DOI: 10.1111/infa.12568

INFANCY THE OFFICIAL JOURNAL OF THE INTERNATIONAL CONGRESS OF INFANT STUDIES WILEY

RESEARCH ARTICLE

# Characterization of children's verbal input in a forager-farmer population using long-form audio recordings and diverse input definitions

**Camila Scaff**[1,2] | **Marisa Casillas**[3] | **Jonathan Stieglitz**[4] | **Alejandrina Cristia**[2]

[1]University of Zurich, Institute of Evolutionary Medicine (IEM), Zurich, Switzerland

[2]PSL University, Laboratoire de Sciences Cognitives et de Psycholinguistique (ENS, EHESS, CNRS, DEC), Paris, France

[3]University of Chicago, Comparative Human Development, Chicago, Illinois, USA

[4]Institute for Advanced Study in Toulouse (IAST), Toulouse, France

**Correspondence**
Camila Scaff.
Email: camillescaff@gmail.com

**Funding information**
James S. McDonnell Foundation, Grant/Award Number: Understanding Human Cognition Scholar Award; H2020 European Research Council, Grant/Award Number: ExELang, Grant agreement No. 101001095; Agence Nationale de la Recherche, Grant/Award Numbers: ANR-14-CE30-0003 MechELex, ANR-16-DATA-0004 ACLEW, ANR-17-EURE-0010, ANR-17-EURE-0017; NWO Veni Innovational Research Scheme, Grant/Award Number: 275-89-033

## Abstract

There is little systematically collected quantitative empirical data on how much linguistic input children in small-scale societies encounter, with some estimates suggesting low levels of directed speech. We report on an ecologically-valid analysis of speech experienced over the course of a day by young children ($N = 24$, 6–58 months old, 33% female) in a forager-horticulturalist population of lowland Bolivia. A permissive definition of input (i.e., including overlapping, background, and non-linguistic vocalizations) leads to massive changes in terms of input quantity, including a quadrupling of the estimate for overall input compared to a restrictive definition (only near and clear speech), while who talked to and around a focal child is relatively stable across input definitions. We discuss implications of these results for theoretical and empirical research into language acquisition.

The editor of this article is Gavin Bremner.

# 1 | INTRODUCTION

Describing the quantity and sources of linguistic input that young children receive is important for theory development (e.g., Cristia, 2020; Loukatou et al., 2022). Here we investigate children's speech environments among the Tsimane', a forager-horticulturalist population in the Bolivian Amazon, employing ecological child-centered long-form audio-recordings, and several definitions of linguistic input. In what follows, we briefly review relevant studies investigating children's language input to contextualize the current study's methodological and analytical approach, and to motivate our research questions.

Although qualitative aspects of input are unambiguously important, in this paper we focus on quantity. The question may appear simple at first: How much speech do children hear? The answer, however, is not simple. Some work suggests that there is high variation in the prevalence of child-directed speech. For example, a recent systematic review estimated that children growing up in urban settings (in post-industrialized, larger, and richer societies, where formal schooling is compulsory) hear ∼8 min/h of vocalization directed to them (henceforth, TCD vocalizations), whereas children in rural settings (which also happened to be from small-scale, subsistence farming or hunter-gatherer societies in the summarized literature) hear one third of that (Cristia, 2022). The review focused on estimates emerging from systematic behavioral observations: The observer follows the infant or child around (for, e.g., 2 h), and notes whether anyone vocalized towards the child during an observation period (e.g., every 15 s). A similar methodology was employed for the only study attempting to estimate input quantity and sources among the Tsimane' (Cristia et al., 2019), resulting in an estimate of about 1 min/h, with no changes as a function of age, and with speech coming mostly from the mother for children under 3 years of age.

Such behavioral observations, however, are not ideal to measure input quantity and sources, as it is possible that non-salient conversations may escape the observer's attention. Instead, the vast majority of studies on language acquisition have employed short video-recordings, in which a researcher films a child for around 30–60 min in their natural environment. In this line of work, we focus on two studies comparing urban against rural settings, which report large (i.e., 3- to 11-fold) differences in the quantity of child-directed speech, one based on a comparison between North American and Yucatec Mayan children (Shneidman & Goldin-Meadow, 2012) and the other comparing children growing up in rural and urban Mozambique (Vogt et al., 2015). Both papers report on the number of utterances directed to children as coded from video-recordings. Although this method may capture child-centered language experiences more precisely than the aforementioned behavioral observations, current evidence suggests that such video-recordings lead to overestimation of North American children's input quantities (see Bergelson, Amatuni, et al., 2019 for discussion) and to an underestimation of Yucatec Mayan children's input quantities (Shneidman et al., 2013). Bergelson, Amatuni, et al. (2019) found that families talked a great deal more when video-recorded than when a long-form recording was used, and that words referring to the recording equipment were among the 10 most frequent words in the former but not the latter. Based on these observations, the authors argue for the greater ecological validity of long-form recordings over short video-recordings, in which families' attention is more drawn to the equipment and the situation, in which families' attention is more drawn to the equipment and the situation. In contrast, Shneidman et al. (2013) counted fewer vocalizations when the investigator video-recorded than when the investigator was counting vocalizations on the fly without a camera, and attributed this to differing comfort level in being observed by an outsider. We therefore focus on the much smaller number of studies using long-form recordings, in which children's verbal input is described based on recordings lasting over 2 hours, and typically over a whole waking day. These studies are more relevant both because they produce less biased estimates of input quantities (and potentially sources), and because this is the technique used in the present study, so that their results can be more easily compared with our own.

The most salient study is one by Bunce and colleagues (Bunce et al., 2021), which contains estimates drawn from five populations (including also data separately published in Casillas et al., 2020; Casillas et al., 2021), with infants learning (mainly or exclusively) English in North America,[1] English in the UK, Spanish in Argentina, Tseltal in Mexico, and Yélî Dnye in Papua New Guinea (for more information on the corpora, see Bunce et al., 2021; Soderstrom et al., 2021). Native speakers of each language helped annotate the data, resulting in estimates of speech directed to the target child (i.e., the child who wears the recording device) from both adults and other children. The English-learning infants had about 3.5 min/h (North America, range 0–10.1) and 3.7 min/h (UK, range 1.2–7.2) of child-directed speech, which was similar to the TCD speech available to Tseltal learners (3.6 min/h, range 0.8–6.6), and somewhat higher than that to Yélî learners (3.1 min/h, range 1.6–6.3). Notice saliently that, counter to evidence from behavioral observations and video-recordings, estimates for urban and rural populations are very similar here. The highest amount of TCD speech was found in the Argentinean sample (4.8 min/h, range 1.4–9.4). However, audio-recordings in the latter sample were 3–4 h in length. The fact that a researcher dropped off the equipment and came back only a couple of hours later may have rendered families more conscious of the recordings than in longer recordings, affecting that estimate.

It is worth mentioning that, in separate publications, Casillas and colleagues report more detailed results for 10 Tseltal children (Casillas et al., 2020) and 10 Yélî children (Casillas et al., 2021). In those publications, TCD speech is broken down as a function of the source, with 80% and 72%, respectively, coming from adults rather than other children.[2] They also provide relatively high estimates of how much speech is overhearable by children (21.1 min/h and 35.9 min/h, respectively, with significant decreases with child age). When reading closely Bunce, Casillas, and colleagues' work, it is clear that their definitions of input were as inclusive of speech experienced by the child as possible: Faint or far away speech was included and even utterances that overlapped across talkers (including with the target child) were counted separately (i.e., if two people talked at the same time for a second, this would result in 2 seconds of speech being considered). What counts towards children's input is defined in a wholly different way in another study using the LENA automated algorithm (Gilkerson et al., 2017), which only considers near and clear speech from adults, discarding faint speech, speech that overlaps with other sounds, and even speech from other children. Gilkerson and colleagues studied 329 North American English learners (aged 2–48 months), and their data suggests children hear an average of 3.5 min/h of speech in total,[3] including both child-directed and overheard speech.[4]

Somewhere in between these two definitions of what counts as input, Weisleder and Fernald (2013) also used near and clear speech from adults from the automated LENA algorithm, but complemented this with exhaustive listening of every 5-min section in every recording from 29 Spanish-learning 19-month-old infants growing up in the USA, to decide whether speech in each 5-min block was mostly child-directed or overheard (Weisleder & Fernald, 2013). On average, children were exposed to about

[1] Another estimation of North American children's input using a subset of the corpora included in Bunce et al. (2021) was published previously (Bergelson, Casillas, et al., 2019). However, the sampling strategy in the earlier paper led to a gross overestimation of children's input (Bunce et al., 2021). We therefore do not discuss it further.

