Exploring variation in infants' preference for infant-directed speech: Evidence from a multi-site study in Africa

Angeline Sin Mei Tsui[1], Alexandra Carstensen[1], George Kachergis[1], Amina Abubakar[2], Mulat Asnake[3], Oumar Barry[4], Dana M. Basnight-Brown[5], Dangkat Bentu[6], Christina Bergmann[7], Evans Binan Dami[6], Natalie Boll-Avetisyan[8], Marguerite de Jongh[9], Yatma Diop[10], Reginald Akuoko Duah[11], Esther Herrmann[12], Chaning Jang[13], Simon Kizito[14], Tilinao Lamba[15], Limbika Maliwichi-Senganimalunje[15], Joyce Marangu[2], Maya Mathur[1], Catherine V. Mbagaya[16], Demeke Mekonnen Mengistie[17], Carmen Milton[9], Febronie Mushimiyimana[18], Mikateko Ndhambi[9], Irene Ngina[13], Eunice Njoroge[2], Paul Odhiambo Oburu[16], Asana Okocha[19], Paul Okyere Omane[8], Anisha Singh[13], Andrew S. Ssemata[14], Juliette Unyuzumutima[18], Henriette Zeidler[20], Casey Lew-Williams[19], & Michael C. Frank[1]

[1] Stanford University

[2] Institute for Human Development, Aga Khan University, Kenya

[3] Addis Ababa University, Ethiopia

[4] Cheikh Anta Diop University - Dakar, Senegal

[5] United States International University-Africa, Kenya

[6] University of Jos, Nigeria

[7] Max Planck Institute for Psycholinguistics, The Netherlands

[8] University of Potsdam, Germany

[9] Sefako Makgatho Health Sciences University, South Africa

[10] Michigan State University, USA

[11] Humboldt-Universität, Berlin and University of Ghana, Legon

[12] University of Portsmouth, UK

[13] Busara Center for Behavioral Economics, Kenya

[14] Makerere University, Uganda

[15] University of Malawi, Chancellor College, Zomba, Malawi

[16] Maseno University, Kenya

[17] St. Peter Specialized Hospital, Ethiopia

[18] University Teaching Hospital of Kigali, Rwanda

[19] Princeton University, USA

[20] Aston University, UK & University of Gothenburg, Sweden

Author Note

Correspondence concerning this article should be addressed to Michael C. Frank, 450 Jane Stanford Way, Stanford, CA 94305-2130, USA. E-mail: mcfrank@stanford.edu

Abstract

Infants show a preference for infant-directed speech (IDS) over adult-directed speech (ADS). This preference has been linked to infants' language processing and word learning in experimental settings, and also correlates with later language outcomes. Recently, the cross-cultural consistency of infants' IDS preference has been confirmed by large-scale, multisite replication studies, but conclusions from these studies were primarily based on participants from North America and Europe. The current study addressed this sampling bias via a large-scale, multisite study of infants (3-15 months) across XYZ communities in Africa. We investigated whether participants showed a preference for IDS over ADS, and if so, whether the magnitude of their preference differs from effects documented in other populations of infants. DESCRIBE RESULTS HERE

Keywords: infant-directed speech; reproducibility; Africa; infants; generalizability

Word count: XYZ

Exploring variation in infants' preference for infant-directed speech: Evidence from a multi-site study in Africa

Adults often speak to infants differently than to other adults, using a speech register known as infant-directed speech (IDS). Infant-directed speech tends to have exaggerated prosodic characteristics, including higher pitch, greater pitch variation, longer pauses, simplified grammatical structure, and shorter and slower utterances as compared to adult-directed speech (ADS; e.g., Fernald et al., 1989, Trainor & Desjardins, 1997). Even very young infants from a variety of language backgrounds have a preference for listening to IDS over ADS (e.g., Cooper & Aslin, 1994; Cooper, Abraham, Berman, & Staska, 1997; Fernald, 1985; Hayashi, Tamekawa, & Kiritani, 2001; Kitamura & Lam, 2009; Newman & Hussain, 2006; Pegg, Werker, & McLeod, 1992; Santesso, Schmidt, & Trainor, 2007; Singh, Morgan, & Best, 2002; Werker & McLeod, 1989). Infants' preference for IDS over ADS has also been demonstrated in a meta-analysis; across 34 studies, IDS preference had a fairly large average effect size with a value of Cohen's d 0.72 (Dunst, Gorman & Hamby, 2012) (Bergmann et al., 2018).

Why do infants prefer IDS? Perhaps IDS is intrinsically salient to infants because of its perceptual characteristics (e.g., higher pitch, greater pitch variability). Or perhaps, as infants are exposed to IDS, familiarity leads to preference. These explanations have different developmental predictions: while the intrinsic view would suggest an early preference (e.g., Cooper & Aslin, 1990), the exposure account would predict developmental increases in preference. Further, these explanations are not mutually exclusive: infants' early preference for IDS may motivate their parents to use more IDS, which in turn could lead infants to show a stronger IDS preference. Regardless of its origins, infants' preference for IDS may benefit their early language development. For example, in experimental studies, infants can segment words better in fluent speech produced in IDS than ADS (Thiessen, Hill & Saffran 2005), show better recognition of words introduced in IDS after a

24-hour delay (Singh, Nestor, Parikh, Yull, 2009), and more successfully learn words from IDS than ADS (Graf Estes & Hurley, 2013).

