Using big data from long-form recordings to study development and optimize societal impact

Margaret Cychosz^{1,2} and Alejandrina Cristia³

¹Department of Hearing and Speech Sciences, University of Maryland, College Park

²Center for Comparative and Evolutionary Biology of Hearing, University of Maryland, College Park

³Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'études cognitives, ENS,
EHESS, CNRS, PSL University, Paris, France

Abstract

Big data are everywhere. In this chapter, we focus on one source: long-form, child-centered recordings collected using wearable technologies. Because these recordings are simultaneously unobtrusive and encompassing, they may be a break-through technology for clinicians and researchers from several diverse fields. We demonstrate this possibility by outlining three applications for the recordings—clinical treatment, large-scale interventions, and language documentation—where we see the greatest potential. We argue that incorporating these recordings into basic and applied research will result in more equitable treatment of patients, more reliable measurements of the effects of interventions on real-world behavior, and deeper scientific insights with less observational bias. We conclude by outlining a proposal for a semi-structured online platform where vast numbers of long-form recordings could be hosted and more representative, less biased algorithms could be trained.

Keywords: big data, wearable technology, algorithm bias, automatic measurement, language, audio recording, children

1 Introduction

Big data are everywhere. In this chapter, we focus on one source: long-form, child-centered recordings collected using wearable technologies. We argue that long-form recordings could be a break-through technology, allowing researchers and practitioners from several different domains to unearth key scientific insights and measure the effects of interventions on real-world behavior. We also consider priorities for policy makers, students of developmental science, as well as researchers who consume literature employing this methodology.

We will not provide here a technical introduction to collecting long-form recordings (see Casillas and Cristia 2019), the ethics of the methodology (see Cychosz et al. 2020), or an argument for the theoretical value of these recordings (see Lavechin, de Seyssel, Gautheron, Dupoux, and Cristia 2021). Instead, we focus on the societal and human applications of long-form recordings, elaborating on three key areas where additional reflection and development could increase the positive impacts of this methodology.

Before proceeding, we note that many of our examples come from the study of language, since most current work on data collected with wearables bears on speech and language. However, our key arguments apply to most human behavior. For this reason, we hope that readers interested in socioemotional, motoric, or cognitive development also heed our experiences and avoid pitfalls related to technical progress, participant recruitment, and scientist diversity that we already see emerging in the speech and language domain.

1.1 Brief overview of long-form recordings

Long-form recordings are typically gathered with a wearable recording device, which collects audio (and sometimes visual) signals over long periods of time, often a whole day. By virtue of their length and unobtrusiveness, these recordings capture a key participant’s naturalistic behaviors and surroundings in a wide range of settings throughout the participant’s day.

For students of child development, recording these behaviors “in the wild” can have many benefits. First, measures derived from long-form recordings represent children’s everyday behaviors. In the context of research or practical applications, this means that we do not need to establish external or ecological validity of outcome measures, like a standardized assessment, because the outcome measures are instead drawn directly from everyday behavior. Second, since the recording device goes everywhere the child goes, the child as well as their family and all other interlocutors (within a certain radius of the wearable) are chronicled. The resulting audio can then be useful when studies or interventions involve not just the key participant but also those around them, such as caregivers or siblings.

Long-form recordings also present challenges. One involves determining the costs and benefits for individual participants, the communities they belong to, and the broader population. For example, the child who wears the recording device may be the “key” participant, but others around the child are also recorded, including those who may not be aware of the recording or its implications (longevity, potential for data sharing, etc.). Moreover, actions and speech captured in the recordings could inappropriately be interpreted as representative of the child’s overall community, with the problematic assumption that the results generalize to all community members. Another, related challenge involves obtaining truly informed consent, particularly among participants who vary in their exposure to this type of technology (Cychosz et al., 2020). These considerations should inform policy and legislation, where definitions of personal data are currently driven by benefits and risks for individuals, rather than communities.

Long-form recordings generate immense amounts of data. Consequently, relying on human coding to quantify behavior in the recordings is costly and time consuming. As a result, automated algorithms are increasingly used to process recordings (Figure 1). For example, current algorithms aim to identify television and electronic sounds in the recordings, as well as speech and non-speech vocalizations made by the key child and

others in the child’s vicinity. Algorithms can also make more detailed classifications such as vocal maturity types for the child. Important progress has recently been made in infant emotion recognition (Yao, Micheletti, Johnson, & de Barbaro, 2020).

Figure 1. Example data processing flows for long-form recordings. Human and automated annotations often complement one another; human annotation sometimes follows automated annotation.

Based on current advances in commercial processing of audio-videos, we expect additional algorithms will emerge in the near future that draw even more information from audio recordings, including some that determine types of conversations (e.g., dyadic, multyadic), the setting of the recording (e.g., home, restaurant, or park), or ongoing

activities (e.g., eating; Bin Morshed et al. 2020).

Additional hardware development permits other types of signal tracking, to capture other behaviors. For instance, a device called Babylogger (Cao et al., 2018), which is currently in the piloting phase contains an accelerometer whose data could be useful for evaluating gross motor behaviors. Turning extant technology (Marchionni, Scalise, Ercoli, & Tomasini, 2013) into wearables would also enable us to collect respiration, heart rate, and skin conductance, and thus study other cognitive phenomena like stress levels, attention phases, and sleep. Current options for head-mounted digital and video cameras have limited battery lives, meaning it is not possible to capture what the child sees (and where the child looks) for whole days, but even shorter video recordings (spanning 6 hours or less) have lent insight into infants' visual input, social interactions, and language development in naturalistic settings (Clerkin, Hart, Rehg, Yu, & Smith, 2017; Fausey, Jayaraman, & Smith, 2016; Laing & Bergelson, 2020). Finally, one could collect data not just from one child but from several participants, fitting each with radio-frequency systems to locate them in 3D space, for instance to perform network analyses and study social interactions more broadly. An overview of the range of possibilities from wearable sensors can be found in de Barbaro (2019).

Although we look forward to these future developments, this chapter will focus mainly on tools applicable to the audio signal and the currently restricted choice of algorithms for processing long-form, naturalistic audio recordings. The most commonly used algorithm, part of the Language ENvironment Analysis (LENA) system, was trained on data from English-learning infants and young children (up to four years old), but in recent years has been used with a much wider range of ages and languages (see Ganek and Eriks-Brophy 2018 for a review). More recently, open-sourced alternatives to LENA have been developed by members of the ACLEW project (Lavechin, Bousbib, Bredin, Dupoux, & Cristia, 2021; Räsänen et al., 2019; Räsänen, Seshadri, Lavechin, Cristia, & Casillas, 2020), including a system to identify speakers and another to count words, syllables, and phones, all trained

on multilingual datasets (henceforth, the ACLEW pipeline). Independent assessments comparing automated annotations against human ones suggest that the LENA and ACLEW algorithm accuracy varies widely across participants (even across English-speaking participants: Cristia, Lavechin, et al. 2020; Lehet, Arjmandi, Dilley, and Houston 2020; Räsänen et al. 2020). Moreover, accuracy against human annotations are only part of the story. Internal algorithmic validity (as established via test-retest reliability) and external validity (such as to clinical tests) are also important and remain understudied. For example, to our knowledge, for internal validity there is only one report that examines the stability of individual variation in measurements from the LENA algorithm (Gilkerson, Richards, & Topping, 2017), and only a handful assessing how LENA metrics relate to standardized language assessments (see the meta-analysis in Wang, Williams, Dilley, and Houston 2020: All except two studies were carried out on English-speaking North American infants). For these and other metrics developed in the future, it would be important to check for the stability in individual variation, particularly among non-English or non-infant participant samples.

For current and future algorithms, we should bear in mind the threat of algorithm bias. Several high-profile examples have recently come to light, including facial recognition software that more often misidentifies non-white than white individuals as potential suspects (Grother, Ngan, & Hanaoka, 2019), a program that recommended harsher judgments for black defendants (Angwin, Larson, Mattu, & Kirchner, 2016), and algorithms that contribute to health care decisions being more favorable to white than non-white individuals (Matheny, Whicher, & Thadaney, 2020). There has been, as of the time of writing, no discussion of the possibility that algorithms employed with long-form recordings could have similar (racial, ethnic, etc.) algorithmic biases, despite this being a crucial topic for any real-world application. Public institutions considering the use of long-form technology should perhaps recommend its evaluation by relevant bodies (e.g., National Institute of Standards and Technology in the United States), as is currently the

case for the other types of algorithms mentioned here.

1.2 Three key areas for future development

In the remainder of this chapter, we elaborate on some unique opportunities and challenges for developmental researchers and others who may consider using long-form recordings. This chapter may also assist individuals who consume research that uses these recordings, for their studies or in a professional context, including to make policy decisions. Our primary goal is to inform readers how long-form recordings can be usefully harnessed, in a number of situations, considering the alternatives. That is, we have a pragmatic and realistic approach: We know that no methodology is perfect, and thus the question is rather *when* and *how* to incorporate long-form audio recordings into one's methodological toolkit.

