Quantifying sources of variability in infancy research using the infant-directed speech preference

DOI:
10.1177/2515245919900809

Document Version
Accepted author manuscript

Citation for published version (APA):
The Many Babies Consortium (2020). Quantifying sources of variability in infancy research using the infant-directed speech preference. *Advances in Methods and Practices in Psychological Science* , *3*(1), 24-52. https://doi.org/10.1177/2515245919900809

Published in:
Advances in Methods and Practices in Psychological Science

Quantifying sources of variability in infancy research using the infant-directed speech preference

The ManyBabies Consortium[1]

[1] See author note

Author Note

The ManyBabies Consortium consists of 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 - CNRS), Elika Bergelson (Duke University), Christina Bergmann (Max Planck Insitute for Psycholinguistics), Alexis K. Black (Haskins Laboratories), Ryan Blything (University of Bristol), Maximilian P. Böhland (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), Mihye Choi (University of Massachusetts Boston), Joan Christodoulou (UCLA), Laura K. Cirelli (University of Toronto Mississauga), Stefania Conte (University of Milano-Bicocca), Sara Cordes (Boston College), Christopher Cox (University of York), Alejandrina Cristia (Laboratoire de Sciences Cognitives et Psycholinguistique, Dept d'Etudes Cognitives, ENS, PSL University, EHESS, CNRS), Rhodri Cusack (Trinity College Dublin), Catherine Davies (University of Leeds), Maartje de Klerk (Utrecht University), Laura de Ruiter (University of Manchester), Claire Delle Luche (University of Essex), Dhanya Dinakar (Western Sydney University), Kate C. Dixon (University of Louisville), Virginie Durier (Université de Rennes 1 - CNRS), Samantha Durrant (University of Liverpool), Christopher Fennell (University of Ottawa), Brock Ferguson (Strong Analytics), Alissa Ferry (University of Manchester), Paula Fikkert (Radboud University), Teresa Flanagan (Franklin & Marshall College), Caroline Floccia (University of Plymouth), Megan Foley (Florida State University), Tom Fritzsche (University of Potsdam), Rebecca L. A. Frost (Max Planck Institute for Psycholinguistics), Anja Gampe (University of Zurich), Judit Gervain (Université Paris Descartes), Nayeli Gonzalez-Gomez (Oxford Brookes University), Anna Gupta (Leiden University), Laura E. Hahn (Radboud University), J. Kiley Hamlin (University of British Columbia), Erin E. Hannon (University of Nevada, Las Vegas), Naomi

Havron (Laboratoire de Sciences Cognitives et Psycholinguistique, Dept d'Etudes Cognitives, ENS, PSL University, EHESS, CNRS), Jessica Hay (University of Tennessee, Knoxville), Mikołaj Hernik (Central European University), Barbara Höhle (University of Potsdam), Derek M. Houston (The Ohio State University), Lauren H. Howard (Franklin & Marshall College), Mitsuhiko Ishikawa (Kyoto University), Shoji Itakura (Kyoto University), Iain Jackson (University of Manchester), Krisztina V. Jakobsen (James Madison University), Marianna Jarto (University of Hamburg), Scott P. Johnson (UCLA), Caroline Junge (Utrecht University), Didar Karadag (Bogazici University), Natalia Kartushina (University of Oslo), Danielle J. Kellier (Stanford University), Tamar Keren-Portnoy (University of York), Kelsey Klassen (University of Manitoba), Melissa Kline (Massachusetts Institute of Technology), Eon-Suk Ko (Chosun University), Jonathan F. Kominsky (Harvard University), Jessica E. Kosie (University of Oregon), Haley E. Kragness (McMaster University), Andrea A.R. Krieger (Saarland University), Florian Krieger (University of Luxembourg), Jill Lany (University of Notre Dame), Roberto J. Lazo (University of Miami), Michelle Lee (University of California, San Diego), Chloé Leservoisier (Université de Rennes 1 - CNRS), Claartje Levelt (Leiden University), Casey Lew-Williams (Princeton University), Matthias Lippold (University of Goettingen), Ulf Liszkowski (University of Hamburg), Liquan Liu (Western Sydney University), Steven G. Luke (Brigham Young University), Rebecca A. Lundwall (Brigham Young University), Viola Macchi Cassia (University of Milano-Bicocca), Nivedita Mani (University of Goettingen), Caterina Marino (Université Paris Descartes), Alia Martin (Victoria University of Wellington), Meghan Mastroberardino (Concordia University), Victoria Mateu (UCLA), Julien Mayor (University of Oslo), Katharina Menn (Radboud University), Christine Michel (Max Planck Institute for Human Cognitive and Brain Sciences), Yusuke Moriguchi (Kyoto University), Benjamin Morris (University of Chicago), Karli M. Nave (University of Nevada, Las Vegas), Thierry Nazzi (Université Paris Descartes), Claire Noble (University of Liverpool), Miriam A Novack (Northwestern University), Nonah M. Olesen (University of Louisville), Adriel John Orena (McGill University), Mitsuhiko Ota

(University of Edinburgh), Robin Panneton (Virginia Tech), Sara Parvanezadeh Esfahani (University of Tennessee, Knoxville), Markus Paulus (Ludwig Maximilian University), Carolina Pletti (Ludwig Maximilian University), Linda Polka (McGill University), Christine Potter (Princeton University), Hugh Rabagliati (University of Edinburgh), Shruthilaya Ramachandran (National University of Singapore), Jennifer L. Rennels (University of Nevada, Las Vegas), Greg D. Reynolds (University of Tennessee, Knoxville), Kelly C. Roth (University of Tennessee, Knoxville), Charlotte Rothwell (Lancaster University), Doroteja Rubez (The Ohio State University), Yana Ryjova (University of Nevada, Las Vegas), Jenny Saffran (University of Wisconsin-Madison), Ayumi Sato (Shimane University), Sophie Savelkouls (Boston College), Adena Schachner (University of California, San Diego), Graham Schafer (University of Reading), Melanie S. Schreiner (University of Goettingen), Amanda Seidl (Purdue University), 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 (Bogazici 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. Trøan (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).

Correspondence concerning this article should be addressed to The ManyBabies Consortium, Department of Psychology, 450 Serra Mall, Stanford, CA 94305. E-mail: mcfrank@stanford.edu

Abstract

The field of psychology has become increasingly concerned with issues related to methodology and replicability. Infancy researchers face specific challenges related to replicability: high-powered studies are difficult to conduct, testing conditions vary across labs, and different labs have access to different infant populations, amongst other factors. Addressing these concerns, we report on a large-scale, multi-site study aimed at 1) assessing the overall replicability of a single theoretically-important phenomenon and 2) examining methodological, situational, 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 were created using semi-naturalistic laboratory-based audio recordings in North American English. Infants' relative preference for IDS and ADS was assessed across 67 laboratories in North America, Europe, Australia, and Asia using the three commonly-used infant discrimination methods (head-turn preference, central fixation, and eye tracking). The overall meta-analytic effect size (Cohen's d) was 0.35 [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 infant-directed speech 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

Word count: 11680

Quantifying sources of variability in infancy research using the infant-directed speech preference

The recent focus on power, replication, and replicability has had important consequences for many branches of psychology. Confidence in influential theories and classic psychological experiments has been shaken by demonstrations that much of the experimental literature is under-powered (Button et al., 2013), that surprisingly few empirical claims have been subject to direct replication (Makel, Plucker, & Hegarty, 2012), and that the direct replication attempts that do occur often fail to substantiate original findings (Open Science Collaboration, 2015). As disturbing as these demonstrations may be, they have already led to important positive consequences in psychology, encouraging scientific organizations, journals, and researchers to work to improve the transparency and replicability of psychological science.

To date, however, researchers in infancy have remained relatively silent on issues of replicability. This silence is not because infant research is immune from the issues raised. Indeed, the statistical power associated with infant psychology experiments 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 infant data – and developmental data 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 varies across individuals, ages, context, cultures, languages, and socioeconomic groups. While these challenges may make replicability in infancy research more 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, multi-lab, pre-registered infant study. This study was inspired by the ManyLabs studies (e.g., Klein et al., 2014), in which multiple laboratories attempt to replicate various social and cognitive psychology studies, and moderators of study replicability are assessed systematically across labs. Given the reasons discussed above, 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 and delineation of decisions concerning data collection and analysis across a large number of labs studying similar phenomena or using similar methods. For this first "ManyBabies" project, we selected a finding that the field has good reason to believe is robust – namely, infants' preference for infant-directed speech over adult-directed speech – 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 infant-directed speech in development, and then discuss our motivations and goals in studying a single developmental phenomenon at scale.

Infant-Directed Speech Preference

Infant-directed speech (IDS) is a descriptive term for the characteristic speech that caregivers in many cultures direct towards infants. Compared to adult-directed speech (ADS), IDS is often higher pitched, with greater pitch excursions, and shorter utterances, among other differences (Fernald et al., 1989). While 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 has observed infants' preference for IDS over ADS using a range of stimuli and procedures. 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 to ADS (Cooper & Aslin, 1994; Cooper, Abraham, Berman, & Staska, 1997; Fernald, 1985; Hayashi, Tamekawa, & Kiritani, 2001; Kitamura & Lam, 2009; Newman & Hussain, 2006; Pegg, Werker, & McLeod, 1992; Santesso, Schmidt, & Trainor, 2007; L. Singh, Morgan, & Best, 2002; Werker & McLeod, 1989). In a meta-analysis by Dunst, Gorman, and Hamby (2012), which included 34 experiments, the IDS preference typically had an effect size of Cohen's $d = 0.67$ [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 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 than with 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 to ADS. Naturalistic observations confirm that the amount of speech directed to US 18-month-olds (which likely bears IDS features), rather than the amount of overheard speech (which is likely predominantly ADS), relates to the efficiency of word processing and expressive vocabulary knowledge at 24 months (Weisleder & Fernald, 2013). Finally, infants show increased neural activity to familiar words in IDS compared to ADS, and also compared to 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 boost social preferences and perhaps attention in general (Schachner & Hannon, 2011), as it even has been reported to improve performance in non-linguistic associative learning (e.g., Kaplan, Jung, Ryther, & Zarlengo-Strouse, 1996).