Further evidence comes from correlational studies, which have found that the amount of IDS in the language environment is positively related to children's language outcomes, such as vocabulary size (e.g., Ramirez-Esparza, Garcia-Sierra and Kuhl, 2014; Shneidman, Arroyo, Levine & Goldin-Meadow, 2013; Shneidman & Goldin-Meadow, 2012; Weisleder & Fernald, 2013; but cf. Casillas, Brown & Levinson, 2020; 2021, who found similar timing of language development milestones even in a population that hears very limited IDS). Together, this work suggests that infants' preference for IDS over ADS may support their language development, which explains why infants' IDS preference continues to be an important topic in the literature on early childhood.

However, it is important to note that almost all prior studies, including the meta-analysis by Dunst and colleagues (2002), have included mainly infants learning English in Western, educated, industrialized, rich, and democratic (WEIRD) societies (Henrich, Heine, & Norenzayan, 2010), with only a few studies extended to non-Western infant populations learning languages other than English (Hayashi et al., 2001; Werker, Pegg, & Mcleod, 1994). As such, there is a large sampling bias in the existing data about infants' preference for IDS, as in many other research topics in developmental psychology (see Nielsen, Haun, Kärtner, & Legare, 2017). This sampling bias is a problem for generalizing findings about infants' IDS preference to infants growing up in different cultures and learning different languages. In light of this generalizability issue – as well as the recent replication crisis in psychology (e.g., Open Science Collaboration, 2015) – infant researchers have begun to collaborate on large-scale, multi-site studies to replicate key developmental findings (Frank et al., 2017).

One of these multi-site projects investigated infants' preference for IDS over ADS: the ManyBabies1 study (MB1; ManyBabies Consortium, 2020). MB1 collected monolingual

data from 67 laboratories, with a total sample of 2329 monolingual infants 3 – 15 months old. The protocol for this experiment was simple: infants listened to alternating audio clips of IDS and ADS while viewing an uninformative visual stimulus (a colored checkerboard). Their looking time was measured over the course of up to 16 trials, 18s each in length (8 IDS and 8 ADS). Notably, all participants in the study listened to stimuli that were constructed from naturalistic speech by North American mothers (speaking either to another adult or to their own infant). The mismatch between the stimuli and the native language of many infants in the study allowed inferences about native language effects and also minimized variability due to differences in the stimuli (a follow-up project now in progress seeks to measure native-language preferences in a subset of MB1 labs). Overall, older infants showed a stronger preference for IDS than younger infants. There was also an effect of infants' language backgrounds: North American infants exhibited a stronger IDS preference than infants who were not exposed to North American English (NAE). Although infants' ages and language backgrounds affected the magnitude of IDS preference, essentially all groups of infants preferred NAE IDS over ADS.

Despite the breadth of its sample relative to previous work, the MB1 study still constitutes a biased sample of infant populations in the world. Most of the data in MB1 were contributed by laboratories in economically-advantaged areas, accessing relatively high socio-economic status participant populations. Further, although this large-scale study had a diverse sample from 17 countries, 60 out of the 67 participating laboratories were from Europe and North America, only a handful of laboratories were from Australia and Asia, and none were from Africa or South America. Thus, the sample studied in MB1 came almost exclusively from Western, educated, affluent populations who heard Indo-European languages, limiting the generalizability of the findings to infants growing up in other cultural and linguistic contexts. This lack of evidence on generalizability of a key finding about infants' preference restricts our ability to build robust developmental theories of language learning across cultural contexts. Our current study takes a step towards

addressing this gap.

We investigate whether infants growing up in a variety of African cultures show an IDS preference, using the paradigm developed by the MB1 study. Our study has both a theoretical goal and a practical goal. Theoretically, we are interested in whether IDS preference is a culturally and linguistically invariant developmental pattern (Neilson et al., 2017). The inclusion of infants across many African cultures (who are acquiring many different languages, see Table 1) provides an important test of generalizability of the IDS preference. Practically, increasing sample diversity also promotes diversity among researchers engaged in developmental science and hopefully increasing exchanges between researchers across cultures. Thus, one goal of our study is building research networks to facilitate further studies with the communities represented in the current study.

Our study builds on a foundation of prior descriptive work investigating the generality of IDS across cultures. Although this work has investigated a variety of different cultures and languages, it can be (and often is) crudely summarized via the distinction between WEIRD and non-WEIRD cultures discussed above. We follow this convention here without endorsing this distinction as necessarily being meaningful in the context of our study, as IDS in WEIRD and non-WEIRD cultures shares similar prosodic properties. For example, Broesch and Bryant (2015) reported that IDS produced by North-American mothers, as well as by Kenyan and Fijian mothers, is produced with higher pitch, greater pitch variation, and is spoken at a slower rate than ADS. This finding is consistent with past work reporting that IDS shares some common exaggerated prosodic features (e.g., higher pitch, larger pitch variation) across diverse languages, which include French, Italian, German, Japanese, British English, American English (Fernald et al., 1989), Mandarin Chinese (Grieser & Kuhl, 1988), Thai, Australian English (Kitamura et al., 2001), Arabic (Farran, Lee, Yoo & Oller, 2016).