To facilitate our arguments, we focus on three emergent case studies for long-form recordings, where we expect long-form recordings can provide the most unique value. In Section 2, we discuss how long-form recordings are relevant for clinical practice in pediatrics and speech-language pathology, for instance to test potential effects of clinical treatment within individuals, and/or measure developmental progression. We then turn to a second case study in Section 3: large-scale interventions. Long-form recordings can be used to measure infants' and caregivers' behaviors in their everyday lives, thus alleviating concerns about ecological validity for the outcome measures of interventions (as well as immediate intervention effects, for interventions aimed at changing caregiver behavior). Finally, because long-form recordings have most frequently been used in the realm of language development and learning, in the final case study we describe the immense potential to use long-form recordings to document and revitalize under-studied languages (Section 4).

Our chapter closes with a discussion of additional, potential future developments (Section 5). We argue for the importance of increasing the geographic and sociocultural diversity of users of this methodology, and we encourage readers to think of the longer-term future, beyond single uses of individual recordings.

2 Clinical relevance

Scientists have been making large-scale, naturalistic observations of children’s behavior since technology permitted the data collection and management. But only recently have the advances from big data methods gained popularity in clinical research or practice. We see enormous promise for large-scale data methods, and long-form recordings in particular, in clinical settings to reliably measure speech-language development and accurately diagnose developmental delays. The use of new, semi-automated tools in the clinic could be timely given the current shortage of speech-language pathologists in many countries, as well as the toll that communication disorders take on life quality and opportunities. For the former, Wylie, McAllister, Davidson, and Marshall (2013) estimated just one speech-language pathologist per 2,500-4,700 people in the United States, United Kingdom, Canada, and Australia¹—an alarming statistic given that approximately 5% of children in the United States have a “noticeable” speech-language disorder or delay by age 6 (NIDCD, 2016). The situation is dire in other geographical regions, with Wylie et al. estimating one speech-language pathologist per 2-4 million people in sub-Saharan Africa. Clearly, tools that can facilitate clinicians’ work are welcome.

Clinicians, such as speech-language pathologists and pediatricians, are responsible for *measuring* children’s speech-language development over time and *diagnosing* speech-language disorders and delays. But these responsibilities are complicated by 1) participant age (it can be difficult, if not impossible, to elicit controlled speech samples from young children and infants), 2) a paucity of data (for delays characterized by limited communication), and 3) a lack of assessment standardization and reliability, including for situations where clinicians are forced to improvise due to child fussiness or discomfort. In the following sections we outline how long-form recordings can help address these issues.

¹ Considering the United States alone, the ratio in 2019 was closer to one speech-language pathologist per 1,807 people (American Speech-Language-Hearing Association, 2019)

2.1 Measurement

Measuring developmental progressions reliably is especially important—and difficult. A clinician may wish to evaluate the communicative inventory of a two-year-old who recently received an ASD diagnosis and who is now receiving an experimental intervention. Accurate measurement requires that the clinician both make consistent evaluations over time, avoiding human-related error in coding and evaluation, and adapt the developmental measure as the child matures—so distinguishing improvements due to age from those due to the intervention is challenging. Another situation where measuring changes is difficult is among the small percentage of children with disorders such as ASD who regress, (reverting to earlier, less mature communicative and social stages (Hansen et al., 2008)). How, then, to reliably and accurately measure development and distinguish it from intervention effects?

Automated measurements using algorithms that are appropriately informed by the child’s characteristics made over long-form recordings can be more objective. In our own cross-cultural work, standardized, automated workflows of long-form recordings allowed us to compare infants’ developmental trajectories in speech and language across multiple geographic locations and cultures, so we could attribute developmental differences not to individual researchers, recording setting, or collection methods, but instead cultural practices (Cychosz et al., 2021; Räsänen et al., 2019).

2.2 Diagnosis

Correct diagnosis. Early screening and diagnosis of neuro-developmental and communicative disorders is key to effective treatment and outcomes. Accordingly, accurate diagnostics, especially those that are appropriate for young infants and children, must be developed. In our discussion of long-form recordings’ potential for accurate diagnosis, we focus mainly on disorders known to manifest via vocal bio-markers in infancy because this is where the the most diagnostic work has been conducted. However, we also briefly mention some longer-term potential to diagnose disorders, such as stuttering or

developmental language disorder, that do not manifest until early childhood.

As it became clear that prelinguistic vocalizations were not meaningless babble but instead strongly predictive of children's later speech and language development (Oller, 2000), measuring the diversity and structure of those vocalizations became a key clinical goal. Furthermore, numerous developmental disorders result in delayed or impaired vocal development. Indeed, one hallmark of infancy research from the past two decades has been the recognition that there are specific vocal bio-markers of disorders such as Autism Spectrum Disorder (ASD), Angelman Syndrome, and childhood apraxia of speech in infancy (Marschik et al., 2017; Sheinkopf, Mundy, Oller, & Steffens, 2000).

There remain, however, several roadblocks to measuring infant vocalizations for diagnosis. For one thing, one characteristic common to infants who eventually receive diagnoses of speech-language or neuro-developmental disorders, such as childhood apraxia of speech, ASD, or neurogenetic syndromes associated with ASD like fragile X syndrome, is the infrequency or near absence of vocalizations in infancy (Belardi et al., 2017; Hamrick, Seidl, & Tonnsen, 2019; Overby, Belardi, & Schreiber, 2020; Patten et al., 2014; Warlaumont, Richards, Gilkerson, & Oller, 2014). Logically, if an infant is not vocalizing, clinicians cannot evaluate the vocalization structure. Or, in the case of infrequent or unreliable vocalizations, huge speech samples may be required to amass sufficient tokens for diagnosis. The latter results in laborious human-coding and annotation (McDaniel, Yoder, Estes, & Rogers, 2020). Such coding techniques are impractical for clinicians with heavy caseloads.

Long-form recordings speak to these issues surrounding clinical diagnosis of language-related delays. As such, this technology may enable healthcare professionals to make earlier, more precise diagnoses. Already the comprehensive nature of long-form recordings means that they document a great deal about the developmental profiles of neurodiverse infants and children (e.g., Swanson et al. 2019). For diagnosis of delays characterized by infrequent vocalizations, long-form recordings are key because they sample

over long periods of time, capturing sufficient tokens of infant behavior that are representative of what the infant actually produces (Oller et al., 2010). Furthermore, current algorithms (the LENA pipeline), as well as those under continued development such as the ACLEW pipeline, allow for effortless characterization of infant vocal quantity and quality (i.e., vocal maturity and complexity) (Oller et al., 2010; Seidl et al., 2018; Yoder, Oller, Richards, Gray, & Gilkerson, 2013). The continued development of automated signal processing techniques will garner additional information about the developmental profiles of neuro-diverse infants and children, hopefully enabling clinicians to make more fine-grained distinctions between disorders, such as ASD, fragile X, and hearing loss, that have similar developmental profiles in infancy (VanDam & Yoshinaga-Itano, 2019), but clearly very different treatment profiles. In addition, the accumulation of long-form datasets may give us enough training data to develop metrics for more nuanced disorders like developmental language disorder and stuttering, for which no infant vocal markers have been postulated (perhaps due to the limited data that can be analyzed with traditional clinical approaches).

Misdiagnosis and overdiagnosis. Another concern in pediatric speech-language pathology is misdiagnosis, and especially overdiagnosis. In contexts where locally standardized norms exist, multilingual children are at heightened risk of over- or mis-diagnosis of speech-language delays because they can develop speech and language at different rates than their monolingual peers—who are the basis of most standardized tests—and because they are a highly heterogeneous group in terms of bilingual language exposure (Byers-Heinlein, 2013; Place & Hoff, 2011). Additionally, many regions and whole countries do not have locally normed tests, so practitioners must adapt tests (e.g., changing a lexical item like “snow” that is irrelevant in a tropical climate) or translate them to a local language on the fly. Regions with high degrees of dialectal and language variation may be particularly challenging, as it is up to the clinician to reflect on how words and structures may be used in the language variant(s) that the child is exposed to. In all of these cases, it is not uncommon for the clinician to apply norms from a different language

variety and population to inform their decision of whether a given child meets expectations or not.

For children speaking a smaller-scale language that may lack documented norms, we cannot be certain that standardized tests designed for majority languages do not increase the chance of mis-diagnosis. For example, a clinician may unwittingly translate a difficult item like “butterfly”—an item that only roughly 50% of American 16-month-olds produce—into a simpler word; or perhaps a clinician fails to adapt a culturally meaningless word like “monkey,” if monkeys are less common in the local culture. In either case, substitutions lead to greater uncertainty when applying norms from other languages, and to less informative diagnoses.

One reason for overdiagnosis in multilingual children is that, initially, some aspects of multilingual development can mimic characteristics of speech-language delay (Fabiano-Smith & Hoffman, 2018). Consider the following: a bilingual three-year-old is producing very few words in one of their languages, scoring in the bottom 10th percentile on a standardized (monolingual) expressive vocabulary assessment. To the examiner, the low score could simply reflect the child’s infrequent exposure to the language of assessment, but it might signify a more profound delay.