[2] Note that in the Yélî sample, a significant increase with age was found for speech from other children, and the percentage of TCD coming from other children was higher for older than younger target children. We also note that the literature distinguishes "directed" from "overheard". One can argue that the latter is misleading, as we cannot be sure that target children really paid attention and overheard what was said, so the term "overhearable" may be preferable. For compatibility with previous work, however, we typically use overheard.

[3] They estimated that children encountered an average of 1025 adult words per hour, which amounts to about 3.5 min per hour (henceforth min/h) of total input (assuming each word is about 200 milliseconds long, based on Supplementary Materials of Cristia et al., 2019).

[4] The authors report that speech quantity is higher for the younger infants in the sample (under 4 months).

2 min/h of directed speech, and 1.3 min/h of overhearable speech.[5] The sum of these two (3.3 min/h) is very similar to Gilkerson's total estimate for English-learning infants. Although the number is also similar to that reported for the North American corpora in Bunce et al., the two papers are based on different input definitions: Bunce and colleagues excluded overheard speech but included overlapping and faint speech as well as speech from any speaker type, whereas Gilkerson and colleagues included overheard speech but counted only near and clear speech from adults.

In sum, the search for an answer to "what is the input to language acquisition?" starts with defining input. However, the field has not established a single definition because of the uncertainty of what actually counts as input from the child's point of view. Diverse definitions can be arrived at depending on a number of factors that we introduce next.

Regarding the sources of input, it has been recently argued that the developmental literature is culturally biased in that it often focuses on mothers' speech (Loukatou et al., 2022), despite the fact that comparative cross-cultural research suggests the mother was the primary caregiver in only 60% of 158 societies (Barry & Paxson, 1971). Research on who talks to the child (and whether this impacts the child) is scarce, but the work of Shneidman and colleagues is important in that it highlights correlations between words available in the input and the children's own vocabulary. According to this research, considering speech from people other than the mother improves prediction of children's vocabulary (Shneidman et al., 2013), suggesting that speech from other adults should be considered as part of children's effective input, whereas evidence is mixed for speech from other children (e.g., Shneidman et al., 2013; Shneidman & Goldin-Meadow, 2012). Another obvious divergence pertains to the addressee of the speech. Correlational studies show stronger positive associations between children's vocabulary skills and adult TCD speech, as opposed to overheard or total verbal input from adults or other children (Shneidman & Goldin-Meadow, 2012; Weisleder & Fernald, 2013). Indeed, some researchers argue that adult TCD verbal input may not just be helpful, but actually necessary for language acquisition (Golinkoff et al., 2015). However, others provide evidence for robustness of learning some aspects of language across wide variation in the quantity of TCD speech (e.g., Cristia, 2020; Ochs & Schieffelin, 2011) and even of learning words from overheard speech (e.g., Akhtar et al., 2001; Foushee & Srinivasan, 2023).

There is also some variation in terms of how the addressee of a given vocalization is determined, with the main divide being whether the content of the speech is used or not. In the former case, human annotators familiar with the language must be involved in the transcription, and they also need to be able to interpret the situation, to which end familiarity with the families may help (Casillas et al., 2020, 2021). This is also only feasible for relatively smaller corpora, for instance, those that rely on extracts from long-form recordings. When such manual annotation is not possible, researchers have turned to contextual cues: In the literature building on LENA software, the distinction is made between measures that include all adult speech and those that only count adult vocalizations that occur temporally close to children's own vocalizations (e.g., Ellwood-Lowe et al., 2022). Adult vocalizations not occurring in close proximity to the child's production are considered as merely overheard. A recent methodological study explored the unique characteristics of TCDS versus overheard speech (Bang et al., 2023). Clips were annotated as TCDS or overheard by human listeners, and these annotations were used to train a classifier. This revealed that, after silence or noise periods, adult-child adjacent vocalizations were the next most important feature distinguishing TCDS from overheard sections. Thus, contextual cues can be informative and indicate a greater likelihood of TCDS being employed (i.e., when child and adult vocalizations alternate, that segment is more likely to contain child-directed speech).

[5] We digitized their Figure 1 to find an average of 595 words per hour directed, and 995 words per hour total, resulting in 400 overheard words, then assumed 200 milliseconds per word to derive the min/h estimation.

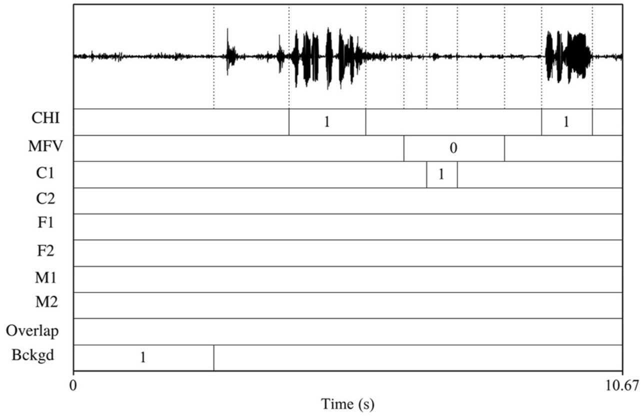

FIGURE 1 Schematic representation depicting a 10-s segment of a Praat textgrid annotation, extracted from a randomly selected one-minute audio clip. CHI stands for the target child; MFV main female voice. Separate tiers are used to encode vocalizations by other children (C), female adults (F) and male adults (M), two for each. Finally, tiers are used to encode overlapping speech that cannot be segmented by two or more speakers (Overlap) and background speech (Bckgd).

An additional dimension is introduced when overheard speech is considered. Some overheard speech is distant or backgrounded, in which case it often proves impossible for an annotator to identify what is being said or even who is talking. Even if an audio recorder cannot pick up a quality signal to enable an annotator to discern speaker identity and speech contents, a target child could nonetheless learn from or attend to such distant speech (e.g., by turning their head toward the speaker to improve signal quality or by using visual cues). This difference in how the infant versus the annotator processes the event may introduce noise in the transcription, casting a shadow of doubt over studies suggesting that overheard speech may not impact vocabulary learning (Shneidman et al., 2013).

The studies summarized above diverge in their treatment of overlapping speech, with some excluding it (Gilkerson et al., 2017; Weisleder & Fernald, 2013) and others including it, and even counting the whole duration multiple times (Casillas et al., 2020; Casillas et al., 2021; see SM4 for table summarizing input quantities in previous work). Some researchers have proposed that capacities for sound source segregation are already functional in infancy (Demany, 1982; Werner, 2002), though to what extent they are developed and actively utilized throughout early childhood remains unclear (see Saffran et al., 2007). Ultimately, when considering whether overheard and overlapping speech "count" as verbal input to children from a cross-cultural perspective, one must consider the circumstances that enable overheard and overlapping speech to occur: Overheard and overlapping speech will be more common in communities where families are larger, community spaces are shared, and where natural materials that promote the transmission of sound are used. Moreover, there is likely cultural variation in terms of how pragmatically acceptable overlapping speech is (Cecil, 2010; Tannen, 2012).

Finally, vocalization type is also relevant in considering what counts as input. Non-linguistic vocalizations (e.g., laughing and crying) are often excluded from verbal input counts (Gilkerson et al., 2017; Shneidman & Goldin-Meadow, 2012). In contrast, because the literature on child-caregiver bonding often considers all vocalizations as being potentially reinforcing, many studies include both non-linguistic and linguistic vocalizations in input counts (Casillas et al., 2020, 2021; Cristia et al., 2019; Mallinckrodt, 1992; Vogt et al., 2015).