IDS can also be recognized as being infant-directed by listeners from non-WEIRD

cultures. Bryant, Liénard and Barrett (2012) reported that Turkanan adults in Kenya can discriminate between NAE IDS and ADS (see similar results in Bryant & Barrett, 2007 for Shuar hunter horticulturists from Amazonian Ecuador). These studies are consistent with findings from the MB1 studies showing that children who are not learning NAE, including children from Singapore and Korea, nonetheless show a preference for NAE IDS over ADS. Taken together, the common acoustic properties of IDS across different languages, and how NAE IDS can be recognized by non-native participants, raise the possibility of infants' IDS preference over ADS being quite consistent across different cultures and languages. However, it is possible that the strength of this preference will nonetheless be influenced by similarity between the test language (English) and the language(s) that each infant is learning, which could bolster the measured preferences to the extent that test and native language are similar (as in the case of infants learning other Indo-European languages with similar phonetic and acoustic properties). If this is the case, we expect that phylogenetic similarity between Indo-European languages and our stimuli would lead to comparable or stronger observed IDS preferences in samples of infants learning Indo-European languages than those learning languages in other families (e.g., Bantu, the language family we expect to be most prevalent in our sample).

Despite evidence for general recognition of and preference for IDS across cultures, the strength of IDS preferences is likely modulated by exposure. Exposure to IDS in the home environment varies widely both within and between cultures (Casillas et al., 2020; 2021; Cristia, Dupoux, Gurven & Stieglitz, 2017; LeVine et al., 1994; Shneidman & Goldin-Meadow, 2012; Vogt, Mastin, & Schots, 2015). Differences in IDS quantity have also been hypothesized to reflect differences in child-rearing practices across cultures. For example, direct verbal interaction between parents and infants can be rare in some societies (Heath, 1983; LeVine et al., 1994; Shneidman & Goldin-Meadow, 2012; Weber, Fernald, & Diop, 2017; LeVine & LeVine, 2016). Children in these societies – which are typically non-WEIRD, though certainly not all non-WEIRD societies can be characterized this way

– are often expected to learn through observation and participation according to their skill levels (see Legare, 2019 for a review). Thus, infants and young children in such societies may hear less IDS directly from their caregivers than those in WEIRD societies in which the norm involves a greater degree of direct address to parents. Of course, variation is also present within as well as across cultures. Within-culture variation has primarily been studied in North American contexts, where children from higher socioeconomic status (SES) families tend to hear more IDS than children from lower SES families (e.g., Hart & Risley, 1995; Hoff, 2003; Huttenlocher, Waterfall, Vasilyeva, Vevea & Hedges, 2010; Rowe, 2012; Shneidman & Goldin-Meadow, 2012; Weisleder & Fernald, 2013).

By virtue of our broad sample of African cultures, we expect that our study will likely capture substantial cultural variation in the average amount of IDS in children's environments. The African sites we sample vary widely in their degree of urbanization, their culture, their parenting values, and the average resources available in children's home environments – all of which have been argued to be meaningful dimensions governing children's early linguistic environment. For example, Keller (2012) suggested three prototypical cultural environments for children based on the degree of urbanization of the families in Western and non-Western societies. In this framework, in Western middle-class urban societies, highly educated parents generally aim to help children develop individual psychological autonomy. In contrast, in non-Western rural subsistence-based societies, parents generally aim to help children develop communal action autonomy, so that children have a strong sense of social responsibility and can contribute to the economic functioning of the family (e.g., farming). Importantly, non-Western middle-class urban societies are a hybrid of non-Western, rural and Western, urban societies, where parents generally want their children to develop more individual autonomy but also emphasize the importance of social responsibility in a large family. Broadly speaking, we expect that the African families in our study will be from the non-Western, urban and non-Western, rural groups in this taxonomy (see Table 1).

The confirmatory analyses of our study are designed to test whether there are differences in the magnitude of IDS preferences measured in this sample and in the prior samples of MB1. Although the average IDS production in the African sites we examine is unknown, consistent differences along this dimension might plausibly lead to variation in the magnitude of IDS preferences between our current study and MB1. In addition, our exploratory analyses will attempt to understand whether variation in IDS preference among infants in our sample of African cultures is explained by demographic proxies related to this taxonomy (e.g., urbanization and/or socioeconomic status). Finally, we will use an exploratory measure of subjective IDS use as a proxy of IDS quantity within families to probe links between parent reported IDS use and infant preference.

Since multilingualism is common in Africa (e.g., Posel & Zeller, 2016; Rosenhouse & Goral, 2008), many African children begin learning two or more different languages during infancy. Will early multilingualism alter infants' preferences for IDS? The ManyBabies1-Bilingual (MB1B) study provides some evidence that bilingual infants showed a similar preference for NAE IDS when compared to monolingual infants (Byers-Heinlein et al., in press). MB1B examined bilingual infants' preference for NAE IDS at 6 to 9 months and 12 to 15 months and found that bilingual and monolingual infants did not differ in terms of the magnitude of their IDS preferences. MB1B also found similar results to MB1, that older bilingual infants and those bilinguals with higher exposure to NAE show stronger IDS preference. However, as in the MB1 study, data collected in MB1B mainly came from laboratories in WEIRD areas, such as North America and Europe, with no laboratories from Africa, so the same caveats of generalizability apply to MB1B as to MB1. Thus, in the current study, we will include both monolingual and multilingual infants, allowing us to assess the generalizability of MB1B's conclusions to our samples in Africa.