The Current Study: Motivations and Goals

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

What is the magnitude of the IDS preference? First and foremost, our study serves as a large-scale, precise measurement of IDS preference across a large number of labs. Based on evidence summarized in a previous meta-analysis (Dunst et al., 2012), we expect that the preference will be non-zero and positive. We suspect, 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 is reason to believe that effect sizes in infancy research are often incorrectly reported; for example, partial eta-squared η_p^2 is often misreported as eta-squared η^2. This confusion is likely to inflate the practical significance of the findings, leading to an overestimation of the statistical magnitude and importance of effects (Mills-Smith, Spangler, Panneton, & Fritz, 2015). Therefore, the mean effect size of 0.67 reported by Dunst et al. (2012) is likely an overestimate of the real effect size.

How does IDS preference vary across age? 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(s), but with experience gain sensitivity to native language distinctions and lose sensitivity to non-native distinctions (Maurer & Werker, 2014). If preference for IDS follows a similar pattern, then we predict that older infants tested in their native language will show a stronger preference for IDS over ADS than infants tested in a non-native language.

Faced with several competing concerns, we made the decision that all infants in our study, regardless of native language, would be exposed to ADS and IDS stimuli in North American English (NAE). This design choice had several practical advantages. Most importantly, every infant was tested with the same stimulus set. Creating different stimulus sets in different languages would add 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 would reduce the time and cost of conducting the study.

There are both design-related advantages and drawbacks to this decision. 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 and/or may have a harder time recognizing IDS features as such when they 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 own age (McRoberts, McDonough, & Lakusta, 2009). Although this is a relevant concern, previous research has documented some IDS preference in the face of language and age mismatches (McRoberts et al., 2009; Werker, Pegg, & McLeod, 1994); and corpus studies suggest 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 does not allow us to disentangle the effects of stimulus language (native vs. non-native) from the effects of infants' cultural background, we can explore how aspects of these factors influence infants' preference for IDS.

After weighing these considerations, we adopted NAE stimuli to provide the maximal chance of recovering a positive effect, ensure that stimuli are not a source of variance across labs, allow comparability with previous work, and also minimize the barriers to entry (i.e., the need to create lab-specific stimuli) for each participating lab. 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 analyzed IDS preferences in bilingual children; Byers-Heinlein et al., accepted pending data collection).

We focused here on three primary methods: 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 infants' looking to an unrelated visual stimulus. In the single screen central fixation method, infants were shown an uninformative image (a checkerboard) on a single, centrally-located monitor, while listening to either IDS or ADS, and looking time to the monitor was manually coded via a closed-circuit video camera. In the eye tracking method, infants saw a similar display, but looking times were measured automatically via a remote corneal-reflection eye tracker. In 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 while IDS or ADS played was manually coded via a closed-circuit video camera (Nelson et al., 1995).

Each lab tested the same phenomenon, using the same stimuli and the same general experimental parameters (including, e.g., trial order, maximum trial length), varying only in the method of measuring preference. We thus can analyze whether this theoretically irrelevant methodological choice influences effect size, helping to 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 been tested on additional studies; measurements from these infants can inform future lab administration practices.

What should our expectations be regarding replicability and statistical power in studies of infancy? Although we are only replicating a single phenomenon, the importance and assumed robustness of the IDS preference means that our study still provides data relevant to developing a more nuanced understanding of replicability and power in infancy research. Because of the large number of participating labs, data from some labs does not support an IDS preference (i.e., yields a small – or even negative – effect size when analyzed individually). Some variability is expected due to the mathematics of estimating an effect at so many independent sites. Nonetheless, we inspect whether there is systematic variability explained by lab effects.

In addition, by providing an unbiased estimate of effect size for an important developmental phenomenon (including estimates of how that effect varies across ages, language backgrounds, and tasks), this work gives a rough baseline for other scientists to use when planning studies. Existing attempts to estimate the statistical power of infant experiments have been contaminated by publication bias, which leads to an overestimation of typical effect sizes in infant research. Such overestimates can lead subsequent studies to be under-powered (expecting to see larger effects than are truly present). Though our report estimates the effect for a particular developmental preference, we can compare our unbiased estimate, calculated both across all three methods and for each method, to the meta-analytic effect extracted from previously published studies. This calculation can provide a rough estimate of the effect size inflation in general, and for each method in particular, at least for this particular phenomenon.

How should we think about the relationships between experimental design, statistical significance, and developmental change? Previous work often employs a contrast between two ages to suggest that a developmental change has taken place; for example, by showing that 7-month-old infants show 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 non-sigificant difference) is not sufficient to show a difference between two time points (Nieuwenhuis, Forstmann, & Wagenmakers, 2011). Even in the case where a significant difference is found between the two age groups, such a result is not sufficient to elucidate the developmental pattern underlying this discrete test. By measuring how effect sizes change over age with a much denser sampling approach, our data and continuous analytic approach illustrate what stands to be gained with a more gradient approach to testing behavior over development.

Summary

This broad replication of IDS preferences helps to answer basic questions about the replicability of developmental psychology findings and will also provide useful benchmarks for how to design infant cognition studies going forward. Just as projects such as ManyLabs have led to important improvements in research practices in cognitive and social psychology, we hope that ManyBabies will play a similar role for developmental cognitive science.

Methods

Participation Details

Time frame. We issued an open call for labs to participate on February 2nd, 2017. Data collection began on May 1st, 2017. Data collection was scheduled to end on April 30th, 2018 (one 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 15th, 2018. Data collection from one laboratory extended beyond this timeframe (see below in Methods Addendum).

Age distribution. Each participating lab was asked to recruit participants in one or more of four age bins: 3;0 - 6;0, 6;1 - 9;0, 9;1 - 12;0, and/or 12;1 - 15;0 months. Each lab was tasked with ensuring that, for each age bin they contributed, the mean age fell close to the middle of the range and the sample was distributed across the bin. We selected three-month bins as a compromise, on the assumption that tighter bins would make recruitment more difficult while 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). This flexibility was necessary because labs differ in their ability to recruit infants of different ages.

Lab participation criterion. 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 value was Cohen's $d = .72$. In a paired t-test, 95% power to detect this effect requires 27 participants, and 80% power requires 17. On the basis of these calculations, we asked participating labs to commit to samples with a minimum of $N = 32$ in a single age group. However, given that for many of our analyses, power across labs is more critical than within a lab (Judd, Westfall, & Kenny, 2017), we allowed labs to contribute a "half sample" of $N = 16$, 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 labs should recruit with respect to the desired demographic characteristics of the study (e.g., full-term infants; see below for full list of exclusion criteria). Given this recruitment strategy, however, we asked that sample Ns be calculated on the basis of the number of total infants tested, not the 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 they were able to achieve the minimum N required for a half-sample in their age bin ($N = 16$) by the end date of testing and if, after exclusions, they contributed 10 or more data points. 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 lab internal decision-making regarding how many infants to recruit/when to stop recruitment.

Participants

Our final sample was comprised of 2329 monolingual infants from 67 labs (mean sample size per lab: 34.76, $SD = 20.33$, range: 10 – 93; 45 contributed data at multiple ages). Demographic exclusions were primarily implemented during recruitment; despite this, additional infants were tested and excluded based on preset criteria (see Exclusions below for percentages). 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. Information about all 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). Many labs collected data in more than one bin. Of the total sample, 1066 infants (from 30 labs) were acquiring NAE, and 1263 infants (from 37 labs) were acquiring a language other than NAE. As discussed above, a separate sample of bilingual children was tested in a parallel investigation, but these data are not reported in the current manuscript.

Table 1

Statistics of the included labs. N refers to the number of infants included in the final analysis. English from the US and Canada are both treated as North American English.

lab	Mean age (days)	*N*	Method	Language	Country
babylabbrookes	255	53	central fixation	English	UK
babylabvuw	224	15	central fixation	English	Australia
babylabyork	268	32	central fixation	English	UK
baldwinlabuoregon	320	16	central fixation	English	US
bchdosu	269	67	central fixation	English	US
bcrlunlv	411	29	central fixation	English	US
bounbcl	411	31	central fixation	Turkish	Turkey
icclbc	222	15	central fixation	English	US
infantcoglablouisville	325	35	central fixation	English	US
ldlottawa	276	59	central fixation	English	Canada
madlabucsd	234	10	central fixation	English	US