In sum, numerous reasonable definitions exist for what counts as verbal input. One can think of variations along each of the dimensions reviewed above (i.e., TCD vs. overheard, near vs. background,

TABLE 1 Two extreme input definitions defining what counts as input (see Methods for details).

| Definition | Restrictive | Permissive |
|---|---|---|
| Source | MFV, other adults, other children | MFV, other adults, other children |
| TCD | Target-child adjacent | Target-child adjacent + monologues |
| Overheard | All other non-target-child vocalizations | Non-target-child vocalizations that alternate with vocalizations by other non-target-child sources |
| Total | TCD + overheard | TCD + overheard |
| Faint or far-away | Excluded | Included |
| Overlap | Excluded | Included |
| Vocalization type | Only speech | Speech and non-speech (cry, laugh) |

*Note*: TCD stands for Target-child-directed. Target-child adjacent refers to vocalizations occurring within 2 seconds of a Target-child vocalization. Monologues refers to vocalizations that do not occur within 2 seconds of vocalizations by others (e.g., MFV vocalization preceded and followed by silence, or preceded and/or followed by MFV vocalizations).

non-overlapping vs. overlapping, linguistic vs. communicative) as defining a range of input definitions. In the present work, we will consider two extreme definitions (see Table 1): the most restrictive input definition, inspired by work employing the LENA software (Gilkerson et al., 2017; Weisleder & Fernald, 2013) and the most permissive input definition, inspired by work employing manual annotations (e.g., Bunce et al., 2021; Casillas et al., 2021). This allows us to do two important things. First, we can provide estimates that are more easily compared with those reported by previous work, subsetting the restrictive estimates to adult speech only to compare against Gilkerson and colleagues, and the permissive estimates to compare against Casillas and colleagues. Second, we provide a first bridge across these two definitions that currently divide the literature, highlighting the variability in potential answers to our overarching question of "what is the input to language acquisition?".

## 1.1 | The present work

Using long-form audio recordings of young children, we seek to answer two key questions, laid out below. We investigate them in a sample of Tsimane' children, a population for which an estimate of TCD speech exists, albeit based on systematic behavioral observations (Cristia et al., 2019), which are insufficient for the reasons we summarized above. In addition to contributing a more accurate estimate of Tsimane' children's language input, we sought to bridge the gap between current extreme definitions of input: one that provides a maximally conservative estimate of input (restrictive) and the other that yields a maximally inclusive estimate (permissive). As noted above, current psycholinguistic literature does not have a final answer regarding which is more appropriate. By considering these two extreme input definitions, we can assess which answers are robust to different input definitions. In the latter case, answers must be considered tentative until children's information uptake is established. The questions to which we seek answers are:

1. How much verbal input is available to young children? We consider not only TCD but also overheard input. For completeness, we report on the proportion of TCD relative to total verbal input, but we do not discuss this variable as its cognitive significance is currently unclear.
2. What are the sources of verbal input (i.e., who speaks to and around the child)?

Regarding question 1, while we obviously expect input quantities to be smaller in the restrictive than the permissive input definitions, we have no ex ante predictions regarding how much smaller. We

TABLE 2 Descriptive statistics for study participants.

| | All (*N* = 24) | | | | Males (*N* = 16) | | | | Females (*N* = 8) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | M | Mdn | Min | Max | M | Mdn | Min | Max | M | Mdn | Min | Max |
| Age (months) | 30.3 | 32.0 | 6.0 | 58.0 | 33.9 | 35.5 | 8.0 | 58.0 | 23.1 | 22.5 | 6.0 | 43.0 |
| Birth order | 3.1 | 3.0 | 1.0 | 7.0 | 3.2 | 3.0 | 1.0 | 6.0 | 2.9 | 2.5 | 1.0 | 7.0 |
| Household size | 8.1 | 7.0 | 3.0 | 13.0 | 8.4 | 8.0 | 4.0 | 13.0 | 7.5 | 6.0 | 3.0 | 13.0 |
| # Siblings | 2.9 | 3.0 | 0.0 | 6.0 | 3.2 | 3.0 | 1.0 | 6.0 | 2.2 | 2.0 | 0.0 | 6.0 |
| Rec. (hours) | 13.8 | 16.0 | 5.5 | 17.4 | 13.4 | 16.0 | 5.0 | 16.0 | 14.5 | 16.0 | 5.5 | 17.4 |
| Annotated mins. | 12.7 | 15.0 | 4.0 | 16.0 | 12.4 | 15.0 | 4.0 | 15.0 | 13.4 | 15.0 | 4.0 | 16.0 |

*Note*: M stands for Mean, Mdn for Median, Min for Minimum, Max for Maximum. "Household size" indicates total number of co-resident individuals. "Rec." indicates recording length in hours. "Annotated mins." shows the number of daytime minutes that were annotated. "# siblings" indicates the number of siblings.

predict similar amounts of TCD input as in previous work among the Tsimane' (Cristia et al., 2019), about 1 min/h. That paper also reported on other-directed and undirected input, which they estimated at about 7 min/h, leading us to a prediction of about 8 min/h total input.

Regarding question 2, we do not expect composition in terms of sources to differ across restrictive versus permissive input definitions. Based on previous similar work (Casillas et al., 2020, 2021; Cristia et al., 2019), we expect adults to be a primary source of TCD input.[6]

## 2 | METHODS

The Tsimane' live in villages ranging in size from 50 to 550 individuals. Villages are composed of extended family clusters, in which the majority of food and labor sharing occurs. Verbal exchanges are usually in the native Tsimane' language, but Spanish may be spoken to non-Tsimane' Bolivians (e.g., merchants), and Spanish is more frequent in villages in closer proximity to the main nearby town of San Borja.

On average, women have their first child by 19 years of age, with an interbirth interval averaging 33.7 months (Stieglitz et al., 2019) and a total fertility rate of about nine births (Kaplan et al., 2015). Infants are kept close to their mothers, and are regularly carried in a sling so that mothers can perform subsistence activities and on-demand breastfeeding. Toddlers are often cared for by older siblings or other kin. Tsimane' mothers provide ~80% of direct child care in the first 6 months of life, and ~70% in the first 6 years (Winking et al., 2009).

Data were collected by the first author in July 2017 from 25 children from 15 families in one village. One child's recording was lost, leading to a total of 24 children (age range: 6–58 months; 33% female; see Table 2 for participant demographics). All resident families with a child under 4 years of age were invited to participate in the study; siblings of target children under 6 years of age were also offered participation.

Institutional IRB approval was granted by the University of New Mexico (HRRC # 17–262). Informed consent was obtained at three levels: (1) the Tsimane' government that oversees research projects, (2) village leadership, and (3) the parent or guardian for each child.

[6] Initial exploration including age did not reveal it as a significant predictor (see Table S3 in the Supplementary Materials for sample results). Although this null result is aligned with the Tsimane' behavioral observations results (no change in input quantity across the first 4 years of life, see Cristia et al. (2019) for more details), and with previous results combining several corpora (Bunce et al., 2021), it may also reveal power limitations.

Families were visited by the first author and a bilingual (Tsimane'-Spanish) research assistant in the morning. Recordings began in participants' own homes to minimize participant burden and influence of the investigator's presence on recorded speech. All co-resident eligible children were recorded on the same day. We used three different brands of recording devices (i.e., LENA, Olympus, and USB Esky recorders; see Supplementary Material Table S1 for more information) because at the time we were unsure which device would perform best in this tropical environment, which had not been considered in previous research at the time. We randomly selected a recorder for each child based on its availability on the day of recording. After explaining study procedures to the parent(s), each child was fitted with a customized T-shirt (for Olympus and USB) or LENA vest, in whose breast pocket we placed a recording device (see SM2). After fitting the customized clothing on a target child and ensuring the child's comfort, the researchers left the residence. The researchers returned to the residence the following day to collect the equipment. The start time of recordings was between 7 a.m. and 11 a.m.

We did not rely on automated annotations from LENA because that would not have allowed us to analyse the recordings of all children, but only the minority recorded using a LENA device. Instead, recordings were manually annotated using Praat acoustic analysis and annotation software (Boersma & Heuven, 2002). We used a periodic sampling scheme to code 1 minute of each recorded hour after discarding the first 30 min of the recording. We bypassed the initial 30 min to allow the researchers sufficient time to leave the residence and to let participants and their families acclimate to the recorder. We do not know how long it takes participants to acclimate, but to the best of our knowledge, previous research on long-form recordings does not exclude an initial portion of the data (despite it likely being most sensitive to observer effects), so a 30-minute warm-up duration appears conservative. After this first 30 min, the first 5 minutes of each recording hour were extracted, and the last minute of each 5-minute clip was manually annotated. The remaining minutes of each 5-minute clip provided context for the annotated minute (e.g., recognizing participants). The decision to annotate 1 minute was motivated by our limited annotation resources and the structure of speech in long-form recordings. Speech is likely to be "bursty" or clumped in time: If a given point in time contains speech, it is likely that neighboring points in time do too, but if a given point does not, then neighboring time points will likely not either. This data type is efficiently processed by extracting shorter rather than longer clips (Marasli & Montag, 2023; Pisani et al., 2021). Therefore, we chose 1 minute every hour as a compromise between sufficient investment in contextually analysing each segment while also maintaining a representative sampling period over the course of the day. It is worthwhile to note that, since our data annotation preceded that done by Bunce and colleagues (Bunce et al., 2021), theirs could be inspired in part by ours—they also selected to sample short audio sections (2 minutes). However, they used random rather than periodic sampling. A subsequent methodological study has revealed that input quantity estimations are more accurate using periodic rather than random sampling (Pisani et al., 2021).

