Typ	42	24–60	5	No
8	Canault 2016+	Fr	3,240	High		Typ	25.5	3–48	18	No
9	Caskey 2014	NAE, Sp	25			Preterm	0	32–36 wg	5	No
10	Cristia 2019	NAE, UKE, Tsimane	1,472	Random	cont		16.8	3–58	49	No
11	D'Apice 2019	UKE	320	High	cont	Typ	33.2	24–48	107	No
12.1	Elo 2016	Finnish	698	Random	cont	Twin	7	0	1	No
12.2	Elo 2016	Finnish	630	Random	cont	Twin	9	0	1	No
13	Ganek 2018+	Viet	100	algo_driven	cont	Mix	30.5	22–42	10	No
14	Gilkerson 2015*+	Chn	330	High	cont	Typ	12.1	3–23	22	No
15.1	Jones 2019+	NAE	120	Random	cont	ASD	78	60–96	8	No
15.2	Jones 2019+	NAE	105	Random	cont	ASD	192	168–216	7	No
15.3	Jones 2019+	NAE	2,210		segm	ASD	96	60–204	36	No
16	Ko 2016	NAE	26	algo_driven	segm	Typ	20.4	12–30	13	Yes
17	Lehet 2019	NAE	734	algo_driven	cont	Mix	20	4–34	23	No
18	McCauley 2011	NAE	150	Random		ASD			5	No
19	Merz 2019	NAE	600	algo_driven		Mix	90	61–119	10	No
20	Oetting 2009+	NAE	510	Random	cont	Low SES	42	24–60	17	No
21	Orena 2019+	Fr, NAE	945	algo_driven	segm	Bilingual	10	10–11	21	No
22	Pae 2016*	Korean	630			Typ	12.5	4–16		No
23	Schwarz 2017	Swedish	240	Random	cont	Typ	30		4	No
24	Seidl 2018	NAE		algo_driven	segm	ASDrisk			10	No
25	Soderstrom 2016	NAE	1,305	Random	segm	Typ	25	12–38	32	Yes
26	VanDam 2016+	NAE	47	algo_driven	segm	Typ	29.1		26	Yes
27	Weisleder 2013	Sp	600	algo_driven		Mix	19	0	10	No
28.1	Xu 2009*	NAE	4,200	High	segm	Mix	19	2–36	70	Yes
28.2	Xu 2009*	NAE	720	Random		Typ	10	0	1	Yes
28.3	Xu 2009*	NAE	720	Random		highL	31	0	1	Yes

Note. Regardless of how many authors a publication has, articles are identified by the first author and year of publication. Articles with an asterisk (*) denote Language Environment Analysis (LENA) affiliation; those with a plus sign (+) were published as evaluations in peer-reviewed journals. "Match" reflects whether it matches the LENA training sample. All numeric predictors have been rounded for this display. Language = native language of participants; Total = total duration of annotated samples (in minutes); Sel = type of selection; Size = size of the sample provided to the human annotator; Participant = characteristics of the participant sample; Age = mean age of participants in months; Range = age range of participants in months (except for Caskey et al. [2014], where it indicates weeks gestation [wg]); *N* = number of children included in the sample; mix = a mixture of any of the above; high = based on high AWC/CVC/CTC; Drisk = at risk for developmental delays; NAE = North American English; segm = segment; typ = typically developing and with none of the other characteristics; random = unrelated to LENA segmentation; cont = continuous; ASD(risk) = diagnosed with (or at risk for) autism spectrum disorder; Fr = French; Sp = Spanish; UKE = United Kingdom English; twin = having a twin; Viet = Vietnamese; Chn = Shanghai Chinese and/or Mandarin Chinese; low SES = family with low socioeconomic status; highL = high language.

be incidentally producing talker identification data for both children and adult categories. Nonetheless, one can design an annotation scheme where this information is not produced incidentally—for instance, if the annotator has to decide on the fly whether there was a child vocalization or not and then only writes down the total number for the clip. Assuming this worst case scenario, the first author revisited the description of the annotation, human reliability, and results reported, to judge what the authors could have calculated LENA validity on (regardless of whether they did or did not ultimately report validity results). We found that, in general, validation study reports are less informative than they could be. Specifically, 21 studies could have calculated talker identification accuracy, but only 12 reported some recall data, and 13 did so for precision. Seventeen could have reported on CVC accuracy, but only seven reported a correlation or an RER. Furthermore, 16 could have reported on CTC accuracy, with only seven providing information on a correlation or an RER. The exception concerned AWC, with all studies evaluating AWC reporting a correlation and/or an error rate.

