Quantifying Sources of Variability in Infancy Research Using the Infant-Directed-Speech Preference

The ManyBabies Consortium*

*All consortium members are listed in the Transparency section at the end of the article.

Abstract
Psychological scientists have become increasingly concerned with issues related to methodology and replicability, and infancy researchers in particular face specific challenges related to replicability: For example, high-powered studies are difficult to conduct, testing conditions vary across labs, and different labs have access to different infant populations. Addressing these concerns, we report on a large-scale, multisite study aimed at (a) assessing the overall replicability of a single theoretically important phenomenon and (b) examining methodological, cultural, and developmental moderators. We focus on infants’ preference for infant-directed speech (IDS) over adult-directed speech (ADS). Stimuli of mothers speaking to their infants and to an adult in North American English were created using seminaturalistic laboratory-based audio recordings. Infants’ relative preference for IDS and ADS was assessed across 67 laboratories in North America, Europe, Australia, and Asia using the three common methods for measuring infants’ discrimination (head-turn preference, central fixation, and eye tracking). The overall meta-analytic effect size (Cohen’s $d$) was 0.35, 95% confidence interval = [0.29, 0.42], which was reliably above zero but smaller than the meta-analytic mean computed from previous literature (0.67). The IDS preference was significantly stronger in older children, in those children for whom the stimuli matched their native language and dialect, and in data from labs using the head-turn preference procedure. Together, these findings replicate the IDS preference but suggest that its magnitude is modulated by development, native-language experience, and testing procedure.

Keywords
language acquisition, speech perception, infant-directed speech, reproducibility, experimental methods, open data, open materials, preregistered

To date, however, infancy researchers have remained relatively silent on issues of replicability. This silence is not because infancy research is immune from the issues raised. Indeed, the statistical power associated with experiments on infant psychology is often unknown (and presumably too low; Oakes, 2017), and the replicability of many classic findings is uncertain. Instead, one reason for the infancy field’s silence is likely related to the set of challenges that come with collecting and interpreting data on infants—and development more generally. For example, it can be quite costly to test
large samples of infants or to replicate past experiments. Another challenge for infancy researchers is that it is often difficult to interpret contradictory findings in developmental populations, given how children's behavior and developmental timing vary across individuals, ages, contexts, cultures, languages, and socioeconomic groups. Although these challenges may make replicability in infancy research particularly difficult, they do not make it any less important.

Indeed, it is of primary importance to evaluate replicability in infancy research (see Frank et al., 2017). But how can this evaluation be done? Here we report the results of a large-scale, multilab, preregistered infant study. This study was inspired by the ManyLabs studies (e.g., Klein et al., 2014), in which multiple laboratories have attempted to replicate various social- and cognitive-psychology studies and have assessed moderators of replicability systematically across labs. For the reasons just discussed, it would be prohibitively difficult to examine the replicability of a large number of infant studies simultaneously. Instead, we chose to focus on what developmental psychology can learn from testing a single phenomenon, assessing its overall replicability, and investigating the factors moderating it. As a positive side effect, this approach leads to the standardization of decisions concerning data collection and analysis across a large number of labs studying similar phenomena or using similar methods. For this first “Many Babies” project, we selected a finding that the field has good reason to believe is robust—namely, infants’ preference for infant-directed speech (IDS) over adult-directed speech (ADS)—and tested it in 67 labs around the world. This phenomenon has the further advantage that it uses a dependent measure, looking time, that is ubiquitous in infancy research. In the remainder of this introduction, we briefly review the literature on the relevance of IDS in development and then discuss our motivations and goals in studying a single developmental phenomenon at a larger scale than is typical in developmental research.

Infant-Directed-Speech Preference

IDS is a descriptive term for the characteristic speech that caregivers in many cultures direct toward infants. Compared with ADS, IDS is typically higher pitched, has greater pitch excursions, and is characterized by shorter utterances, among other differences (Fernald et al., 1989). Although caregivers across many different cultures and communities use IDS, the magnitude of the difference between IDS and ADS varies (Englund & Behne, 2006; Farran, Lee, Yoo, & Oller, 2016; Fernald et al., 1989; Newman, 2003). Nevertheless, the general acoustic pattern of IDS is readily identifiable to adult listeners (Fernald, 1989; Grieser & Kuhl, 1988; Katz, Cohn, & Moore, 1996; Kitamura & Burnham, 2003).

A substantial literature reporting studies using a range of stimuli and procedures has demonstrated that infants prefer IDS over ADS. For example, Cooper and Aslin (1990), using a contingent visual-fixation auditory preference paradigm, showed that infants fixate on an unrelated visual stimulus longer when hearing IDS than when hearing ADS, even as newborns. Across a variety of ages and methods, other studies have also found increased attention to IDS compared with ADS (Cooper, Abraham, Berman, & Staska, 1997; Cooper & Aslin, 1994; 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). In a meta-analysis by Dunst, Gorman, and Hamby (2012), which included 34 experiments, the IDS preference had an effect size (Cohen’s $d$) of 0.67, 95% confidence interval (CI) = [0.57, 0.76]—quite a large effect size for an experiment with infants (Bergmann et al., 2018).

The evidence suggests that IDS augments infants’ attention to speakers (and presumably what speakers are saying) because of its highly salient acoustic qualities, such as frequency modulation (Cusack & Carlyon, 2003). In addition, it is hypothesized that the IDS preference plays a pervasive supporting role in early language learning. For example, young infants are more likely to discriminate speech sounds when they are pronounced with typical IDS prosody rather than ADS prosody (Karzon, 1985; Trainor & Desjardins, 2002). There are also reports that infants show preferences for natural phrase structure in narratives spoken in IDS but not in ADS (cf. Fernald & McRoberts, 1996; Hirsh-Pasek et al., 1987). In addition, word segmentation (Thiessen, Hill, & Saffran, 2005) and word learning (Graf Estes & Hurley, 2013; Ma, Golinkoff, Houston, & Hirsh-Pasek, 2011) are reported to be facilitated in IDS compared with ADS. Naturalistic observations confirm that the amount of speech directed to U.S. 18-month-olds (which likely bears IDS features), rather than the amount of speech they overhear (which is likely predominantly ADS), relates to the efficiency of their word processing and to their expressive vocabulary at age 24 months (Weisleder & Fernald, 2013). Finally, infants show increased neural activity in response to familiar words in IDS compared with the same words in ADS, and also compared with unfamiliar words in either register (Zangl & Mills, 2007). From a theoretical perspective, the IDS register has been claimed to trigger specialized learning mechanisms (Csibra & Gergely, 2009), as well as to boost social preferences and perhaps attention in general (Schachner & Hannon, 2011), as it
has been reported to improve even nonlinguistic associative learning (e.g., Kaplan, Jung, Ryther, & Zarleno-Strouse, 1996).

The Current Study: Motivations and Goals

Despite the large body of research on infants’ preference for IDS and the positive effects of IDS on the processing of linguistic and nonlinguistic stimuli, a number of open questions remain. This study was designed to answer some of these IDS-specific questions, as well as questions about methods for assessing infants’ cognition, including questions about the interaction between statistical power and developmental methodologies. In this section, we describe the key questions for our study (as well as our predictions, where applicable), in rough order of decreasing specificity, highlighting methodological decisions that followed from particular goals.

What is the magnitude of the IDS preference?

First and foremost, our study was intended to provide a large-scale, precise measurement of IDS preference across a large number of labs. Given the evidence summarized in a previous meta-analysis (Dunst et al., 2012), we expected that the preference would be nonzero and positive. We suspected, however, that this phenomenon, like many others, suffers from a file-drawer effect, in which studies with low effect sizes (or large p values) often do not get published. Also, there was reason to believe that effect sizes in infancy research are often incorrectly reported; for example, partial eta-squared is often misreported as eta-squared. This confusion is likely to inflate the practical significance of findings, leading to an overestimation of the statistical magnitude and importance of effects (Mills-Smith, Spangler, Panneton, & Fritz, 2015). Therefore, we believed that the mean effect size of 0.67 reported by Dunst et al. (2012) was likely an overestimate of the real effect size.

How does IDS preference vary across ages?

We could plausibly predict that, all else being equal, older infants can more effectively process ADS than younger infants, and so the attraction of IDS over ADS might attenuate with age (Newman & Hussain, 2006). On the other hand, older infants might show a stronger preference for IDS over ADS, given that older infants have had more opportunity to experience the positive social interactions that likely co-occur with IDS, including, but not limited to, eye contact, positive facial expressions, and interactive play.

How does IDS preference vary with linguistic experience and language community?

Preference for IDS might be affected by infants’ language experience. Across many areas of language perception, infants show a pattern of perceptual narrowing. They begin life as “universal listeners” ready to acquire any language, but with experience gain sensitivity to distinctions in their native language and lose sensitivity to nonnative distinctions (Maurer & Werker, 2014). If preference for IDS follows a similar pattern, then older infants would be expected to show a stronger preference for IDS over ADS in their native language than in a nonnative language.

Faced with several competing concerns, we made the decision that all infants in our study, regardless of their native language, would be exposed to ADS and IDS stimuli in North American English (NAE). This design choice had several practical advantages. Most important, it allowed every infant to be tested with the same stimulus set. Creating different stimulus sets in different languages would have added methodological variability across labs that would be statistically indistinguishable from lab identity and language environment. Further, creating a single high-quality stimulus set shared across labs reduced the time and cost of conducting the study.

This decision had both advantages and drawbacks. A limitation of our design is that NAE stimuli are unfamiliar to infants from other language or dialect communities; thus, these infants might show less interest for NAE speech overall or might have a harder time recognizing IDS features as such when those features differ from those used in their native language or dialect. In fact, previous work even suggests that infants’ IDS preference depends on the characteristics of the type of IDS addressed to children their age (McRoberts, McDonough, & Lakusta, 2009). Although this was a relevant concern, previous research had documented some IDS preference in the face of language and age mismatches (McRoberts et al., 2009; Werker, Pegg, & McLeod, 1994), and corpus studies suggested that, if anything, the distinction between IDS and ADS is more salient in NAE than in other linguistic variants (e.g., Fernald et al., 1989; Shute, 1987). Further, although this design did not allow us to disentangle the effects of stimulus language (native vs. nonnative) from the effects of infants’ cultural background, we were able to explore how aspects of these factors influence infants’ preference for IDS.
After weighing these considerations, we adopted NAE stimuli in order to have the maximal chance of recovering a positive effect, ensure that stimuli were not a source of variance across labs, allow comparability with previous work, and also minimize participating labs’ barriers to entry (i.e., the need to create labspecific stimuli). So as to be able to assess children’s language background at the group level, we also chose to focus our primary analyses on monolingual infants (a separate effort focused on IDS preferences in bilingual children; Byers-Heinlein et al., 2020).

**What is the impact of methodology on measurement of infants' preference?**

We focused on three primary methods for assessing infants’ interest: single-screen central fixation, eye tracking, and the head-turn preference procedure (HPP). All three methods are widely used in the field of infant language acquisition and yield measurements of preference for a given type of auditory stimulus, indexed by looking to an unrelated visual stimulus. In labs using the single-screen central-fixation method, infants were shown an uninformative image (a checkerboard) on a single, centrally located monitor, while they listened to either IDS or ADS, and the time they spent looking toward the monitor was manually coded via a closed-circuit video camera. In labs using the eye-tracking method, infants saw a similar display, but looking times were measured automatically via a remote corneal-reflection eye tracker. In labs using the HPP method, infants saw an attractor visual stimulus (often a flashing light bulb) appear to either their left or their right, and the duration of their head turn toward the attractor while IDS or ADS played was manually coded via a closed-circuit video camera (Nelson et al., 1995). Labs using the eye-tracking and central-fixation methods (and some of those using the HPP) employed an attention getter between trials to orient the infant’s gaze toward the display.