minddevlabbicocca	158	15	central fixation	Italian	Italy
udssaarland	332	43	central fixation	German	Germany
unlvmusiclab	138	20	central fixation	English	US
weescienceedinburgh	213	32	central fixation	English	UK
wsigoettingen	274	88	central fixation	German	Germany
infantcogubc	165	39	central fixation, eye tracking	English	Canada
lancaster	326	42	central fixation, eye tracking	English	UK
babylablangessex	289	27	eye tracking	English	UK
babylablmu	368	62	eye tracking	German	Germany
babylabshimane	195	28	eye tracking	Japanese	Japan
babylabuclajohnson	408	22	eye tracking	English	US
babylabumassb	308	30	eye tracking	English	US
babylingoslo	227	31	eye tracking	Norwegian	Norway
callab	369	30	eye tracking	English	US
cdcceu	272	27	eye tracking	Hungarian	Hungary
cfnuofn	298	15	eye tracking	English	Australia
childlabmanchester	269	26	eye tracking	English	UK
cogdevlabbyu	161	29	eye tracking	English	US
dcnlabtennessee	345	19	eye tracking	English	US
earlysocogfm	310	35	eye tracking	English	US
escompicbsleipzig	159	14	eye tracking	German	Germany
ethosrennes	187	90	eye tracking	French	France
irlconcordia	310	37	eye tracking	English	Canada
jmucdl	340	17	eye tracking	English	US
kokuhamburg	305	25	eye tracking	German	Germany
kyotobabylab	281	30	eye tracking	Japanese	Japan
labunam	302	36	eye tracking	Spanish	Mexico
lcdfsu	354	23	eye tracking	English	US
lcduleeds	413	14	eye tracking	English	UK
lllliv	302	36	eye tracking	English	UK
lscppsl	404	14	eye tracking	French	France
pocdnorthwestern	409	30	eye tracking	English	US
socialcogumiami	131	19	eye tracking	English	US

weltentdeckerzurich	414	30	eye tracking	German	Switzerland
nusinfantlanguagecentre	337	21	eye tracking, central fixation	Mandarin	Singapore
babylabkingswood	312	32	HPP	English	Australia
babylabkonstanz	235	15	HPP	German	Germany
babylableiden	319	15	HPP	Dutch	Netherlands
babylabnijmegen	279	49	HPP	Dutch	Netherlands
babylabparisdescartes1	403	16	HPP	French	France
babylabplymouth	332	34	HPP	English	UK
babylabprinceton	307	24	HPP	English	US
babylabutrecht	276	61	HPP	Dutch	Netherlands
bllumanitoba	281	79	HPP	English	Canada
chosunbaby	313	77	HPP	Korean	Korea
infantlanglabutk	323	65	HPP	English	US
infantllmadison	316	93	HPP	English	US
infantstudiesubc	228	20	HPP	English	Canada
islnotredame	411	28	HPP	English	US
isplabmcgill	411	11	HPP	French	Canada
langlabucla	250	63	HPP	English	US
lppparisdescartes2	241	30	HPP	French	France
musdevutm	229	31	HPP	English	Canada
purdueinfantspeech	355	58	HPP	English	US
trainorlab	241	24	HPP	English	Canada
babylabpotsdam	306	46	HPP, central fixation	German	Germany

Materials

Visual stimuli. For labs using central fixation or eye tracking methods, a brightly colored static checkerboard was used as the fixation stimulus, and a small engaging video (an animation of colorful rings decreasing in size) as an attention-getter. For labs using HPP, we asked labs to use their typical visual stimulus, which varied considerably across laboratories. Some labs used flashing lights as the visual fixation stimulus (the original protocol that was developed in the 1980s), while 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 speaking to their infants. To do so, we gathered a set of recordings of mothers speaking to their infants and to experimenters, selected a subset of individual utterances from these (see below), and then constructed stimulus items from this subset. All other characteristics of the recordings besides register (IDS vs. ADS) were as balanced as possible across clips. Based on our intuitions and the data from the norming ratings described below, we consider these stimuli to be representative of naturally produced IDS and ADS across middle- and high-SES mothers in North America. Although future studies could attempt to 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 here. 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 – 250 days. The same recording procedures were followed in both laboratories. Recordings were collected in an infant-friendly greeting area/testing room using a simple lapel clip-on microphone connected to a smartphone (iPhone 5s or 6s), with the "Voice Record" or "Voice Record Pro" apps (Dayana Networks Ltd.) in the Canadian lab, and the "Voice Memos" app (Apple Inc.) in the US 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 mothers used consistent labels, a small sticker was affixed to each object showing its name.

Each object was taken out of the bag one at a time and the mother was asked to talk about the object, either to her baby (for the IDS samples) 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. Order of IDS and ADS recording was counterbalanced across participants. A total of 11 mothers were recorded in Canada and four in the United States.

There were a total of 179 unedited minutes of recording from Canada and 44 from the United States. A first-pass selection of low-noise IDS and ADS samples yielded 1281 utterances, for a total of 4479 s. From this first pass, 238 utterances were selected that were considered to be the best examples of IDS and ADS and met other basic stimulus selection criteria (e.g., 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 about each variable so as to identify utterances that were clearly anomalous in some respect and exclude them. In each case, a set of naïve, North American English-speaking adults recruited from Amazon Mechanical Turk (MTurk) listened to all 238 of the utterances and rated them on a 7-point Likert scale. Raters were assigned randomly to one of the five variables, with the number of participants assigned to a particular rating task ranging between eight and 18 due to variability in random assignment. Affect and IDS ratings were made using low-pass filtered recordings (a 120-Hz filter with standard rolloff was applied twice using the `sox` software package). These ratings were intended to give us a principled basis on which to exclude clips that were outliers on particular dimensions (such as having odd affect or background noise). In general, with the exception of IDS-ness, ratings were not highly variable across clips (the largest *SD* was .85, for noise ratings).

Ratings from the tasks were then used to produce a set of utterances such that accent was rated similar to "standard English" (ratings < 3, with 1 being completely standard), naturalness was rated high (> 4, with 7 being completely natural), noisiness was rated low (< 4, with 1 being noiseless), and IDS and ADS clips were consistently distinguished (with IDS having ratings > 4 and ADS having ratings < 4, with 7 being clearly directed at a baby or child). This procedure resulted in 163 total 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. To do so, we assembled utterances from our filtered set. All clips were root mean square amplitude-normalized to 70 dB sound pressure level (SPL) before assembly, and then the final stimuli were amplitude-renormalized to 70 dB SPL. We assembled the final stimuli considering the following issues:

- *Identity.* Audio stimuli were constructed using clips from more than one mother. The number of different mothers included in a given stimulus was matched across 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 presence of object labels in the clips across 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 declaratives over questions in the IDS sample. The final stimulus set contained 47% questions in the IDS samples and 3% questions in the ADS samples. We felt that retaining this naturally-occurring difference in 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 on duration or number of clips. Because there were more clips per stimulus in the IDS samples, there were also more utterances boundaries. This property is consistent with 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 > 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 F0 extraction. For Figure 1, F0 values for voiced portions of the stimuli were collapsed into a series of logarithmically-spaced bins spanning the algorithm's F0 search range of 32-650 Hz.

Table 3 provides a comparison of our stimuli to a sample of others that have been used previously in the IDS preference literature. Across studies, the only statistic that was reported reliably across papers was the mean pitch (F0) for IDS and ADS and even this one was only reported in about half the studies we sampled. Various measures of variability were reported in some studies (e.g., range within each sample, range across samples, standard deviation), but due to variation in the length and number of different samples used in each study, and a lack of systematicity in reporting, it was difficult to compare directly. Numerically, the average IDS/ADS pitch difference in our materials was less extreme than that found in previous studies.

To confirm that our composite IDS and ADS stimuli were rated as natural and that the more limited pitch difference between registers still led to the stimuli being categorized

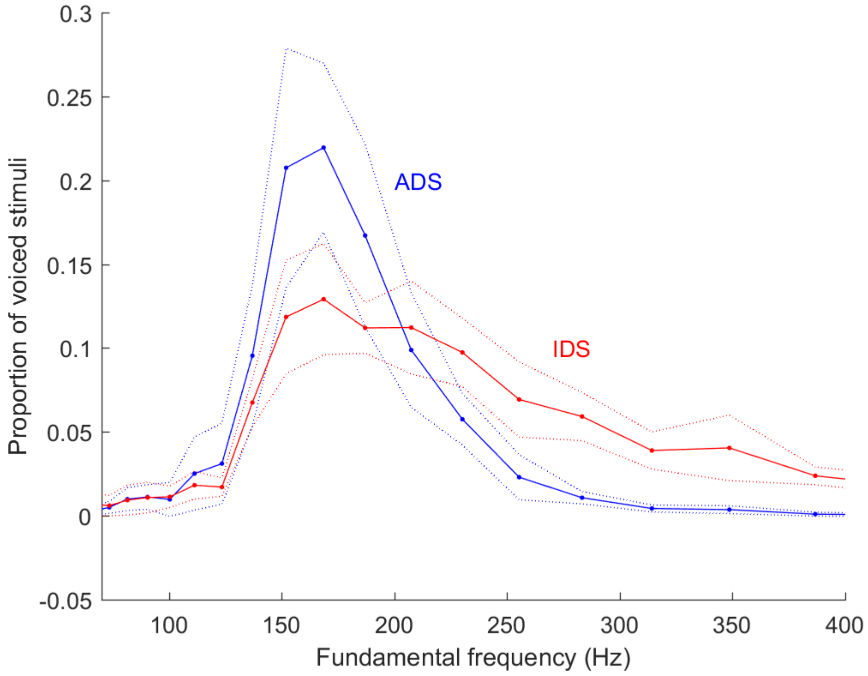

Figure 1. The distribution of F0 values for IDS and ADS is displayed as the proportion of voiced segments that fell in each F0 bin. Dashed lines show mean plus or minus one standard error across stimuli.

differently, we conducted another norming study. Using the same basic paradigm as above, we collected a new sample of judgments from MTurk participants. Raters were randomly assigned to listen to all 16 stimuli and judge either whether they were directed at infants/children or adults ($N = 22$) or else whether the stimuli sounded natural ($N = 27$). All IDS clips were judged extremely likely to be directed at infants or children ($M = 6.74$, $SD = .09$, on a $1 - 7$ rating scale), while all ADS clips were judged highly likely to be directed to adults ($M = 2.12$, $SD = .38$). Both were judged to be relatively natural, with the ADS, if anything, slightly more natural ($M = 5.18$, $SD = .19$) than the IDS ($M = 4.47$, $SD = .31$). In sum, because our stimuli were created from naturalistic productions from a wide range of mothers, they were less extreme in their intonation, but they were judged as natural and were easily identified as infant-directed.