In sum, there are three primary (confirmatory) goals for the current study. First, we aim to measure infants' preference for North-American English IDS across a range of cultural and linguistic contexts in Africa. Second, we seek to measure developmental

changes in this preference. As we found that older infants show stronger IDS preferences than younger infants in both MB1 and MB1B, we will evaluate whether participants in our study show the same developmental increases in IDS preference. Finally, we will investigate whether there are differences in IDS preferences between infants in Africa in our study and those in Europe and Asia in MB1 and MB1B. As an exploratory aim, we also will examine relationships between parents' demographics, their responses to survey items regarding subjective use of IDS, and their child's IDS preference.

Methods

Participation Details

Time-frame. On July 23, 2018, we issued an open call for participation by African researchers via listservs and social/professional networks. In total, XYZ laboratories agreed to participate (See Table 1 for target sample characteristics of each site). Our participating laboratories will recruit infants living in eastern (e.g., Kenya), western (e.g., Senegal) and southern (e.g., South Africa) regions of Africa. We also note that many of our participating laboratories are located in East Africa, thus East African participants are disproportionately represented in our sample. Given the current state of the COVID-19 outbreak, all of our participating sites are currently under lockdown. So, we plan to start data collection as soon as the outbreak is controlled, and the participating sites reopen. We expect data collection to be finished after 12 months. Data collection began on DATE and ended one year later, on DATE. Further, we expect to complete data analysis and write up the whole manuscript for Stage 2 submission 4 months after data collection.

Age distribution. Each participating laboratory was asked to recruit participants in two age bins: 3;0 – 9;0 and 9;1 – 15;0 months. Similar to MB1, each laboratory was asked to collect data spanning the age bin window, but aiming for the mean of the age bin.

Sample size determination. We estimated the effect size of infants' IDS preference on the basis of the data from MB1. We used data from laboratories in MB1 that used the single-screen central visual-fixation preference procedure (which we also use here: see below) and that tested infants with no exposure to North American English (similar to our population of interest). In a mixed-effects model, we examined the effect of test trial type (IDS vs. ADS) on infants' looking time (log-transformed seconds), while controlling for normally-distributed random intercepts by infant and laboratory. The intercept, representing infants' average log-looking time across ADS trials, was 1.91; the variances of the random intercepts were 0.074 and 0.022 at the infant and laboratory levels respectively. The fixed-effect coefficient representing infants' preference for IDS over ADS was 0.080 and the residual variance was 0.33.

In the first power analysis, we simulated datasets based on the above coefficient estimates and variances. Using the simr package in R (Green & MacLeod, 2016), we ran a power analysis for a mixed-effect analysis with the above-mentioned simulated datasets (number of simulations = 1000). We were uncertain exactly how many labs to assume but settled on 10, given the likelihood of some later signups as well as some lab attrition. Assuming that we had 240 infants across 10 laboratories in each simulated dataset and an alpha level of 0.05, we found that the average power was 99.40% [95% confidence interval: 98.70% – 99.78%] to detect the fixed ADS vs. IDS coefficient of 0.08. This first power analysis was based on very small random-effect variances estimated from MB1 and MB1B datasets. Given that most of the laboratories that participated in MB1 and MB1B had more resources and more extensive experience in running infancy studies in comparison to the participating laboratories in Africa, we planned for potentially higher variances in the data collected in the current project. Thus, we ran a conservative second power analysis by doubling the values of the random intercept and residual variances reported in the datasets from MB1 and MB1B, while holding constant the intercept and the fixed-effect coefficient representing infants' preference for IDS over ADS. With larger variances, the average

power estimate dropped to 87.20% [95% confidence interval: 84.97% – 89.21%] for a total sample of 240 infants. The power analysis can be found at https://osf.io/jgr79.

Prior to submission of the Stage 1 report, we had 11 laboratories committed to collecting data for this project. Given that MB1 reported around 15% data excluded in the final analysis, we expect the exclusion rate for our project is around 15% to 20%. Thus, each laboratory agreed to contribute a minimum of 32 infants (16 infants in each age bin), including infants tested but excluded for reasons not related to the demographic and age inclusion criteria (e.g., fussiness). Further, we encouraged each laboratory to contribute additional data beyond that minimum. We propose that our projected sample size of 352 would have sufficient power, as 80% of this sample size exceeds our targeted final sample size (n = 240) based on the power analysis described above.

Ethics. All laboratories collected data under their own independent IRB protocol. Videos of individual infant participants during the experiment were recorded and stored at each laboratory. However, these videos were not shared with the central data analysis team. Laboratories were instead asked to only submit de-identified data for central data analyses.

Exclusion Criteria

All data collected for the study (i.e., every infant for whom a data file was generated, regardless of how many trials were completed) were uploaded to a central database for data analysis. Every laboratory followed the protocol to report any infants who were tested in this study, including those who were excluded from the analysis. Furthermore, each laboratory followed the protocol to make note of the reasons that infants were excluded from the study.