Each lab tested the same phenomenon, using the same stimuli and the same general experimental parameters (including, e.g., trial order, maximum trial length); only the method of measuring preference varied. We therefore could analyze whether this theoretically irrelevant methodological choice influenced the observed effect size, which could help guide future decision making.

**What are the effects of testing infants in multiple experiments during a single lab visit?**

Labs vary in whether each infant visiting the lab completes a single experiment only, or whether some infants participate in a second study as well. These “second session” experiments are thought by some researchers to yield greater dropout rates and less reliable measurements, but the existence and magnitude of a second-session effect has not been tested, to our knowledge. In our study, a number of participating labs ran the IDS-preference study with some infants who had already participated in an additional study; measurements from these infants could inform future lab administration practices.

**What should expectations be regarding replicability and statistical power in studies of infancy?**

Although we replicated only a single phenomenon, the importance and assumed robustness of the IDS preference meant that our study would provide data relevant to developing a more nuanced understanding of replicability and power in infancy research. Because of the large number of participating labs, it was expected that data from some labs would not support an IDS preference (i.e., the data would yield a small, or even negative, effect size when analyzed individually). Some variability was to be expected given the mathematics of estimating an effect at so many independent sites. Nonetheless, we inspected whether there was systematic variability explained by lab effects.

In addition, by providing an unbiased estimate of effect size for an important developmental phenomenon (plus estimates of how that effect varies across ages, language backgrounds, and methods), this project would set a rough baseline for other scientists to use when planning studies. Existing attempts to estimate the statistical power of experiments with infants have been contaminated by publication bias, which has led to an overestimation of typical effect sizes in infancy research. Such overestimates can lead subsequent studies to be underpowered (researchers expect to see larger effects than are truly present). Though we planned to estimate the effect size for a particular developmental preference, we also planned to compare our unbiased estimate, calculated both across all three methods and for each method, with the meta-analytic effects extracted from previously published studies.

**How should researchers think about the relationships among experimental design, statistical significance, and developmental change?**

Previous work has often employed a contrast between two ages to suggest a developmental change, for example, by showing that 7-month-old infants exhibit a statistically reliable preference in a task, but 5-month-old
infants do not. Such a finding (the pairing of a significant difference and a nonsignificant difference) is not sufficient to show a difference between two time points (Nieuwenhuis, Forstmann, & Wagenmakers, 2011). Moreover, even when a significant difference between two age groups is found, such a result is not sufficient to elucidate the developmental pattern underlying this discrete test. By measuring how effect sizes changed with age using a much denser sampling approach, we aimed to illustrate what stands to be gained with a more gradient approach to testing behavior over development.

Summary

This broad replication of the IDS preference was aimed at helping to answer basic questions about the replicability of developmental-psychology findings and also at providing useful benchmarks for how to design infant cognition studies going forward. Projects such as ManyLabs have led to important improvements in research practices in cognitive and social psychology, and we hoped that ManyBabies would play a similar role for developmental cognitive science.

Disclosures

Preregistration

Prior to data collection, our manuscript was reviewed, and we registered our instructions and materials on the Open Science Framework (OSF; see https://osf.io/gf7vh/).

Data, materials, and online resources

All materials, data, and analytic code are available on OSF (https://osf.io/re95x/). The specific code and data required to render our submitted manuscript are also available there (https://osf.io/zaewn/).

Reporting

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Ethical approval

All the labs collected data under their own independent ethical approval via the appropriate governing body for their institution. Central data analyses used exclusively de-identified data. Identifiable video recordings of individual infant participants were coded and archived locally at each lab. Where institutional review boards' protocols permitted, video recordings were also uploaded to Databrary, a central controlled-access database accessible to other researchers (http://databrary.org).

Method

Participation details

Time frame. We issued an open call for labs to participate in our study, which we designated “ManyBabies 1,” on February 2, 2017. Data collection began on May 1, 2017. Data collection was scheduled to end on April 30, 2018 (1 year later). In order to allow labs to complete their sample, however, a 45-day extension was granted, and data collection officially ended on June 15, 2018. Data collection from one laboratory extended beyond this time frame (see Deviations From the Registered Protocol, later in the Method section).

Age distribution. Each participating lab was asked to recruit participants in one or more of four age bins: 3 months 0 days through 6 months 0 days, 6 months 1 day through 9 months 0 days, 9 months 1 day through 12 months 0 days, and 12 months 1 day through 15 months 0 days. Each lab was tasked with ensuring that, for each age bin contributed, the mean age fell close to the middle of the range and the sample was distributed across the bin. We selected 3-month bins as a compromise, on the assumption that tighter bins would make recruitment more difficult whereas broader bins would lead to more variability and would blur developmental trends (i.e., by introducing possible interactions between age and lab-specific effects, for instance, if a particular method turned out to be most appropriate for a subset of the ages tested). The ability to contribute to one or more age bins at the laboratory’s discretion was necessary because labs differ in their ability to recruit infants of different ages.

Criterion for lab participation. During study planning, we used data from MetaLab (Bergmann et al., 2018) to compute the meta-analytic mean effect size for IDS preference; the resulting Cohen’s $d$ was 0.72. A power analysis indicated that 95% power to detect this effect in a paired-samples $t$ test would require 27 participants, and 80% power would require 17. On the basis of these calculations, we asked participating labs to commit to a minimum sample of 32 in each age group they targeted. However, given that for many of our analyses, power across labs was more critical than power within a lab (Judd, Westfall, & Kenny, 2017), we allowed labs to contribute a “half sample” of 16, with the assumption that this would increase the number of laboratories capable of participating and allow more laboratories to contribute
samples from multiple age bins. We specified that in their recruitment efforts, labs should specifically target infants with the desired demographic characteristics outlined in the study’s protocol (e.g., full-term infants; the full list of inclusion criteria is discussed later in the Method section). Given this recruitment strategy, however, we asked that sample Ns be calculated on the basis of the total number of infants tested, not the number of infants retained after exclusions (which were performed centrally as part of the broader data analysis, not at the lab level).

We included data from a lab in our analysis if the lab achieved the minimum N (16) required for a half-sample in an age bin by the end date of testing and if, after exclusions, the lab contributed data for 10 or more infants. If a lab collected more than their required sample, we included the extra data as well. Laboratories were cautioned not to consider the data (e.g., whether a statistically significant effect was evident) in their internal decision making regarding how many infants to recruit and when to stop recruitment.

Participants

Our final sample comprised 2,329 monolingual infants from 67 labs (mean sample size per lab = 34.76, SD = 20.33, range: 10–93). Demographic exclusions were implemented primarily during recruitment; despite this, additional infants were tested and excluded on the basis of preset criteria (see Inclusion Criteria for percent-ages). In addition, 2 labs registered to participate but failed to collect data from at least 10 included infants, and so their data were not included in analyses. Information about all the included labs is given in Table 1.

The mean age of infants included in the study was 291.99 days (range: 92–456). There were 310 infants in the 3- to 6-month-old bin (23 labs), 772 infants in the 6- to 9-month-old bin (49 labs), 554 infants in the 9- to 12-month-old bin (35 labs), and 693 infants in the 12- to 15-month-old bin (42 labs). Forty-five labs collected data in more than one bin. Of the total sample, 1,066 infants (from 30 labs) were acquiring NAE, and 1,263 infants (from 37 labs) were acquiring a language other than NAE. As noted earlier, a separate sample of bilingual children was tested in a parallel investigation, but these data are not reported here.

Materials

Visual stimuli. In labs using the central-fixation or eye-tracking method, a brightly colored static checkerboard was the fixation stimulus, and a small engaging video (an animation of colorful rings decreasing in size) was the attention getter. Labs using the HPP were asked to use their typical visual stimulus, which varied consider-ably across labs. Some used flashing lights as the visual fixation stimulus (as in the original protocol developed in the 1980s), whereas others used a variety of other visual displays on video screens (e.g., a looming circle).

Speech stimuli. The goal of our stimulus-creation effort was to construct a set of recordings of naturalistic IDS and ADS gathered from a variety of mothers. To do so, we made recordings of mothers speaking to their infants and to experimenters, selected a subset of individual utterances from these recordings, and then constructed stimulus items from this subset. All characteristics of the recordings other than register (IDS vs. ADS) were as bal-anced as possible across clips. On the basis of our intuitions and the data from the norming ratings described later, we consider these stimuli to be representative of naturally produced IDS and ADS across middle- and high-socioeconomic-status mothers in North America. Although future studies could vary particular aspects of the IDS systematically (e.g., age of the mother, age of the infant being spoken to, dialect), we did not do so in the current study. Our stimulus-elicitation method was designed to meet the competing considerations of laboratory control and naturalism.

Source recordings were collected in two laboratories, one in central Canada and one in the northeastern United States. The recorded mothers had infants whose ages ranged from 122 to 250 days. The same recording procedures were followed in the two laboratories. In an infant-friendly greeting area or testing room, recordings were collected using a simple lapel clip-on microphone connected to a smartphone (iPhone 5s or 6s), with the Voice Record or Voice Record Pro app (Dayana Networks Ltd.) in the Canadian lab and the Voice Memos app (Apple Inc.) in the U.S. lab. The targets for conversation were objects in an opaque bag: five familiar objects (a ball, a shoe, a cup, a block, a train) and five unfamiliar objects (a sieve, a globe, a whisk, a flag, and a bag of yeast). To ensure that the mothers used consistent labels, we affixed to each object a small sticker showing its name. Each object was taken out of the bag separately, and the mother was asked to talk about the object, either to her baby (for the IDS sam-ples) or to an experimenter (for the ADS samples), until she ran out of things to say; at this point, the next object was taken out of the bag. Recording stopped when all the objects had been removed from the bag and had been talked about. The order of IDS and ADS recording was counterbalanced across mothers. A total of 11 mothers were recorded in Canada and 4 in the United States.

There were a total of 179 unedited minutes of recording from Canada and 44 from the United States. A
Table 1. Summary of the Included Labs and Their Samples