Table 2

Characteristics of the IDS and ADS stimuli, with standard deviations computed across stimuli.

Measurement	IDS Mean	IDS *SD*	ADS Mean	ADS *SD*
Number of mothers speaking per stimulus	4.00	0.00	3.75	0.46
Number of clips per stimulus	6.88	1.13	4.50	0.76
Number of objects mentioned per stimulus	2.75	0.71	2.75	0.71
Mean F0 (Hz) per stimulus	206.90	19.50	174.90	13.20
10th percentile F0 (Hz) per stimulus	131.40	26.10	139.00	17.70
90th percentile F0 (Hz) per stimulus	340.00	21.50	232.00	13.80
Mean number of utterances per stimulus	7.75	1.04	6.63	0.92
Mean duration (sec) of utterances	1.58	0.74	2.12	1.41
Mean inter-utterance interval (sec)	0.75	0.30	0.59	0.33

Table 3

Comparison of our study's stimuli to those of previous studies on infant-directed speech preferences.

Study	Mean Ages (Months)	Context of Recording	Quantity of Stimuli	Mean IDS F0 (Hz)	Mean ADS F0 (Hz)	IDS-ADS (Hz)	IDS/ADS
Present Study	3 – 15	semi-structured, 4-8 month old child present	8 full trial lengths' worth for each type	206.90	174.90	32.00	1.18
Cooper & Aslin (1990)	0, 1	read speech, no infant present	4 sentences produced in each type	315.88	259.58	56.30	1.22
Newman & Hussain (2006)	4.5, 9, 13	read speech, no infant present	4 passages produced in each type	225.70	189.65	36.05	1.19
Thiessen et al. (2005)	7	nonsense strings of syllables, no infant present	12 sentences in each style	292.00	230.00	62.00	1.27
Cooper et al. (1997)	1, 4	naturalistic speech to own infants	20s of each style	219.30	184.30	35.00	1.19
Schachner & Hannon (2011)	5	elicited speech, with speaker looking at a picture	1 min long videos, 2 in each style	273.00	224.70	48.30	1.21

Procedure

Basic Procedure. Each lab used the testing paradigm(s) with which they were most familiar, among variants of three widely-used measurement methods: 20 laboratories used the HPP, 16 used the single-screen central visual-fixation preference procedure (CF), and 27 used single-screen central visual fixation with fixations recorded by a corneal-reflection eye tracker (ET); four labs contributed data using two different methods. All procedural instructions to participant labs can be found at 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 due to variation in the software and procedures used in each laboratory and were documented for future analysis.

1st vs. 2nd 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 1st (and possibly only) or 2nd 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. Upon fixation, this event was followed by a visual stimulus (a checkerboard for CF and ET, a light or a similar video for HPP). The stimulus appeared to the left or right of the infant in HPP setups and in the center in CF and ET setups.

Trials. At the beginning of the session, there were two warm-up trials that familiarized infants with the general procedure. The auditory stimulus for warm-up trials was an 18-second clip of piano music, and the visual stimulus was identical to the test trials.

These trials familiarized infants to the general experimental setup and highlighted the contingency between looking at the visual display and the onset of the auditory stimulus. We did not analyze data from these trials. Training trials were then followed by up to 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 used during analysis for inclusion of a trial. The 2-s minimum trial time was chosen after discussion across laboratories regarding typical standards of practice on minimum trial length, which varied considerably across laboratories. This criterion was selected to ensure that the infant 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, corresponding to the duration of each sound file. For labs whose software could implement infant-controlled trial lengths, the 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 pseudo-random trial orders were created. Each order contained four blocks, with each block containing two IDS and two ADS trials in alternating order. Two blocks in each order began with IDS and the other two began with ADS. To facilitate analyses of preference scores by item, the same IDS and ADS stimuli were always paired with one another.

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 dB SPL level with and without a white noise reference audio clip playing, though not all contributing labs reported these measurements ($N = 47$). 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. We created a custom blend of instrumental music and a pastiche of stimulus materials triggered at random times and with random amplitude (available as part of the study materials). This masking stimulus was played to the caregiver over noise-attenuating headphones, to mask the IDS/ADS 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 infant eye gaze: online coding by an experimenter via button press during the experimental session, offline coding of a video after the experimental session, or automatic coding collected by an eye tracker. In the case that we received online and offline coding data, we used the offline coding.

Minimizing experimenter bias. Experimenters making online coding decisions (in CF and HPP methods) were blind to the particular stimulus presented during testing trials, as they were either located in a different room from the infant, or were in the same room but were wearing noise-attenuating headphones and hearing the same masking stimuli as the infant's caregiver. Offline coding was conducted without direct access to the auditory stimuli.

Demographics. All labs were instructed to collect a set of basic participant demographic information: sex, date of birth, estimated proportion language exposure for the language(s) that they hear in their daily life, race/ethnicity (using categories appropriate for the cultural and geographic context), preterm/fullterm status, history of ear infections, known hearing or visual impairments, and known developmental concerns (e.g., developmental disorders). Parents were also asked to report information about themselves (gender, level of education, and native language/languages) and the child's siblings (sex/gender and date of birth). A standard recommended participant questionnaire was distributed to participating labs as part of the instructions, although labs were permitted to use their own forms as long as they gathered the necessary information. In addition, a subset of participating laboratories provided extensive additional information about infants and testing circumstances (not analyzed here), for use in planned followup projects.

General Lab Practices

Training of research assistants. Each lab was responsible for maintaining good experimenter training practices, and was expected to use the same rigor with the ManyBabies study as with any other study in their laboratory. Laboratories reported on which research assistant ran each infant using pseudonyms or numerical codes. Each laboratory completed a questionnaire regarding their training practices, the experience and academic status of each experimenter, and their basic participant greeting practices.

Reporting of technology mishaps and infant/parent behavior. Laboratories were asked to note relevant concerns, anomalies and comments according to their standard lab practices and these were provided along with the looking time data and converted to a standardized form during the central analysis. Examples of relevant concerns included the infant crying during testing, parents intervening in a way that would affect their infant's looking behavior (e.g., talking or pointing), or technical problems that prevented the normal presentation of experimental stimuli.

Videos

All laboratories provided a "walk-through" video that detailed their basic processes including greeting, consent and data collection and showing the physical characteristics of their laboratory. (In our preregistration we stated that further procedural documentation would be available, but standardized reporting for procedural decision-making proved difficult to develop and deploy.) In addition, we strongly encouraged laboratories to collect and share video recordings of their data collection according to what was permissible given their ethics approval and participant consent. If labs could not provide participant videos, they were asked to provide a video showing a run-through of their procedure and/or pictures and information regarding the study setup. A number of laboratories contributed these video recordings to Databrary, where they can be found by searching for “ManyBabies 1.”

Exclusion Criteria

All data collected for the study (i.e., every infant for whom a data file was generated, regardless of how many trials were completed) were given to the analysis team for confirmatory analyses. Participants were only included in analysis if they met all of the criteria below. All exclusion rules are applied sequentially, and percentages reflect this sequential application to an initial sample prior to exclusions of 2754. N.B.: the first three criteria preemptively prevent participation (except in case of erroneously running the experiment with children outside of the inclusion guidelines).

- *Monolingual.* Monolingual infants of any language background were included in the sample. Monolingual was defined as 90% parent-reported exposure to the native language. This cutoff score struck a balance between including most infants who are typically considered monolingual in infant language studies, while excluding those who might be considered bilingual (Byers-Heinlein, 2015). 162 (5.88%) infants were tested but did not meet this criterion.

- *Full-term.* We defined full term as gestation times greater than or equal to 37 weeks. Of the remaining sample, 62 (2.39%) infants were tested but 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. Due to concerns about the accuracy of parent reports, we did not exclude infants based on parent-reported ear infections unless parents reported medically-confirmed hearing loss.

- *Contributed usable data.* A child must have contributed non-zero looking time on a pair of test trials (i.e., one trial each of IDS and ADS from a particular stimulus pair), after trial-level exclusions were applied, to be included in the study. Of the remaining sample, 41 (1.65%) infants were tested but did not meet these criteria. 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 we were able to include in the initial analysis, and our paradigm was, by design, less customized for any particular age group (and hence likely to produce greater data loss, especially for older children, who tend to habituate more quickly). In the exploratory analyses below, we consider how exclusion decisions affected our effect size estimates.

After these exclusions were applied, participants could also be excluded for analysis based on session-level errors, including: equipment error (e.g., no sound or visuals on the first pair of trials), experimenter error (e.g., an experimenter was unblinded in setups where infant looking was measured by live button press), or evidence of consistent parent/outside interference noted by participating labs (e.g., talking or pointing by parents, construction noise, sibling pounding on door). 78 (3.18%) infants for whom we had other reported data were dropped from analysis due to session-level error. This number is likely an underestimate, however. 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.

We further excluded individual trials that were reported as having issues (e.g., fussiness, incorrect stimulus, single instance of parent or sibling interference). A total of 4471 (10.61%) trials were affected by such errors. As with session level errors, classification of these was inconsistent across participating labs, but the most common source of trial-level errors was infant fussiness.

Based on our trial-length minimum, we also excluded 6027 (16.13%) trials with total looking times shorter than 2 s. These trials are analyzed as "missing" in our planned analysis below.

As discussed above, 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 10 or more data points. 11 (0.47%) infants from 2 labs were not included in the final sample because of this criterion.