Acknowledging some of these issues facing clinicians, professional speech and hearing associations stress the need for culturally sensitive speech-language assessment tools (e.g., for the United States: American Speech-Language-Hearing Association 2017). In North America, bilingual children are increasingly assessed in both of their languages—so their competencies are additive—or they are evaluated using more language-neutral tools (Ortiz, 2021; Seymour, Roeper, & de Villiers, 2005). The same cannot be said of standard practice even in other high income countries, like France. Moreover, appropriate evaluation is challenging in all cosmopolitan areas where multilingual children are exposed to a variety of languages. For example, in a survey conducted in three Parisian daycares, 55 of 107 toddlers’ parents reported that their child was exposed to one or more languages besides

French. Even though parents did not always specify the language varieties (e.g., answering Arabic or Chinese, although some variants of each are mutually unintelligible), 26 different languages were represented in the sample. Clinicians everywhere agree that more detailed information about multilingual children's language exposure, dominance, and practice would help them distinguish between typical multilingual development and speech-language delay. Long-form recordings could help address these needs, alleviating some cultural biases inherent to many assessments.

We envision a socioculturally-sensitive practice where clinicians could quantify the home language environment of bilingual children using long-form recordings. Using one or more recordings, the clinician could quantify what percentage of the child's everyday language exposure was in both languages, including how many words were spoken to the child in each language by any number of interlocutors, as well as what language dominated the child's multimedia exposure and screen time. From these algorithmically-derived estimates of language exposure, the clinician in the example of the bilingual three-year-old who performed poorly on the vocabulary assessment might find that only 15% of the child's language exposure was in the assessment language—and that most of that exposure originated from only one speaker. So, the clinician might determine that the child's low score is unlikely to be attributable to a language disorder (although the clinician could further evaluate this possibility by conducting parental interviews about the child's communicative behaviors or assessing how quickly the child learns novel words, (de Villiers, 2017))². The clinician would then be making treatment decisions backed by hours of (automated) observation of the child's actual language experience. We also see long-form recordings being used in contexts like these to encourage language transmission within multilingual families. Many clinicians who work with multilingual families already support bilingual language use and enrichment strategies, such as book reading in both languages, for multilingual families (Kohnert, Yim, Nett, Kan, & Duran, 2005). Using long-form

² Thank you to Kathleen Oppenheimer for suggesting this assessment protocol

recordings, clinicians could also help families set quantitative language-use goals for their languages to, for example, ensure inter-generational language transmission if a family desired that.

Currently, the multilingual language environment is quantified in basic research studies, but estimates of exposure to each language can only partially be automated (Cychosz, Villanueva, & Weisleder, *to appear*; Orena, Byers-Heinlein, & Polka, 2020). Language discrimination algorithms, especially for naturalistic speech, are still under active development. However, long-form recordings still represent an improvement over current practices—often simply a written language use questionnaire—for measuring families’ multilingual practices. Furthermore, we stress that the development of any discrimination algorithm will require large amounts of training data. And the largest goal of all—one or more language-neutral discrimination algorithms—will require data from many diverse multilingual environments. In short: the lack of a high-performing discrimination algorithm *right now* should not discourage the collection of multilingual data, since it is possible to quantify the proportion of languages within recordings relatively easily via smart sampling techniques and since these multilingual data could contribute greatly to basic science and clinical practice.

3 Large-scale interventions

Large-scale interventions were central to the development of child-centered, long-form recording methods, pioneered by the LENA Foundation. Founded in 2009, the LENA Foundation developed hardware and software tools—updated versions of which are still prevalent today—and, even more importantly, demonstrated that the collection and automated analysis of long-form audio recordings at scale was a feasible way to study child language development. Over the years, the LENA Foundation has increasingly invested effort in its mission, to change caregiver behavior to facilitate children’s linguistic and cognitive development: in 2019 alone, interventions including this technology reached over

10,000 children in North America.

The Foundation’s work is aligned with a rising interest in early childhood interventions. Among researchers of cognitive and language development, recent examples include Ferjan Ramírez, Lytle, Fish, and Kuhl (2019), who developed and evaluated an intervention on the role of child-directed speech and parentese among American infants (see also Ferjan Ramírez and Kuhl 2017, for a trial on second language learning); and Suskind et al. (2016) who carried out a spoken language intervention among families of children from low socioeconomic backgrounds and/or with hearing loss.

In this section, however, we would like to turn to much larger interventions, which are typically carried out by developmental economists, who aim to measure return on investment. We first discuss why such interventions may be interesting. We then elaborate on why measuring their effects is so expensive and difficult, and lay out the role that long-form recordings can play in their execution.

3.1 Early childhood interventions

From the point of view of a country (or an NGO, or any other such organization) that aims to improve life, developmental, or health outcomes, a key question is: How to allocate one’s limited budget? Since any action or program (e.g., creating and maintaining a daycare or rolling out a nutrition program) involves a cost, it is often the case that these organizations need to measure the costs of programs and actions, and the effects of the programs and actions on the individuals who receive them, as well as society at large. The costs and benefits of these various programs and actions are often measured using randomized control trials where one group of individuals is randomly assigned to receive an intervention (e.g., access to a daycare) while another (the control group) does not. Based on the cost of the program and its potential benefit, one can calculate a “return on investment,” via a cost-benefit analysis, which quantifies the positive outcome (e.g., reductions in health costs and jail time or increases in earnings) society gets per unit of

investment.

The Nobel prize winner James Heckman is among the leading voices arguing for governments and other managerial bodies to invest in early childhood interventions, which reputedly have the largest returns on investment, with positive impacts demonstrated throughout a child’s academic career into adulthood (Heckman, 2006). Early childhood interventions span a broad spectrum, from “body-driven” programs based on nutritional supplementation, deworming, and frequent health checks; to “mind-driven” programs, such as providing information on child cognitive stimulation, support for families, and daycare. A recent systematic review of randomized control trials aimed at improving people’s outcomes throughout the lifespan supported Heckman’s intuitions, in particular recommending parental interventions in early childhood (Busso, Cristia, Hincapié, Messina, & Ripani, 2017).³ It should be noted that these parental programs are a great deal more holistic than the language-centered interventions mentioned above, providing caregivers with other types of support and information than just increasing child-directed speech (e.g., Gertler et al. 2014). These more holistic early childhood interventions appear to have excellent returns on investment (e.g., Walker et al. 2015).

If the evidence for early childhood interventions (both general and language-centered) is so tantalizing, why aren’t more of these programs being developed and evaluated? Conducting these trials is challenging for a number of reasons, starting with the scale: Trials ideally involve hundreds, if not thousands, of children. Additionally, it is hard to measure trial effects. There are very few accurate and informative measures of children’s language or cognition, particularly before three years of age. For instance, before age three, researchers need to decide whether to measure a child’s vocabulary size based on parental report (as in the MacArthur-Bates Early Communicative Development Inventory [MB-CDI], Fenson et al. 2007), or to opt for a more holistic developmental measure, like

³ Evidence for this nevertheless remains preliminary: only a handful of randomized control trials have been done and none were pre-registered to avoid outcome selection and other biases.

the Bayley Scales Test for Infant and Toddler Development (Albers & Grieve, 2007). Neither measure has a test-retest reliability above .7 across all countries that could consider using them. Additionally, these metrics represent a biased view of what is “important”, “desirable”, or common developmentally, which may not be shared across all cultural groups (Heath, 1983; Ochs & Schieffelin, 2001).

In view of such difficulties, the intervention evaluation team has to make several difficult choices. They could use suboptimal measurements to evaluate the short-term effects of the intervention; but these may mis-represent effects, especially in infancy and early childhood when language and cognition are difficult to measure. Alternatively, the team could track the participant sample longitudinally up to school-age, when assessment is easier; but then shorter-term effects of the intervention go undocumented, wasting intervention resources (also, in the absence of short-term effects, it may be difficult to garner resources for longitudinal tracking).

For interventions that aim to change parental or caregiver (e.g., daycare personnel) behaviors in order to impact the child, assessing caregiver behavior changes post-intervention introduces additional challenges. Researchers can measure caregiver behavior via a short observation of an interaction but this takes time (to train observers, collect the data, and annotate it). Furthermore, despite their best efforts, an observer’s presence also affects the observation, and caregivers’ behavior in the treatment versus the control group may well reflect their greater knowledge of what the observer is expecting. The same goes for self-report measurements.

3.2 Scaling and adapting child language and caregiver behavior measures

In light of these challenges, it becomes clearer why many large-scale trials are turning to long-form recordings. Long-form recordings capture children’s and their interlocutors’ natural behavior with minimal observer effects. Algorithms to process the recordings hold the annotator constant, potentially leading to more objective measures. The same audio

data can be investigated along multiple dimensions, including metrics that reflect the child’s environment and experiences, in addition to the child’s own behaviors. Current metrics are also sensitive to development from zero to three years of age, so researchers do not need to wait many years to measure language effects.