<table>
<thead>
<tr>
<th>Lab</th>
<th>Mean age (days)</th>
<th>N</th>
<th>Language</th>
<th>Method</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>babylabbrookes</td>
<td>255</td>
<td>53</td>
<td>English</td>
<td>Central fixation</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>babylabvuw</td>
<td>224</td>
<td>15</td>
<td>English</td>
<td>Central fixation</td>
<td>New Zealand</td>
</tr>
<tr>
<td>babylabyork</td>
<td>268</td>
<td>32</td>
<td>English</td>
<td>Central fixation</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>baldwinlabuoregon</td>
<td>320</td>
<td>16</td>
<td>English</td>
<td>Central fixation</td>
<td>United States</td>
</tr>
<tr>
<td>bchdosa</td>
<td>269</td>
<td>67</td>
<td>English</td>
<td>Central fixation</td>
<td>United States</td>
</tr>
<tr>
<td>bcrluny</td>
<td>411</td>
<td>29</td>
<td>English</td>
<td>Central fixation</td>
<td>United States</td>
</tr>
<tr>
<td>bounbc</td>
<td>411</td>
<td>31</td>
<td>Turkish</td>
<td>Central fixation</td>
<td>Turkey</td>
</tr>
<tr>
<td>icclbc</td>
<td>222</td>
<td>15</td>
<td>English</td>
<td>Central fixation</td>
<td>United States</td>
</tr>
<tr>
<td>infantcogablouisville</td>
<td>325</td>
<td>35</td>
<td>English</td>
<td>Central fixation</td>
<td>United States</td>
</tr>
<tr>
<td>ldlottawa</td>
<td>276</td>
<td>59</td>
<td>English</td>
<td>Central fixation</td>
<td>Canada</td>
</tr>
<tr>
<td>madlabucd</td>
<td>234</td>
<td>10</td>
<td>English</td>
<td>Central fixation</td>
<td>United States</td>
</tr>
<tr>
<td>minddevlabbicocca</td>
<td>158</td>
<td>15</td>
<td>Italian</td>
<td>Central fixation</td>
<td>Italy</td>
</tr>
<tr>
<td>udsssaarland</td>
<td>332</td>
<td>43</td>
<td>German</td>
<td>Central fixation</td>
<td>Germany</td>
</tr>
<tr>
<td>unlvmusicalab</td>
<td>138</td>
<td>20</td>
<td>English</td>
<td>Central fixation</td>
<td>United States</td>
</tr>
<tr>
<td>weescienceedinburgh</td>
<td>213</td>
<td>32</td>
<td>English</td>
<td>Central fixation</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>wsiogettingen</td>
<td>274</td>
<td>88</td>
<td>German</td>
<td>Central fixation</td>
<td>Germany</td>
</tr>
<tr>
<td>infantcogubc</td>
<td>165</td>
<td>39</td>
<td>English</td>
<td>Central fixation, eye tracking</td>
<td>United States</td>
</tr>
<tr>
<td>lancaster</td>
<td>326</td>
<td>42</td>
<td>English</td>
<td>Central fixation, eye tracking</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>babylabllangssex</td>
<td>289</td>
<td>27</td>
<td>English</td>
<td>Eye tracking</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>babylablmu</td>
<td>308</td>
<td>62</td>
<td>German</td>
<td>Eye tracking</td>
<td>Germany</td>
</tr>
<tr>
<td>babylabalshimane</td>
<td>195</td>
<td>28</td>
<td>Japanese</td>
<td>Eye tracking</td>
<td>Japan</td>
</tr>
<tr>
<td>babylabucljohnson</td>
<td>408</td>
<td>22</td>
<td>English</td>
<td>Eye tracking</td>
<td>United States</td>
</tr>
<tr>
<td>babylabumassb</td>
<td>308</td>
<td>30</td>
<td>English</td>
<td>Eye tracking</td>
<td>United States</td>
</tr>
<tr>
<td>babylalingoslo</td>
<td>227</td>
<td>31</td>
<td>Norwegian</td>
<td>Eye tracking</td>
<td>Norway</td>
</tr>
<tr>
<td>callab</td>
<td>369</td>
<td>30</td>
<td>English</td>
<td>Eye tracking</td>
<td>United States</td>
</tr>
<tr>
<td>cdccceu</td>
<td>272</td>
<td>27</td>
<td>Hungarian</td>
<td>Eye tracking</td>
<td>Hungary</td>
</tr>
<tr>
<td>cfnoufn</td>
<td>298</td>
<td>15</td>
<td>English</td>
<td>Eye tracking</td>
<td>Australia</td>
</tr>
<tr>
<td>childlabmanchester</td>
<td>269</td>
<td>26</td>
<td>English</td>
<td>Eye tracking</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>cogledevlabbyu</td>
<td>161</td>
<td>29</td>
<td>English</td>
<td>Eye tracking</td>
<td>United States</td>
</tr>
<tr>
<td>dcnlabtennessee</td>
<td>345</td>
<td>19</td>
<td>English</td>
<td>Eye tracking</td>
<td>United States</td>
</tr>
<tr>
<td>earlyzosocgfm</td>
<td>310</td>
<td>35</td>
<td>English</td>
<td>Eye tracking</td>
<td>United States</td>
</tr>
<tr>
<td>escompichsleipzig</td>
<td>159</td>
<td>14</td>
<td>German</td>
<td>Eye tracking</td>
<td>Germany</td>
</tr>
<tr>
<td>ethosrennes</td>
<td>187</td>
<td>90</td>
<td>French</td>
<td>Eye tracking</td>
<td>France</td>
</tr>
<tr>
<td>irlconcordia</td>
<td>310</td>
<td>37</td>
<td>English</td>
<td>Eye tracking</td>
<td>Canada</td>
</tr>
<tr>
<td>jmucll</td>
<td>340</td>
<td>17</td>
<td>English</td>
<td>Eye tracking</td>
<td>United States</td>
</tr>
<tr>
<td>kokuhamburg</td>
<td>305</td>
<td>25</td>
<td>German</td>
<td>Eye tracking</td>
<td>Germany</td>
</tr>
<tr>
<td>kyotobabylab</td>
<td>281</td>
<td>30</td>
<td>Japanese</td>
<td>Eye tracking</td>
<td>Japan</td>
</tr>
<tr>
<td>labunam</td>
<td>302</td>
<td>36</td>
<td>Spanish</td>
<td>Eye tracking</td>
<td>Mexico</td>
</tr>
<tr>
<td>lcdlsu</td>
<td>354</td>
<td>23</td>
<td>English</td>
<td>Eye tracking</td>
<td>United States</td>
</tr>
<tr>
<td>lcduleeds</td>
<td>413</td>
<td>14</td>
<td>English</td>
<td>Eye tracking</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>llliv</td>
<td>302</td>
<td>36</td>
<td>English</td>
<td>Eye tracking</td>
<td>United Kingdom</td>
</tr>
<tr>
<td>lscppl</td>
<td>404</td>
<td>14</td>
<td>French</td>
<td>Eye tracking</td>
<td>France</td>
</tr>
<tr>
<td>pocdnorthwestern</td>
<td>409</td>
<td>30</td>
<td>English</td>
<td>Eye tracking</td>
<td>United States</td>
</tr>
<tr>
<td>socialcogumiami</td>
<td>131</td>
<td>19</td>
<td>English</td>
<td>Eye tracking</td>
<td>United States</td>
</tr>
<tr>
<td>wellentdeckerzurich</td>
<td>414</td>
<td>30</td>
<td>German</td>
<td>Eye tracking</td>
<td>Switzerland</td>
</tr>
<tr>
<td>nusinfantlanguagecentre</td>
<td>337</td>
<td>21</td>
<td>Mandarin</td>
<td>Eye tracking, central fixation</td>
<td>Singapore</td>
</tr>
<tr>
<td>babylabkingswood</td>
<td>312</td>
<td>32</td>
<td>English</td>
<td>HPP</td>
<td>Australia</td>
</tr>
<tr>
<td>babylabkonstanz</td>
<td>235</td>
<td>15</td>
<td>German</td>
<td>HPP</td>
<td>Germany</td>
</tr>
</tbody>
</table>

(continued)
first-pass selection of low-noise IDS and ADS samples yielded 1,281 utterances, with a total duration of 4,479 s. From this first pass, we selected 238 utterances that were considered to be the best examples of IDS and ADS and met other basic stimulus selection criteria (e.g., they did not contain laughter or the baby’s name). This library of 238 utterances was then normed on five variables: accent, affect, naturalness, noisiness, and IDS-ness. The goal of this norming was to gather intuitive judgments so that we would have a principled basis for identifying utterances that were clearly anomalous in some respect (e.g., odd affect or background noise) and excluding them. Naive, NAE-speaking adults recruited from Amazon Mechanical Turk listened to all 238 utterances and rated them on 7-point Likert scales. Raters were assigned randomly to one of the five variables; the number of raters assigned to a particular task ranged from 8 to 18 because of variability in random assignment. Affect and IDS ratings were made using low-pass-filtered recordings (a 120-Hz filter with standard roll-off was applied twice using the sox software package, available at http://sox.sourceforge.net). In general, with the exception of IDS-ness, ratings were not highly variable across clips (the largest SD was 0.85, for noise ratings).

These ratings were used to produce a set of utterances in which accent was rated similar to “standard English” (ratings < 3; 1 = completely standard), naturalness was rated high (ratings > 4; 7 = completely natural), noisiness was rated low (ratings < 4; 1 = noiseless), and IDS and ADS clips were consistently distinguished (ratings > 4 for IDS clips and ratings < 4 for ADS clips; 7 = clearly directed at a baby or child). This procedure resulted in a total of 163 utterances that met our inclusion criteria.

Our next goal was to create eight IDS and eight ADS stimuli that were exactly 18 s in length, each containing utterances from the set we created. First, the amplitudes of all the clips were root mean square normalized to 70 dB sound-pressure level (SPL). Stimuli were then assembled from the normalized clips, and finally, the amplitudes of the stimuli were renormalized to 70 dB SPL. We assembled the final stimuli considering the following issues:

- **Identity**: Each audio stimulus was constructed using clips from more than one mother. The number of different mothers included in a given stimulus was matched across the IDS and ADS stimuli. In addition, multiple clips from the same mother...
were grouped together within a given stimulus in order to match the number of “mother transitions” across registers.

- **Lexical items**: We matched the frequency of object labels in the clips across the IDS and ADS contexts. We also ensured an even distribution of the order in which each particular word was presented across stimuli and registers (ADS vs. IDS).

- **Questions**: IDS tends to include a much higher proportion of questions compared with ADS (Snow, 1977; Soderstrom, Blossom, Foygel, & Morgan, 2008). However, because the nature of the recording task may have served to inflate this difference, we preferentially selected declarative sentences over questions in the IDS sample. In the final stimulus set, 47% of the utterances in the IDS samples and 3% of the utterances in the ADS samples were questions. We felt that retaining this naturally occurring difference between IDS and ADS within our stimuli was more appropriate than precisely and artificially controlling for utterance type across registers.

- **Duration of individual clips**: As expected, the utterances in IDS were much shorter than those in ADS, so it was not possible to match the IDS and ADS stimuli on duration or number of clips. Because there were more clips per stimulus in the IDS samples, there were also more utterance boundaries. This property is consistent with what has been reported in the literature on the natural characteristics of IDS (Martin, Igarashi, Jincho, & Mazuka, 2016).

- **Total duration**: We fixed all stimuli to have a total duration of 18 s by concatenating individual utterance files into single audio files that were more than 18 s in length, trimming these down to 18 s, and fading the audio in and out with 0.5-s half-cosine windows.

Table 2 and Figure 1 provide additional details regarding the final stimulus set. Measurements were made using STRAIGHT (Kawahara & Morise, 2011), using default values for extraction of the fundamental frequency (F0). Table 3 provides a comparison of our stimuli with a sample of others that have been used in previous IDS-preference studies. Across studies in the broader literature, the only measure that we found to be reported consistently was F0 for IDS and ADS, and even this statistic was reported for only about half the studies we examined (those studies listed in Table 3). Various measures of variability in F0 were reported in some cases (e.g., range within each sample, range across samples, standard deviation), but because of variation in the length and number of samples used, and a lack of systematicity in reporting, it was difficult to compare studies directly. Numerically, the average pitch difference between IDS and ADS in our materials was less extreme than that in previous studies.