Post-Data Collection Methods Addendum

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

These protocol deviations include the following:

- Before labs had commenced data collection, we altered our attention-getter stimulus to be a precessing annulus accompanied by chimes (to address the concern that a laughing baby might be more associated with infant-directed speech); some labs used the old stimulus.

- Variation in trial length beyond the assumed maximum of 18 s emerged due to deviations in lab's protocols for a variety of reasons. In all cases, looking times on these trials were truncated to 18 s.
- A number of labs provided data from infants that were within the 3–15 month age range, but outside of the submitting lab's pre-registered age bin. These infants were included in the analyses.
- Many labs deviated from their pre-registered sample size due to 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 as session-level errors for reasons other than equipment error, experimenter error or outside interference.

This last point bears further discussion. Some labs marked participants as exclusions at the participant level for trial-level errors (e.g. infant fussy, parental interference), even though there was sufficient trial-level data available for analysis. Similarly, individual trials were sometimes marked as errors for reasons related to participant-level issues. All trial-level and participant-level errors were reviewed centrally by at least two coders using all available information in the spreadsheet to determine whether a trial-level or participant-level error was appropriate. Specific information about each trial or participant error coding that was changed during this process can be found by reviewing metadata within the data analysis codebase.

In total, 313 participants from 50 labs previously marked as participant-level exclusions were retained for further processing and analysis. Participants originally coded as having session-level errors were recoded for the following reasons: when the participant-level exclusion was based solely on the existence of trial-level errors (190 infants), when exclusion was based on a different exclusion criterion (e.g., participants were out of the age range or were preterm) (93 infants), or if an issue identified by the lab at the participant level was deemed acceptable by the central analysis team (e.g., if a lab implemented a slightly different look-away criterion, see below) (30 infants). Note that many of the retained participants were subsequently excluded at other points in the analysis pipeline because, although they did not meet the criteria for session-level errors, they did meet the conditions for other exclusion criteria (e.g., participants did not contribute enough useable trials or were excluded based on language exposure).

In addition to recoding session-level errors, we also corrected the coding of trial-level errors where appropriate. 778 total trial-level errors from 62 participants in 16 different labs were recoded. The majority of trials were corrected when labs coded a participant-level error (e.g. age exclusion) on the trial level (584 trials) or coded a trial-level error on the participant level (e.g., if labs marked a participant as a session-level error for fussiness on a specific trial, but did not code the affected trials as errors) (133 trials). Other trials were corrected when subsequent investigation of lab notes and discussion with lab members revealed that the original trial-level error code needed to be changed (61 trials).

In addition, a variety of errors were found (e.g., pilot participants not properly excluded but noted in the comments) and fixed within the spreadsheets. Video data were not reviewed centrally, although in some cases where a question arose, the laboratory reviewed their own video in-house in order to respond. The entire process has been carefully documented and can be accessed upon request, but because in some cases this included identifiable information about participants, it is not possible to share it publicly.

Other reported protocol deviations included: No preregistration form submitted (1 lab); trial look-away time set to 3 s for some participants (1 lab); lab temporarily moved location during data collection (1 lab); minor protocol technical changes after start of data collection (2 labs); alternated left-right presentation and tested skin conduction during procedure (1 lab); procedural differences related to high-chair usage (1 lab); attention-getter deviation (4 labs); use of a pinwheel rather than checkerboard as the main visual fixation stimulus in HPP (1 lab).

We also detected a large number of data submission errors (typographical or otherwise) as a result of the comprehensive checking process in analysis. These were resolved when necessary by contacting the original lab. In general, we were inclusive of data with minor protocol deviations, and erred on the side of excluding data, when necessary, at the trial rather than participant level. A few demographic variables required greater central scrutiny than originally anticipated. Most notably, there was considerable variability in the interpretation of preterm and bilingual designations (despite centrally-dictated standards). When necessary, we recoded lab data so as to conform to the original protocol definitions.

There was an ambiguity in our lab-level exclusion criteria between whether labs would be included if they contributed 10 or more datapoints, or more than 10 datapoints. We chose the more liberal of these two criteria.

Finally, two labs submitted data after the deadline. In one case this was due to a communication error; in the other case, the lab continued data collection, resulting in 8 additional infants being tested. Both datasets are included in the final analysis here.

Results

Confirmatory Analyses

Data processing and analytic framework. All planned analyses were pre-registered in our initial registered report submission (available at https://osf.io/vd789/). Our primary dependent variable of interest was looking time (LT). Looking time was defined as time spent fixating the screen (for central fixation and eye tracking methods, and some HPP set-ups) or light (HPP) during test trials; LT scores did not count any time spent looking away from the screen, even if looks away were below the threshold for terminating a trial. Since looking times are non-normally distributed, following Csibra, Hernik, Mascaro, Tatone, and Lengyel (2016), we log-transformed all looking times prior to statistical analysis (we refer to this transformed variable as "log LT").

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 result of this analysis 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 that have been estimated in previous meta-analyses. However, the standard random-effects meta-analytic model is designed for a case where 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 across all labs allows us to use mixed-effects models, which account for the nesting and crossing of random effects (e.g., subjects nested within labs, items crossed across labs), and can provide more accurate estimates of the main effect and moderators. Both analyses were therefore included to allow for the most comprehensive understanding of the variance in the data.

Our meta-analyses were conducted as follows. The datasets provided by each lab were considered as separate "studies." For each lab's dataset, we first computed individual infants' IDS preference by 1) subtracting looking times to each IDS trial from its paired ADS trial (excluding trial pairs with missing data) and 2) computing a mean difference score (across trial pairs). Then we computed a group IDS preference for each lab and infant age group using dz, 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 REML with the `metafor` package (Viechtbauer, 2010).

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

Our mixed effects models, fit to the entire dataset collected from the 67 labs, were specified as:

$$DV \sim IV_1 + IV_2 + ... + (...|\text{subject}) + (...|\text{item}) + (...|\text{lab})$$

The goal of this framework was to examine effects of the independent variables (notated 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") based on relevant grouping units (subjects, items, and labs). The use of mixed-effects models also allowed us to move away from using difference scores as the dependent variable of interest. While difference scores simplify the process of calculating effect sizes for the meta-regression, their use requires that trials be paired, so some collected data (i.e., unpaired trials) cannot be analyzed. In the mixed effects framework, in contrast, looking time on individual trials is the dependent measure, ensuring that all trials can be included.

In our mixed-effects models, we planned a maximal random effects structure (Barr, Levy, Scheepers, & Tily, 2013), which entails specifying all random effects that are appropriate for the experimental design (e.g., IDS/ADS trial type can be nested within subjects – since each infant heard stimuli in both conditions — but cannot be nested within items since each item is 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 removing random slopes nested within subjects and then labs. We then removed random intercepts from groupings in the same order, retaining effects of trial type until last since these were of greatest theoretical interest. We fit all models using the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015) and computed *p* values using the `lmerTest` package (Kuznetsova, Brockhoff, & Christensen, 2017).

IDS preference. What was the overall magnitude of the IDS preference we observed? This question is answered within the cross-lab meta-analysis by fitting the main effect model specified by $dz \sim 1$ to the 108 separate group means and variances (after aggregating by lab and age group). The mean effect size estimate was 0.35 (CI = [0.29 - 0.42], $z = 10.67$, $p < .001$). A forest plot for this meta-analysis is shown in Figure 2. Further, 1373/2329 infants (58.95%) showed a numerical preference for IDS.

Independent relationship of IDS preference to moderating variables. We next fit 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 contributed sample as the moderator value. Labs that contributed samples from two age bins had values added separately for each age (because of the small number of these, we did not model this dependency between labs). For ease of interpretation, we centered age in this analysis. The age-moderated model, $dz \sim 1 + \text{age}$, yielded an estimated main effect of 0.35 (CI = [0.29 - 0.41], $z = 11.47$, $p < .001$) and an age effect of 0.05 (CI = [0.03 - 0.07], $z = 4.89$, $p < .001$).