A trained phonetician segmented vocalizations separated by at least 300 ms or uttered by different people (Fernald, 1992), and assigned them to different speaker sources. In Praat, this is done by the use of "tiers". Figure 1 represents the annotation for about 10 s of audio, with one or two tiers per each source type: Target child, main female adult, other female adults, male adults, other children, overlapping speech, and background speech. We explain each in turn.

One speaker source was the target child (i.e., the child wearing the recording device). Another speaker source included speech from the most common female voice (hereafter "Main Female Voice," MFV). Separating the MFV from other females allows us to better connect with Cristia et al. (2019), where speech from mothers was reported separately from other women's speech. The recognition of these two individuals (target child and MFV) is based on the annotator's judgment, as he did not know the participants and was unfamiliar with the Tsimane' language. While Casillas and colleagues (Casillas et al., 2020, 2021) relied on a local research assistant who knew the families, and could thus identify individuals, it was not logistically possible for us to do the same. We turned instead to a phonetically trained annotator with experience in child-centered recordings collected in multiple languages.

We had two tiers per type for segmenting vocalizations by other (i.e., “non-man”) female adults, male adults, and other children (i.e., not the target child). Specifically, there were two tiers for each broad age-gender category (i.e., two child tiers, two male adult tiers, and two female adult tiers) so that the annotator could tag individuals who were speaking simultaneously. We determined that two tiers per type were ideal given visual crowding (more tiers take more space, leaving less for the signal) and task difficulty. Indeed, the task of identifying each individual speaker was overly difficult for our non-native annotator, and it was not necessary to answer our research questions. Note that individual speaker identity was not constant throughout the same age-gender tiers. For example, in one clip the grandmother's voice may occur first (F1) and then the aunt's voice second (F2), but in another clip, the voice-tier association might be swapped. Similarly, if in a clip three males spoke but their voices did not overlap, their vocalizations may have been annotated all in a single tier (M1). The decision as to whether someone was male or female, or child or adult, was made purely on the basis of acoustic evidence by the annotator, and thus we cannot be certain of the gender or precise age of the speakers.

For all of the above, each segment was also categorized into two vocalization “types”: linguistic (including speaking, babbling) and non-linguistic (crying or laughing, Franklin et al., 2014). Vegetative sounds (e.g., respiration, swallowing) were not segmented.

Other segment types included stretches of the signal that could not be attributed to different speaker types because either two or more people were speaking simultaneously such that the beginning and end of spoken turns was unidentifiable (source type: “two or more speakers talking at the same time”), or because the speech sounded distant (source type: “background speech”). In order to decide whether to use the “background speech” tier, the annotator listened to the relevant section three times; if he was unable to decide whether the person was male or female, child or adult, then this section of speech was considered as “background”.[7]

There are a total of 3576 segments, corresponding to an average of 149 segments per recording (range: 70–286). This includes vocalizations produced by the target child and all other speakers, as well as segments in the “two or more speakers” and “background” tiers. We aggregated for each child all vocalizations of the same speaker source (normalized for annotation length and multiplied by 60 to get an estimation of min/h). Specifically, we pooled quantities from the MFV, from other females and male adults (summed into “Other adults”), and from non-target children (i.e., “Other children”).

The restrictive input definition is based on previous work (Gilkerson et al., 2017; Weisleder & Fernald, 2013) that depends on the LENA software (see Cristia et al., 2020 for a meta-analysis of LENA's accuracy). This input definition includes only vocalizations that were human-labeled as being near and non-overlapping, and only speech in quality (i.e., crying and laughing were excluded). In contrast, the permissive input definition follows Casillas and colleagues (Casillas et al., 2020, 2021). This input definition includes vocalizations that were labeled as being near and distant, overlapping and non-overlapping, and speech and non-speech in quality, and thus nothing is excluded.

Next, we classified each segment as containing TCD verbal input or overheard. As discussed above, some previous work used content and context to identify when speech is directed to children, classifying an utterance as TCD verbal input because of what is said and/or the child's preceding and following behavior (e.g., Weisleder & Fernald, 2013). However, more recent methodological research suggests that lower-level co-occurrence is a good predictor of whether speech is child-directed or overheard

[7] The annotator was asked to code music, radio, tool noise, etc. into a “Noise” tier, so as to use this information when developing automatic analysis software. At the time of our visit, some households had a radio, but it was not used frequently during the day. Given our focus on linguistic input, we have removed the “Noise” tier from consideration altogether.

(Bang et al., 2023). Therefore, we implemented a proxy based on a bottom-up definition of context, which was more appropriate because our annotator was not a Tsimane' speaker. We implemented two alternative definitions. In the restricted input definition, we counted as child-directed any vocalization that was temporally close to a vocalization attributed to the target child. We used 2 seconds as the maximum inter-turn interval as a compromise between the length used in LENA studies (5 s, Ford et al., 2008) and the overwhelming evidence that inter-turn intervals are much shorter in caregiver-child conversations (a meta-analysis in Nguyen et al., 2022). In the permissive input definitions, we additionally counted as TCD a given speaker's vocalizations that were not followed or preceded by vocalizations by other speakers. The intuition here is that if the infant is too young to respond, or not encouraged to do so, people may talk to them and this may look like a monologue. In the context of discussions regarding what the minimum TCD verbal input allows for language acquisition, we thought this decision was appropriate as it may lead to an overestimation of TCD, but cannot lead to an underestimation. Vocalizations that alternate with speakers other than the key child thus count towards Overheard verbal input.

All speech segments in the "two or more speakers" tier were considered to be "in overlap." This tier was used for speech that contained such rapid turn transitions and/or overlap among similar-sounding speakers that our annotator was unable to segment out individual turns. Whenever he could, however, he would segment speech segments by attributing them to the different speaker types. Thus, all speech segments that partly overlapped with speech segments marked in another tier were also considered to be "in overlap." Overlapping segments are only counted for the permissive input definitions.

All analyses were conducted in R (R Core Team, 2017); figures were generated using the ggplot2 package (Wickham, 2009). Scripts are available on OSF (https://osf.io/sfhza/).

## 3 | RESULTS

For restrictive and permissive input definitions we present descriptive statistics of TCD and overheard input, and the proportion of TCD input relative to total input (research question 1). We then report quantities and distribution by input source (i.e., MFV, other children, adults; research question 2).

**Research question 1.** How much verbal input is available to young children?

Table 3 provides summary statistics for TCD, overheard, total input (TCD + overheard), and the percentage of TCD, under both restrictive and permissive input definitions. Higher quantities are observed for the permissive input definitions in comparison to the restrictive, with over 4-fold increases (see SM6 for convergent evidence). Figure 2 shows that estimates for TCD do not overlap across restrictive and permissive definitions, suggesting that the decision of whether apparent monologues are considered as being addressed to the target child has a massive impact on our estimates. Although the distribution of overheard verbal input is more similar across definitions, it is still the case that permissive estimates are much larger than the restrictive ones (see SM6 for convergent evidence), highlighting the importance of whether overlapping and background speech is considered as contributing to children's input. The compound effect of these decisions is apparent in Figure 3, where individual children's input estimates are provided for both input definitions.