To confirm that our composite IDS and ADS stimuli were perceived to be natural and that the pitch difference between registers was sufficient to lead to the two sets of stimuli being categorized differently, we conducted another norming study. Using the same basic paradigm as in the first norming study, we collected a new sample of judgments from Mechanical Turk participants. They were randomly assigned to listen to all 16 stimuli and judge either whether they were directed at infants or children or at adults (N = 22) or else whether they sounded natural (N = 27). All IDS clips were judged extremely likely to be directed at infants or children (M = 6.74, SD = 0.09), on a rating scale from

**Table 2.** Characteristics of the Infant-Directed Speech (IDS) and Adult-Directed Speech (ADS) Stimuli

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>IDS</th>
<th>ADS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of mothers speaking per stimulus</td>
<td>4.00 0.00</td>
<td>3.75 0.46</td>
</tr>
<tr>
<td>Number of clips per stimulus</td>
<td>6.88 1.13</td>
<td>4.50 0.76</td>
</tr>
<tr>
<td>Number of objects mentioned per stimulus</td>
<td>2.75 0.71</td>
<td>2.75 0.71</td>
</tr>
<tr>
<td>F0 per stimulus (Hz)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>206.90 19.50</td>
<td>174.90 13.20</td>
</tr>
<tr>
<td>10th percentile</td>
<td>131.40 26.10</td>
<td>139.00 17.70</td>
</tr>
<tr>
<td>90th percentile</td>
<td>340.00 21.50</td>
<td>232.00 13.80</td>
</tr>
<tr>
<td>Mean number of utterances per stimulus</td>
<td>7.75 1.04</td>
<td>6.63 0.92</td>
</tr>
<tr>
<td>Mean duration (s) of utterances</td>
<td>1.58 0.74</td>
<td>2.12 1.41</td>
</tr>
<tr>
<td>Mean interutterance interval (s)</td>
<td>0.75 0.30</td>
<td>0.59 0.33</td>
</tr>
</tbody>
</table>

Note: F0 = fundamental frequency.
Table 3. Comparison of the Present Study and Previous Studies on Infant-Directed-Speech Preferences

<table>
<thead>
<tr>
<th>Study</th>
<th>Age group or groups targeted (months)</th>
<th>Nature of the speech recorded</th>
<th>Number and description of stimuli</th>
<th>Mean F0 (Hz)</th>
<th>Difference in mean F0 (IDS – ADS; Hz)</th>
<th>Ratio of IDS F0 to ADS F0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present study</td>
<td>3–15</td>
<td>Semistructured speech directed toward a 4- to 8-month-old child or an adult experimenter</td>
<td>Eight full trials for each type of speech</td>
<td>207</td>
<td>175</td>
<td>32</td>
</tr>
<tr>
<td>Cooper and Aslin (1990)</td>
<td>0, 1</td>
<td>Scripted speech read with no infant present</td>
<td>Four sentences in each type of speech</td>
<td>316</td>
<td>260</td>
<td>56</td>
</tr>
<tr>
<td>Newman and Hussain (2006)</td>
<td>4.5, 9, 13</td>
<td>Scripted speech read with no infant present</td>
<td>Four passages in each type of speech</td>
<td>226</td>
<td>190</td>
<td>36</td>
</tr>
<tr>
<td>Thiessen, Hill, and Saffran (2005)</td>
<td>7</td>
<td>Nonsense strings of syllables read with no infant present</td>
<td>Twelve strings in each type of speech</td>
<td>292</td>
<td>230</td>
<td>62</td>
</tr>
<tr>
<td>Cooper, Abraham, Berman, and Staska (1997)</td>
<td>1, 4</td>
<td>Mothers' naturalistic speech directed toward their own infants</td>
<td>Twenty seconds of speech in each type of speech</td>
<td>219</td>
<td>184</td>
<td>35</td>
</tr>
<tr>
<td>Schachner and Hannon (2011)</td>
<td>5</td>
<td>Speech elicited by asking adults to describe a picture they were looking at</td>
<td>Two 1-min videos in each type of speech</td>
<td>273.0</td>
<td>224.7</td>
<td>48.3</td>
</tr>
</tbody>
</table>

Note: F0 = fundamental frequency; IDS = infant-directed speech; ADS = adult-directed speech.

Fig. 1. Distribution of fundamental frequency (F0) values in the stimulus set. The graph shows the proportion of voiced segments of infant-directed speech (IDS) and adult-directed speech (ADS) that fell in each bin of a logarithmically spaced series of F0 bins within a search range of 32 through 650 Hz. The solid and dotted lines show means ±1 SE across stimuli.
1 through 7), whereas all ADS clips were judged highly likely to be directed at adults ($M = 2.12, SD = 0.38$). Both sets of clips were judged to be relatively natural; if anything, the ADS clips were rated as slightly more natural ($M = 5.18, SD = 0.19$) than the IDS clips ($M = 4.47, SD = 0.31$). In sum, because our IDS stimuli were created from naturalistic productions from a wide range of mothers, they were less extreme in their intonation than the stimuli used in previous studies, but they were judged as natural and were easily identified as infant directed.

**Procedure**

**Basic procedure.** Each lab used the testing paradigm (or paradigms) with which they were most familiar, among variants of three widely used paradigms: Twenty laboratories used the HPP, 16 used the single-screen central-fixation preference procedure, and 27 used single-screen central-fixation with fixations recorded by a corneal-reflection eye tracker; 4 labs contributed data using two different methods. All procedural instructions to participant labs can be found on OSF (https://osf.io/s3jca/).

To minimize researcher degrees of freedom, we asked participating labs to adhere to our instructions closely. Deviations from the basic protocol for each paradigm were necessary in some cases because of variation in the software and procedures used in each laboratory and were documented for future analysis.

**First versus second test session.** In some laboratories, infants were sometimes tested in an unrelated experiment during their visit, either prior to or following the IDS-preference experiment. Each lab noted whether infants completed the IDS-preference experiment as their first (and possibly only) or second test session.

**Onset of each trial.** At the beginning of each trial, a centrally positioned visual stimulus (typically the study's standard attention getter, or a light in some HPP labs) was used to attract the infant's attention. When the infant fixated on this attention getter, a visual stimulus appeared (a checkerboard on a screen in the central-fixation and eye-tracking paradigms, a flashing light or fixation stimulus on a screen—e.g., a flashing circle—in the HPP paradigm). The stimulus appeared to the left or right of the infant in HPP setups and in the center in the other two setups.

**Trials.** At the beginning of each session, there were two warm-up trials that familiarized the infant with the general procedure. The auditory stimulus for warm-up trials was an 18-s clip of piano music, and the visual stimulus was identical to that shown on the test trials. In addition to familiarizing the infant with the general experimental setup, the warmup trials highlighted the contingency between looking at the visual display and the onset of the auditory stimulus. We did not analyze data from these trials. The warmup trials were followed by 16 test trials presenting the IDS and ADS auditory stimuli.

**Minimum looking time.** There was no minimum required looking time during data collection (i.e., trials were never repeated). A minimum looking time of 2 s was the criterion for inclusion of a trial in analysis. This 2-s minimum was chosen after the laboratories discussed their typical standards of practice regarding minimum trial length, which varied considerably. The 2-s criterion was selected to ensure that the infants had sufficient time to hear enough of the stimulus to discriminate IDS from ADS.

**Maximum looking time.** On each test trial, infants could hear speech for a maximum of 18 s (i.e., the duration of one sound file). At labs using software that could implement infant-controlled trial lengths, a trial ended if the infant looked away from the visual stimulus for two consecutive seconds. Otherwise, the trial continued until the stimulus ended.

**Randomization.** Four pseudorandom trial orders were created. Each order contained four blocks, and each block contained two IDS and two ADS trials, in alternation. Two blocks in each order began with IDS, and the other two began with ADS. The same IDS and ADS stimuli were always paired with one another, to facilitate analyses of preference scores by item.

**Volume.** Each lab was asked to use a stimulus volume level that was consistent with their general lab practices; this decision was not standardized across labs. Labs were instead instructed to measure and report their average volume (dB SPL) with and without a white-noise reference audio clip playing, though not all contributing labs reported these measurements (measurements were reported by 47 labs). From these values, we calculated a signal-to-noise ratio for each lab, $M = 1.95, SD = 0.43$, range: 1.25–3.30.

**Minimizing caregiver bias.** A custom masking stimulus containing a blend of instrumental music and a pastiche of speech stimulus materials triggered at random times and with random amplitude was available as part of the study materials. This masking stimulus was played to the caregiver over noise-attenuating headphones, to mask the stimuli that the infant was hearing via external loudspeakers. Experimenters were instructed to play the masking music at a high (but comfortable and safe) volume.

**Coding.** Coding of looking times was conducted via the standard procedure in each lab. There were three methods
of coding infants’ eye gaze: on-line coding by an experi-
menter via button press during the experimental session, 
off-line coding of a video after the experimental session, 
and automatic coding by an eye tracker. When both on-
line and off-line coding were reported, we used the off-
line coding.

Minimizing experimenter bias. Experimenter mak-
ing on-line coding decisions (in the central-fixation and 
HPP methods) were blind to the particular stimulus pre-
sented during each test trial, as they either were located 
in a room separate from the infant or were in the same 
room but were wearing noise-attenuating headphones 
playing the same masking stimulus presented to the 
infant’s caregiver. Off-line coding was conducted without 
direct access to the auditory stimuli.

Demographics. All labs were instructed to collect basic 
demographic information about participants: sex, date of 
birth, estimated proportion of language exposure for 
each language heard in their daily life, race-ethnicity 
(using categories appropriate for the cultural and geo-
graphic context), preterm-versus-full-term status, history 
of ear infections, known hearing or visual impairments, 
and known developmental concerns (e.g., development-
mental disorders). Parents were also asked to report informa-
tion about themselves (sex, level of education, and native 
language or languages) and the infant’s siblings (sex and 
date of birth). A standard recommended questionnaire 
was distributed to participating labs as part of the instruc-
tions, although labs were permitted to use their own forms 
as long as they gathered the necessary information. A sub-
set of laboratories provided extensive additional informa-
tion about participants and testing circumstances (not 
analyzed here), for use in planned follow-up projects.

General lab practices

Training of research assistants. Each lab was respon-
sible for maintaining good practices for training exper-
imenter and was expected to use the same rigor with 
the ManyBabies study as with any other study in their lab. 
Laboratories reported on which research assistant ran each 
infant, using pseudonyms or numerical codes. Each lab-

oratory completed a questionnaire regarding its training 
practices, the experience and academic status of each 
experimenter, and its basic practices for greeting partici-
pants.

Reporting of technology mishaps and infants’ and 
parents’ behavior. Laboratories were asked to record 
relevant concerns, anomalies, and comments according 
to their standard lab practices, and these were converted 
to a standardized form during the main analysis. Relevant 
concerns included infants crying during testing, parents 
to intervene in a way that would affect infants’ looking 
behavior (e.g., talking or pointing), and technical prob-
lems that prevented the normal presentation of experi-
mental stimuli.

Videos

All laboratories provided a walk-through video that 
detailed their basic processes, including procedures for 
greeting participants and caregivers, obtaining consent, 
and collecting data, and that showed the physical char-
acteristics of their laboratory. (Our preregistration stated 
that further procedural documentation would be col-
lected and made available, but standardized reporting 
for procedural decision making proved difficult to 
develop and implement.) In addition, labs were strongly 
encouraged to collect and share video recordings of 
their data collection, within the limits of what was per-
missible given their ethics approval and participants’ 
consent. If labs could not provide videos of partici-
pants, they were asked to provide a video showing a 
run-through of their procedure, pictures and informa-
tion regarding the study setup, or both. A number of 
laboratories contributed these video recordings to Data-
brary, where they can be found by searching for 
“ManyBabies.”

Inclusion criteria for participants

All data collected for the study (i.e., data for every 
infant for whom a data file was generated, regardless 
of how many trials were completed) were given to the 
analysis team for confirmatory analyses. Participants 
were included in analysis only if they met all the criteria 
described in this section. All exclusion rules were 

applied sequentially, and the reported percentages of 
excluded infants reflect this sequential application to 
an initial sample (prior to exclusions) of 2,754. The 
following list describes the inclusion criteria, in the 
order of their application. Note that the first three cri-
teria preemptively prevented participation (although 
some infants were run erroneously).

- Monolingual: Only monolingual infants, of any 
language background, were included in the final 
sample. Monolingualism was defined as a mini-
umum of 90% exposure to the native language, as 
reported by the infant’s parent. This cutoff score 
struck a balance between including most infants 
who would typically be considered monolingual 
in infant language studies and excluding those who
might be considered bilingual (Byers-Heinlein, 2015). Of the initial sample, 162 (5.88%) infants did not meet this criterion.

- **Full-term:** We defined full term as gestation time greater than or equal to 37 weeks. Of the remaining sample, 62 (2.39%) infants did not meet this criterion.

- **No diagnosed developmental disorders:** We excluded infants with parent-reported developmental disorders (e.g., chromosomal abnormalities) or diagnosed hearing impairments. Of the remaining sample, 2 (0.08%) infants were tested but did not meet this criterion. Because of concerns about the accuracy of parent reports, infants whose parents reported that they had experienced ear infections were not excluded unless the parents also reported medically confirmed hearing loss.