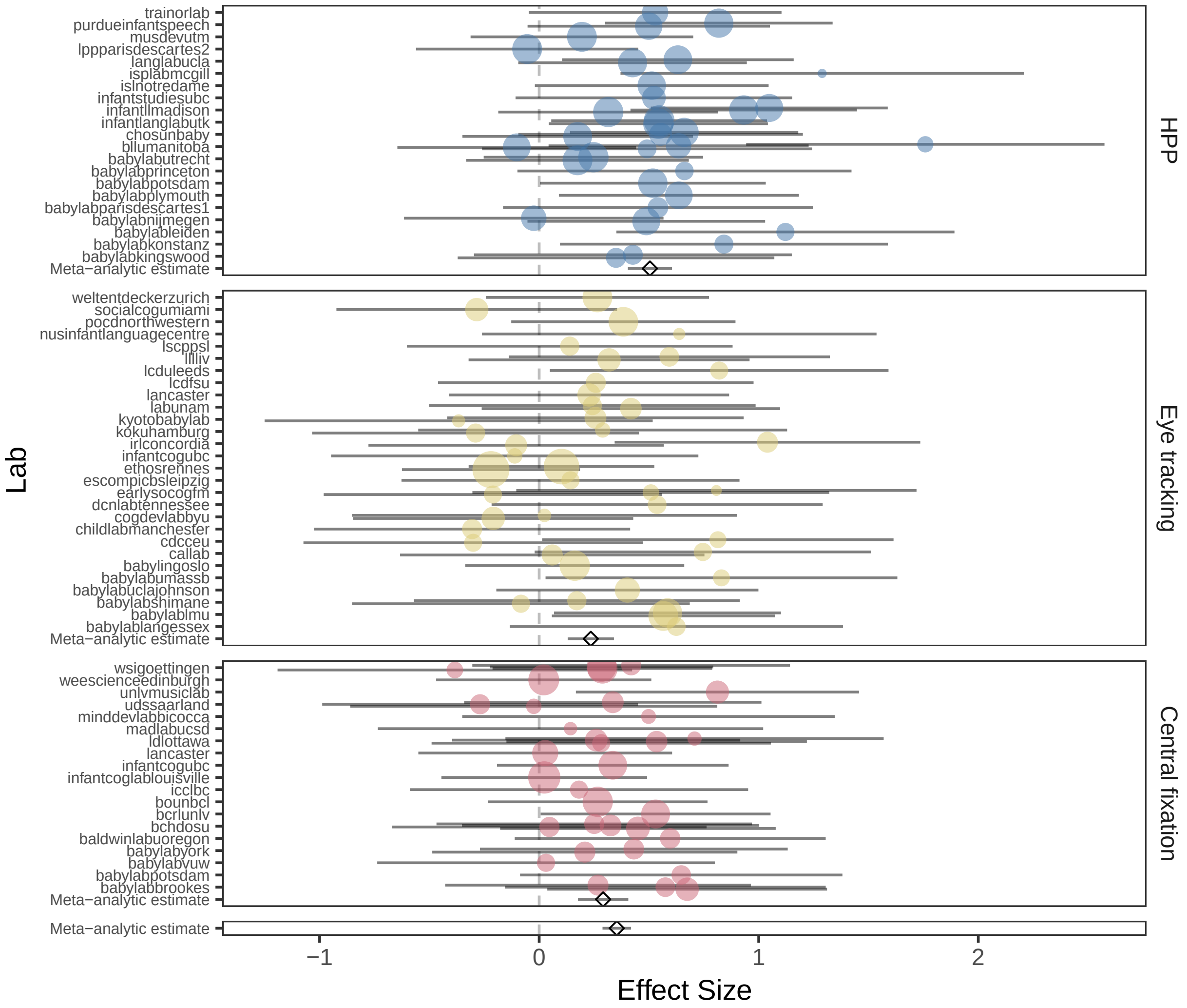

Figure 2. Forest plot. Standardized effect sizes are shown for each lab, with error bars showing 95% confidence intervals. Labs are grouped by method. Points are scaled by inverse variance and colored by experimental method. In each panel, the diamond and associated interval represents the meta-analytic estimate from the method-moderated model and its 95% confidence interval. The bottom panel shows the global meta-analytic estimate from the unmoderated model.

This positive age coefficient indicated that the measured IDS preference was on average larger for older children. Age trends are plotted in Figure 3.

We next investigated effects of experimental method, with method dummy-coded using

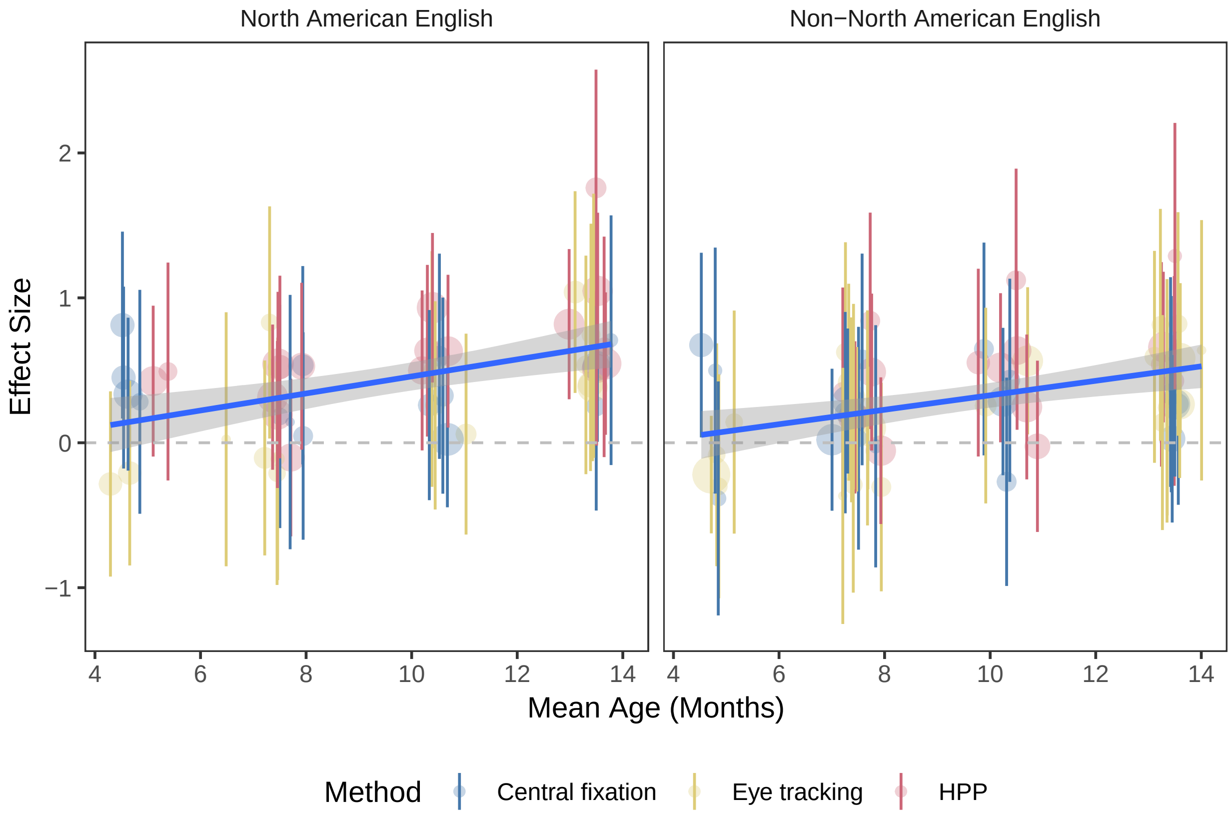

Figure 3. Lab effect size estimates plotted by age and method. Subplots show language groups. Standardized effect sizes are shown for each lab, with error bars showing 95% confidence intervals. Points are scaled by number of participants and colored by experimental method; they are slightly transparent to avoid overplotting.

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 (CI = [0.18 - 0.41], $z = 4.98$, $p < .001$), reflecting the mean effect size for single-screen presentation. The HPP yielded an additional effect of 0.21 (CI = [0.06 - 0.37], $z = 2.74$, $p = .006$), indicating a substantial gain in measured IDS preference for those labs using HPP as compared with single-screen central fixation. In contrast, eye-tracking yielded an effect of -0.06 (CI = [-0.21 - 0.10], $z = -0.71$, $p = .479$), indicating a slight, non-significant decrease in measured effect size for eye-tracking relative to single-screen central fixation.

The language-moderated model ($dz \sim 1 + \text{language}$) was fit with language group coded as a categorical variable indicating whether infants were tested in a lab in which NAE was the standard language (e.g., in the United States or Canada). The reference level effect (i.e., not NAE) was 0.29 (CI = [0.20 - 0.37], $z = 6.56$, $p < .001$), while for infants in North American labs, the effect was increased by 0.15 (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.

Joint relationship of IDS preference to moderating variables. Because infant age, 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 subject-level age effects and lab-level method effects. To help visualize the spread of subject-level effects, Figure 4 shows IDS preferences for individual participants.

Our main model was:

$$\begin{aligned}
\text{log lt} \sim & \text{trial type} * \text{method} + \text{trial type} * \text{trial num} + \text{age} * \text{trial num} + \\
& \text{trial type} * \text{age} * \text{language} + \\
& (\text{trial type} * \text{trial num} \mid \text{subid}) + \\
& (\text{trial type} * \text{age} \mid \text{lab}) + \\
& (\text{method} + \text{age} * \text{language} \mid \text{item})
\end{aligned} \tag{1}$$

Trial type, language, and method were dummy-coded (with ADS trials, non-NAE, and single-screen method) as the reference level; thus, coefficients are interpretable such that e.g., positive effects of trial type indicate longer looking to IDS. To increase the interpretability of coefficients, age (in months) was centered and trial number was coded 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 show larger IDS preferences) and trial type with trial number (modeling the possibility of faster habituation to ADS) as well as 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 across age and language group). Both developmental effects and trial effects are treated linearly in this model; although both likely have non-linear effects, adding quadratic or other effects would have substantially increased model complexity. After pruning random effects for non-convergence,[1] our final model specification was:

$$\begin{aligned}\log \text{lt} \sim & \text{trial type} * \text{method} + \text{trial type} * \text{trial num} + \text{age} * \text{trial num} + \\ & \text{trial type} * \text{age} * \text{language} + \\ & (1 \mid \text{subid}) + \\ & (1 \mid \text{lab}) + \\ & (1 \mid \text{item}).\end{aligned} \tag{2}$$

Table 4 shows coefficient estimates from this model.

Overall, the fitted coefficients of the mixed effects model were consistent with the results of the individual meta-analyses. Within the structure of the mixed effects model, IDS preferences are shown by positive coefficients on the IDS predictor (reflecting greater looking times to IDS stimuli). The fitted model shows a significant positive effect of IDS stimuli,

---

[1] Pruning was done using models fitted with ‘lme4‘ version 1.1-21.

Table 4

Coefficient estimates from a linear mixed effects model predicting log looking time.