Typically, participants were only included in the analysis if they met all of the criteria below. However, we allowed parents to choose not to answer some of the questions (e.g., about full-term gestation and developmental disorders) because disclosures might

violate cultural norms in some areas of Africa. Thus, participating laboratories may have included infants who did not fully meet the inclusion criteria defined here:

Full-term. We defined full term as gestation times greater than or equal to 37 weeks. XYZ (%XYZ) of infants tested did not meet this criterion, and were excluded from further analysis. To maximize parents' comfort in participating in the experiment, they were given the option of not responding to questions about gestation.

No developmental disorders or hearing loss. We excluded infants with parent-reported developmental disorders (e.g., chromosomal abnormalities, etc.) or diagnosed hearing impairments. Developmental disorders and delays are stigmatized in some cultures in Africa (e.g., negative attitudes towards children with disorders or delays), therefore some parents may decline to answer the question about children's developmental disorders. In this case, we still tested the infants and included the infants' data in the analysis. This inclusion criterion was chosen to allow us to retain as much data as possible while ensuring our questionnaire accommodates cultural norms. Further, we noted that only 2 participants (i.e. less than 0.1%) in MB1 were excluded based on parents' report of developmental disorders. Accordingly, we do not expect that including children whose parents decline to answer this question will lead to an inclusion of large numbers of children with developmental disorders that could potentially skew the results in the study. XYZ (%XYZ) of the infants tested did not meet this criterion. (We did not plan exclusions based on self-reported ear infections unless parents reported medically-confirmed hearing loss.)

Trial-level and session-level errors. Following MB1 and MB1B, we adopted a relatively liberal inclusion criterion for this study. To be included in the study, a child must have contributed non-zero looking time on at least one pair of test trials (i.e., one trial each of IDS and ADS from a particular stimulus pair). We asked laboratories to identify two different types of errors when uploading their data: trial-level errors and session-level errors. Trial-level exclusions were based on whether we could use infants' data from a particular test trial. For example, if an infant only completed the first six test trials of the

experiment, we entered this infant's data from the first six trials and discarded data from all other trials. In this case, laboratories would identify this infant's data from the first to sixth trials as "no trial errors" and any trials from the seventh trial onwards would be identified as "trial errors". In contrast, session-level errors were errors that occurred when running a particular participant. This type of error is different from the trial-level error exclusions because it indicates that errors occurred which affected an entire session (e.g., failure to save data in the experiment). If a laboratory indicated a session-level error for a particular infant, all data from this infant was excluded from the analysis. In sum, infants who can contribute at least one pair of test trials (i.e., one IDS trial and one ADS trial) would have some data excluded at the trial level whereas infants who cannot contribute one pair of test trials would be excluded at the session level. In general, errors included the following: equipment error (e.g., no sound or visuals on the first pair of trials), experimenter error (e.g., an experimenter was unblinded in setups where infant looking was measured by live button press), or evidence of parent interference or other types of interference (e.g., talking or pointing by parents, construction noise, sibling pounding on door), and infants being uncooperative or fussy (e.g., crying, not willing to do the experiment). Overall, for trial-level exclusions, XYZ trials (%XYZ of all trials) were excluded. For session-level exclusions, XYZ (XYZ% in the final sample) infants were dropped from analysis due to session-level errors.

Participants

Final sample. Our final sample included XYZ infants (XYZ% female; see Table XYZ for more specific sample demographic information) from XYZ laboratories (mean sample size per laboratory: XYZ, SD: XYZ, range: XYZ – XYZ). The mean age of infants included in the study was XYZ days (range: XYZ – XYZ). There were XYZ infants in the 3- to 9-month-old bin, XYZ infants in the 9- to 15-month-old bin. An additional XYZ infants were tested but excluded (see the full details on exclusions above).

As mentioned in the Introduction, multilingualism is common in Africa. Thus, many infants in the final sample are likely to have been exposed to more than one language. To assess infants' language backgrounds, each laboratory completed a family questionnaire with the participating parents (see materials in linked repository: https://osf.io/jgr79/?view_only=5ee43f58762742daaa2caa21b85e3780). Our family language background questionnaire was created based on the family language background questionnaire in the MB1 and MB1B studies, and included questions asking parents to estimate the number of hours that their infants heard different languages. We calculated the percentage of time that infants were exposed to a given language as the number of hours they hear that language (per day) divided by the total number of hours the infant hears any language each day. This method is simpler than the traditional interview method used in assessing bilingual infants' language exposure (Byers-Heinlein et al., 2019), but in order to minimize the burden on participating laboratories and families, we decided to use a short questionnaire method to assess infants' language backgrounds.

In this paper, we define bilingualism following the criteria established in MB1B (Byers-Heinlein et al., 2021). Monolingual infants are defined as those who have a minimum of 90% exposure to one language. Simultaneous bilingual infants are defined using the following criteria: (i) infants are regularly exposed to two or more languages beginning within the first month of life; (ii) they have a minimum of 25% exposure to each of their languages. In other words, bilingual infants are exposed to two languages between 25% to 75% of their time. Based on these criteria, it is possible that bilingual infants in our paper were exposed to multiple languages. For example, an infant with 45% English, 45% French, and 10% Spanish exposure would be regarded as a bilingual infant. Infants who did not meet the bilingual or monolingual criteria were designated as "other language background." All infants were included in the main, confirmatory analyses regardless of language background. Language background groupings were treated as a covariate in the analyses.

Based on the above-mentioned criteria, XYZ infants were classified as monolingual

infants, XYZ infants were classified as bilingual infants, and XYZ infants were classified as other.