- **Contributed usable data:** To be included in the study, an infant was required to have contributed nonzero looking times for at least one pair of test trials (i.e., IDS and ADS trials from the same stimulus pair), after trial-level exclusions were applied. Of the remaining sample, 78 (3.09%) infants did not meet this criterion. We adopted this relatively liberal inclusion criterion even though it is at variance with the more stringent standards that are typically used in infancy research. We were interested in maximizing the amount of data from each lab to be included in the initial analysis, and our paradigm was, by design, less customized for any particular age group than previously used paradigms were (and hence likely to produce greater data loss, especially for older infants, who tend to habituate more quickly). In the exploratory analyses we report, we considered how exclusion decisions affected our effect-size estimates.

After these rules were applied, participants could also be excluded from analysis because of session-level errors, including equipment error (e.g., no sound or visuals on the first pair of trials), experimenter error (e.g., an unblinded experimenter, if looking time was measured by live button press), or reported evidence of consistent outside interference (e.g., talking or pointing by parents, construction noise, sibling pounding on the door). Session-level error resulted in 78 (3.18%) infants for whom we had other reported data being dropped from analysis. This number is likely an underestimate, however, because many participating labs did not provide data for all children with session-level errors. In addition, session-level errors were not classified consistently across labs, so an accurate classification of the proportion of different types of errors was not possible.

### Trial exclusions

We excluded individual trials for which issues were reported (e.g., fussy infant, incorrect stimulus, single instance of parent or sibling interference). A total of 4,471 (10.61%) trials were affected by such errors, which resulted in the complete removal of 29 infants (1.22%). As with session-level errors, classification of these concerns was inconsistent across participating labs, but the most common source of trial-level errors was infant fussiness.

Our trial-length minimum of 2 s of looking time resulted in the exclusion of an additional 6,027 (16.13%) trials. These trials were analyzed as missing in our planned analysis. This trial-level exclusion led to the exclusion of 3 additional infants who had no usable trial pairs (0.13%).

### Inclusion of a lab’s data

We included a lab’s data if they were able to achieve the minimum $N$ required for a half-sample and if, after exclusions, they contributed data from 10 or more infants. This criterion led to 11 (0.47%) infants from two labs being excluded from the final sample.

### Deviations from the registered protocol

Given that this was the first experimental cross-laboratory infant study of such a large scale, there were a number of unanticipated issues that arose during data collection within individual labs and at the study level, and these issues resulted in deviations from our registered protocol. Necessary decisions were made without consideration of their impact on the results, and all such cases were documented. Fuller documentation can be found accompanying our shared data in the OSF repository; here we summarize the nature and extent of these deviations. Note that some were the result of typical within-laboratory protocol deviations (experimenter error, etc.), whereas others stemmed from the additional challenges inherent in harmonizing methodology and data format across such a large number of laboratories with different internal protocols and standards.

The protocol deviations include the following:

- Before labs had commenced data collection, we altered our attention-getter stimulus to be a processing annulus accompanied by chimes (to address the concern that a laughing baby, our
original attention getter, might be a factor in observing a preference for IDS rather than ADS, e.g., via presenting an infant as part of the content of the experiment); however, some labs used the original stimulus.

- For a variety of reasons, labs’ protocols deviated from the registered protocol in ways that resulted in trials longer than the assumed maximum of 18 s. In all cases, looking times on these trials were truncated to 18 s.
- A number of labs provided data from infants who were within the 3- to 15-month age range, but outside the individual lab’s preregistered age bin. These infants were included in the analyses.
- Many labs deviated from their preregistered sample size because of constraints on testing resources. We included these labs provided they met the minimum inclusion criteria for the study as a whole. All such labs certified that they did not make decisions regarding sample size on a data-dependent basis.
- A number of laboratories marked participants for exclusion due to session-level errors for reasons other than equipment error, experimenter error, or outside interference.

Deviations regarding exclusions bear further discussion. Some labs marked participants for exclusion because of trial-level errors (e.g., fussiness, parental interference), even though sufficient trial-level data were available for analysis. Similarly, individual trials were sometimes marked as error trials for reasons related to session-level issues. All trial-level and session-level errors were reviewed centrally by at least two coders using all available information in the spreadsheet to determine whether the error was more appropriately categorized as a trial-level or session-level error. Specific information about each error coding that was changed during this process is available in the metadata directory within the data-analysis codebase, available on our OSF project page.

In total, 313 participants (from 50 labs) who had been marked for session-level exclusions were retained for further processing and analysis, for the following reasons: The session-level exclusion was based solely on the existence of trial-level errors (190 infants), the exclusion was based on a different exclusion criterion (e.g., the participant was out of the age range or was preterm; 93 infants), or the central analysis team decided that the issue identified by the lab did not warrant exclusion (e.g., the lab implemented a look-away criterion slightly different from the preregistered one; 30 infants). Note that many of these retained participants were subsequently excluded at other points in the analysis pipeline because, although they did not meet the criteria for exclusion on the basis of session-level errors, they did meet other conditions for exclusion (e.g., as noted, some of these participants were out of the age range or were preterm).

In addition to recoding session-level errors, we corrected the coding of trial-level errors when appropriate. In total, 778 such errors, involving 62 participants in 16 different labs, were recoded. The majority of these cases involved a lab coding a session-level error (e.g., outside the age range) as a trial-level error (584 trials) or coding a trial-level error as a session-level error (e.g., a lab recorded a session-level error when an infant was fussy on a specific trial, but did not code the affected trial as an error trial; 133 trials). Other trials were corrected when subsequent investigation of lab notes and discussion with lab members revealed that the original code needed to be changed (61 trials).

A variety of additional errors were found (e.g., participants who were identified as having been run in a lab’s pilot test but who were not properly excluded) and fixed within the spreadsheets. Video data were not reviewed centrally, although in some cases when a question arose, the laboratory reviewed its own video in order to respond. Our corrections of trial- and session-level errors have been carefully documented, and this information can be accessed upon request, but because the documentation in some cases includes identifiable information about participants, it is not possible to share it publicly.

Other reported protocol deviations included failing to submit a preregistration form (one lab), setting the trial-end (look-away) criterion to 3 s rather than 2 s for some participants (one lab), temporarily changing location during data collection (one lab), making minor technical changes in the protocol after the start of data collection (two labs), alternating left- and right-side presentation of stimuli and testing skin conduction during the procedure (one lab), implementing procedural deviations related to high-chair usage (one lab), using attention getters other than the preregistered stimulus (four labs), and using a pinwheel rather than a checkerboard as the main visual fixation stimulus in the HPP (one lab).

We also detected a large number of other kinds of errors in the submitted data as a result of the comprehensive checking process conducted during analysis. These typographical and other errors were resolved when necessary by contacting the submitting lab. In general, we included data obtained with minor protocol deviations, and erred on the side of excluding data, when necessary, at the trial rather than session level. A
Quantifying Sources of Variability in Infancy Research

few demographic variables required greater central scrutiny than originally anticipated. Most notably, there was considerable variability in the interpretation of the preterm and bilingual designations (despite centrally dictated standards). When necessary, we recoded lab data so as to conform to the definitions in the original protocol.

The instructions in our registered protocol were ambiguous as to whether the inclusion criterion for labs was contributing data for 10 or more infants or for more than 10 infants. The more liberal of these two criteria was used.

Finally, two labs submitted data after the deadline. In one case, this was because of a communication error; in the other case, the lab continued data collection after the deadline, and 8 additional infants were tested. Both data sets were included in the final analysis reported here.

Results

Confirmatory analyses

Data processing and analytic framework. All planned analyses were preregistered in our initial Registered Report submission (available at https://osf.io/vd789/). Our primary dependent variable of interest was looking time during test trials. Looking time was defined as time spent fixating the screen (central-fixation and eye-tracking methods, some HPP setups) or light (other HPP setups); looking time did not include any time spent looking away from the screen, even when that time was below the threshold for terminating a trial. Because looking times are nonnormally distributed, we followed Csibra, Hernik, Mascaro, Tatone, and Lengyel’s (2016) recommendation and log-transformed all looking times prior to statistical analysis.

We adopted two complementary analytic frameworks: meta-analysis and mixed-effects regression. In the meta-analytic framework, we conducted standard analyses within each lab and then estimated variability in the results of these analyses across labs. The meta-analytic approach has a number of advantages over the mixed-effects approach, including the use of simple within-lab analyses, the ability to estimate cross-lab variability directly, and the possibility of making direct comparisons with the standardized effect sizes estimated in previous meta-analyses. However, the standard random-effects meta-analytic model is designed for cases in which the raw data are unavailable and procedures and data types are not standardized. In contrast, in our situation, procedures and data were standardized across labs, and relevant moderators were recorded. The availability of trial-by-trial data for all labs allowed us to use mixed-effects models, which account for the nesting and crossing of random effects (e.g., participants nested within labs, items crossed across labs) and can provide more accurate estimates of the main effect and moderators. We report both analyses to allow for the most comprehensive understanding of the variance in the data.

Our meta-analyses were conducted as follows. Each lab’s data set was considered a separate study. For each such study, we computed individual infants’ IDS preference by (a) subtracting the looking time on each IDS trial from the looking time on its paired ADS trial (excluding trial pairs with missing data) and (b) computing a mean difference score across trial pairs. Then we computed a group IDS preference for each lab and age group within that lab using $d_z$, a version of Cohen’s standard $d$ statistic, computed as the average of infants’ IDS preference scores divided by the standard deviation of those scores. We then used standard random-effects meta-analysis fit using restricted maximum likelihood (REML) with the metafor package (Viechtbauer, 2010).

Although we did not anticipate this in our initial analysis plan, a large number of labs collected data outside of their planned samples. For example, many labs contributed data from a sample of children within a specific age bin as well as several children outside of that age bin, or from a sample of children using one method and from a handful of children using another. Although we included these children in the mixed-effects analyses described next, we worried that the inclusion of many unplanned samples of just one or a few infants in the meta-analytic models would excessively increase lab-level variance. Thus, for the meta-analyses, we included only samples (e.g., age, language, or method groups) with 10 or more infants.

Our mixed-effects models, fit to the entire data set collected from the 67 labs, were specified as follows:

\[
DV \sim IV_1 + IV_2 + \ldots + (\ldots | participant) + (\ldots | item) + (\ldots | lab)
\]

The goal of this approach was to examine effects of the independent variables (IV) on the dependent variable (DV), while controlling for variation in both the DV (random intercepts) and the relationship of the IV to the DV (random slopes) due to relevant grouping units (participants, items, and labs). The use of mixed-effects models also allowed us to move away from using a difference score as the dependent variable of interest. Although difference scores simplified the process of calculating effect sizes for the metaregression, their use required that trials be paired, so some collected data...
(i.e., from unpaired trials) could not be analyzed. In the mixed-effects framework, in contrast, looking time on individual trials was the dependent measure, so all trials could be included.

In our mixed-effects models, we planned a maximal random-effects structure (Barr, Levy, Scheepers, & Tily, 2013), which entailed specifying all random effects that were appropriate for the experimental design (e.g., IDS/ADS trial type could be nested within participants—because each infant heard stimuli in both conditions—but could not be nested within items because each item was unique to its trial type). In cases of mixed-effects models that failed to converge, we pursued an iterative pruning strategy. We began by removing random slopes nested within items (as that grouping was of least theoretical interest) and next removed random slopes nested within participants and then random slopes nested within labs. We then removed random intercepts from groupings in the same order, retaining effects of trial type until last because these were of greatest theoretical interest. We fit all models using the lme4 package (Kuznetsova, Brockhoff, & Christensen, 2017).