**Research question 2.** Who speaks to and around children?

Table 4 shows the percentage of input that is attributed to the different potential sources for TCD and overheard input, under both restrictive and permissive input definitions. For TCD verbal input,

TABLE 3 Estimates of target-child-directed (TCD), overheard input, total (TCD + overheard) and percent (pc) TCD (TCD/Total), from both restrictive and permissive input definitions.

| | Restrictive | | | | Permissive | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean | Median | Min | Max | Mean | Median | Min | Max |
| TCD | 0.7 | 0.6 | 0.2 | 2.2 | 10.6 | 9.6 | 2.8 | 27.1 |
| Overheard | 3.5 | 3.2 | 1.0 | 10.4 | 6.2 | 4.2 | 1.6 | 32.7 |
| Total | 4.2 | 3.8 | 1.1 | 12.6 | 16.9 | 13.8 | 4.5 | 59.8 |
| pc TCD | 17.1 | 15.5 | 14.9 | 17.4 | 63.1 | 69.5 | 63.6 | 45.3 |

*Note*: All except for percent TCD are expressed in minutes per hour. See Table 1 for the definition of restrictive and permissive.

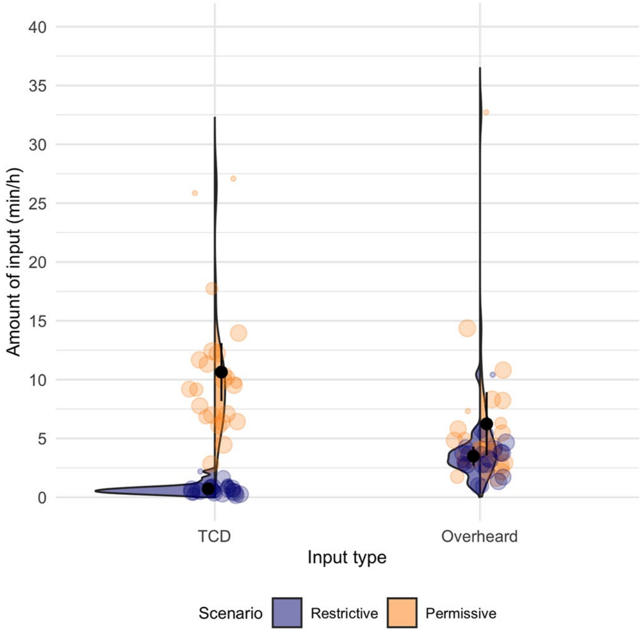

FIGURE 2 Verbal input (min/h) of Tsimane' children ($N = 24$) as a function of input type (TCD = target-child-directed input, Overheard) and definition. Split-violins show distributions and mean with CI of each input definition. Each semi-transparent point represents a child, and point size is proportional to the number of minutes annotated (range = 4–16).

other children constitute the main source, with MFV being a close second, and other adults contributing less than a fifth of input, and this applies in both input definitions. For overheard verbal input, trends are more variable across input definitions, but MFV and other children continue to constitute the main source of input compared to other adults. Individual data is presented in Figure 4, where it becomes obvious that the difference in definitions across input definitions plays an important role. For instance, 6 children had no TCD input from MFV in the restrictive input definition, whereas this only occurs for one child in the permissive input definition. Figure 4 also highlights the impact of considering background verbal input (defined as that which was so soft that the annotator couldn't decide on the speaker's age/gender after listening to the section three times), and which constitutes most of the overheard verbal input for most children.

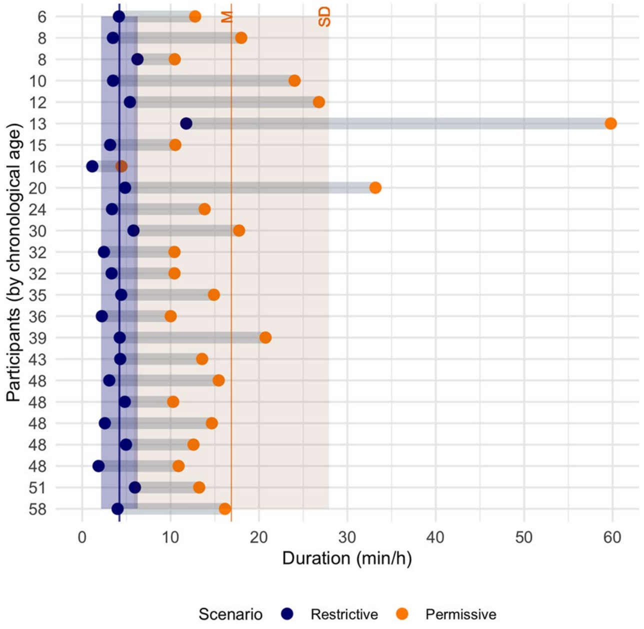

**FIGURE 3** Total verbal input (min/h) of Tsimane' children ($N = 24$). The left (restrictive) and right (permissive) solid vertical lines show the mean and shaded areas standard deviation of each input definition. Each horizontal line represents a child. The number on the left indicates the child's age in months.

**TABLE 4** Estimates of the mean (M) and standard deviation (SD) of the percent (pc) input as a function of source for target-child-directed (TCD) and overheard input, in both restrictive and permissive input definitions.

| | **Restrictive** | | | | **Permissive** | | | |
|---|---|---|---|---|---|---|---|---|
| | **TCD** | | **Overheard** | | **TCD** | | **Overheard** | |
| | **pc M** | **SD** | **pc M** | **SD** | **pc M** | **SD** | **pc M** | **SD** |
| MFV | 28 | 27 | 42 | 24 | 21 | 14 | 21 | 12 |
| Other children | 58 | 33 | 36 | 25 | 27 | 14 | 22 | 16 |
| Other adults | 14 | 24 | 22 | 16 | 9 | 7 | 16 | 11 |
| Overlap | NA | NA | NA | NA | 8 | 10 | 12 | 14 |
| Background | NA | NA | NA | NA | 35 | 17 | 28 | 23 |

Abbreviation: NA, not applicable.

## 4 | DISCUSSION

We sought to address two key questions: 1. How much verbal input is available to young children? 2. Who speaks to and around target children? Previous work led us to couch our replies based on two substantially different operationalizations of the input, which we have termed the restrictive and permissive input definitions. These are also the extremes in terms of input quantities, which allows us to draw conclusions that can be generalized to intermediate input definitions. In what follows, we summarize our answers and integrate them with previous literature before turning to a more general discussion.

For our first question we asked how much verbal input Tsimane' children encounter in their environment. In a nutshell, results are heavily influenced by input definition. Using the restrictive input

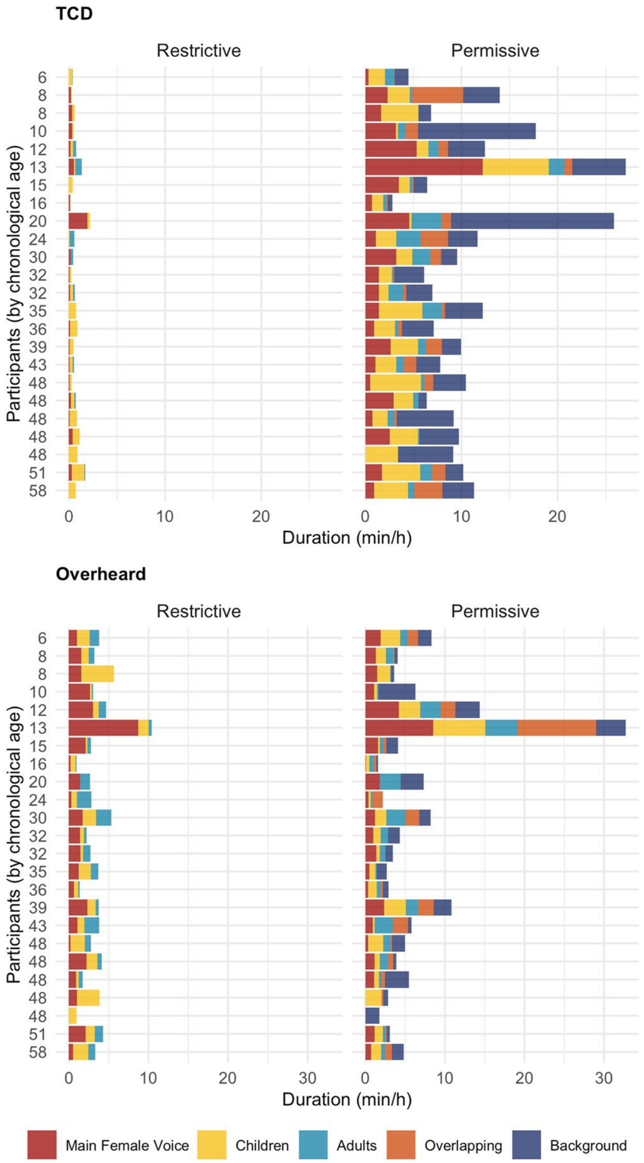

**FIGURE 4** Quantity of input (min/h) by input type, input definition, and source. Each bar represents a target child, sorted by age (in months).

definition, we find estimates for TCD verbal input comparable with those extrapolated from Cristia et al. (2019) despite differences between our study and theirs, including the use of a different method and observer/annotator training, the times of day sampled (7a.m. to 9p.m. here, vs. 7a.m.-7p.m. in the previous study), and how TCD is detected (defined via a temporal proxy here, vs. including only one-on-one conversation in Cristia et al., 2019). Using the permissive input definition, we find estimates well above that study of the Tsimane'. Next, we integrate our results with previous literature.