Regarding capturing effects on caregiver behavior, language algorithms currently available for long-form recordings reliably estimate the quantity of language in the recording. It is straightforward to complement this measure with what appear to be more qualitative measures, such as the structure of adult vocalizations that are preceded or followed by a child vocalization; or the pitch or fundamental frequency in adult conversational turns. Nevertheless, we still stress that most of the available measures represent a specific conception of how the child learns language and what are desirable behaviors in and around the child. For instance, in some cultures, it is more common for toddlers to spend time with other children than with adults, so when only interactions with adults are considered, these children may appear to engage in quantitatively fewer interactions. As for prosodic characteristics, although there is a tendency in many cultures to raise one’s pitch when talking to infants, some work suggests cross-cultural variation in acoustic features, and thus this is not a universal “quality” feature either (Ratner & Pye, 1984).

Fortunately, long-form recordings can help us get out of these biases: By capturing not only the behavior of the key child, but also that of others around them, the stage is set to extract communication “targets” that are community-specific, so the child’s development can be benchmarked against older children and adults in the child’s *own* community, rather than a normative sample of children who may not be raised with the same practices, values, or goals. Similarly, the audio data can be interrogated to assess which caregiver behaviors in that particular community are associated with faster growth, and use these within-community metrics to evaluate caregiver input quality. Hypotheses regarding what these behaviors may be can be elicited via qualitative interviews with representatives of

these communities (Crago, Annahatak, & Ningiuruvik, 1993; Keller, Yovsi, & Voelker, 2002).

3.3 Large-scale interventions in a different region

In this section, we want to take a moment to highlight some actions that should increase the positive impact of interventions that transcend regional (i.e., linguistic, cultural, or legal) borders. We are assuming that readers are already familiar with best practices in cross-cultural research, including community engagement and avoiding ethical dumping (Appiah, 2020; Broesch et al., 2020; Chatfield et al., 2021).

A key question to ask when considering the use of long-form recordings for cross-regional interventions is how benefits of this research will be shared with the communities involved. Some actions that can be taken include direct payment to participating families, and the promise of scaling up the intervention should initial trials suggest positive effects. With long-form recordings, there is also the possibility of creating a large-scale knowledge database that can represent local customs and the local language(s), where local communities of linguists and information technology or machine learning experts could engage (see Corbett and Kulchyski (2009) for an example from the Hul’q’umi’num’ language in Canada). Creating such a database entails unique interdisciplinary opportunities for experts in child development. Additionally, such collaboration could eventually result in discoveries and commercialization (which should be shared with the involved communities; see Section 5.2), for instance by stimulating the development of natural language processing and speech technology tools, which can then be used in web and smartphone applications. Given such potential outcomes, it is worthwhile to think carefully about how to word consent forms, to give participants the option to allow their data to be reused and ensure that benefits (financial or otherwise) will find their way back to participants and/or to the communities from which those participants were sampled.

In this context, we think, to the extent that it is possible, that researchers should be mindful of potential changes in regulation, making sure their research practices not only comply with current legislation in their own home country and the participants' home country, but also with future legislation that may arise due to advances in technological or ethical understanding. For instance, the Nagoya Protocol on Access and Benefit-sharing proposes that biological samples, as well as local knowledge concerning such samples, constitute intellectual property of the local population. One can make the case that language samples and the local knowledge those samples represent should eventually be covered by a similar regulation (see Carroll et al. 2020).

4 Language documentation and revitalization

Thus far, we have talked about two areas in which developmental science intersects with societal impact, and highlighted how long-form recordings, typically analyzed through automated means, may be incorporated into one's toolkit. We now turn to another area of potential societal impact, the documentation and revitalization of languages, where most of the current technology does not allow automated analyses (see Table 1 for contrasting examples). Although long-form recordings could be useful for cultural and anthropological research in general, we will discuss here the use of long-form recordings specifically for a studies where language acquisition, documentation, and revitalization are more intertwined.

Researchers of child language acquisition are often asked how best to foster first and second language learning and development. Advances in language acquisition could contribute to the preservation of local languages and, more broadly, maintenance of the rich linguistic diversity present today. One difficulty, however, is that there is scant documentation for thousands of the over 7,000 languages currently spoken. Long-form recordings of children's learning environments can uniquely help address the formidable challenges of language documentation and revitalization to promote adult and child language learning for revitalization and to help meet a central goal in linguistics to

document the full breadth of human languages.

4.1 What is language documentation?

Language documentation is the practice of making a lasting record of the language practices of a speech community.⁴ There are many different goals of documentation projects, but the documentary process typically involves a linguist working with one or more language users to compile a description of a language and make it accessible to current and future audiences. Language documentarians strive for complete descriptions of a language that account for the community’s wants and needs and that are accessible to a wide array of individuals, including language speakers, educators, and learners.

Language documentation involves two data structures: structured linguistic elicitation and observable linguistic behavior (Himmelman, 2006). Both constitute vital components of any documentary project, although traditional structural linguistics has focused primarily on the former (Bower, 2015; Hermes & Engman, 2017). In structured linguistic elicitation, a researcher, often with specialized training in linguistic description and annotation, works one-on-one or in small groups with speakers (“consultants”) to derive metalinguistic knowledge about the language. Consultants may work with the researcher in tightly-controlled settings to, for example, translate word lists, provide grammaticality judgments, or give monolingual definitions of words. The short-term result of these efforts may be a searchable lexical database with morphological glosses. Longer-term results may be descriptive grammars, bilingual dictionaries, or a writing system.

In contrast, observable linguistic behavior refers to a record of language use in context, usually in the form of audio or video records of native speakers interacting, recounting oral literature, or gesturing in tandem with spoken language. Written forms also exist, such as

⁴ A speech community is composed of people who speak and interpret a language variety in similar ways. It can be many things, including a family, a town, or a nation. Sometimes, in the case of small-scale languages, a speech community may be a single speaker.

an individual's journal entries or records of a child's first words. Observable linguistic behavior data are usually spontaneous, although staged interactions between interlocutors are not uncommon. As a general goal, researchers collecting these data try to minimize observation biases and may only ask speakers to comment on particular interactions at a later date, during transcription, for example. The results of observable linguistic behavior data are books and recordings of conversations, folktales, and rituals as well as eventual pedagogical grammars and textbooks for language teaching and revitalization purposes (some of which can also stem from structured linguistic elicitation).

4.2 Children's long-form recordings and language documentation

The data collected via long-form recordings clearly reflect observable linguistic behavior. Consequently, long-form recordings are proving to be important vehicles for language documentation (Nee, 2021), especially documentation of child language behavior and learning because the thoroughness of the recordings facilitates comprehensive, accessible documentary materials that could serve community goals for language revitalization.

Comprehensive documentation. *Comprehensive language documentation*, or documenting language use in a variety of situations, is essential for language documentation because language structure is context-dependent. For example, speakers use different words in casual speech between friends than in more formal, ritualized settings or in speech directed to an infant. Consequently, it is not sufficient to only record isolated words or phrases from a single adult consultant. A complete language record will include words used by a number of interlocutors, like caregivers and their infants or small children, not typically chronicled during traditional one-on-one elicitation.

But comprehensive language documentation—particularly for child language development—is challenging. Researchers are, after all, limited by resources of time, money, and methodological expertise. Many documentarians, trained in traditional

elicitation methods with adult consultants, feel ill-equipped to commence a research project with infants or children, or are unsure how to solicit approval from an ethics review board for this type of data collection (for assistance with this see the ethics repository we constructed at <https://osf.io/u3tfv/> that includes sample informed consent forms and scripts used for the collection of long-form recordings in a number of speech communities). Young children are also unlikely to be cooperative one-on-one language consultants for the documentation of their speech registers and patterns. And linguists often cite the need for basic grammatical descriptions of a language before collecting or analyzing any observable linguistic data, unaware that observable linguistic data can be informative even before extensive grammatical descriptions are available.

Long-form recording methodologies, however, speak directly to the challenges of comprehensive language documentation. The collection of long-form recordings requires minimal time investment, particularly at a small scale. Linguistic fieldworkers can easily collect a single recording from the children of 10-20 families during a 6-8 week fieldstay by visiting with the family in the morning, prior to structured elicitation with an adult consultant, and again visiting the family in the evening or the following morning to collect the recording device. In this way, little structured elicitation time with adults is lost and the researcher does not need to invest significant time in learning a new child-friendly experimental paradigm. Long-form recording collection likewise requires minimal hardware expertise. While some hardware forms (i.e., the LENA system) require substantial monetary investments, the number and frequency of developmental labs employing long-form recording methods means that the recording devices are now easily shared. And promising steps are being taken to document at least basic linguistic structure (e.g., phoneme inventory) using unsupervised algorithms on raw audio recordings (Conneau, Baevski, Collobert, Mohamed, & Auli, 2020).

Finally, multilingualism is common amongst child speakers of under-documented languages (Child Language Research and Revitalization Working Group, 2017). As we

noted in Section 2.2, the entirety of a child’s multilingual language exposure and use can be estimated using long-form recordings. So some additional uses of this methodology for communities interested in language revitalization and reclamation could be to record bilingual practices to see if and how inter-generational language transmission occurs (see Table 1). Researchers interested could likewise estimate children’s dual language exposure to evaluate mechanisms of contact-induced language change or questions of learnability given access to two or more languages.