Moreover, there is considerable variability in methods used. The best example is probably RERs, which can be calculated in many ways, for instance, by summing counts across all clips and participants or by calculating relative errors for each clip and participant separately and then averaging this. By and large, if a system is unbiased, the former will produce lower error rates than the latter. Moreover, the interpretation is not the same; if one calculates an error rate from the sum of words transcribed over a whole

recording day, the generalization will be across recording days, whereas if the error rates are calculated in smaller time units, then this error rate will generalize more readily to other small time units than to the whole day. Of the 16 studies in which some RER could be derived (collapsing across AWC, CVC, and CTC), we calculated eight based on reported mean counts, and the remaining eight were reported on in four articles. Two articles did not explain how they had estimated the error rate (although, in one case, it seems likely that it is the mean of RERs calculated separately for each clip and child). One study reported RER separating each of 3 full days of audio, whereas another also analyzed 2 full days of audio, but first calculated RER per hour, and then applied the mean to this. In the discussion, we suggest that data sharing may be the best solution to the problem of the blooming diversity in calculation methods.

Quantitative Analyses: Central Tendencies

As mentioned in the qualitative overview, very few studies provided data at the ideal level of granularity. In light of this sparsity, we made the analytic decision to calculate a global recall and precision metric, defined as averages of whatever recall/precision were reported within a study. This allowed us to consider 12 studies (36% of all studies) for recall and 13 (39%) for precision; that is, fewer than half of studies could contribute data to this analysis even with this lax definition. Problematically, authors may only report the categories for which they obtained relatively high recall and/or precision. If so, this could introduce outcome selection bias into our results. With that caveat in mind, we report that overall recall and precision estimates are quite high (recall: $N = 12$, $M = 59\%$; $N = 12$, weighted mean = 62%, where weights are based on the total cumulated duration of annotated data; precision: $N = 13$, $M = 68\%$; $N = 11$, weighted mean = 64%).

Regarding LENA's numeric output (AWC, CVC, CTC), evaluations are based on correlations (r) and/or some form of RER. As a reminder, most of these RERs come from taking the sum or average count for the system minus the same for human annotators divided by the latter (multiplied by 100 to have a percentage), such that a positive number indicates that the LENA overestimates and a negative number indicates that it underestimates. The Pearson r for AWC was very high ($N = 13$, $M = .79$; $N = 13$, weighted mean = .79), with relatively low RERs suggesting a slight tendency to overestimate AWC ($N = 14$, $M = 13.76\%$; $N = 13$, weighted mean = 4.48%).

For CVC, the Pearson r was also quite high ($N = 5$, $M = .77$; $N = 5$, weighted mean = .74), but RERs were numerically larger and indicated an underestimated CVC ($N = 6$, $M = -24.17\%$; $N = 6$, weighted mean = −40.48%, which would mean CVCs are nearly halved). For CTC, the Pearson r was quite low ($N = 6$, $M = .36$; $N = 6$, weighted mean = .36), and the RERs also indicated a rather strong underestimation tendency ($N = 5$, $M = -34\%$; $N = 4$, weighted mean = −50%).

Quantitative Analyses: Participant and Methodological Moderators

We next intended to carry out a meta-regression with moderators in order to assess whether participant-specific or methodological factors impacted recall, precision, or AWC estimate accuracy. As a first approach, we provide readers with figures summarizing trends in the data. These figures highlight the fact that, with such a small number of points, the variability is large enough that no differences can be detected, with low confidence even about whether line slopes are positive or negative for continuous variables.

Figure 2 displays our three dependent measures as a function of the infants' native language (NAE or other), population sample (matching the sample in the LENA training or not), and age (mean and range within each study). Figure 3 depicts our dependent measures split by methodological variants: whether the clips were selected at random or using the LENA algorithm (to identify periods with high word counts or in any other way), how long clips were, and how much data were annotated.