	Estimate	*SE*	*t*	*p*
Intercept	2.180	0.051	43.100	0.000
IDS	0.099	0.036	2.740	0.010
Eye-tracking	-0.265	0.046	-5.790	0.000
HPP	-0.052	0.051	-1.020	0.308
Trial #	-0.038	0.002	-25.000	0.000
Age	-0.035	0.004	-7.950	0.000
NAE	-0.016	0.049	-0.335	0.738
IDS * Eye-tracking	-0.009	0.017	-0.548	0.584
IDS * HPP	0.034	0.015	2.270	0.023
IDS * Trial #	-0.003	0.002	-1.370	0.172
Trial # * Age	0.001	0.000	3.140	0.002
IDS * Age	0.012	0.003	4.300	0.000
IDS * NAE	0.039	0.013	3.060	0.002
Age * NAE	0.001	0.006	0.198	0.843
IDS * Age * NAE	0.004	0.004	1.050	0.292

consistent with a global IDS preference. Consistent with the age- and language-moderated meta-analyses, there were significant and positive two-way interations of IDS with age and with NAE, suggesting greater IDS preferences for older children and for children in NAE contexts. Further, there was a positive interaction with the HPP method, consistent with the method-moderated model. There was not a significant three-way interaction of IDS, age, and NAE, however, suggesting that there was not a reliable differential change in IDS preference for older children in NAE contexts over and above that expected based on each of these factors alone.

In addition to these results, a number of other factors were significant predictors of looking time. Looking time decreased across trials, and did so especially for older children, generally confirming that all infants habituated to our experimental stimuli and older infants did so more quickly. Further, eye-tracking led to lower looking times overall across stimulus classes.

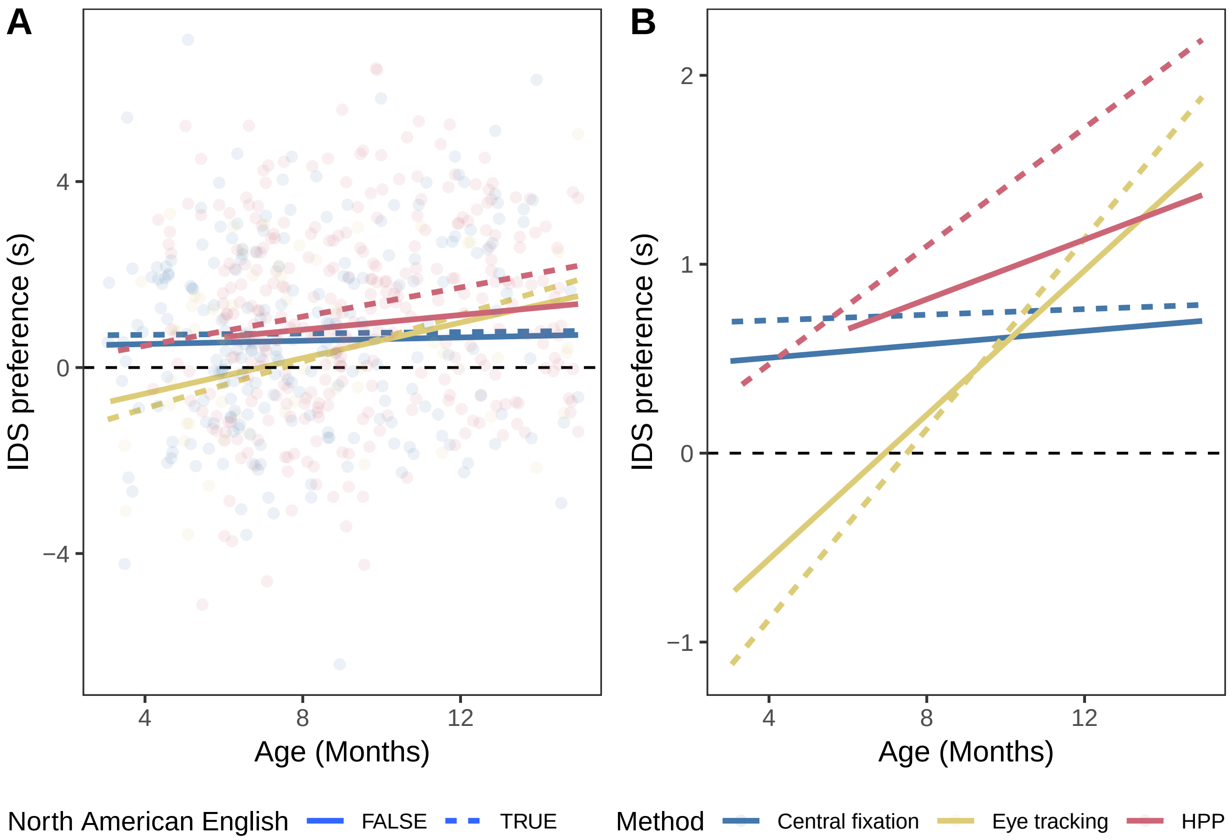

Figure 4. Simple linear trends for IDS preference by age and language group, plotted (A) with individual participants' preferences and (B) without individual participants' preferences to show trends more effectively.

Effects of second-session testing on IDS preference. We preregistered an analysis of whether second-session infants showed a different pattern of infant-directed speech preference. Only 6 labs contributed second-session infants, however, with a total of only 0 infants represented. Thus, we did not fit the full, pre-registered mixed-effects model for this variable as we did not have enough variability on the important covariates to estimate this variable. As an exploratory analysis, we note that 19/41 second-session infants (46.30% [31.60 - 61.30]) showed a numerical preference for IDS. This number was numerically different but not distinguishable statistically from the 58.95% of IDS preferences in the first-session infants, likely due to the small sample of second-session infants.

Sex and IDS preference. In order to investigate effects of biological sex on IDS preference, we fit the model specified above with the addition of a sex main effect and trial type by sex interaction.[2] Female was coded as the reference level, so effects are stated in terms of changes for male infants. The main effect of sex $\beta = 0.01$ ($SE = 0.02$, $p = 0.67$) and the interaction with trial type was $\beta = -0.01$ ($SE = 0.01$, $p = 0.56$). These predictors were small and nonsignificant, suggesting 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 just the amount of looking time we recorded, but 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 (e.g., no looking recorded, a look under 2 s, or no looking because the infant had already terminated the experiment) and false otherwise. We fit a logistic version mixed-effects model with all two-way interactions between method, age, and trial number, using the specification:

---

[2] Because this model did not converge, following our protocol, we pruned random effects of item.

$$\begin{aligned}\text{missing} \sim & \text{method} * \text{age} + \text{method} * \text{trial num} + \text{age} * \text{trial num} + \\ & (1 \mid \text{subid}) + \\ & (\text{trial num} * \text{age} \mid \text{lab}) + \\ & (\text{method} + \text{age} \mid \text{item}).\end{aligned} \tag{3}$$

After pruning for non-convergence, our final model specification was:

$$\begin{aligned}\text{missing} \sim & \text{method} * \text{age} + \text{method} * \text{trial num} + \text{age} * \text{trial num} + \\ & (1 \mid \text{lab}).\end{aligned} \tag{4}$$

Table 5 shows 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. Older children and later trials had greater amounts of missing data, 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 (we return to this issue in the general discussion). 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. When a meta-analysis shows heterogeneity, that finding indicates the presence of unexplained variance in effect size over and above that due to sampling variation; the τ^2 provides an estimate of the total heterogeneity in our models. We further assess heterogeneity using the I^2 statistic (Higgins, Thompson, Deeks, &

Table 5

Coefficient estimates from a linear mixed effects model predicting whether an observation was missing.

	Estimate	SE	z	p
Intercept	-1.090	0.152	-7.140	0.000
Eye-tracking	0.167	0.130	1.290	0.198
HPP	-0.178	0.195	-0.913	0.361
Age	0.356	0.038	9.380	0.000
Trial #	0.663	0.030	22.100	0.000
Eye-tracking * Age	-0.238	0.047	-5.090	0.000
HPP * Age	-0.059	0.051	-1.150	0.251
Eye-tracking * Trial #	0.068	0.036	1.850	0.064
HPP * Trial #	0.046	0.040	1.130	0.257
Trial # * Age	-0.003	0.014	-0.208	0.835

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; when this test is statistically significant, that 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, $\tau^2 = 0.01\%$, $I^2 = 12.39\%$, and $Q(107) = 122$, $p = 0.15$. In the language-moderated model, $\tau^2 = 0.01\%$, $I^2 = 7.76\%$, and $Q(106) = 116.18$, $p = 0.23$. In the age-moderated model, $\tau^2 = 0\%$, $I^2 = 0\%$, and $Q(106) = 98.06$, $p = 0.70$. Finally, in the method-moderated model, $\tau^2 = 0\%$, $I^2 = 3.20\%$, and $Q(105) = 106.78$, $p = 0.43$. In none of these 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 meta-analytic results above), these moderators were quite small in magnitude relative to the sampling variation in individual lab effect size estimates (because of the small median sample size within each lab).

Exclusion criteria. Because our criterion for including infants in the analysis was so liberal (infants needed to contribute data from only two trials to be included), we next conducted an exploration of the effects of different inclusion rules on the results we reported above. In particular, we calculated the meta-analytic effect size with 4 trials and 8 trials as minimum inclusion criteria. For a minimum of 4 trials, the effect size was 0.42 (CI = [0.35 - 0.48], $z = 12.05$, $p < .001$) and for a minimum of 8 trials the effect size was 0.48 (CI = [0.40 - 0.57], $z = 11.23$, $p < .001$). In comparison, our original results showed a meta-analytic effect size of 0.35 (CI = [0.29 - 0.42], $z = 10.67$, $p < .001$). Furthermore, we computed effect sizes for each method for each of these additional exclusion criteria (see Table 6). Overall, more stringent inclusion criteria yielded substantially larger effects, although they also led to substantial data loss (especially for eye-tracking labs).