**IDS preference.** What was the overall magnitude of the IDS preference we observed? This question was answered in the cross-lab meta-analysis by fitting the main-effect model specified by \( d_z \sim 1 \) to the 108 separate group means and variances (after aggregating the data by lab and age group). The mean effect-size estimate was 0.35 (95% CI = [0.29, 0.42], \( z = 10.67, \ p < .001 \)). A forest plot for this meta-analysis is shown in Figure 2. Further, 1,373 of the 2,329 infants (58.95%) showed a numerical preference for IDS.

**Independent relationship of IDS preference to moderating variables.** We next fitted a set of moderated meta-analytic models. We began by examining the relationship of IDS preferences to age, using the average age in months for each lab’s sample as the moderator value. Labs that contributed samples from two age bins had a separate value added for each age (because of the small number of such cases, we did not model this dependency between labs). For ease of interpretation, we centered age in this analysis. The age-modernated model, \( dz \sim 1 + \text{age} \), yielded an estimated main effect of 0.35 (95% CI = [0.29, 0.41], \( z = 11.47, \ p < .001 \)) and an age effect of 0.05 (95% CI = [0.03, 0.07], \( z = 4.89, \ p < .001 \)). This positive age coefficient indicated that the measured IDS preference was on average larger for older children. The age trends for the NAE and non-NAE samples are plotted in Figure 3.

We next investigated effects of experimental method, with method dummy coded using single-screen central fixation as the reference level. The method-moderated model (\( dz \sim 1 + \text{method} \)) yielded a reference-level intercept of 0.29 (95% CI = [0.18, 0.41], \( z = 4.98, \ p < .001 \)), which is the mean effect size for single-screen presentation. The HPP yielded an additional effect of 0.21 (95% CI = [0.06, 0.37], \( z = 2.74, \ p = .006 \)), a substantial gain in measured IDS preference for those labs using the HPP as compared with single-screen central fixation. In contrast, eye tracking yielded an effect of –0.06 (95% CI = [–0.21, 0.10], \( z = –0.71, \ p = .479 \)), thus, there was a slight, nonsignificant decrease in measured effect size when the method used was eye tracking, rather than single-screen central fixation.

The language-moderated model (\( dz \sim 1 + \text{language} \)) was fitted with language group coded as a categorical variable indicating whether infants were tested in a lab in which NAE was the standard language (i.e., in the United States or Canada). The reference-level effect (i.e., not NAE) was 0.29 (95% CI = [0.20, 0.37], \( z = 6.56, \ p < .001 \)). For infants in North American labs, the effect was increased by 0.15 (95% CI = [0.02, 0.27], \( z = 2.26, \ p = .024 \)). Thus, measured IDS preferences were higher in those infants for whom the stimuli were native-language congruent.

**Moderating variables’ joint relationships with IDS preference.** Because age group, language, and method were confounded across labs (labs with particular methods also chose specific sample age ranges, and these choices were not independent), we next turn to the mixed-effects modeling framework to estimate participant-level age effects and lab-level method effects. Figure 4a shows the spread of participant-level IDS preference, and Figure 4b shows estimated trends in IDS preference. The mean looking time across all trials was 8.21 s for the IDS condition and 7.38 s for the ADS condition.

Our main model was as follows:

\[
\log \text{looking time} \sim \text{trial type} \ast \text{method} + \text{trial type} \\
\ast \text{trial number} + \text{age} \ast \text{trial number} + \text{trial type} \\
\ast \text{age} \ast \text{language} + (\text{trial type} \ast \text{trial number}|\text{participant}) \\
+ (\text{trial type} \ast \text{age}|\text{lab}) \\
+ (\text{method} \ast \text{age} \ast \text{language}|\text{item})
\]

Trial type, language, and method were dummy coded with ADS trials, non-NAE community, and single-screen method as the reference levels; thus, coefficients are interpretable such that positive effects of trial type indicate longer looking on IDS trials, positive effects of language indicate longer looking in NAE communities,
Fig. 2. Forest plot of the meta-analytic results for infant-directed-speech preference. The standardized effect size is shown for each lab; error bars indicate 95% confidence intervals. Labs are grouped into separate panels by method: head-turn preference procedure (HPP), eye tracking, or central fixation. Larger plotted points correspond to greater inverse variance. In each panel, the diamond and associated error bars represent the meta-analytic estimate from the method-moderated model and the estimate's 95% confidence interval. The bottom panel shows the global meta-analytic estimate and 95% confidence interval from the unmoderated model.
and positive effects of method indicate longer looking with eye tracking and the HPP. To increase the interpretability of coefficients, we centered age (in months) and coded trial number with Trial 1 as the reference level.

We specified this model to minimize higher-order interactions but preserve theoretically important interactions. We included main effects of trial type, method, language, age, and trial number, capturing the basic effects of each on looking time (e.g., longer looking times for IDS, shorter looking times on later trials). In addition, we included two-way interactions of trial type with method (modeling the possibility that some methods would show larger IDS preferences than others) and trial type with trial number (modeling the possibility of faster habituation to ADS), as well as an interaction of age and trial number (modeling faster habituation for older children). We also included two- and three-way interactions of age, trial type, and language (modeling possible developmental changes in IDS preference and developmental differences in IDS preference across language groups). Both developmental effects and trial effects were treated linearly in this model; although they likely had nonlinear effects, adding quadratic or other effects would have substantially increased the model’s complexity. After pruning random effects for nonconvergence, our final model specification was

\[
\text{log looking time} \sim \text{trial type} \times \text{method} + \text{trial type} \times \\
\text{trial number} + \text{age} \times \text{trial number} + \text{trial type} \times \text{age} \times \\
\text{language} + (1|\text{participant}) + (1|\text{lab}) + (1|\text{item})
\]

Table 4 shows the coefficient estimates from this model.

Overall, the fitted coefficients of the mixed-effects model were consistent with the results of the individual meta-analyses. Given the structure of the mixed-effects model, positive coefficients for the IDS predictor indicated greater IDS preference (i.e., greater looking times on IDS trials). The fitted model showed a significant
positive effect of IDS stimuli, consistent with a global IDS preference. Results were also consistent with the age- and language-moderated meta-analyses, as there were significant and positive two-way interactions of IDS with age and with NAE; preferences for IDS were greater among older children and children in NAE contexts. Further, there was a positive interaction of IDS with the HPP method, consistent with the method-moderated model. There was not a significant three-way interaction of IDS, age, and NAE, however; in other words, there was not a reliable differential change in IDS preference for older children in NAE contexts over and above that expected given each of these factors alone.

In addition, a number of other factors were significant predictors of looking time. Looking time decreased across trials and was shorter among older children. This result generally confirmed that all infants habituated to our experimental stimuli, and older infants did so more quickly. Further, eye tracking led to shorter looking times overall across stimulus classes.

**Effect of second-session testing on IDS preference.** We preregistered an analysis of whether the pattern of IDS preference was different for second-session infants than for first-session infants (i.e., those who completed the IDS-preference experiment as their first, and possibly only, session). Only six labs contributed data for second-session infants, however, and only 41 infants were represented. Thus, we did not fit the full, preregistered mixed-effects model for this variable, as we did not have enough variability on the important covariates. We note that 46.34%
(19/41) of second-session infants (95% CI = [31.65%, 61.30%]) showed a numerical preference for IDS. This percentage was numerically different, but not distinguishable statistically, from the 58.95% of first-session infants who showed an IDS preference, likely because of the small sample of second-session infants.

**Sex and IDS preference.** In order to investigate effects of biological sex on IDS preference, we fitted the mixed-effects model specified earlier with the addition of a main effect of sex and a trial-type-by-sex interaction. Female was coded as the reference level. The main effect of sex was not significant, $\beta = 0.01$ ($SE = 0.02$, $p = .63$), and neither was the interaction of sex with trial type, $\beta = -0.01$ ($SE = 0.01$, $p = .51$). These small, nonsignificant coefficients suggest that sex was not a strong determinant of measured IDS preferences in our data.

**Moderator effects on missing data.** One further question regarding our data was whether particular moderator variables affected not only the amount of looking time recorded, but also whether children looked at all during a trial. To test for effects of moderators on the presence of missing data, we constructed a categorical variable (missing), which was true if a trial had no included looking time (i.e., no looking recorded, a look under 2 s, or no looking because the experiment was already terminated) and false otherwise. We fitted a logistic mixed-effects model with all two-way interactions between method, age, and trial number, using the following specification:

$$
\text{missing data point} \sim \text{method} \times \text{age} \times \text{method} \times \text{trial number} + (1|\text{participant}) + (\text{trial number} \times \text{age}|\text{lab}) + (\text{method} + \text{age}|\text{item})
$$

After pruning for nonconvergence, our final model specification was

$$
\text{missing data point} \sim \text{method} \times \text{age} \times \text{method} \times \text{trial number} + (1|\text{lab})
$$

Table 5 shows the coefficient estimates from this model. To aid convergence, we centered and scaled age and trial number, and set single-screen presentation as the reference level. Positive coefficients indicate a higher probability of missing data. More data were missing for older children and later trials, a pattern consistent with the idea that all children habituated to the stimuli, but that older children habituated faster. There was also a significant negative interaction of age and eye tracking, suggesting that data loss for eye tracking was substantially greater in younger children and lower in older children. The other coefficients were relatively small and nonsignificant.

### Exploratory analyses

**Meta-analytic heterogeneity.** One question of interest was whether we observed any meta-analytic heterogeneity in the data. A finding of meta-analytic heterogeneity indicates the presence of unexplained variance in effect size over and above that due to sampling variation. We calculated $\tau^2$ as an estimate of the total heterogeneity in

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.090 (0.152)</td>
<td>-7.140</td>
<td>.000</td>
</tr>
<tr>
<td>Eye tracking</td>
<td>0.167 (0.130)</td>
<td>1.290</td>
<td>.198</td>
</tr>
<tr>
<td>HPP</td>
<td>-0.052 (0.051)</td>
<td>-1.020</td>
<td>.308</td>
</tr>
<tr>
<td>Trial number</td>
<td>-0.038 (0.002)</td>
<td>-25.000</td>
<td>.000</td>
</tr>
<tr>
<td>Age</td>
<td>-0.035 (0.004)</td>
<td>-7.950</td>
<td>.000</td>
</tr>
<tr>
<td>NAE</td>
<td>-0.016 (0.049)</td>
<td>-0.335</td>
<td>.738</td>
</tr>
<tr>
<td>IDS * Eye Tracking</td>
<td>-0.009 (0.017)</td>
<td>-0.548</td>
<td>.584</td>
</tr>
<tr>
<td>IDS * HPP</td>
<td>0.034 (0.015)</td>
<td>2.270</td>
<td>.023</td>
</tr>
<tr>
<td>IDS * Trial Number</td>
<td>-0.003 (0.002)</td>
<td>-1.370</td>
<td>.172</td>
</tr>
<tr>
<td>Trial Number * Age</td>
<td>0.001 (0.000)</td>
<td>3.140</td>
<td>.002</td>
</tr>
<tr>
<td>IDS * Age</td>
<td>0.012 (0.003)</td>
<td>4.300</td>
<td>.000</td>
</tr>
<tr>
<td>IDS * NAE</td>
<td>0.039 (0.013)</td>
<td>3.060</td>
<td>.002</td>
</tr>
<tr>
<td>Age * NAE</td>
<td>0.001 (0.006)</td>
<td>0.398</td>
<td>.738</td>
</tr>
<tr>
<td>IDS * Age * NAE</td>
<td>0.004 (0.004)</td>
<td>1.050</td>
<td>.292</td>
</tr>
</tbody>
</table>

Note: Numbers in parentheses are standard errors. HPP = head-turn preference procedure; NAE = North American English.