4.3 Accessible language documentation

Accessible language documentation, or ensuring that language records are organized and available for future audiences, is likewise essential for any documentary project. Archivists have made great gains structuring digital language repositories with accompanying basic metadata (e.g., The Archive of the Indigenous Languages of Latin America, California Language Archive, Databrary, Talkbank). Making language records accessible to community members and future researchers is critical so that individuals 1) heed communities’ intellectual property rights and informed consent practices and 2) allow for their documentary work to propagate and grow. If language records are accessible, documentarians can ensure that they may be used even if the researcher cannot anticipate all future uses for the data, or technological changes that may render certain data formats obsolete. Furthermore, records of child language behavior in smaller linguistic communities shifting towards majority languages may be the only documentation of first language acquisition or child-directed speech registers.

However, data accessibility goes beyond organization and openness; documentarians also need to ensure that their records are interpretable to diverse audiences of educators and language learners. A typical archival bundle might include a 20-minute elicitation session recording between researcher and consultant to, for example, flesh out irregular verb inflections. This bundle would also include the researcher’s written notes, such as lists

of verb paradigms. This is invaluable documentary information. Nevertheless, no description of the verbs' use in context is included. There is no indication of how verb irregularity varies within the community, by age, gender, or multilingual experience (for an example of this consider the propensity for younger, but not necessarily older, English speakers to alternate between *sneaked* and *snuck*). Nor is there information about how children initially over-regularize verbs during early language learning. The researcher's notes are comprehensive but annotated with specialist terms like 'zero expression.' In short, the data are not truly accessible to a broad audience because they are not contextualized, representative, or easily interpretable.

Long-form recordings easily satisfy the requirement that archival materials be accessible. The data are, by nature, highly contextualized (although they are usually limited to an auditory medium so still lack visual context). The recordings should be made in generalized, lossless formats, such as .wav, that are unlikely to be discontinued for many years to come. And while researchers can annotate or gloss portions of a long-form recording—and doing so may make the data more accessible to language learners who need records of the language's structure—there are few specialist or terminological barriers to data re-use because the recordings merely capture participants going about their lives.

4.4 Community member involvement in documentation

A final goal of language documentation is community member involvement. This could mean that a community's wishes for language products like a children's book in the target language or bilingual dictionary result from the documentation project. Or a community shifting to a majority language may wish to document child language learning to promote language maintenance or revitalization. Community members should be involved throughout the documentation project, from design, to data collection, and/or linguistic transcription and archiving (Leonard & Haynes, 2010). In conjunction with traditional one-on-one elicitation sessions and spontaneous language data from adult

speakers, long-form recordings can contribute to each of these potential community goals. The diverse, large-scale data of long-form recordings easily lend themselves to re-use. They also capture a variety of settings. Consider the following examples:

1. There is a community with infrequent inter-generational target language transmission. They wish to document the language practices of a handful of families who continue to speak the target language to their children. Specifically, the community wishes to know the factors promoting transmission as well as the contexts of the families' target language use. Some of this information can be gleaned from qualitative interviews with caregivers or questionnaires. But the documentation of causes and contexts of intergenerational transmission may be limited by observation biases, cultural differences, and the children's ages. For non-literate communities, extensive language use questionnaires may be impractical and interviews may be prohibitively time consuming. Also, facing broader societal stigma for target language use (or non-use if the broader community "expects" an individual to speak a language based on appearance or ethnicity), some caregivers may be reticent to report all contexts of target language use or may over-report them.
 - How long-form recordings contribute: One or several recordings of the home environment could elucidate what events promote target language use (e.g., presence of grandparents, radio versus T.V.). They could complement a shorter, oral questionnaire so literacy in the target or contact language is not a requisite for study participation. Finally, the children's role in the families' target language use would be recorded; for example, there would be a record of the child's code-switching patterns which could predict transmission.
2. A community with robust inter-generational transmission and a newly-developed orthography wishes to create a series of primary school textbooks in the target language. A number of linguistic details are required for the books, some of which

can be derived from traditional records of children’s language behavior (1-2 hour recordings of child-caregiver behavior in a standardized, lab-like setting) while others are more easily derived from long-form recordings of the children’s environments.

- How long-form recordings contribute: The books need to reflect children’s language experience. Naturalistic, long-form recordings could comprehensively document which words *adults* continue to use in the target language, but *children* now use in the contact language, especially with other children—the type of peer exchanges infrequently captured in traditional caregiver-child interactions. Likewise, if authors wish to create books that gradually increase in difficulty, they could benefit from information on word frequency statistics and default word order patterns in child and child-directed speech in the target language. Instead of deriving these statistics from shorter, semi-structured child-caregiver interactions, the same amount of data could be sampled from different times of the day in a maximally naturalistic long-form recording and then annotated which would make the books more reflective of actual language patterns in the community.

4.5 Additional contributions of long-form recordings

Long-form recordings contribute to one additional, often overlooked goal of documentary linguistics: cross-site comparison. Documentarians rely on standardized methods of linguistic elicitation, like the approximately 200-item Swadesh word list, to compare between languages, trace historical changes, and establish phylogenetic patterns. Ideally, developmental science would have a similar set of standardized methods but few child or infant behavior methods have the same cross-site compatibility. Novelty preference procedures in infancy research, for example, are notoriously idiosyncratic between labs with settings that are nearly impossible to replicate (The Many Babies Consortium, 2020). However, the relative ease of collecting long-form recordings, combined with

straightforward, standardized hardware, permits large-scale, comparative developmental research unlike any we have seen before. During the collection of long-form recordings, large amounts of data are quickly amassed, participants engage in naturalistic behaviors, and the researcher is, by design, absent—all factors that help ensure cross-site differences are attributable to participant behavior and not observational biases, sample sizes, or researcher degrees of freedom (Cychosz et al., 2021; Räsänen et al., 2019).

Even using long-form recording methods, challenges to cross-site comparison remain: automatic speech processing algorithms are often trained on majority languages, introducing potential biases to data analysis (i.e. target language similarity to English or German could result in stronger algorithmic performance). Data quality could be influenced by similarly superficial environmental components such as the proportion of the recording occurring outside (resulting in wind interference) or the average number of family members present (resulting in overlapping, difficult-to-attribute speech). Individuals and communities may also differ in how naturalistically they act in the presence of the recorder or the types of activities that they choose not to record. Nevertheless, we envision long-form recordings making significant contributions to language documentation and revitalization in the coming years.

5 Discussion

In this chapter, we discussed three case studies to illustrate key opportunities and challenges of long-form recordings for societal good. We outlined how developmental scientists can make the most of the opportunities offered by this technology while tackling the challenges inherent to it.

5.1 Thinking beyond box-ticking

On privacy. Despite the many advantages of long-form recordings, we do want to raise the issue of privacy for their use in projects with a societal impact, including clinical practice, large-scale interventions, and language documentation. The ethical concerns of

long-form recordings (data security, privacy, research equity) are especially relevant for the families of multilingual children, children with language delays, participants in interventions, and speakers of underdocumented languages. For example, while any research project or intervention should bear in mind power differentials, we need to be particularly careful when using long-form recordings in these contexts because they may be viewed as a form of surveillance.

One specific concern that families voice about long-form recordings is privacy. Recorded families may fear that third parties could become privy to intimate details of their everyday lives. To this, it is important to stress that automation is at the heart of many long-form recording workflows: many specialists using this technique do not listen to the recording audio and instead rely on quantitative reports of children’s vocalization quantity per hour, for example. Sometimes, recordings are algorithmically sampled, to complement automated with qualitative metrics. Here again, the specialist would never listen to the whole recording, and only to small, interspersed portions. Are participating families assuaged by this proposal? It is important to ask, rather than deciding for them.

The problem with prioritizing privacy above all else. To address privacy concerns, some colleagues propose deleting recordings after automated analyses or clinical treatments are completed. Privacy is an important value in western ideals, and a basic component of an ethical research framework. For instance, in a piece on long-form recordings ethics, we wrote of the importance of “respect of autonomy, respect for privacy and confidentiality, beneficence, justice, and protection of vulnerable populations” (p. 1953, Cychosz et al., 2020). However, we also believe that these principles alone cannot determine the best course of action in all situations for all communities (Carroll et al., 2020; MacAskill, 2018). To this end, we complement a value-based approach with a *consequentialist framework*, one that emphasizes both the positive and negative repercussions of data collection and analysis practices.

The history of science is replete with scientists who had good intentions, and followed

all ascribed rules, but nevertheless failed to ensure that their research benefited participants and communities (see examples in Crooks, Donenberg, and Matthews 2021). Many negative outcomes follow from such errors (Callier, Husain, & Simpson, 2014): Communities can feel used and either stop participating or ask that their data be accessible only to the person who collected it. Both of these results mean that the community’s data will not be represented in machine learning training sets (leading to further algorithmic bias), and that less work, which could be useful to the community, will be conducted on the data.