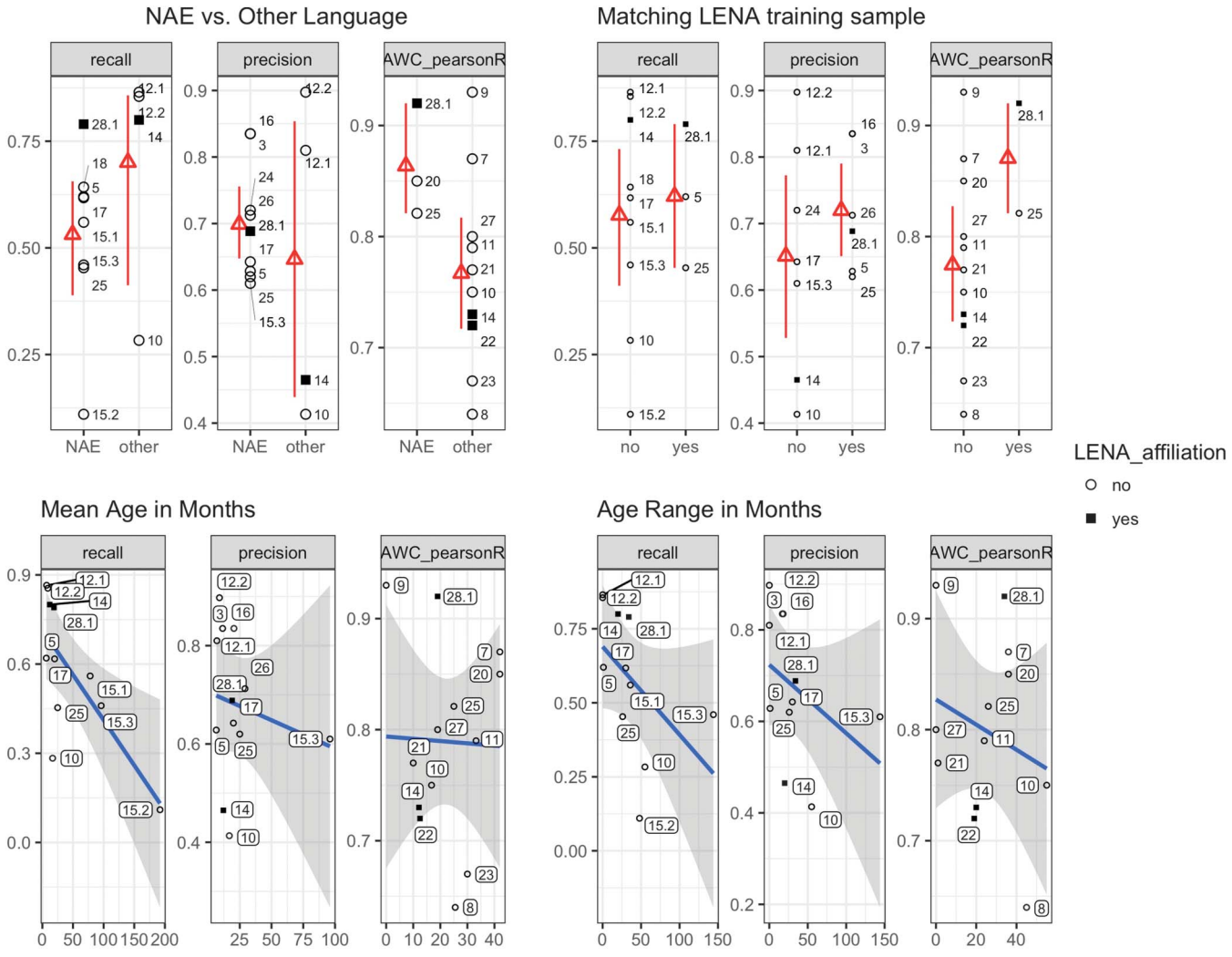
For both sets of graphs, in the top panels, the central tendency is depicted via mean and 95% nonparametric bootstrapped confidence intervals; in the bottom panels, a linear fit is plotted in blue, using R's predict() function to draw a gray 95% confidence band around it. Notice that confidence intervals are wide and overlap across the relevant variables; the confidence bands are also consistent with positively or negatively sloping lines. This suggests that the level of variability masks any potential differences as a function of participant and methodological variables.

We have decided to omit inferential statistics because of several concerns about their interpretation and validity. First, not all data points are mutually independent, which violates basic assumptions of most tests. Second, as the figures highlight, many of the comparisons would be based on 3–5 data points in each cell, which may lead to false negatives (i.e., if we conclude there is no difference in recall/precision/AWC for a potential moderator but, in reality, the test lacked power to detect it) and false positives (i.e., if we conclude there is a difference in recall/precision/AWC for a potential moderator but this is due to a chance drawing of extreme values, which have an overblown effect given very small sample sizes per cell). We encourage readers who would like to perform analyses, which we feel are likely premature, to download the data, available from our online Supplemental Materials (Cristia, Bulgarelli, & Bergelson, 2019).

Discussion

The number of publications using LENA has grown steadily in the preceding decade, with many of these studies including statements that the method has been successfully validated for the American English early childhood sample it has been developed for as well as for other samples. Indeed, a handful of validation studies are scattered in the literature, sometimes as the primary goal of an article, and sometimes as an appendix or footnote. The present

Figure 2. Outcomes by participant moderators. Top left panel: infant language (within panel: The left side shows North American English [NAE], while the right depicts other languages). Top right panel; match of infant population to LENA training sample (within panel: matching samples on the right, mismatching on the left). Bottom panels: infant mean age (bottom left) and infant age range (bottom right). Each point indicates one study; numbers indicate study identity (see Table 1). Filled square points indicate authors affiliated with LENA. *y*-axes indicate the scale for the variable indicated in the panel title (e.g., precision). N. B. axes values vary since different studies may be included across panels, depending on what articles are reported. Red lines indicate bootstrapped confidence intervals (CIs); gray bands in the bottom panel indicate 95% CIs from a linear fit to the data. See text for details and interpretive caveats. AWC = adult word count; LENA = Language Environment Analysis.