Materials

Visual stimuli. All visual stimuli were the same as those used in the MB1 study. We used a brightly colored static checkerboard as the fixation stimulus, and an animation with shrinking concentric multi-colored circles to ensure infants were attending to the screen at the start of each trial. All of the stimuli can be found at https://osf.io/wh7md/.

Auditory stimuli. All auditory stimuli were identical to those used in the MB1 study. The stimuli were recordings of North-American English mothers either speaking with experimenters (ADS) or with their infants whose ages ranged from 122 to 250 days in a laboratory setting. Mothers were provided with a set of objects and were asked to talk about the objects with the experimenters and their infants in separate recording sessions. In total, two sets of auditory stimuli were created: one set consisted of 8 IDS stimuli and the other set consisted of 8 ADS stimuli. Each stimulus lasted for 18 seconds. The details of stimulus creation can be found in the report of MB1 (ManyBabies Consortium, 2020).

Volume. Each laboratory measured stimulus volume level using a smartphone app (e.g., the Android app "Sound Meter"). Labs kept the stimulus volume close to 63 – 65 dB SPL. According to the protocol, labs would measure and report the background noise level and the stimulus level. XYZ labs provided these data. The average background level was XYZ dB SPL (SD: XYZ, range: XYZ – XYZ) and the average stimulus level was XYZ dB SPL (SD: XYZ, range: XYZ – XYZ).

Procedure

Apparatus. Each laboratory used a laptop computer that had the experiment programmed in Habit 2.26 (Oakes, Sperka, DeBolt & Cantrell, 2019). Moreover, each

laboratory used a computer monitor to present the visual stimuli, a speaker for audio stimuli, a webcam for the experimenter to observe and record infants' performance, curtains/room dividers that separated the experimenter from the infant and parent during the experiment, and two sets of headphones: one for the experimenter and one for the parent.

Experimental procedure. The procedure was identical to the single-screen central visual fixation preference procedure reported in the MB1 study (ManyBabies Consortium, 2020). Using the single-screen central fixation method, researchers measured in real time the duration of infants' looking time to the computer monitor while they listened to the audio recordings. Infants' looking time to the computer monitor indicated their preference for the audio recordings (i.e., IDS/ADS). Each laboratory followed procedural instructions closely (based on pre-recorded videos illustrating the procedures, which were shared with all participating laboratories) to maintain the consistency of the experimental procedure across laboratories.

The experimenter explained the study to the parent and obtained consent from the parent before running the experiment. After completing the consent form, the experimenter led the participant to the testing room. To minimize distraction, the experimenter was separated from the infant and parent by curtains or a room divider. During the experiment, the infant sat on the parent's lap. To minimize any bias introduced by the experimenter or parent hearing the stimuli, each of them wore headphones and heard masking music during the experiment.

Parents were instructed not to speak to the infant during the experiment and not to point to the screen. Infants' performance was recorded by a webcam that was placed in front of and below the computer monitor. Infants' looking time to each trial was measured online by the experimenter, who observed the infant's behavior via the webcam. At the beginning of each trial, a short video of a colorful circle was presented to orient the infant's

attention to the screen. Once the infant fixated on the screen, the experimenter started the trial. The first two trials of the session were warm-up trials that accustomed infants with the procedure of the experiment, so the infant's looking time during warm-up trials was not analyzed. The auditory stimuli for the warm-up trials was piano music that lasted 18 seconds on each trial and the visual stimulus was the same as in the test trials (i.e., a colorful checkerboard). After the first two warm-up trials, the infant was tested with 16 trials presenting the IDS and ADS stimuli. Each infant was randomly assigned to one of four pseudo-random orders to counterbalance the order of presentation of IDS and ADS stimuli. Within each order, there were four blocks and each block presented 2 IDS and 2 ADS trials in alternating order. The presentation of the trials within each block were counterbalanced such that two blocks started with an IDS trial, and the other two blocks started with an ADS trial. On each trial, the auditory stimulus would continue to play until the infant looked away for 2 consecutive seconds or reached the maximum length of the auditory stimulus (18 seconds). Experimenters used the Habit program to record all looking time for every trial. There was no minimum looking time per trial that was required for continuation of the experiment. However, as in the MB1 study, any looking time that was less than 2 seconds was not analyzed. We excluded XYZ (XYZ%) trials that had less than 2 seconds looking time in total.

After the main looking-time task, the parents answered questions from the experimenter about participant and family demographic information, such as infant sex, date of birth, language exposure, and preterm/full term status. The questionnaire was translated into the appropriate language(s) for participants from each data collection site. See supplementary materials for the English template and adaptations.

General Lab Practices

Training of the experimenters. Three of the authors conducted a 2-day training workshop in Nairobi, Kenya on January 28 – 29, 2020, which was attended by lead

researchers from 8 of the participating laboratories. The training session provided an overview of the experimental procedure, advice on setting up the apparatus at the researcher's institution, and training, instructions and guidelines for running the experiment. Further, the first author sent instructions for experiment set-up and the workshop materials to all participating laboratories, and kept close contact with all lead researchers in the participating laboratories to provide technical support for the experiment.