While data destruction may be the only option in some contexts, we seldom recommend it for recordings of children’s learning environments as these data present innumerable opportunities for re-use and could contribute—even anonymously—to the World Behavior Bank that we propose in Section 5.2. Destroying the data, while protecting participants’ privacy, also means participating families would not be able to benefit from insights from improved or new algorithms. Recall that algorithms involve a specific view on desirable behaviors, and further reflect properties of their training. Indeed, for all of the case studies we discussed here, algorithmic bias is of particular concern, because the training data for most algorithms available today will be primarily derived from normo-typical, urban, mainstream participants; and even when they are not, the training data may not be a representative sample of a community. This is why we believe it is crucial to involve relevant communities from the very onset of technique development, including training the algorithms and developing metrics, to the very final stages, including long-term archiving of these data (see Section 5.2 on our World Behavior Bank proposal). Doing so will improve algorithmic accuracy, reduce bias, and increase the metrics that can be measured automatically, further protecting participant privacy.

Even when communities may have been open to partnerships that include long-term archiving, they are often not given the chance, with researchers and other agents making the choice *for* the community by agreeing with their ethics review board that they will

delete the data upon project completion. The same goes for the European General Data Protection Regulation, which was written with the aim of protecting individuals, but whose interpretation often leads researchers and other agents to agree to quickly delete data. Doing so actually goes against the conceptual principle of informed consent: The ethics board or regulation pushes for a short-sighted decision that reduces friction but whose consequences, in the long term, are negative to the community (and in some cases, the participants), not to mention to society at large, via the funding of additional bouts of the same data collection which would have been unnecessary had the original data been securely shared. Ideally, regulations and laws for the protection of individuals and groups should take into account positive *and* negative impacts, both in the short and longer term (Bu, 2021).

We think these issues are relevant to developmental scientists, policy makers, educators, clinicians, pediatricians, and students of all of these disciplines that come across this chapter, and finding the best solutions will require bold interdisciplinary reflection. To give an example of how progressive thinking in this framework can lead to useful insights, we put forward the idea of a World Behavior Bank.

5.2 A proposal for a World Behavior Bank

Long-form recordings go hand in hand with machine learning tools: The more the latter are developed, the greater the potential use of long-form recordings as more data are incorporated into measurements and diagnostics. Previous developments in machine learning, ranging from face recognition to genomics research, have been plagued by biased training sets that were not representative of all the diverse samples that the algorithms would later need to classify. This is why a key call in this chapter is to establish a *World Behavior Bank*, which can harness the power of large amounts of naturalistic data, collected from a host of communities, to help current and future generations avoid errors of bias in developmental science.

The idea of the World Behavior Bank is to set up a secure repository where practitioners, researchers, and others who collect long-form, audio(-visual) recordings can deposit their data for additional, future uses. Note that since these are digital data, the collections are not (and in fact should not) all be in the same physical location: different states and countries should have local servers for their data. However, algorithms for certain operations, such as systems to recognize different language sounds or distinguish between voices, could be trained across all or many datasets held in the bank.

Establishing a World Behavior Bank does not grant unlimited access to individuals. It doesn't even mean that users would ever listen to the hosted recordings. Appropriate securities can be put in place so that there is no illicit access or extraction of data against individuals' or communities' wills. For instance, to ensure data security, recordings could be tagged with the level of sharing that individuals and communities have agreed to. Some individuals and communities may wish for their data to enter the public domain immediately. Others may decide that such data should only become publicly available after the people recorded are no longer living, 50 or 100 years later. Yet other communities may feel that it is inappropriate for the recordings to ever be public—for instance, if the data contain sacred songs or rituals that are not meant to be experienced by outside members.

Although the World Behavior Bank would facilitate different levels of data sharing at different points in time, even data that are tagged to remain fully, permanently private, could be queried. These queries could then be monitored and limited so that identifying or sensitive information in the recordings is not retrieved (e.g., Bater et al. 2017). As for fully public data, a system of online annotation would allow re-users to define new categories to train novel classifiers, which could then be used to extract statistics from both public and private datasets.

The World Behavior Bank would facilitate a number of goals, including: the development of norms based on a great deal more (and more varied) data, which would help with clinical uses (Section 2); the development of metrics for behaviors targeted in

interventions but for which no classifier currently exists (e.g., detecting breastfeeding versus eating for a nutrition intervention; Section 3); and the novel study of language/dialect use based on everyday speech to look at the prevalence of different typological patterns like word orders (Section 4). Any commercial development that uses the data would require not only upfront payment, but also a percentage of the commercial gains (see Riley 2000 for an egregious example of commercialization gone awry). This financial model not only helps ensure that the bank’s longevity, but also that benefits are redistributed back among the original participants and/or communities.

Of course, there remain many outstanding questions about the bank: What would the governance structure of the bank be? How would the bank be maintained over many decades, to ensure no one sells or capitalizes on the data? How can data be protected against other malicious uses? These are important questions which require serious thinking, and it would also be ideal to attempt a cost-benefit analysis of the whole enterprise before attempting it.

5.3 Increasing diversity

We feel confident that the recommendations provided here are informed by current best practices and research models. Nevertheless, we cannot be sure that our recommendations won’t change as audio-visual processing technologies advance and societal norms around privacy and data management evolve. One specific issue with both best practices and current research models is that these norms were produced by a sliver of humanity where diverse views were not necessarily represented. Consider the World Behavior Bank: The bank would probably benefit future generations, but its construction is costly for individuals developing it today. Within an individualist framework, which is often the basis of even declarations of human rights, there is little motivation to impose negative outcomes on the individuals for the promise of group gains, especially future group gains.

To overcome this hurdle, we believe that greater diversity among developmental scientists, policy makers, educators, clinicians, pediatricians, and others reading this chapter is needed. We know most about non-representativity in developmental science, and long-form recording research in particular. Figure 2 displays some of these patterns: The middle and bottom panels draw from two systematic reviews of long-form data (Cristia, Bulgarelli, & Bergelson, 2020; Ganek & Eriks-Brophy, 2018). We are thrilled to hear calls for increased diversity in developmental science (Moriguchi, *to appear*), and we look forward to other similar proposals that actually attack the problem at its root, and improve representation.

6 Conclusion

In conclusion, this chapter aimed to highlight opportunities and challenges facing developmental scientists and how the technique of long-form recordings can help us have the most positive impact in the world. There are other domains in which long-form recordings could be useful (e.g., for continuous tracking of developmental disorders, assessment of teacher training, etc.), and we hope other work similarly lays out how to increase the chances that the net impact is positive for these domains as well.

7 Acknowledgements

AC is grateful for the institutional and financial support of Agence Nationale de la Recherche (ANR-17-CE28-0007 LangAge, 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 European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (ExELang, Grant agreement No. 101001095). MC was supported by National Institute on Deafness and Other Communication Disorders Grants T32DC000046 and F32DC019539. The authors thank Ginny Dawson, Lisa Hamrick, Julia Nee, and Kathleen Oppenheimer for their feedback on previous versions of this chapter, as well as Sara Pisani for inputting data

for Figure 2. AC particularly acknowledges many helpful discussions with Emmanuel Dupoux and the Echolalia team. Both authors contributed equally to this work.

Table 1

Extreme examples of potential uses for long-form recordings.

Broad and large-scale	Specific and qualitative
<p>Aera is a developmental economist carrying out early childhood interventions in South Korea. In one intervention, parents receive home visits where they are encouraged to talk frequently to their child. In another, the same funding for the visits is instead given to families in an unconditional cash transfer. A third group of families serves as a control. There are 500 families in each group. Aera plans to assess intervention outcomes using long-form recordings. Specifically, she plans to count (1) adult words in the recordings of the visit and control groups to assess whether the family visits modified parental behavior and (2) child vocalization counts in the recordings of the visit and cash transfer groups to compare if both interventions equally improved children’s speech-language outcomes. Given the intervention’s scale, all data analysis is automated. Aware of the fact that a previous validation study in Korean suggested acceptable correlations for algorithmically-derived adult word counts, but not child vocalization counts, Aera is working with a local machine learning expert to train a classifier to count child vocalizations in the recordings more accurately.</p>	<p>Julie is a field linguist who has been working together with speakers of Tiwa to document their language, currently spoken in a handful of small villages. She recorded 5 families where all household members are native Tiwa speakers. She then used an automated voice classification algorithm to identify sections of the recording that contained interactions between the key child, who wore the recording device, and others. She then played back these sections of the recording to each of the families, and they discussed what was happening and which languages were being used. For each recording section they also discussed why the families thought they used their native language versus the lingua franca. Together, Julie and the families found that the lingua franca was often used when discussing current affairs, whereas the native language was often used during story telling. The 5 families agreed to contribute some of these stories to a baby-friendly audio-book that Julie will produce and bring back on her next visit.</p>

Figure 2. Non-representative sampling of continents' populations in long-form recording research. Top: Children actually growing up in different continents (source: statista.com, 2020). Middle: Child samples in papers from two systematic reviews. Bottom: Affiliation of first authors in the same papers. The same color coding is used throughout.