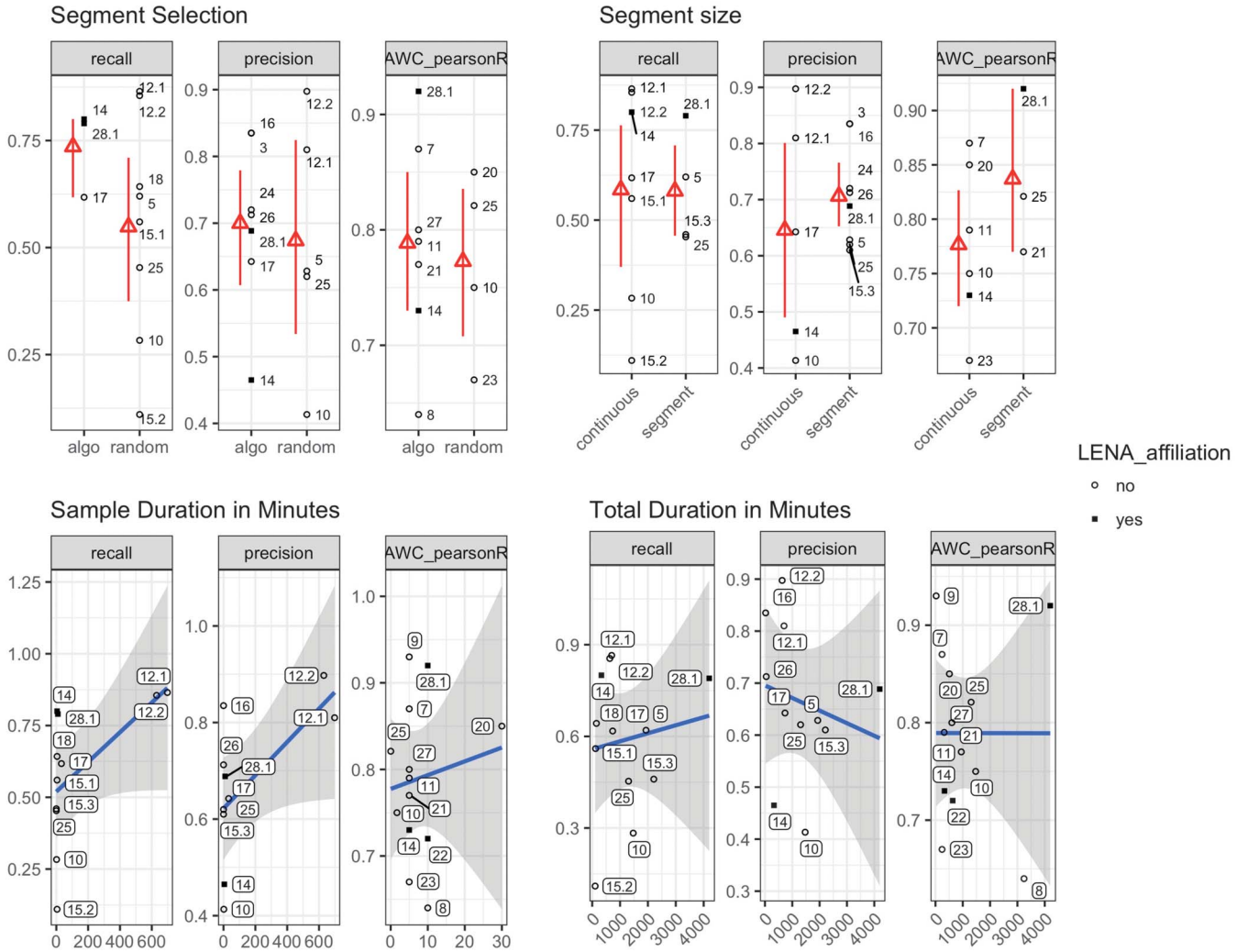


article sought to systematically review and evaluate those studies in order to inform the field on the strength of the evidence for appropriate validation of LENA metrics. Additionally, we used quantitative integration to estimate overall accuracy of the LENA system, and we intended to explore the possibility that its accuracy is moderated by participant and/or methodological factors.

We found a sizable number of studies purporting to report how well the LENA system fared relative to human annotators in terms of classification precision and recall, and/or derived AWCs, CTCs, and CVCs. There were 33 studies appearing in 28 different articles. A systematic review of these data revealed that few studies fully reported all relevant classification information and/or derived counts (with as few as 5 nonindependent points for CVCs). Moreover, few report the full information necessary to assess how exactly the validation took place. For instance, full confusion matrices are virtually never provided, making it impossible to know exactly what was analyzed. That is, authors may have altogether excluded stretches labeled as Silence and Overlap and all "far" labels by LENA, which therefore means that any confusion of the speech categories with these other categories is not counted against the algorithm's accuracy.