### Table 4. Coefficient Estimates From the Linear Mixed-Effects Model Predicting Log Looking Time

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.180 (0.051)</td>
<td>43.100</td>
<td>.000</td>
</tr>
<tr>
<td>IDS</td>
<td>0.099 (0.036)</td>
<td>2.740</td>
<td>.010</td>
</tr>
<tr>
<td>Eye tracking</td>
<td>-0.265 (0.046)</td>
<td>-5.790</td>
<td>.000</td>
</tr>
<tr>
<td>HPP</td>
<td>-0.052 (0.051)</td>
<td>-1.020</td>
<td>.308</td>
</tr>
<tr>
<td>Trial number</td>
<td>-0.038 (0.002)</td>
<td>-25.000</td>
<td>.000</td>
</tr>
<tr>
<td>Age</td>
<td>-0.035 (0.004)</td>
<td>-7.950</td>
<td>.000</td>
</tr>
<tr>
<td>NAE</td>
<td>-0.016 (0.049)</td>
<td>-0.335</td>
<td>.738</td>
</tr>
<tr>
<td>IDS * Eye Tracking</td>
<td>-0.009 (0.017)</td>
<td>-0.548</td>
<td>.584</td>
</tr>
<tr>
<td>IDS * HPP</td>
<td>0.034 (0.015)</td>
<td>2.270</td>
<td>.023</td>
</tr>
<tr>
<td>IDS * Trial Number</td>
<td>-0.003 (0.002)</td>
<td>-1.370</td>
<td>.172</td>
</tr>
<tr>
<td>Trial Number * Age</td>
<td>0.001 (0.000)</td>
<td>3.140</td>
<td>.002</td>
</tr>
<tr>
<td>IDS * Age</td>
<td>0.012 (0.003)</td>
<td>4.300</td>
<td>.000</td>
</tr>
<tr>
<td>IDS * NAE</td>
<td>0.039 (0.013)</td>
<td>3.060</td>
<td>.002</td>
</tr>
<tr>
<td>Age * NAE</td>
<td>0.001 (0.006)</td>
<td>0.398</td>
<td>.738</td>
</tr>
<tr>
<td>IDS * Age * NAE</td>
<td>0.004 (0.004)</td>
<td>1.050</td>
<td>.292</td>
</tr>
</tbody>
</table>

Note: Numbers in parentheses are standard errors. IDS = infant-directed speech; HPP = head-turn preference procedure; NAE = North American English.

### Table 5. Coefficient Estimates From the Linear Mixed-Effects Model Predicting Whether an Observation Was Missing

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Estimate</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.090 (0.152)</td>
<td>-7.140</td>
<td>.000</td>
</tr>
<tr>
<td>Eye tracking</td>
<td>0.167 (0.130)</td>
<td>1.290</td>
<td>.198</td>
</tr>
<tr>
<td>HPP</td>
<td>-0.178 (0.195)</td>
<td>-0.913</td>
<td>.361</td>
</tr>
<tr>
<td>Age</td>
<td>-0.035 (0.004)</td>
<td>-7.950</td>
<td>.000</td>
</tr>
<tr>
<td>NAE</td>
<td>-0.016 (0.049)</td>
<td>-0.335</td>
<td>.738</td>
</tr>
<tr>
<td>IDS * Eye Tracking</td>
<td>-0.009 (0.017)</td>
<td>-0.548</td>
<td>.584</td>
</tr>
<tr>
<td>IDS * HPP</td>
<td>0.034 (0.015)</td>
<td>2.270</td>
<td>.023</td>
</tr>
<tr>
<td>IDS * Trial Number</td>
<td>-0.003 (0.002)</td>
<td>-1.370</td>
<td>.172</td>
</tr>
<tr>
<td>Trial Number * Age</td>
<td>0.001 (0.000)</td>
<td>3.140</td>
<td>.002</td>
</tr>
<tr>
<td>IDS * Age</td>
<td>0.012 (0.003)</td>
<td>4.300</td>
<td>.000</td>
</tr>
<tr>
<td>IDS * NAE</td>
<td>0.039 (0.013)</td>
<td>3.060</td>
<td>.002</td>
</tr>
<tr>
<td>Age * NAE</td>
<td>0.001 (0.006)</td>
<td>0.398</td>
<td>.738</td>
</tr>
<tr>
<td>IDS * Age * NAE</td>
<td>0.004 (0.004)</td>
<td>1.050</td>
<td>.292</td>
</tr>
</tbody>
</table>

Note: Numbers in parentheses are standard errors. HPP = head-turn preference procedure.
Table 6. Meta-Analytic Effect Size ($d_z$) and Percentage of Included Participants for Three Different Inclusion Criteria

<table>
<thead>
<tr>
<th>Method</th>
<th>Minimum number of trials</th>
<th>2 trials</th>
<th>4 trials</th>
<th>8 trials</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Estimate</td>
<td>Included participants (%)</td>
<td>Estimate</td>
</tr>
<tr>
<td>Central fixation</td>
<td></td>
<td>0.29 (0.06)</td>
<td>98</td>
<td>0.34 (0.06)</td>
</tr>
<tr>
<td>Eye tracking</td>
<td></td>
<td>0.24 (0.06)</td>
<td>85</td>
<td>0.33 (0.06)</td>
</tr>
<tr>
<td>HPP</td>
<td></td>
<td>0.51 (0.06)</td>
<td>98</td>
<td>0.56 (0.06)</td>
</tr>
</tbody>
</table>

Note: Numbers in parentheses are standard errors. HPP = head-turn preference procedure.

our models. In addition, we assessed heterogeneity using the $I^2$ statistic (Higgins, Thompson, Deeks, & Altman, 2003), which quantifies the proportion of total variation in estimates that is due to heterogeneity. We also report the results of a standard hypothesis test for heterogeneity, the Cochran $Q$ test; a statistically significant $Q$ test indicates that the null hypothesis of homogeneity of variance can be rejected (Huedo-Medina, Sanchez-Meca, Marin-Martinez, & Botella, 2006).

In our primary, intercept-only meta-analytic model, we found nonsignificant heterogeneity, $\tau^2 = 0.01$, $I^2 = 12.39\%$, and $Q(107) = 122$, $p = .15$. In the language-moderated model, heterogeneity was also nonsignificant, $\tau^2 = 0.01$, $I^2 = 7.76\%$, and $Q(106) = 116.18$, $p = .23$. In the age-moderated model, heterogeneity was even lower, $\tau^2 = 0.00$, $I^2 = 0.00\%$, and $Q(106) = 98.06$, $p = .70$. Finally, in the method-moderated model, heterogeneity was also low, $\tau^2 = 0.00$, $I^2 = 3.20\%$, and $Q(105) = 106.78$, $p = .43$. For none of these models could we reject the null hypothesis of no heterogeneity beyond sampling variation, and in no case was the magnitude of observed heterogeneity large. Although there were reliable moderators (see the meta-analytic results reported earlier), the effects of these moderators were quite small in magnitude relative to the sampling variation in individual labs’ effect-size estimates (because of the small median sample size within each lab).

**Usable-data inclusion criterion.** Because our criterion for including infants in the analysis was so liberal (i.e., infants needed to contribute data from only two trials to be included), we next explored the effects of different inclusion rules on the results. In particular, we calculated the meta-analytic effect size with four trials and eight trials as the minimum inclusion criterion. With a minimum of four trials, the effect size was $0.42$ ($95\%$ CI = $[0.35$, $0.48]$), $z = 12.05$, $p < .001$), and with a minimum of eight trials, the effect size was $0.48$ ($95\%$ CI = $[0.40$, $0.57]$), $z = 11.23$, $p < .001$). In comparison, our original analysis yielded a meta-analytic effect size of $0.35$ ($95\%$ CI = $[0.29$, $0.42]$), $z = 10.67$, $p < .001$). Furthermore, we computed the effect size for each method for each of these alternative inclusion criteria (see Table 6). Overall, more stringent inclusion criteria yielded substantially larger effects, although they also led to substantial data loss (especially for labs using the eye-tracking method).

**General Discussion**

We designed a large-scale, multilab study of infants’ preference for IDS and invited infancy researchers to participate. Our call for participation resulted in contributions from 69 labs, representing a total of 2,845 infants from 16 countries. The final sample used for analysis included 2,329 infants (see Table 1). We believe that this is the largest laboratory study of infancy to date. We begin our discussion by summarizing the principal results of the study with respect to four analytic questions and then discuss limitations of the study as well as future directions.

**Summary of findings**

Our first goal was to address the issue of the replicability of infants’ preference for IDS over ADS by conducting a preregistered study to obtain an unbiased measure of the magnitude of this preference. We expected to replicate prior demonstrations of the existence of an IDS preference in infant listeners, and our study indeed confirmed this expectation. Our overall meta-analytic mean was smaller in size than the effect found in a preceding meta-analysis of the literature, however (Bergmann et al., 2018; Dunst et al., 2012).

Although one possible interpretation of this difference in magnitude is that previously obtained effect sizes were inflated by publication bias, there are other possible explanations as well. In an individual laboratory, the methodology is tailored to the specific research
question and the age range and other characteristics of the infants tested (or, conversely, research questions are tailored to the existing methodological expertise of the laboratory). The approach used here, namely, applying multiple methodologies to the same research question across diverse age ranges and samples of infants, including non-English-learning infants, may have led to an underestimate of the true effect size (i.e., because the ideal presentation details that would maximize effect sizes might differ across methods and ages). Further, several of our methodological decisions might have decreased the effect size. For example, our stimuli had less extreme acoustic characteristics than the stimuli used in previous work and were generated by multiple speakers. In addition, our criterion for including participants in the final sample was less stringent than the criteria used previously.

Our second goal was to examine possible age effects in the preference for IDS. As did the prior published meta-analysis (Dunst et al., 2012), we found an increase in IDS preference across development. This trend is consistent with the idea that preference for IDS grows in response to experience with positive social interactions, but contrasts with some other reports in the literature (e.g., Hayashi et al., 2001; Newman & Hussain, 2006; Segal & Newman, 2015). Further, the magnitude of the positive developmental change we observed was considerable, at 0.05 standard deviations per month. This finding suggests that the preference for IDS is at a minimum modulated by experience, maturation, or both.

Any developmental trend, however, may be driven by changes in factors other than the underlying construct. As we discuss later, the stimuli we used may have been best suited for the older age ranges in our study. In addition, stronger effects among older infants may result from a more robust or more measurable behavioral response, independently of an underlying preference. Some evidence in favor of this possibility can be found in the data in MetaLab, an online data bank for meta-analysis in infancy research: Most extant meta-analyses show an increase in absolute effect size as infants mature, regardless of the research question (see, e.g., Bergmann et al., 2018).

Our third goal was to examine how the preference for IDS varies with the differing linguistic experiences of infants growing up in different linguistic communities. We found a preference for NAE IDS over NAE ADS even among participants whose native language or dialect was not NAE. This finding replicates previous work (Werker et al., 1994). However, in our study, NAE-exposed infants showed the strongest preference. Note that our findings do not support the idea of a simple attentional effect (i.e., greater attention to speech overall when presented in the native language): The effect of language background on overall (as opposed to preferential) looking times was not large in our regression models.

There are several possible interpretations of the native-language effect we observed. One possibility is that as infants become experts in their native language’s phonology and begin to acquire word meanings, they listen to speech in their own language differently, starting to process what is being said not just as “speech” or “register” per se, but as meaningful language (Gervain & Mehler, 2010; Johnson, 2016). For infants hearing a foreign language or even a foreign dialect of their native language, the ability to listen in this deeper, or more predictive, way is not available. Another possibility is that processing speech in an unfamiliar language requires more attentional resources, leaving fewer attentional resources to process some of the characteristics that may differentiate IDS and ADS. Regardless of which of these possibilities is true, preference for IDS may depend in part on the similarity of the IDS to one’s native-language experiences with IDS. This idea is somewhat supported by the age effect we observed; however, we did not observe a three-way interaction among age, stimulus type, and language background, which this interpretation would predict. Companion data are currently being collected in several non-NAE language communities using native-language stimuli created using the ManyBabies1 protocol, and these data may shed further light on this issue.