8 References

- Albers, C. A., & Grieve, A. J. (2007). Test review: Bayley, n.(2006). bayley scales of infant and toddler development—third edition. san antonio, tx: Harcourt assessment. *Journal of Psychoeducational Assessment*, 25(2), 180–190.
- American Speech-Language-Hearing Association. (2017). *Issues in ethics: Cultural and linguistic competence*. Author.
- American Speech-Language-Hearing Association. (2019). *Profile of ASHA members and affiliates, year-end 2019* (Annual Demographic & Employment Data).
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*, May, 23(2016), 139–159. Retrieved from <http://perma.cc/3VTJ-K5GP>
- Appiah, R. (2020). Community-based participatory research in rural african contexts: Ethico-cultural considerations and lessons from ghana. *Public Health Reviews*, 41(1), 1–13.
- Bater, J., Elliott, G., Eggen, C., Goel, S., Kho, A. N., & Rogers, J. (2017). Smcql: Secure query processing for private data networks. *Proc. VLDB Endow.*, 10(6), 673–684.
- Belardi, K., Watson, L. R., Faldowski, R. A., Hazlett, H., Crais, E., Baranek, G. T., . . . Oller, D. K. (2017). A Retrospective Video Analysis of Canonical Babbling and Volubility in Infants with Fragile X Syndrome at 9 – 12 Months of Age. *Journal of autism and developmental disorders*, 47(4), 1193–1206. doi: 10.1007/s10803-017-3033-4
- Bin Morshed, M., Kulkarni, S. S., Li, R., Saha, K., Roper, L. G., Nachman, L., . . . Abowd, G. D. (2020). A Real-Time Eating Detection System for Capturing Eating Moments and Triggering Ecological Momentary Assessments to Obtain Further Context: System Development and Validation Study. *JMIR mHealth and uHealth*, 8(12), e20625. doi: 10.2196/20625
- Bowern, C. (2015). *Linguistic fieldwork: A practical guide* (Second ed.). New York, NY: Palgrave Macmillan.

- Broesch, T., Crittenden, A. N., Beheim, B. A., Blackwell, A. D., Bunce, J. A., Colleran, H., ... others (2020). Navigating cross-cultural research: methodological and ethical considerations. *Proceedings of the Royal Society B*, 287(1935), 20201245.
- Bu, Q. (2021). The global governance on automated facial recognition (afr): ethical and legal opportunities and privacy challenges. *International Cybersecurity Law Review*, 2(1), 113–145.
- Busso, M., Cristia, J., Hincapié, D., Messina, J., & Ripani, L. (2017). *Learning better: Public policy for skills development*. Inter-American Development Bank.
- Byers-Heinlein, K. (2013). Parental language mixing: Its measurement and the relation of mixed input to young bilingual children's vocabulary size. *Bilingualism: Language and Cognition*, 16(1), 32–48. doi: 10.1017/S1366728912000120
- Callier, S., Husain, R., & Simpson, R. (2014). Genomic data-sharing: what will be our legacy? *Frontiers in genetics*, 5, 34.
- Cao, X.-N., Dakhliya, C., del Carmen, P., Jaouani, M.-A., Ould-Arbi, M., & Dupoux, E. (2018). Baby Cloud, a technological platform for parents and researchers. In *Proceedings of the Language Resources and Evaluation Conference*. Miyazaki, Japan.
- Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., ... Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal*, 19, 43. doi: 10.5334/dsj-2020-043
- Casillas, M., & Cristia, A. (2019). A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings. *Collabra: Psychology*, 5(1), 24. doi: 10.1525/collabra.209
- Chatfield, K., Schroeder, D., Guantai, A., Bhatt, K., Bukusi, E., Adhiambo Odhiambo, J., ... Kimani, J. (2021). Preventing ethics dumping: the challenges for kenyan research ethics committees. *Research Ethics*, 17(1), 23–44.
- Child Language Research and Revitalization Working Group. (2017). *Language Documentation, Revitalization and Reclamation: Supporting Young Learners and*

- Their Communities* (Tech. Rep.). Waltham, MA:EDC.
- Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711), 20160055. doi: 10.1098/rstb.2016.0055
- Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). Unsupervised Cross-lingual Representation Learning for Speech Recognition. *arXiv:2006.13979 [cs, eess]*.
- Corbett, J., & Kulchyski, T. (2009). Anti-social computing: Indigenous language, digital video, and intellectual property. *Participatory Learning and Action*, 59(1), 52–58.
- Crago, M. B., Annahatak, B., & Ningiuruvik, L. (1993). Changing patterns of language socialization in inuit homes. *Anthropology & Education Quarterly*, 24(3), 205–223.
- 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. doi: 10.1044/2020_JSLHR-19-00017
- Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., . . . Bergelson, E. (2020). A thorough evaluation of the Language Environment Analysis (LENA) system. *Behavior Research Methods*. doi: 10.3758/s13428-020-01393-5
- Crooks, N., Donenberg, G., & Matthews, A. (2021). Ethics of research at the intersection of covid-19 and black lives matter: a call to action. *Journal of Medical Ethics*, 47(4), 205–207.
- Cychosz, M., Cristia, A., Bergelson, E., Casillas, M., Baudet, G., Warlaumont, A. S., . . . Seidl, A. (2021). Vocal development in a large-scale crosslinguistic corpus. *Developmental Science*. doi: 10.1111/desc.13090
- Cychosz, M., Romeo, R. R., Soderstrom, M., Scaff, H., Camila Ganek, Cristia, A., Casillas, M., . . . Weisleder, A. (2020). Longform recordings of everyday life: Ethics for best

- practices. *Behavior Research Methods*, 52, 1554–3528. doi: 10.3758/s13428-020-01365-9
- Cychosz, M., Villanueva, A., & Weisleder, A. (to appear). Efficient estimation of children’s language exposure in two bilingual communities. *Journal of Speech Language and Hearing Research*.
- de Barbaro, K. (2019). Automated sensing of daily activity: A new lens into development. *Developmental Psychobiology*, 61(3), 444–464. doi: 10.1002/dev.21831
- de Villiers, J. G. (2017). Unbiased language assessment: Contributions of linguistic theory. *Annual Review of Linguistics*, 3, 309–330.
- Fabiano-Smith, L., & Hoffman, K. (2018). Diagnostic Accuracy of Traditional Measures of Phonological Ability for Bilingual Preschoolers and Kindergarteners. *Language, Speech, and Hearing Services in Schools*, 49(1), 121–134. doi: 10.1044/2017_LSHSS-17-0043
- Fausey, C. M., Jayaraman, S., & Smith, L. B. (2016). From faces to hands: Changing visual input in the first two years. *Cognition*, 152, 101–107. doi: 10.1016/j.cognition.2016.03.005
- Fenson, L., Marchman, V., Thal, D. J., Dale, P., Reznick, J., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories User’s Guide and Technical Manual* (2nd Edition ed.). San Diego, CA: Singular.
- Ferjan Ramírez, N., Lytle, S. R., Fish, M., & Kuhl, P. K. (2019). Parent coaching at 6 and 10 months improves language outcomes at 14 months: A randomized controlled trial. *Developmental Science*, 22(3), e12762. doi: 10.1111/desc.12762
- Ferjan Ramírez, N., & Kuhl, P. (2017). Bilingual baby: Foreign language intervention in madrid’s infant education centers. *Mind, Brain, and Education*, 11(3), 133–143.
- Ganek, H., & Eriks-Brophy, A. (2018). Language ENvironment analysis (LENA) system investigation of day long recordings in children: A literature review. *Journal of Communication Disorders*, 72, 77–85. doi: 10.1016/j.jcomdis.2017.12.005

- Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeersch, C., Walker, S., ...
Grantham-McGregor, S. (2014). Labor market returns to an early childhood stimulation intervention in jamaica. *Science*, *344*(6187), 998–1001.
- Gilkerson, J., Richards, J. A., & Topping, K. (2017). Evaluation of a LENA-Based Online Intervention for Parents of Young Children. *Journal of Early Intervention*, *39*(4), 281–298. doi: 10.1177/1053815117718490
- Grother, P., Ngan, M., & Hanaoka, K. (2019). *Face recognition vendor test (fvrt): Part 3, demographic effects*. National Institute of Standards and Technology.
- Hamrick, L. R., Seidl, A., & Tonnsen, B. L. (2019). Acoustic properties of early vocalizations in infants with fragile X syndrome. *Autism Research*, *12*(11), 1663–1679. doi: 10.1002/aur.2176
- Hansen, R. L., Ozonoff, S., Krakowiak, P., Angkustsiri, K., Jones, C., Deprey, L. J., ...
Hertz-Picciotto, I. (2008). Regression in Autism: Prevalence and Associated Factors in the CHARGE Study. *Ambulatory Pediatrics*, *8*(1), 25–31. doi: 10.1016/j.ambp.2007.08.006
- Heath, S. B. (1983). *Ways with words: Language, life and work in communities and classrooms*. cambridge university Press.
- Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*(5782), 1900–1902.
- Hermes, M., & Engman, M. M. (2017). Resounding the clarion call: Indigenous language learners and documentation. *Resounding the clarion call: Indigenous language learners and documentation*, *14*, 59–87.
- Himmelman, N. P. (2006). Language documentation: What is it and what is it good for? In J. Gippert, N. P. Himmelman, & U. Mosel (Eds.), *Essentials of Language Documentation*. Berlin, New York: Mouton de Gruyter. doi: 10.1515/9783110197730.1
- Keller, H., Yovsi, R. D., & Voelker, S. (2002). The role of motor stimulation in parental