The restrictive input definition allows us to compare our manually annotated Tsimane' data to that derived from typical LENA-based studies (Gilkerson et al., 2017) in that it excludes overlapping, background, and non-linguistic vocalizations. Using this restricted definition, we find that Tsimane' children experience about 4.21 min of total verbal input and about 43 s of TCD verbal input. For total input, about 42% came from adults, leading to a conservative estimate of total input from adults of 1.75 min/h for Tsimane'.[8] This estimate is lower than that of Gilkerson and colleagues (Gilkerson et al., 2017) for long-form recordings among North American children, which estimated 3.5 min/h of total verbal input from adults. This definition of input would lead to the conclusion that young Tsimane' children receive less overall input than young North American children.

The permissive input definition, which includes overlapping, background, and non-linguistic vocalizations from all speakers (not just adults), is useful for comparing our data to that derived from manually annotated long-form recordings (Bunce et al., 2021; Casillas et al., 2020, 2021). Our permissive input definition estimated TCD at 10.64 min/h, which is more directed input than children in all other long-form corpora studied previously. However, our TCD permissive definition is likely even more inclusive than the most permissive studies: Using a temporal proxy for TCD, we may include verbal input that happens to temporally coincide with the target child's vocalizations, and we further include all monologues. In contrast, if overall input is considered (and not just TCD), Tsimane' children hear a total of 16.87 min/h, which is 30% less than Tseltal children in Mexico (total of 24 min/h, Casillas et al., 2020) and 57% less than Yélî Dnye children in Papua New Guinea (total of 39 min/h, Casillas et al., 2021).

Obviously, how input is operationalized heavily influences input estimates, rendering comparisons across papers with diverse methodologies fraught. Additional work employing the same definitions to (re-)analyse data from multiple cultures (in line with Bunce et al., 2021) will be crucial for appropriately establishing the extent of cross-cultural variation in both TCD and overheard verbal input.

Our results show more similar patterns across restrictive and permissive input definitions in terms of who talks to the child. To begin with, the most common sources of non-background TCD verbal input is other children, then MFV, then other adults in both input definitions. Moreover, the contribution of MFV is twice that of other adults in both. That said, the consideration of background speech radically alters input composition for both TCD and overheard speech, constituting the most common source within the permissive input definition.

Turning now to a broader discussion of implications, we see that overall estimates of input quantity vary dramatically across input definitions: estimates increase 15-fold (TCD) or 1.79-fold (overheard) in the permissive input definition compared to the restrictive one. Put otherwise, were we to follow Gilkerson et al. (2017)' definition and count only near, non-overlapping speech from adults, this would entail excluding background speech (35%), overlapping speech (8%), and vocalizations by other children (27%; see Table 4). As a result, we would neglect 70% of input. But does this 70% actually affect young children's language learning?

Only a handful of studies have addressed this question, yielding inconsistent results. On the one hand, Akhtar et al. (2001) report that American children aged 18–24 months make use of overheard speech for novel word learning in experimental contexts, and recent evidence suggests Tseltal infants demonstrate knowledge of words that they could have only learned from overheard speech (Foushee & Srinivasan, 2023). On the other hand, only TCD speech, not overheard or total speech, has been found to be associated with vocabulary development in American and Yucatec samples (Shneidman et al., 2013; Shneidman & Goldin-Meadow, 2012; Weisleder & Fernald, 2013). One limitation of previous work is that it has focused on lexical development. Given the likely diversity of learning mechanisms across language levels (e.g., Heinz & Idsardi, 2013), future research should investigate whether there are associations between TCD and/or total speech with respect to phonological (Cristia, 2020) and morphosyntactic development.

[8] For comparability with previous work, this estimate is computed using the mean total input estimate per input definition and the mean percentage of speech attributed to MFV as well as other adults.

Verbal input produced by other children was found to be prevalent in both this and other similar work summarized in the Introduction. This is unsurprising given the prominent role of children as allocare providers in many societies (e.g., Weisner & Gallimore, 1977; including among the Tsimane': Stieglitz et al., 2013). Therefore, studies on the quality of input from other children would be advisable. Previous studies have argued that other children may be less able to speak with referential clarity, or focus on the locus of interest of another immature speaker (Dunn & Kendrick, 1982; Hoff-Ginsberg & Krueger, 1991; Sachs & Devin, 1976; Tomasello & Mannle, 1985), all the while adapting to their addressee (Loukatou et al., 2022). The effect of other children on target children's language development is uncertain. Some find that a greater proportion of verbal input from other children is linked to delayed language learning (Nelson, 1973) — potentially due to the additional children competing with the target children for adult attention in cultures where adults are the primary caregivers (as argued for birth order effects, Havron et al., 2019). Others document that exposure to peer speech has positive associations with language abilities (Justice et al., 2011; Mashburn et al., 2009). Similarly, results regarding whether target children's vocabulary size is better predicted by taking into account other children's TCD speech are inconsistent (cf. Shneidman et al., 2013; Shneidman & Goldin-Meadow, 2012).

When considering what types of verbal input support native language development, it is important to distinguish the types of experiences that are (a) necessary and sufficient, versus (b) beneficial for specific linguistic knowledge and skills. Descriptions of what is beneficial are tainted by the fact that they are based almost exclusively on children in urban, typically North American settings.[9] For instance, literacy skills are now extremely important for thriving in many contemporary situations, and current results suggest that literacy skills are strongly predicted by early childhood language skills (Dickinson et al., 2010). However, language learning during the first years of life is a process independent of literacy, and has been for the vast majority of human history. What is sufficient and necessary for language to be acquired is thus a different question than what features make one better at the language skills valued by a given culture. Therefore, different backgrounds have different expectations and the child may accommodate local linguistic demands (Hymes, 1972; Ochs & Schieffelin, 2011).

Before closing, we would like to acknowledge a few limitations of this study. A clear limitation of the paper is its focus on verbal, in some cases specifically verbal linguistic, input captured by recordings, which excludes non-verbal forms of communication such as gestures. Moreover, the paper does not give a definitive answer about the role of non-linguistic cues like laughing or crying, overheard, overlapping, and other types of input in the language acquisition process and whether they should be considered as part of verbal input. This omission is intended to provoke thought and discussion: this paper presents ranges and scenarios to encourage readers to critically examine these choices. Determining the "final" definition of input goes beyond the scope of this paper and invites additional research and attention. Another weakness of our data is that decisions about gender and age category for speakers in the recordings were based on acoustic judgment by a non-native speaker. Moreover, TCD labeling was based on a temporal proxy, rather than using content. Importantly, these are issues that affect not only our manual annotations (performed by a non-native annotator) but also algorithmic approaches such as LENA. Both types of methodology offer benefits in remote areas with limited resources, where recruiting and training native speaker annotators may be challenging. However, the accuracy of such estimates should be interpreted with caution and considered in light of this potential uncertainty. To address this concern in our data specifically, we re-annotated 10% of samples with a native speaker, who corrected the voice attributions and decided on directedness based

[9] A recent systematic review of this literature does not even code for country of origin of the participants in the 190 studies they discovered (Walker et al., 2020).

on both context and content. The re-annotation suggests that input quantities are captured accurately (median relative error rate was 0% for all estimates; mean relative error rates, which are affected by outliers, were 4% for permissive total, CDS, and overheard each, and 14%, −2%, and 6% for restrictive total, CDS, and overheard respectively; positive numbers indicate over-estimation and negative ones under-estimation; see SM5 and SM6 for more information). This accuracy may be specific to our non-native annotator, who was highly experienced and multilingual. Future work should carry out a more in-depth analysis of the effects of using restrictive and permissive annotations on the accuracy of estimates, for both human (native and non-native) and automated annotations. Our data are also limited in terms of sample size (which was nonetheless comparable to that found in previous relevant work) and the fact that only one Tsimane' village was sampled. Although Cristia et al. (2019) found little variation in TCD input between less ($N = 5$ villages) versus more market-integrated villages ($N = 1$), it would be worthwhile to systematically evaluate linguistic experiences in heterogeneous villages in the future (e.g., Cristia et al., 2023; Padilla-Iglesias et al., 2021). Expanding the sample would also allow for a more balanced gender distribution than analysed here. We particularly hope to see additional work with larger samples in varied populations, which would allow us to better represent the richness of variation within and across populations (Kline et al., 2018).