A fourth goal was to examine whether the measured experimental effect varied with the methodological approach. We found a stronger effect with the HPP than with the central-fixation or eye-tracking approaches. One potential interpretation of this finding is that the greater effort on the part of the infant in the HPP (i.e., turning the head, as opposed to making small eye movements) leads to stronger engagement in the task and therefore to stronger effects.

It is important to keep in mind, however, that methodology was not randomly assigned to laboratories, and the characteristics of laboratories probably varied systematically with their methodological choices. It may well be, for example, that laboratories with more expertise in investigating infants’ language acquisition were more likely to use the HPP. Furthermore, these findings should not be interpreted as suggesting that the HPP would be best suited for all research questions. Instead, a more modest interpretation is simply that a theoretically irrelevant variable related to laboratories and their methodological decisions appears to have a substantial and systematic effect on measured effect size (see Bergmann et al., 2018, for a similar conclusion based on meta-analytic data). We hope to undertake secondary
analyses of our data set to better understand factors that may have covaried with methodological choices. Moreover, further large-scale projects that include methodological contrasts of this type—perhaps with random assignment—may allow more specific conclusions about the sources of methodological variability and their interactions with the investigated phenomenon and participants’ age.

Another methodological contribution of this project was our investigation of how different infant-level inclusion criteria affect the magnitude of the obtained effect size. For our main analysis, we included all infants who completed at least one IDS and one ADS trial. This is something of a departure from the norm, as most participating labs reported using a stricter inclusion criterion in their own independent work. Our meta-analytic effect size was 0.35 when we included all infants with a minimum of two trials, grew to 0.42 when the minimum was raised to four trials, and increased further to 0.48 when the minimum was eight trials. Moreover, the effect of age on the amount of missing data was substantially larger in the eye-tracking paradigm compared with the other methods. The amount of missing data increased across the length of the experiment; this increase was numerically most prevalent for eye tracking. Setting stricter inclusion criteria necessarily decreases final sample size if the total sample tested remains constant, but at the same time, stricter criteria appear to lead to more robust effects in this paradigm.

**Challenges and limitations**

As with any study, the current experiment required specific methodological choices, several of which influence the generalizability of our results. Two aspects of the decision making regarding the stimuli in particular are worth further discussion. The first is the choice to use NAE (as opposed to, say, the native language or dialect for each infant group tested). This choice was based on the need to use consistent stimuli across laboratories, in order to limit cross-lab variation and ensure feasibility of the overall project, and to use stimuli from a language in which there was robust evidence of a strong IDS-preference effect, both in native and non-native settings. However, our design necessarily complicates the interpretability of our findings from laboratories outside of North America. They confound effects of native language or dialect (infants prefer listening to their native language) and true cultural variation in IDS preference. Further, substantial diversity in the non-NAE samples was obscured in our preregistered analyses. Together with the previously mentioned native-language follow-up study using the ManyBabies 1 protocol, further analyses of our data set focusing on specific subsamples with sufficient sample size (e.g., French, German, Dutch, British English) will shed additional light on how the differences between the North American and other infants in the current study should be interpreted.

The second challenging decision hinged on the elicitation of the IDS stimuli. Stimuli used in previous IDS-preference studies range from scripted speech with no infant present (e.g., Cooper & Aslin, 1990; Newman & Hussain, 2006), which maximizes control over the experimental stimuli, to more naturalistic samples collected in free-play, unscripted contexts (e.g., Hayashi et al., 2001; Werker et al., 1994), which maximize generalizability to real-world contexts. We opted for a relatively naturalistic approach, with an elicitation protocol using real mothers and their infants and centering around concrete objects. It is likely that this approach led to the observed reduction in the distinctiveness of the acoustic characteristics of our IDS samples, and it limited our ability to fully control the characteristics of the samples. Other aspects of our elicitation approach are important to keep in mind in interpreting our finding, including our developmental effects. In particular, our speech stimuli were elicited from mothers speaking to 4- to 8-month-old infants in the context of an objects-focused task (which was likely best suited to infants at the older range of our age bins). The extent to which these age-related characteristics of the IDS stimuli affected the magnitude of infants’ IDS preference across development merits further inquiry. Further, the use of multiple speakers in the stimuli may have increased the processing load for infants.

As the first collaboration of its kind, ManyBabies 1 revealed a number of important challenges in conducting multilab infancy research. As any researcher who has tested infant participants knows, data collection with infants is slow and labor intensive. In the current project, over a period of approximately 13 months, 69 labs were able to collect data from 2,845 infants. In contrast, in ManyLabs 1, a similar initiative with adult participants (Klein et al., 2014), data were collected from more than 6,000 participants tested in 36 labs over just a handful of months. Moreover, whereas adults can often be tested in multiple studies in a single session, this option is very limited for infants.

We expected challenges in implementing a standardized data-collection procedure across labs, but the depth of these challenges, and the diversity of methodological implementation across laboratories, was surprising. Laboratories that conduct infancy research are highly diverse in both the software and the hardware they have available to implement testing methods. We planned flexibility in the specific setup (eye
tracking, HPP, central fixation) given this known vari-
ability, but despite this flexibility, several labs were
directed to deviate from aspects of the protocol, for
equation, because of limitations in how stimuli could
be presented (e.g., the ability to implement infant-
controlled trial lengths, software settings for repeating
trials). One important conclusion from our work, as
evidenced in the walk-through videos laboratories pro-
vided to illustrate their protocols (see the next para-
graph), is that typical Method sections fail to provide
the detail necessary to capture the methodological
diversity in the ways that different labs implement a
paradigm with the same name.

Additional benefits of large-scale
collaboration

Although our primary goals were empirical, the
ManyBabies 1 project offers numerous additional ben-
efits to both individual researchers and the field at
large. All of the questionnaires, how-tos, and stimuli
(e.g., attention getters) used in the project are freely
available for reuse in future studies. Each participating
lab created a walk-through video that showed the lab
and study setup. These videos provide an unprece-
dented peek "behind the curtain" of other infancy
labs—something that previously was possible only
through visiting labs in person. Such information could
be a particularly helpful resource for investigators set-
ting up an infancy lab for the first time. It also provides
a unique data set whereby infancy researchers can
begin to understand the variety of lab setups and study
implementations.

This large-scale collaborative effort also has provided
broader benefits for the field. It created a strong col-
laborative network of infancy researchers. Informal
“ManyBabies” gatherings are now organized at devel-
opmental conferences, enabling researchers who have
previously collaborated only virtually to meet in person.
This project also was many participating researchers’
introduction to open-science practices and tools, such
as preregistration and OSF.

Finally, ManyBabies 1 has launched several "knock-
on" projects. For example, ManyBabies Bilingual (Byers-
Heinlein et al., 2020) is comparing bilingual infants’
preference for IDS with our results for monolingual
infants. Other projects will examine the test-retest reli-
ability of infants’ IDS preference, examine whether IDS
preference predicts vocabulary size at 18 and 24 months
(Soderstrom et al., 2020), and test whether lab-specific
variables affect infants’ performance and attrition. We
believe that these additional benefits are not unique to
infancy research, and that other scientific communities
embarking on large-scale collaborative projects will
gain similar benefits.

Conclusion

Replication research can go far beyond simply asking
whether an effect is present: It can allow for an assess-
ment of how an effect varies and how it changes over
development. We observed a robust and statistically
significant preference for IDS over ADS, confirming
previous observations in the literature. Yet the value
of our experiment lies not purely in this binary result—
or even in the quantitative estimate of the overall mag-
nitude of the IDS preference—but also in the further
theoretical and methodological opportunities that the
data afford. By measuring the relationship of IDS pref-
ence to age and language community, this experi-
ment provides a starting point for developing a more
nuanced theory of how IDS preference relates to chil-
dren’s language experiences. Further, by revealing the
substantial contributions of methodological decision
making to observed effect sizes, our study points the
way toward developing best-practices templates in fur-
ther infancy work of this kind. In sum, we hope our
work reported here illustrates the power of large-scale
collaboration for the study of developmental variation
and change.

Transparency

Action Editor: Daniel J. Simons
Editor: Daniel J. Simons
Consortium Members
Michael C. Frank, Stanford University
Katherine Jane Alcock, Lancaster University
Natalia Arias-Trejo, Universidad Nacional Autónoma de
México (UNAM)
Gisa Aschersleben, Saarland University
Dare Baldwin, University of Oregon
Stéphanie Barbu, Université de Rennes 1 and Centre
national de la recherche scientifique
Elika Bergelson, Duke University
Christina Bergmann, Max Planck Institute for Psycho-
linguistics, Nijmegen, The Netherlands
Alexis K. Black, Haskins Laboratories, New Haven,
Connecticut
Ryan Blything, University of Bristol
Maximilian P. Böhlender, Technische Universität Dresden
Petra Bolitho, Victoria University of Wellington
Arielle Borovsky, Purdue University
Shannon M. Brady, UCLA
Bettina Braun, University of Konstanz
Anna Brown, University of Liverpool
Krista Byers-Heinlein, Concordia University
Linda E. Campbell, University of Newcastle, Australia
Cara Cashon, University of Louisville
Mohinish Shukla, University of Massachusetts Boston
Elizabeth A. Simpson, University of Miami
Leher Singh, National University of Singapore
Barbora Skarabela, University of Edinburgh
Gaye Soley, Boğaziçi University
Megha Sundara, UCLA
Anna Theakston, University of Manchester
Abbie Thompson, University of Notre Dame
Laurel J. Trainor, McMaster University
Sandra E. Trehub, University of Toronto Mississauga
Anna S. Troan, University of Oslo
Angeline Sin-Mei Tsui, University of Ottawa
Katherine Twomey, University of Manchester
Katie Von Holzen, Université Paris Descartes
Yuanyuan Wang, The Ohio State University
Sandra Waxman, Northwestern University
Janet F. Werker, University of British Columbia
Stephanie Wermelinger, University of Zurich
Alix Woolard, University of Newcastle, Australia
Daniel Yurovsky, University of Chicago
Katharina Zahner, University of Konstanz
Martin Zettersten, University of Wisconsin–Madison
Melanie Soderstrom, University of Manitoba

**Author Contributions**


**Declaration of Conflicting Interests**

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

**Funding**

Data collection was supported by a grant from the Laura and John Arnold Foundation through the Association for Psychological Science. Individual participating labs further acknowledge funding support from the Natural Sciences and Engineering Research Council of Canada (Grants 12R81103, 2018-05823, and 402470-2011); the Social Sciences and Humanities Research Council of Canada (Insight Grant 12R20580); the United Kingdom’s Economic and Social Research Council (Grants ES/L008955/1 and ES/ N005635/1); the Agence Nationale de la Recherche (Grant ANR-17-EURE-0017); a European Research Council Synergy Grant (N005635/1); the Agence Nationale de la Recherche (Grant ANR-17-EURE-0017); a European Research Council Synergy Grant (SOMICS, Grant 609819); the Alvin V. Jr. and Nancy C. Baird Professorship; the Korean National Research Fund (Grant NRF-2016S1A2A2912606); the U.S. National Institutes of Health (Grants R03 HD079779 and R37 HD037466); Leibniz ScienceCampus Primate Cognition seed funds; The Science Academy, Turkey, Young Scientist Award Program (BAGEP); Research Manitoba, University of Manitoba; and Children’s Hospital Research Institute of Manitoba.
Open Practices
Open Data: https://osf.io/re95x/
Open Materials: https://osf.io/re95x/
Preregistration: https://osf.io/gf7vh

All data and materials have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/re95x/. The design and analysis plans were preregistered at the Open Science Framework and can be accessed at https://osf.io/gf7vh. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/25152456. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/25152456. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/25152456. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iD
Michael C. Frank https://orcid.org/0000-0002-7551-4378

Prior Versions
A previous version of this manuscript was posted at https://psyarxiv.com/s98ab/.

References