- ethnotheories: The case of cameroonian nso and german women. *Journal of cross-cultural psychology*, 33(4), 398–414.
- Kohnert, K., Yim, D., Nett, K., Kan, P. F., & Duran, L. (2005). Intervention With Linguistically Diverse Preschool Children: A Focus on Developing Home Language(s). *Language, Speech, and Hearing Services in Schools*, 36, 251–263.
- Laing, C., & Bergelson, E. (2020). From babble to words: Infants’ early productions match words and objects in their environment. *Cognitive Psychology*, 122, 101308. doi: 10.1016/j.cogpsych.2020.101308
- Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., & Cristia, A. (2021). An open-source voice type classifier for child-centered daylong recordings. *arXiv:2005.12656 [eess]*. (Comment: accepted to Interspeech 2020)
- Lavechin, M., de Seyssel, M., Gautheron, L., Dupoux, E., & Cristia, A. (2021). *Reverse-engineering language acquisition with child-centered long-form recordings* (Preprint). PsyArXiv. doi: 10.31234/osf.io/pt9xq
- Lehet, M., Arjmandi, M. K., Dilley, L., & Houston, D. (2020). Circumspection in using automated measures: Talker gender and addressee affect error rates for adult speech detection in the Language ENvironment Analysis (LENA) system. *Behavior Research Methods*.
- Leonard, W. Y., & Haynes, E. (2010). Making “collaboration” collaborative An examination of perspectives that frame linguistic field research. *Language Documentation*, 4, 26.
- MacAskill, W. (2018). Understanding effective altruism and its challenges. In *The palgrave handbook of philosophy and public policy* (pp. 441–453). Springer.
- Marchionni, P., Scalise, L., Ercoli, I., & Tomasini, E. (2013). An optical measurement method for the simultaneous assessment of respiration and heart rates in preterm infants. *Review of Scientific Instruments*, 84(12), 121705.
- Marschik, P. B. ., Pokorny, F. B., Peharz, R., Zhang, D., O’Muircheartaigh, J., Roeyers,

- H., ... Kaufmann, W. E. (2017). A Novel Way to Measure and Predict Development: A Heuristic Approach to Facilitate the Early Detection of Neurodevelopmental Disorders. *Current Neurology and Neuroscience Reports*, 17(5), 43. doi: 10.1007/s11910-017-0748-8
- Matheny, M. E., Whicher, D., & Thadaney, S. (2020). Artificial intelligence in health care: A report from the national academy of medicine. *Journal of the American Medical Association*, 323(6), 509–510.
- McDaniel, J., Yoder, P., Estes, A., & Rogers, S. J. (2020). Predicting Expressive Language From Early Vocalizations in Young Children With Autism Spectrum Disorder: Which Vocal Measure Is Best? *Journal of Speech, Language, and Hearing Research*, 63(5), 1509–1520. doi: 10.1044/2020_JSLHR-19-00281
- Moriguchi, Y. (*to appear*). Beyond bias to Western participants, authors and editors in developmental science. *Infant and Child Development*.
- Nee, J. (2021). Understanding the effects of language revitalization workshops using long-format speech environment recordings. *Proceedings of the Linguistic Society of America*, 6(1), 213. doi: 10.3765/plsa.v6i1.4967
- NIDCD. (2016). *Quick statistics about voice, speech, language* (Online Report). Bethesda, MD: National Institute on Deafness and Other Communication Disorders.
- Ochs, E., & Schieffelin, B. (2001). Language acquisition and socialization: Three developmental stories and their implications. In A. Duranti (Ed.), *Linguistic anthropology: A reader* (p. 263-301). Blackwell.
- Oller, D. K. (2000). *The emergence of the speech capacity*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., ... Warren, S. F. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences*, 107(30), 13354–13359. doi: 10.1073/pnas.1003882107

- Orena, A. J., Byers-Heinlein, K., & Polka, L. (2020). What do bilingual infants actually hear? Evaluating measures of language input to bilingual-learning 10-month-olds. *Developmental Science*, 23(2), 1–14. doi: 10.1111/desc.12901
- Ortiz, J. (2021). Using nonword repetition to identify language impairment in bilingual children: A meta-analysis of diagnostic accuracy. *American Journal of Speech-Language Pathology*.
- Overby, M., Belardi, K., & Schreiber, J. (2020). A retrospective video analysis of canonical babbling and volubility in infants later diagnosed with childhood apraxia of speech. *Clinical Linguistics & Phonetics*, 34(7), 634–651. doi: 10.1080/02699206.2019.1683231
- Patten, E., Belardi, K., Baranek, G. T., Watson, L. R., Labban, J. D., & Oller, D. K. (2014). Vocal Patterns in Infants with Autism Spectrum Disorder: Canonical Babbling Status and Vocalization Frequency. *Journal of Autism and Developmental Disorders*, 44(10), 2413–2428. doi: 10.1007/s10803-014-2047-4
- Place, S., & Hoff, E. (2011). Properties of Dual Language Exposure That Influence 2-Year-Olds’ Bilingual Proficiency: Dual Language Exposure and Bilingual Proficiency. *Child Development*, 82(6), 1834–1849. doi: 10.1111/j.1467-8624.2011.01660.x
- Räsänen, O., Seshadri, S., Karadayi, J., Riebling, E., Bunce, J., Cristia, A., . . . Soderstrom, M. (2019). Automatic word count estimation from daylong child-centered recordings in various language environments using language-independent syllabification of speech. *Speech Communication*, 113, 63–80. doi: 10.1016/j.specom.2019.08.005
- Räsänen, O., Seshadri, S., Lavechin, M., Cristia, A., & Casillas, M. (2020). ALICE: An open-source tool for automatic measurement of phoneme, syllable, and word counts from child-centered daylong recordings. *Behavior Research Methods*. doi: 10.3758/s13428-020-01460-x
- Ratner, N. B., & Pye, C. (1984). Higher pitch in BT is not universal: Acoustic evidence

- from Quiche Mayan. *Journal of Child Language*, 11(3), 515–522. doi: 10.1017/S0305000900005924
- Riley, R. (2000). Recovering collectivity: Group rights to intellectual property in Indigenous communities. *Cardozo Arts & Entertainment Law Journal*, 18(1), 175–226.
- 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 : JSLHR*, 61(6), 1369–1380. doi: 10.1044/2018_JSLHR-S-17-0287
- Seymour, H., Roeper, T., & de Villiwers, J. (2005). *Diagnostic evaluation of language variation - Norm Referenced (DELV_NR)*. San Antonio, TX: The Psychological Corporation.
- Sheinkopf, S., Mundy, P., Oller, D. K., & Steffens, M. (2000). Vocal Atypicalities of Preverbal Autistic Children. *Journal of Autism and Developmental Disorders*, 30, 345–54. doi: 10.1023/A:1005531501155
- Suskind, D. L., Graf, E., Leffel, K. R., Hernandez, M. W., Suskind, E., Webber, R., . . . Nevins, M. E. (2016). Project ASPIRE: Spoken Language Intervention Curriculum for Parents of Low-socioeconomic Status and Their Deaf and Hard-of-Hearing Children. *Otology & Neurotology*, 37(2), e110-e117. doi: 10.1097/MAO.0000000000000931
- Swanson, M. R., Donovan, K., Paterson, S., Wolff, J. J., Parish-Morris, J., Meera, S. S., . . . Piven, J. (2019). Early language exposure supports later language skills in infants with and without autism. *Autism Research*, 12(12), 1784–1795. doi: 10.1002/aur.2163
- The Many Babies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed speech preference. *Advances in Methods and Practices in Psychological Science*, 3(1), 24–52.

- VanDam, M., & Yoshinaga-Itano, C. (2019). Use of the lena autism screen with children who are deaf or hard of hearing. *Medicina*, 55(8), 495.
- Walker, S., Powell, C., Chang, S., Baker-Henningham, H., Grantham-McGregor, S., Vera-Hernandez, M., & López Bóo, F. (2015). Delivering parenting interventions through health services in the caribbean.
- Wang, Y., Williams, R., Dilley, L., & Houston, D. M. (2020). A meta-analysis of the predictability of lena™ automated measures for child language development. *Developmental Review*, 57, 100921.
- Warlaumont, A. S., Richards, J. A., Gilkerson, J., & Oller, D. K. (2014). A Social Feedback Loop for Speech Development and Its Reduction in Autism. *Psychological Science*, 25(7), 1314–1324. doi: 10.1177/0956797614531023
- Wylie, K., McAllister, L., Davidson, B., & Marshall, J. (2013). Changing practice: Implications of the world report on disability for responding to communication disability in under-served populations. *International journal of speech-language pathology*, 15(1), 1–13.
- Yao, X., Micheletti, M., Johnson, M., & de Barbaro, K. (2020). *Detection of infant crying in real-world home environments using deep learning* [arXiv Preprint].
- Yoder, P. J., Oller, D. K., Richards, J. A., Gray, S., & Gilkerson, J. (2013). Stability and Validity of an Automated Measure of Vocal Development From Day-Long Samples in Children With and Without Autism Spectrum Disorder. *Autism Research*, 6(2), 103–107. doi: 10.1002/aur.1271