## 4.1 | Conclusion

Integrating our results with those of other studies based on long-form recordings, we estimate that, according to some input definitions, Tsimane' children experience similar levels of target-child-directed verbal input compared to North American and other small scale settings, but not according to other definitions. Our paper highlights the importance of input definition. For this reason, we call for further work with identical recording and annotation procedures to more accurately measure similarities and differences. More stability across input definitions was observed for target-child-directed input sources, with the most common source of non-background speech being other children. Our results provide important evidence regarding the estimate of target-child-directed verbal input afforded to Tsimane' children, and contribute much needed data to ongoing debates on the quantity and composition of early input.

## ACKNOWLEDGMENTS

We are grateful to the Tsimane' community, including saliently our research assistants, Maria Eugenia Maito and Dyersin Lero; to Mustapha Mardi, who patiently annotated these data; and to William Havard, Arnulfo Cari Ista, and Manuel Roca Moye for the data reannotation. JS acknowledges IAST funding from the French National Research Agency (ANR) under grant ANR-17-EURE-0010 (Investissements d'Avenir program). MC's contributions were supported by an NWO Veni Innovational Research Scheme (275-89-033). AC acknowledges funding from Agence Nationale de la Recherche (ANR-16-DATA-0004 ACLEW, ANR-14-CE30-0003 MechELex, ANR-17-EURE-0017); the J. S. McDonnell Foundation Understanding Human Cognition Scholar Award; and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ExELang, Grant agreement No. 101001095).[1]

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest with regard to the funding source for this study.

## ORCID

*Camila Scaff* https://orcid.org/0000-0002-7546-9538
*Alejandrina Cristia* https://orcid.org/0000-0003-2979-4556

## ENDNOTES

[1] Data accessibility

All relevant computer code for variable definitions and statistical analysis is downloadable from the following OSF component: https://osf.io/sfhza/. Data are available for reproducibility/verification purposes by requesting access in the following OSF component: https://osf.io/dvzfw/. Data availability for reuse is restricted for ethical reasons. Tsimane Health and Life History Project (THLHP)'s highest priority is the safeguarding of human subjects and minimization of risk to study participants. The THLHP adheres to the CARE Principles for Indigenous Data Governance, which assure that the Tsimane: (i) have sovereignty over how data are shared; (ii) are the primary gatekeepers determining ethical use; (iii) are actively engaged in the data generation; and (iv) derive benefit from data generated and shared use whenever possible. The THLHP is also committed to the FAIR Principles to facilitate data reuse. Requests for data reuse should take the form of an application that minimally details the exact uses of the data and the research questions to be addressed, procedures that will be employed for data security and individual privacy, potential benefits to the study communities and procedures for assessing and minimizing stigmatizing interpretations of the research results (see the following webpage for links to the data sharing policy and data request forms: https://tsimane.anth.ucsb.edu/data.html). Requests for data reuse will require institutional IRB approval (even if exempt) and will be reviewed by an Advisory Council composed of tribal leaders, tribal community members, Bolivian scientists, and the THLHP leadership.

## REFERENCES

Akhtar, N., Jipson, J., & Callanan, M. A. (2001). Learning words through overhearing. *Child Development*, *72*(2), 416–430. https://doi.org/10.1111/1467-8624.00287

Bang, J. Y., Kachergis, G., Weisleder, A., & Marchman, V. A. (2023). An automated classifier for periods of sleep and target-child-directed speech from LENA recordings. *Language Development Research*, *3*(1), 211–248.

Barry, H., & Paxson, L. M. (1971). Infancy and early childhood: Cross-cultural codes 2. *Ethnology*, *10*(4), 466–508. https://doi.org/10.2307/3773177

Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., & Tor, S. (2019). Day by day, hour by hour: Naturalistic language input to infants. *Developmental Science*, *22*(1), e12715. https://doi.org/10.1111/desc.12715

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.e12724

Boersma, P., & Heuven, V. V. (2002). Praat, a system for doing phonetics by computer. *Glot International*, *5*, 341–347.

Bunce, J., Casillas, M., Bergelson, E., Rosemberg, C., Rowland, C., Warlaumont, A. S., & Soderstrom, M. (2021). A cross-cultural examination of child-directed speech across development. Retrieved from https://psyarxiv.com/723pr/

Casillas, M., Brown, P., & Levinson, S. C. (2020). Early language experience in a Tseltal Mayan village. *Child Development*, *91*(5), 1819–1835. https://doi.org/10.1111/cdev.13349

Casillas, M., Brown, P., & Levinson, S. C. (2021). Early language experience in a Papuan community. *Journal of Child Language*, *48*(4), 792–814. https://doi.org/10.1017/s0305000920000549

Cecil, M. J. (2010). *Cross-linguistic variation in turn taking practices: A computational study of the callhome corpus (PhD thesis)*. University of Colorado at Boulder.

Cristia, A. (2020). Language input and outcome variation as a test of theory plausibility: The case of early phonological acquisition. *Developmental Review*, *57*, 100914. https://doi.org/10.1016/j.dr.2020.100914

Cristia, A. (2022). *A systematic review suggests marked differences in the prevalence of infant-directed vocalization across groups of populations*. Developmental Science.e13265

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*, *63*(4), 1093–1105. https://doi.org/10.1044/2020_jslhr-19-00017

Cristia, A., Dupoux, E., Gurven, M., & Stieglitz, J. (2019). Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child Development*, *90*(3), 759–773. https://doi.org/10.1111/cdev.12974

Cristia, A., Gautheron, L., & Colleran, H. (2023). *Vocal input and output among infants in a multilingual context: Evidence from long-form recordings in Vanuatu*. Developmental Science.e13375

Demany, L. (1982). Auditory stream segregation in infancy. *Infant Behavior and Development*, *5*(2–4), 261–276. https://doi.org/10.1016/s0163-6383(82)80036-2

Dickinson, D. K., Golinkoff, R. M., & Hirsh-Pasek, K. (2010). Speaking out for language: Why language is central to reading development. *Educational Researcher*, *39*(4), 305–310. https://doi.org/10.3102/0013189x10370204

Dunn, J., & Kendrick, C. (1982). The speech of two-and three-year-olds to infant siblings: 'Baby talk' and the context of communication. *Journal of Child Language*, *9*(3), 579–595. https://doi.org/10.1017/s030500090000492x

Ellwood-Lowe, M. E., Foushee, R., & Srinivasan, M. (2022). What causes the word gap? Financial concerns may systematically suppress child-directed speech. *Developmental Science*, *25*(1), e13151. https://doi.org/10.1111/desc.13151

Fernald, A. (1992). *Meaningful melodies in mothers' speech to infants*. Nonverbal Vocal Communication Comparative and Developmental Approaches (p. 262).

Ford, M., Baer, C. T., Xu, D., Yapanel, U., & Gray, S. (2008). *The LENA language environment analysis system: Audio specifications of the DLP-012 (No. LTR-03-2)*. LENA Foundation.