Training of research assistants. Each laboratory was responsible for maintaining good experimenter training practices. We extended an invitation for the training workshop to one research assistant in each laboratory, so that the researcher primarily responsible for data collection could receive training directly as well. Following the MB1 study, each laboratory reported on which research assistant ran each infant using pseudonyms or numerical codes. After data collection, each laboratory completed a questionnaire regarding their training practices, the experience and academic status of each experimenter, and their basic participant greeting practices.

Results

Confirmatory Analyses

Data processing and analytic framework. Our primary dependent variable of interest was infants' looking time (LT). Infants' looking time was defined as time spent fixating on the computer screen during test trials. We did not count LT when infants looked away from the screen, though the trial was discontinued if an infant looked away and did not look back to the screen within 2 seconds. Following MB1 and MB1B, we log-transformed looking times prior to statistical analysis (Csibra, Hernik, Mascaro, Tatone, & Lengyel, 2016). We made this decision because we wanted to compare the data of the current study with those in MB1 and MB1B.

We tested our research questions via general linear mixed effects models. We fit all models using a maximal random effects structure (Barr, Levy, Scheepers, & Tily, 2013). Under this approach, we first specified all random effects that are appropriate for the experimental design (e.g., IDS/ADS trial type varied within subjects in our experimental design, thus it can be specified as a random effect by subject; see below for the full list of effects considered). If any of these mixed-effects models failed to converge, we used an iterative pruning strategy: first removing random slopes nested within subjects, next removing random slopes nested within labs, and finally removing random intercepts from groupings in the same order, retaining effects of trial type as these were of greatest theoretical interest. Following MB1 and MB1B, we fit all models using the lme4 package with the bobyqa optimizer, version XYZ (Bates, Maechler, Bolker, & Walker, 2015) and computed confidence intervals and p values using the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017).

In addition to the mixed-effect models, we assessed the reliability of measurement in our study by reporting the reliability of the infants' looking time difference to the IDS vs ADS stimuli across different trials. Following Byers-Heinelin et al. (under review), we reported the intraclass correlation coefficient (ICC) as our reliability measure. The ICC was computed using the psych package in R (Reville, 2018). We reported an ICC3k measure, on the basis of a two-way random effects model, a mean-rating of 8 (i.e., we had 8 pairs of IDS and ADS trials) and consistency agreement (Koo & Mae, 2016, Parson et al., 2019). The estimated ICC was _____, 95% CI [_____, _____]. We will compare the estimated ICC in our study with the ICC in MB1(B); we expect that the ICC between our study and that in MB1(B) to be similar in magnitude.

Below is a description of variables in our mixed-effect models:

- Log_lt: Dependent variable. Log-transformed looking time in seconds.
- Trial_type: a dummy coded variable with two levels: ADS (reference) and IDS. A

positive coefficient means that infants look longer to IDS trials compared to ADS trials.

- Age_months: a continuous variable measuring the infant's age in months (centered).
- Trial_num: An index for the current trial (1-16 for infants who completed the experiment). Excluded trials were reflected as missing trial numbers.
- Language_background: this consisted of two dummy coded variables that represented infants from three different language backgrounds: monolinguals (>= 90% exposure to one's native language); bilinguals (>=25 % to each of their languages); other (any infants who were not categorized as monolinguals or bilinguals). Using monolinguals as the reference level, the two dummy-coded variables are: (i) bilingual – infants who were categorized as bilinguals would be coded as 1 and all other infants would be coded as 0; (ii) Other (any infants who are not monolinguals or bilinguals) – infants who were categorized as other would be coded as 1 and all other infants would be coded as 0. In this case, monolingual infants would be coded as 0 in the above-mentioned dummy-coded variables.
- Infant_ID: a dummy coded variable with two levels, representing infants living in Africa in our current study (coded as 1) and infants living in Europe, Australia and Asia who were not hearing North American English, with data from MB1(B) (coded as 0).

As a reminder, we examined the following research questions in our paper: (1) IDS preference: whether infants in our multi-site African sample showed a preference for IDS and what is the corresponding effect size of this preference; (2) Age effect: whether there were changes in the infants' IDS preference across different ages; (3) Population comparison: examine whether the magnitude of infants' IDS preference in our study differed from infants in MB1 and in MB1B (comparing only infants in these three samples who were not exposed to North American English).

Research questions 1 and 2: Infants' IDS preference and age effect. We addressed our first two research questions using only the data collected in the current paper from laboratories in Africa. We specified the following model: log_lt ~ trial_type + trial_num + age_months + trial_type * trial_num + age_months * trial_num + age_months * trial_type + (trial_type * trial_num | subid) + (trial_type | lab)

The fixed-effects structure of this model included main effects of trial type (IDS vs ADS), age, and trial number. This structure controls for the effects of each independent variable on infants' looking time (e.g., longer looking times for IDS, shorter looking times on later trials). In addition, we included several two-way interaction terms: trial type interacting with trial number to model the possibility of infants' faster habituation to ADS, age interacting with the trial type to model the developmental trajectory of infants' IDS preference, and age interacting with trial number to model faster habituation for older children. The random effects structure of the model controlled for subject-level and lab-level grouping. For subject-level grouping, we added random intercepts and random effects of trial type, trial number, and their interaction to model the possibility that each infant may have different rates of habituation for IDS and ADS trials. For lab-level grouping, we added a random effect trial type to model differences in IDS preferences across labs.