To a certain extent, lack of information is likely a side effect of these validations being done in service of a specific research question and thus reported on very briefly,

Figure 3. Outcomes by methodological moderators. Top left panel: segment selection by algo(rithm) (left) versus randomly (right). Top right panel: segment size (continuous [left] vs. single segment [right]). Bottom left panel: duration of individual samples. Bottom right panel: total cumulative annotated data. Each point indicates one study; numbers indicate study identity (see Table 1). Filled square points indicate authors affiliated with LENA. *y*-axes indicate the scale for the variable indicated in the panel title (e.g., precision). N. B. axes values vary since different studies may be included across panels, depending on what articles are reported. Red lines indicate bootstrapped confidence intervals (CIs); gray bands in the bottom panel indicate 95% CIs from a linear fit to the data. See text for details and interpretive caveats. AWC = adult word count; LENA = Language Environment Analysis.



sometimes not in the main manuscript at all. In fact, only 36% of the articles had been peer reviewed as a validation study, and even these did not report all information needed to interpret results in the context of daylong recordings at large. To make this more concrete, consider studies reporting precision and recall. Only four were published in a peer-reviewed journal as validation studies (in Table 1: Bulgarelli & Bergelson, 2019; Gilkerson et al., 2015; Jones et al., 2019; VanDam & Silbert, 2016). To pick on our own work, in Bulgarelli and Bergelson (2019), only segments containing concrete nouns were tagged for speaker identity, leading to precision and recall estimates that may be hard to generalize to daylong data at large, particularly for the key children (who were 6–7 months of age). Specifically, since infants this age are not saying any words, any section that was tagged as them by the LENA system could only count as a false positive; no child utterances that the LENA system was successful in finding would count in its favor.

A more common issue is collapsing across LENA categories (Gilkerson et al., 2015; VanDam & Silbert, 2016), for example, reporting on accuracy of the algorithm when distinguishing children (collapsing between CHN and CXN) from adults (collapsing between FAN and MAN). By collapsing in this way, any confusion within the collapsed categories is not penalized, again inflating accuracy estimations. Moreover, some LENA end users may not realize that this level of accuracy cannot be expected in samples where there are many other children.

A third issue we would like to point out relates to the segmentation, that is, how stretches of the acoustic signal are labeled. It is very common for annotators to hear LENA segments, in (Bulgarelli & Bergelson, 2019; Gilkerson et al., 2015; Jones et al., 2019) or out (VanDam & Silbert, 2016) of context. In the latter case, any interstitial silences and even talker overlap may be ignored by the coders. In the former, it may be easier for annotators to notice that one subsection of the segment is actually the end of the utterance by another speaker, but it is not always the case that annotators can actually signal this and correct for it. For instance, Gilkerson et al. (2015) state in the Method section that coders could correct the LENA segmentation, but in the results, they do not mention how often this was necessary. That said, in their Study 1, Jones et al. (2019) did report that resegmentation was rarely needed.

Moreover, since it is not standard to report all methodological choices and validation results, it is possible that authors (present company included) may have reported only "relevant" analyses or determined whether accuracy was "good enough" for a given purpose, especially when the validation served as a stepping stone to a further question of interest (e.g., Bergelson et al., 2018; Seidl et al., 2018). As in other fields, bias may be deepened by conflicts of interest involving authors and funders, but here, bias could also seep in simply if authors need to justify to reviewers their use of this technology, making it advantageous or more streamlined to combine categories or skip troublesome regions. This falls within the umbrella of "outcome selection bias" in systematic review terms. To be clear, we do not claim any authors are acting maliciously or deceptively, simply that by underreporting validation information, it becomes incredibly difficult to evaluate the evidence base underlying the accuracy of LENA output.

We interpret this evidence as a clear signal that authors publishing validation studies have adopted underinformative habits, as elucidated by our critical consideration of the literature here. We hope this systematic review will lead to a course correction in future work.

With these general considerations in mind, we turn with tempered enthusiasm to a discussion of the otherwise good results we saw in our quantitative integration of reported data: recall and precision higher than 58.80% based on up to 13 nonindependent studies, strong correlations for AWC (mean $r = .79$, on $N = 13$, and a mean RER = 13.76, on $N = 14$), and a similarly high correlation for CVC (mean $r = .77$, on $N = 5$, with a mean RER = −24.17, on $N = 6$). The exception to this general trend toward good performance was CTC, with a mean correlation of $r = .36$, on $N = 6$, and a mean RER = −34.20, on $N = 5$. Given the importance that current theories and descriptions give to LENA measures of conversational turns (e.g., Merz et al., 2019; Romeo et al., 2018), it is important to collect more validation data on CTCs, if possible using unbiased data selection methods (i.e., not using the LENA algorithm segmentation as a starting point) and larger samples than currently used (the mean number of infants in CTC validation studies is 20.60, each represented by 10.28 min of audio). The data on RERs were extremely variable, in part due to variability in implementation.