Foushee, R., & Srinivasan, M. (2023). Infants who are rarely spoken to nevertheless understand many words. Retrieved from.https://psyarxiv.com/g84pw

Franklin, B., Warlaumont, A. S., Messinger, D., Bene, E., Nathani Iyer, S., Lee, C.-C., Lambert, B., & Oller, D. K. (2014). Effects of parental interaction on infant vocalization rate, variability and vocal type. *Language Learning and Development*, *10*(3), 279–296. https://doi.org/10.1080/15475441.2013.849176

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H. L., & Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, *26*(2), 248–265. https://doi.org/10.1044/2016_ajslp-15-0169

Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh-Pasek, K. (2015). (Baby) talk to me: The social context of infant-directed speech and its effects on early language acquisition. *Current Directions in Psychological Science*, *24*(5), 339–344. https://doi.org/10.1177/0963721415595345

Havron, N., Ramus, F., Heude, B., Forhan, A., Cristia, A., Peyre, H., Group, E. M.-C. C. S., Bernard, J. Y., Botton, J., Charles, M. A., Dargent-Molina, P., de Lauzon-Guillain, B., Ducimetière, P., De Agostini, M., Foliguet, B., Fritel, X., Germa, A., Goua, V., & Thiebaugeorges, O. (2019). The effect of older siblings on language development as a function of age difference and sex. *Psychological Science*, *30*(9), 1333–1343. https://doi.org/10.1177/0956797619861436

Heinz, J., & Idsardi, W. (2013). What complexity differences reveal about domains in language. *Topics in Cognitive Science*, *5*(1), 111–131. https://doi.org/10.1111/tops.12000

Hoff-Ginsberg, E., & Krueger, W. M. (1991). Older siblings as conversational partners. *Merrill-Palmer Quarterly*, *37*(3), 465–482.

Hymes, D. (1972). On communicative competence. In *Sociolinguistics* (pp. 269–285). Penguin Books.

Justice, L. M., Petscher, Y., Schatschneider, C., & Mashburn, A. (2011). Peer effects in preschool classrooms: Is children's language growth associated with their classmates' skills? *Child Development*, *82*(6), 1768–1777. https://doi.org/10.1111/j.1467-8624.2011.01665.x

Kaplan, H., Hooper, P. L., Stieglitz, J., & Gurven, M. (2015). The causal relationship between fertility and infant mortality. In *Population in the human sciences: Concepts, models* (pp. 361–376). Evidence.

Kline, M. A., Shamsudheen, R., & Broesch, T. (2018). Variation is the universal: Making cultural evolution work in developmental psychology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1743), 20170059. https://doi.org/10.1098/rstb.2017.0059

Loukatou, G., Scaff, C., Demuth, K., Cristia, A., & Havron, N. (2022). Child-directed and overheard input from different speakers in two distinct cultures. *Journal of Child Language*, *49*(6), 1173–1192. https://doi.org/10.1017/s0305000921000623

Mallinckrodt, B. (1992). Childhood emotional bonds with parents, development of adult social competencies, and availability of social support. *Journal of Counseling Psychology*, *39*(4), 453–461. https://doi.org/10.1037/0022-0167.39.4.453

Marasli, Z., & Montag, J. L. (2023). Optimizing random sampling of daylong audio. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Mashburn, A. J., Justice, L. M., Downer, J. T., & Pianta, R. C. (2009). Peer effects on children's language achievement during pre-kindergarten. *Child Development*, *88*(3), 686–702. https://doi.org/10.1111/j.1467-8624.2009.01291.x

Nelson, K. (1973). Structure and strategy in learning to talk. *Monographs of the Society for Research in Child Development*, *38*(1/2), 62. https://doi.org/10.2307/1165788

Nguyen, V., Versyp, O., Cox, C., & Fusaroli, R. (2022). A systematic review and Bayesian meta-analysis of the development of turn taking in adult–child vocal interactions. *Child Development*, *93*(4), 1181–1200. https://doi.org/10.1111/cdev.13754

Ochs, E., & Schieffelin, B. B. (2011). The theory of language socialization. In A. Duranti, E. Ochs, & B. B. Schieffelin (Eds.), *The handbook of language socialization* (pp. 1–22). John Wiley; Sons, Ltd.

Padilla-Iglesias, C., Woodward, A. L., Goldin-Meadow, S., & Shneidman, L. A. (2021). Changing language input following market integration in a yucatec mayan community. *PLoS One*, *16*(6), e0252926. https://doi.org/10.1371/journal.pone.0252926

Pisani, S., Gautheron, L., & Cristia, A. (2021). Long-form recordings: From a to z. Retrieved from.http://bookdown.org/alecristia/exelang-book/

R Core Team. (2017). R: A language and environment for statistical computing. Retrieved from.https://www.R-project.org/[

Sachs, J., & Devin, J. (1976). Young children's use of age-appropriate speech styles in social interaction and role-playing. *Journal of Child Language*, *3*(1), 81–98. https://doi.org/10.1017/s030500090000132x

Saffran, J. R., Werker, J. F., & Werner, L. A. (2007). *The infant's auditory world: Hearing, speech, and the beginnings of language*. John Wiley; Sons, Ltd.

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*(3), 672–686. https://doi.org/10.1017/s0305000912000141

Shneidman, L. A., & Goldin-Meadow, S. (2012). Language input and acquisition in a Mayan village: How important is directed speech? *Developmental Science*, *15*(5), 659–673. https://doi.org/10.1111/j.1467-7687.2012.01168.x

Soderstrom, M., Casillas, M., Bergelson, E., Rosemberg, C., Alam, F., Warlaumont, A. S., & Bunce, J. (2021). Developing a cross-cultural annotation system and MetaCorpus for studying infants' real world language experience. *Collabra: Psychology*, *7*(1), 23445. https://doi.org/10.1525/collabra.23445

Stieglitz, J., Gurven, M., Kaplan, H., & Hooper, P. L. (2013). Household task delegation among high-fertility forager-horticulturalists of lowland Bolivia. *Current Anthropology*, *54*(2), 232–241. https://doi.org/10.1086/669708

Stieglitz, J., Trumble, B. C., Finch, C. E., Li, D., Budoff, M. J., Kaplan, H., & Gurven, M. D. (2019). Computed tomography shows high fracture prevalence among physically active forager-horticulturalists with high fertility. *Elife*, *8*, e48607. https://doi.org/10.7554/elife.48607

Tannen, D. (2012). Turn-taking and intercultural discourse and communication. In C. Paulston, S. Kiesling, & R. ES (Eds.), *The handbook of intercultural discourse and communication* (pp. 135–157). Blackwell.

Tomasello, M., & Mannle, S. (1985). *Pragmatics of sibling speech to one-year-olds* (pp. 911–917). Child Development.

Vogt, P., Mastin, J. D., & Schots, D. M. (2015). Communicative intentions of child-directed speech in three different learning environments: Observations from The Netherlands, and rural and urban Mozambique. *First Language*, *35*(4–5), 341–358. https://doi.org/10.1177/0142723715596647

Walker, D., Sepulveda, S. J., Hoff, E., Rowe, M. L., Schwartz, I. S., Dale, P. S., Peterson, C. A., Diamond, K., Goldin-Meadow, S., Levine, S. C., Wasik, B. H., Horm, D. M., & Bigelow, K. M. (2020). Language intervention research in early childhood care and education: A systematic survey of the literature. *Early Childhood Research Quarterly*, *50*, 68–85. https://doi.org/10.1016/j.ecresq.2019.02.010

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143–2152. https://doi.org/10.1177/0956797613488145

Weisner, T. S., Gallimore, R., Bacon, M. K., Barry, H., Bell, C., Novaes, S. C., Edwards, C. P., Goswami, B. B., Minturn, L., Nerlove, S. B., Koel, A., Ritchie, J. E., Rosenblatt, P. C., Singh, T. R., Sutton-Smith, B., Whiting, B. B., Wilder, W. D., & Williams, T. R. (1977). My brother's keeper: Child and sibling caretaking. *Current Anthropology*, *18*(2), 169–190. https://doi.org/10.1086/201883

Werner, L. A. (2002). Infant auditory capabilities. *Current Opinion in Otolaryngology and Head and Neck Surgery*, *10*(5), 398–402. https://doi.org/10.1097/00020840-200210000-00013

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from.http://ggplot2.org

Winking, J., Gurven, M., Kaplan, H., & Stieglitz, J. (2009). The goals of direct paternal care among a South Amerindian population. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, *139*(3), 295–304. https://doi.org/10.1002/ajpa.20981

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Scaff, C., Casillas, M., Stieglitz, J., & Cristia, A. (2024). Characterization of children's verbal input in a forager-farmer population using long-form audio recordings and diverse input definitions. *Infancy*, *29*(2), 196–215. https://doi.org/10.1111/infa.12568