After pruning for non-convergence, our final model specification was: To be filled in after data analysis.

As in MB1 and MB1B, the fixed effect estimate for trial type corresponds to the predicted infant-directed speech preference effect in units of log looking time (research question 1). The fixed effect estimate for the interaction of trial type and age indicates the estimated age-related change in infant-directed speech preference in log seconds per month (research question 2).

We will report metrics that directly describe the heterogeneous distribution of effects

across labs (Mathur & VanderWeele, in press; Mathur & VanderWeele, 2019). Specifically, we will use the R package MetaUtility to estimate (Mathur & VanderWeele, 2020): (1) the percentage of true population effects (i.e, infants' preference for IDS over ADS) across labs that are greater than 0; (2) the percentage that are greater than an effect size of by Cohen's d = 0.2; and (3) the percentage of effects in the unexpected direction that are less than d = -0.2 (i.e., representing infants preferring ADS over IDS). We will present these metrics only if the random slopes of trial type by lab are included in the final pruned mixed model, if these random slopes appear heterogeneous across labs (i.e., their estimated variance is greater than 0), and if, as anticipated, at least 10 labs contribute data (i.e., to achieve valid statistical inference).

Research question 3: Population comparison. In this analysis, we compare the data collected from the laboratories in Africa to data collected in MB1 and MB1B in COUNTRIES outside North America. We selected the subset of data from MB1 and MB1B that was collected using central fixation procedures (to match methods across studies) and from infants who were not exposed to North American English (non NAE) (to match stimulus un-familiarity due to language background). While we could have controlled the methodological and demographic variables statistically (and hence included all data from MB1 and MB1B in the full model), we believed that the increase in model complexity – and comparable decrease in interpretability – outweighed the benefits of this strategy.

We examine whether our sample of infants' IDS preference is different from those in MB1 and MB1B with the following model: log_lt ~ trial_type + trial_num + age_months + infant_ID + language_background + trial_type * trial_num + age_months * trial_num + age_months * trial_type + trial_type * infant_ID + trial_num * infant_ID + trial_type * language_background + (trial_type * trial_num | subid) + (trial_type | lab)

In this mixed-effects model, the fixed-effects included main effects of trial type,

language background, age, trial number, infants in our study/non NAE infants in MB1(B) and language background. In addition, we included several two-way interaction terms in the fixed effects structure: (i) trial type interacted with trial number, modeling the possibility of infants' faster habituation to ADS, (ii) age interacted with trial number, modeling faster habituation for older children, (iii) age interacted with trial type, modeling the developmental trajectory of infants' IDS preference, (iv) trial type interacted with infants in our sample, modeling the possible difference in IDS preference between infants in Africa and infants tested in MB1 and MB1B, (v) trial num interacted with infants in our sample, modeling the possible difference in habituation between our sample of infants and infants tested in MB1 and MB1B, and (vi) trial type interacted with language background, modeling the possible difference in IDS preference from infants with different language backgrounds. We adopted the same baseline random effects as in the previous model.

After pruning for non-convergence, our final model specification was: To be filled in after data analysis. The fixed effect estimate corresponding to our research question is the trial_type * infant_ID, which captures differences in measured IDS preference between the current data and data from MB1/MB1B in units of log seconds of looking time. ## Exploratory Analyses

TO BE WRITTEN AFTER THE STUDY IS COMPLETED

Urban vs rural areas. Prior studies (e.g., Keller, 2012; Vogh et al., 2015) have found that parents in non-WEIRD contexts sometimes speak to their infants differently across urban and rural areas. For example, parents in urban areas of Mozambican communities in Southeastern Africa tend to speak more to their children relative to parents in rural areas of Mozambican communities (Vogh et al., 2015). In turn, this could potentially lead to differences in IDS preference, with infants from urban areas, because of their higher input, showing a larger IDS preference, compared to infants from rural areas. We plan to explore this possibility by examining whether possible demographic differences in language input affect infants' preference for IDS in our sample.

Socio-economics status (SES). Previous research in North America (e.g., Hart & Risley, 1995; Hoff, 2006; Weisleder & Fernald, 2013) has shown that the quantity and quality of child-directed speech vary across families with different SES backgrounds. These differences in language input may drive differences in infants' preference for IDS. Thus, we plan to explore how SES affects infants' preference for IDS. SES will be measured by mothers' formal education (number of years), and the MacArthur Scale of Subjective Social Status (MacSSS). We will enter both mothers' formal education and the MacSSS as two separate SES variables in the regression model, after checking the assumption about the collinearity between variables. We note that SES is likely to be positively correlated with whether the family lives in an urban area. However, we propose that our measure of SES most likely can provide more information about the demographic backgrounds of the families in addition to the binary measure of whether the family lives in an urban or rural setting. Thus, this analysis may be more fine-grained, and could allow us to detect differences in infants' IDS preference across different demographic contexts.

General Discussion

TO BE WRITTEN AFTER THE STUDY IS COMPLETED

We summarize our findings with respect to three research questions in the paper. What is the magnitude of the IDS preference from our sample of infants? We found a magnitude of XYZ.

Does IDS preference vary across age in our sample? We found XYZ...

Is there any difference in IDS preference between infants in our sample and those in previous MB samples? We found XYZ...

`The general discussion will include caveats around over-interpretation of any demographi`

References