One important limitation of the present analysis is that recall and precision could only be inspected at a gross level, by averaging across any accuracies that were reported in each study. Thus, while the LENA system could potentially give us insight on many important aspects of naturalistic input, we are not yet in a position to make firm conclusions about its accuracy. For instance, it is extremely relevant to separate input produced by the key child (wearing the recorder) versus other children in the environment for many basic science and clinical applications (including analyzing speech in day cares or families with more than one child). However, only eight nonindependent studies report recall and precision on the CHN label. Moreover, even though some of them omit the confusions between the CHN category and nonspeech categories (such as Silence and Overlap), which likely inflate accuracy, current results show a mean recall of 56% and a mean precision of 55.88%. Similarly, categories such as Silence, Overlap, and Electronics and subcategories within the CHN (speech vs. vegetative sounds and crying) are so rarely included in validation efforts as to preclude their inclusion in this systematic review altogether.

The results above can also help guide our field's approach in uncovering the stability of LENA accuracy across participant moderators. Specifically, the missingness we note leads to several clear recommendations of specific areas in which validation data are sparse. In examining accuracy as a function of whether the sample matches the sample LENA algorithms were trained on, we found that more work is needed particularly in terms of language. There were very few data points on labels' recall and precision for samples that were not exposed to NAE (non-NAE; e.g., for recall, there were four non-NAE vs. eight NAE). The opposite occurs for AWC, with only three studies on NAE as opposed to 10 studies on non-NAE. We therefore recommend more studies evaluating non-NAE samples for recall and precision of speech classes and more NAE samples evaluating AWC accuracy. Regarding mean age, there is a trend for lower recall for data sampled from older children, with no such tendency for precision or AWC. Most data points, however, cluster in the first 2 years of life. Therefore, we recommend continued efforts to assess LENA accuracy with older children.

For authors who are inspired by the wide space of still-important validation to be done, and for those who are already conducting validation as part of their research, we would like to recommend that they follow standardized procedures to facilitate comparisons across articles and reduce the impact of algorithm and researcher bias. Some recommendations have been laid out in other work (Cristia, Lavechin, et al., 2019), including (a) a preference for coding continuous clips rather than individual segments, (b) sampling semirandomly to ensure that selection is not biased by the algorithm, and (c) using the DARCLE Annotation Scheme to increase comparability of the annotation (Casillas et al., 2017).

Regardless of how data are sampled and annotated, we strongly encourage authors of validation work to make the

raw validation data publicly available. Data containing automated and manual categorical and numerical information are intrinsically deidentified and thus bypass most of the usual privacy and ethics-based concerns that apply to the audio data itself. The simplest way to share them is through websites such as osf.io, which in a matter of minutes allows you to create a public-facing project (with the option of private components as well), containing a digital object identifier that can then be readily cited by others. This provides the original authors with "credit" for their validation data, which are separate from the specific research questions addressed in their original empirical reports. In addition to being beneficial to authors, sharing such data would allow others to revisit them and calculate statistics and metrics not considered by the original authors. For instance, as mentioned above, validating CTC requires identifying adults separately from children, and thus data originally created to validate CTC can be re-used to validate the system's categorization into child and adult voices. The current systematic review already benefited from being able to calculate metrics not reported in original articles, thanks to authors' posting their data and code (Bergelson et al., 2018; Bulgarelli & Bergelson, 2019). Moreover, such data sharing is also useful for measures that can be computed in different ways, such as error rates. Shared data would allow meta-analysts to explore various implementations of such error rates with no effort or time required from the original authors.

Finally, as reviewers and editors, we should give articles reporting nothing but a validation effort our full attention: If the field continues to use the LENA technology for both intervention and basic science (e.g., Sosa, 2016; Suskind et al., 2016), it becomes crucial to independently establish its validity. In addition, acceptance should be based on the validation approach and not on the ensuing results, to ensure that studies showing low validity with high-quality methods have a chance of entering the literature. We hope, in particular, that this systematic review serves as an opportunity for those using LENA to take our freely and openly available data, code book, and scripts to make sure their validity experiments using the automated LENA measures or other analyses of naturalistic recordings can be integrated into meta- or mega-analyses. In the future, another systematic review that is preregistered in an available database would then be in a better position to present the most unbiased estimation of the LENA algorithms and other similar systems that may emerge.

Acknowledgments

This work was supported by a Trans-Atlantic Platform "Digging into Data" collaboration grant (Analyzing Child Language Experiences Around The World), with the support of Agence Nationale de la Recherche (ANR-16-DATA-0004 Analyzing Child Language Experiences Around The World; in addition to ANR-17-CE28-0007, ANR-14-CE30-0003 MechELex, and ANR-17-EURE-0017) and the National Endowment for the Humanities (HJ-253479-17), as well as funding from the J. S. McDonnell Foundation (awarded to A. C.) and National Institutes of Health Grant DP5-OD019812 (awarded to E. B.). The funders had no impact on this systematic review. We are grateful to members of the DARCLE Consortium for comments on previous versions of this work.

References

Adams, K. A., Marchman, V. A., Loi, E. C., Ashland, M. D., Fernald, A., & Feldman, H. M. (2018). Caregiver talk and medical risk as predictors of language outcomes in full term and preterm toddlers. *Child Development, 89*(5), 1674–1690. https://doi.org/10.1111/cdev.12818

Berends, C. (2015). *The LENA system in parent–child interaction in Dutch preschool children with language delay* (Master's thesis). Utrecht University, Utrecht, the Netherlands.

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2018). What do North American babies hear? A large-scale cross-corpus analysis. *Developmental Science, 22*(1), 1–12. https://doi.org/10.1111/desc.12724

Bredin-Oja, S. L., Fielding, H., Fleming, K. K., & Warren, S. F. (2018). Clinician vs. machine: Estimating vocalizations rates in young children with developmental disorders. *American Journal of Speech-Language Pathology, 27*(3), 1066–1072. https://doi.org/10.1044/2018_AJSLP-17-0016

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

Burgess, S., Audet, L., & Harjusola-Webb, S. (2013). Quantitative and qualitative characteristics of the school and home language environments of preschool-aged children with ASD. *Journal of Communication Disorders, 46*(5–6), 428–439. https://doi.org/10.1016/j.jcomdis.2013.09.003

Busch, T., Sangen, A., Vanpoucke, F., & van Wieringen, A. (2018). Correlation and agreement between Language ENvironment Analysis (LENA^{TM}) and manual transcription for Dutch natural language recordings. *Behavior Research Methods, 50*(5), 1921–1932. https://doi.org/10.3758/s13428-017-0960-0

Canault, M., Le Normand, M.-T., Foudil, S., Loundon, N., & Thai-Van, H. (2016). Reliability of the Language ENvironment Analysis system (LENATM) in European French. *Behavior Research Methods, 48*(3), 1109–1124. https://doi.org/10.3758/s13428-015-0634-8

Casillas, M., Bergelson, E., Warlaumont, A. S., Cristia, A., Soderstrom, M., VanDam, M., & Sloetjes, H. (2017, August 20–24). *A new workflow for semi-automatized annotations: Tests with long-form naturalistic recordings of childrens language environments* [Conference proceeding]. Paper presented at Interspeech 2017, Stockholm, Sweden. https://doi.org/10.21437/Interspeech.2017-1418

Caskey, M., Stephens, B., Tucker, R., & Vohr, B. (2014). Adult talk in the NICU with preterm infants and developmental outcomes. *Pediatrics, 133*(3), e578–e584. https://doi.org/10.1542/peds.2013-0104

Cristia, A., Bulgarelli, F., & Bergelson, E. (2019). *Accuracy of the language environment analysis system: A systematic review—Supplemental materials*. https://osf.io/4nhms/

Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C. F., Räsänen, O., Bunce, J., & Bergelson, E. (2019). *A thorough evaluation of the Language Environment Analysis (LENA) system. 1–30*. https://doi.org/10.1017/CBO9781107415324.004

D'Apice, K., Latham, R. M., & von Strumm, S. (2019). A naturalistic home observational approach to children's language, cognition, and behavior. *Developmental Psychology, 55*(7), 1414–1427. https://doi.org/10.1037/dev0000733

Elo, H. (2016). *Acquiring language as a twin: Twin children's early health, social environment and emerging language skills* (Dissertation). Tampere University Press. http://urn.fi/URN:ISBN:978-952-03-0296-2

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. https://doi.org/10.1177/1525740117705094

Gilkerson, J., Coulter, K. K., & Richards, J. A. (2008). *Transcriptional analyses of the LENA natural language corpus* (Technical Report LTR-06-2). LENA Foundation.

Gilkerson, J., Zhang, Y., Xu, D., Richards, J. A., Xu, X., Jiang, F., Harnsberger, J., & Topping, K. (2015). Evaluating language environment analysis system performance for Chinese: A pilot study in Shanghai. *Journal of Speech, Language, and Hearing Research, 85*(2), 445–452. https://doi.org/10.1044/2015_JSLHR-L-14-0014

Greenwood, C. R., Thiemann-Bourque, K., Walker, D., Buzhardt, J., & Gilkerson, J. (2011). Assessing children's home language environments using automatic speech recognition technology. *Communication Disorders Quarterly, 32*(2), 83–92. https://doi.org/10.1177/1525740110367826

Jones, R. M., Plesa Skwerer, D., Pawar, R., Hamo, A., Carberry, C., Ajodan, E. L., Caulley, D., Silverman, M. R., McAdoo, S., Meyer, S., Yoder, A., Clements, M., Lord, C., & Tager-Flusberg, H. (2019). How effective is LENA in detecting speech vocalizations and language produced by children and adolescents with ASD in different contexts? *Autism Research, 12*(4), 628–635. https://doi.org/10.1002/aur.2071

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. https://doi.org/10.1017/S0305000915000203

Lehet, M., Arjmandi, M. K., Dilley, L. C., & Houston, D. (2019). *Accuracy of the Language ENvironment Analysis (LENA) system for quantifying adult speech experienced by infants in naturalistic settings* [Manuscript submitted for publication]. Department of Communicative Sciences and Disorders, Michigan State University.

McCauley, A., Esposito, M., & Cook, M. (2011). *Language environment analysis of preschoolers with autism: Validity and application*. Poster session presented at LENA Users Conference 2011, Denver, CO, United States.

Merz, E. C., Maskus, E. A., Melvin, S. A., 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*. Advance online publication. https://doi.org/10.1111/cdev.13239

Oetting, J. B., Hartfield, L. R., & Pruitt, S. L. (2009). Exploring LENA as a tool for researchers and clinicians. *The ASHA Leader, 14*(6), 20–22. https://doi.org/10.1044/leader.FTR3.14062009.20

Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., Yapanel, U., & 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*(30), 13354–13359. https://doi.org/10.1073/pnas.1003882107

Orena, A. J., Byers-Heinlein, K., & Polka, L. (2019). Reliability of the language environment analysis recording system in analyzing French–English bilingual speech. *Journal of Speech, Language, and Hearing Research, 62*(7), 2491–2500. https://doi.org/10.1044/2019_JSLHR-L-18-0342

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*(6), 549–569. https://doi.org/10.1177/0142723716649273

Richards, J. A., Gilkerson, J., Paul, T. D., & Xu, D. (2008). *The LENA^{TM} automatic vocalization assessment* (Technical Report LTR-08-1). LENA Foundation.

Romeo, R. R., Segaran, J., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Yendiki, A., Rowe, M. L., & Gabrieli, J. D. E. (2018). Language exposure relates to structural neural connectivity in childhood. *The Journal of Neuroscience, 38*(36), 7870–7877. https://doi.org/10.1523/JNEUROSCI.0484-18.2018

Schwarz, I.-C., Botros, N., Lord, A., Marcusson, A., Tidelius, H., & Marklund, E. (2017). *The LENA^{TM} system applied to Swedish: Reliability of the adult word count estimate*. Paper presented at Interspeech 2017. https://doi.org/10.21437/Interspeech.2017-1287

Seidl, A., Cristia, A., Soderstrom, M., Ko, E.-S., Abel, E. A., Kellerman, A., & Schwichtenberg, A. J. (2018). Infant–mother acoustic–prosodic alignment and developmental risk. *Journal of Speech, Language, and Hearing Research, 61*(6), 1369–1380. https://doi.org/10.1044/2018_JSLHR-S-17-0287

Soderstrom, M., & Franz, W. (2016). *Comparing human- and machine-annotated language input across childcare settings* [Conference session]. Paper presented at ICIS 2016, Dublin, Ireland.

Sosa, A. V. (2016). Association of the type of toy used during play with the quantity and quality of parent-infant communication. *JAMA Pediatrics, 170*(2), 132–137. https://doi.org/10.1001/jamapediatrics.2015.3753

Suskind, D. L., Leffel, K. R., Graf, E., Hernandez, M. W., Gunderson, E. A., Sapolich, S. G., Suskind, E., Leininger, L., Goldin-Meadow, S., & Levine, S. C. (2016). A parent-directed language intervention for children of low socioeconomic status: A randomized controlled pilot study. *Journal of Child Language, 43*(02), 366–406. https://doi.org/10.1017/S0305000915000033

Suskind, D. L., Leffel, K. R., Hernandez, M. W., Sapolich, S. G., Suskind, E., Kirkham, E., & Meehan, P. (2013). An exploratory study of "quantitative linguistic feedback". *Communication Disorders Quarterly, 34*(4), 199–209. https://doi.org/10.1177/1525740112473146

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), Article e12456. https://doi.org/10.1111/desc.12456

van Alphen, P., Meester, M., & Dirks, E. (2017). *LENA onder de loep [LENA under scrutiny]*. VHZ Artikelen, 2017(April), 14-19.

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

Weisleder, A., & Fernald, A. (2013). Talking to children matters. *Psychological Science, 24*(11), 2143–2152. https://doi.org/10.1177/0956797613488145

Wood, C., Diehm, E. A., & Callender, M. F. (2016). An investigation of language environment analysis measures for Spanish–English bilingual preschoolers from migrant low-socioeconomic-status backgrounds. *Language, Speech, and Hearing Services in Schools, 47*(2), 123–134. https://doi.org/10.1044/2015_LSHSS-14-0115

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