Table 6

Meta-analytic effect size (dz), standard error (SE) and percentage of included participants for three different exclusion criteria

	2 Trials			4 Trials			8 Trials		
method	estimate	SE	%	estimate	SE	%	estimate	SE	%
Central fixation	0.29	0.06	0.98	0.34	0.06	0.88	0.40	0.06	0.73
Eye tracking	0.24	0.06	0.85	0.33	0.06	0.59	0.41	0.10	0.36
HPP	0.51	0.06	0.98	0.56	0.06	0.92	0.63	0.07	0.78

General Discussion

We designed a large-scale, multi-lab 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 2845 infants from 16 countries, 2329 of which were included in the final sample used for analysis (see Table 1). We believe that the resulting dataset represents the largest laboratory study of infancy to date. We begin our discussion by summarizing the principal results of the study with respect to four critical 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 replicability by providing a pre-registered, unbiased measure of the magnitude of infants' preference for IDS over ADS. We expected to replicate prior demonstrations of the existence of an IDS preference in infant listeners, and our study indeed confirms the expected effect. Our overall meta-analytic mean is smaller in size than the effect found in a preceding meta-analysis of the literature, however (Bergmann et al., 2018; Dunst et al., 2012).

While one possible interpretation of this finding is that previous effect sizes were inflated by publication bias, there are other possible explanations as well. In an individual laboratory, the methodology would be tailored to the specific research question, age range and other characteristics of the infants tested (or conversely, research questions would be 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-native English learning infants, may have led to an underestimate of the true effect size (i.e., because an ideal choice of presentation details that would maximize effect sizes might differ between methods and across ages, versus the compromise protocol used here). Further, our protocol included several decisions that might have decreased effect size, including both our stimuli's relatively less extreme acoustic characteristics, the use of multiple speakers, and our less stringent participant inclusion criteria (both discussed below).

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

As with any other developmental trend, however, age-related change may be driven by changes in factors other than the underlying construct. First, as we will discuss in detail below, characteristics of the stimuli may be best suited for an older age range. Second, stronger effects may result from a more robust or more measurable behavioral response on the part of older infants, independent of an underlying preference. Some evidence in favour of this possibility stems from examining the data in MetaLab, an online databank for meta-analysis in infant research: most meta-analyses show an increase in absolute effect size as infants mature, independent of the research question (see e.g., Bergmann et al., 2018).

Our third goal was to examine how the preference for IDS varies based on the differing linguistic experiences of infants growing up across different linguistic communities. We found a preference for North American English IDS over North American English ADS even for participants for whom this was not their native language or dialect. This finding replicates previous work (Werker et al., 1994). However, in our study, North American English-exposed infants showed the strongest preference. Note that our findings do not support the idea of a simple attentional effect (infants attending more to speech overall when presented in their native language): The effect of language background on overall (as opposed to preferential) looking times is 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 phonology and begin to acquire word meanings, they listen to speech in their own language differently, starting to process what's 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 dialect, the ability to listen in this "deeper" or more predictive way is not available. Another possibility is 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. In either situation, preference for IDS may depend in part on the similarity 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 between age, stimulus type, and language background, which would have been a prediction of this interpretation. Companion data in several non-North American English language communities using native language stimuli created using the ManyBabies 1 protocol are currently under development and may shed further light on this issue.

Our fourth and final goal was to examine differences across methodological approaches in the measured experimental effect. We found a stronger effect when using HPP than central fixation or eye-tracking approaches. One potential interpretation of this finding is that the greater effort on the part of the infant in HPP (i.e., a turning of the head, as opposed to 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 infant language acquisition research were more likely to use HPP. Furthermore, these findings should not be interpreted as suggesting that 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 also Bergmann et al., 2018 for a similar conclusion based on meta-analytic data). We hope to undertake future secondary analyses of our dataset to better understand factors that may have covaried with methodological choices. Moreove, further large-scale projects that include methodological contrasts of this type – perhaps with random assignment – may allow us to draw more specific conclusions about the sources of methodological variability, and their interactions with phenomenon and participant 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 somewhat a departure from the literature using this paradigm, as most participating labs reported using a stricter inclusion criterion in their own independent work. Our original meta-analytic effect size was 0.35 when we included all infants with a minimum of two trials, grew to 0.42 with a minimum of four trials, and 0.48 with a minimum of eight trials. Moreover, there was substantially more missing data from younger infants in the eye-tracking paradigm compared with the other methods. While missing data increased across the length of the experiment, this increase was particularly prevalent for eye tracking. Setting stricter inclusion criteria necessarily decreases sample size with the same number of total infants tested, 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 North American English (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 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 a native and non-native setting. However, our design necessarily complicates the interpretability of our findings from laboratories outside of North America. They confound native-language/dialect effects (infants prefer listening to their native language) and true cultural variation in IDS preference. Further, there is substantial diversity in the non-North American English samples that is obscured in our pre-registered analyses. Together with the previously-mentioned native-language follow-up studies using the ManyBabies 1 protocol, further analyses of our dataset on specific sub-samples 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 around the elicitation of the IDS stimuli. Stimuli used in previous IDS preference literature range from scripted speech with no infant present (e.g., Cooper & Aslin, 1990; Newman & Hussain, 2006), which maximizes experimental stimulus control, to more naturalistic samples collected from free-play, unscripted contexts (e.g., Hayashi et al., 2001; Werker et al., 1994), which maximizes generalizability to real-world contexts. We opted for a relatively naturalistic approach, with an elicitation protocol using real mothers and their infants centred around concrete objects. It is likely that this approach may have led to the reduction in the distinctiveness of the acoustic characteristics of the IDS samples that we observed, 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 findings such as our developmental effects – namely the age range of the "target" infants (4-8 months) and the objects-focused nature of the task (something likely best suited to infants at the older range of our age bins). The extent to which these age-related characteristics of IDS affect the magnitude of infants' IDS preference across development merits further inquiry. Further, and as noted above, 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 infant collaborations. As any lab that has tested infant participants knows, data collection is slow and labour intensive. Over a period of approximately 13 months, 69 labs were able to collect data from 2845 infants. In contrast, ManyLabs 1, a similar initiative with adults participants (Klein et al., 2014), was able to collect data from more than 6000 participants tested in 36 labs over just a handful of months. Moreover, while 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 infant labs, but the depth of these challenges, and the diversity of methodological implementation across laboratories, was surprising. Infant laboratories are highly diverse in both the software and hardware they have available to implement experimental infant testing methods. We planned flexibility in the specific setup (eyetracking, HPP, central fixation) due to known variability, but despite this several labs were forced to deviate from aspects of the protocol, for example due to limitations of how stimuli could be presented (e.g., the ability to implement infant-controlled trial lengths, software settings for repeating trials, etc.). One important conclusion from our work, as evidenced in the "walk through videos" laboratories provided to illustrate their protocols (see below), is the extent to which a typical methods section fails to capture this methodological diversity.

Additional Benefits of Large-Scale Collaboration

While our primary goal was an empirical one, the ManyBabies 1 project had numerous additional benefits to both individual researchers as well as the field at large. All of the questionnaires, and how-tos, and stimuli (e.g., attention getters) used in the project are freely available for re-use in future studies. Each participating lab created a walkthrough video that showed their lab and study setup. These videos provide an unprecedented peek “behind the curtain” of other infancy labs, which was previously only possible through visiting labs in person. Such information could be a particularly helpful resource for investigators setting up an infant lab for the first time. It also provides a unique dataset whereby the field of infant research can begin to understand the variety of lab setups and study implementations.

This large-scale collaborative effort also had broader benefits for the field. It created a strong collaborative network of infancy researchers. Informal “ManyBabies” gatherings are now organized at developmental conferences, enabling researchers who have previously collaborated only virtually to meet in person. It also was many researchers’ introduction to open and cumulative science practices and tools, such as pre-registration and the Open Science Framework.

Finally, ManyBabies 1 has launched several “knock-on” projects. For example, ManyBabies Bilingual (Byers-Heinlein et al., accepted pending data collection) is comparing bilingual infants’ preference for infant directed speech with our results from monolinguals. Other projects will examine the test-retest reliability of infants’ IDS preference, examine whether IDS preference predicts vocabulary size at 18 and 24 months (Soderstrom et al., accepted pending data collection), and test whether lab-specific variables affect infant 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 garner similar benefits.

Conclusion

Replication research can go far beyond simply asking whether an effect is present: it can allow for an assessment of how an effect varies and how it develops. 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 magnitude of the IDS preference – but in the further theoretical and methodological opportunities that the data afford. By measuring the relationship of IDS preferences to age and language community, this experiment provides a starting point for developing a more nuanced theory of how IDS preferences relate to children's language experiences. Further, by revealing the substantial contributions of methodological decision-making to effect size, our study points the way towards developing best-practices templates in further infancy work of this kind. In sum, we hope our work here illustrates the power of large-scale collaboration for the study of developmental variation and change.

Author Contributions

Author contribution initials reflect authorship order. MCF, EB, CB, KBH, BF, JG, JKH, MK, CL, CLW, CM, TN, RP, HR, AS, MS contributed to the study concept. MCF, CB, KBH, CF, JG, NGG, JKH, EEH, MK, CLW, TN, RP, HR, JLR, SW, DY, MS contributed to the study design. MCF, RC, CF, DJK, KK, CLW, RP, MS, MS contributed to stimulus creation. NGG, JKH, DJK contributed to piloting. MCF, CB, RB, KBH, LR, CDL, BF, IJ, MK, JFK, MM, KT, DY contributed to the final protocol. MCF, CB, KBH, JG, MK, CLW, MM, MS contributed to study documentation. MCF, CB, KBH, RLAF, JKH, MK, CLW, KT, MS contributed to study management. KJA, NAT, GA, DB, SB, AKB, MPB, PB, AB, SMB, BB, AB, KBH, LEC, CC, MC, JC, LKC, SC, SC, CC, AC, CD, MK, LR, CDL, DD, KCD, VD, SD, CF, AF, PF, TF, CF, MF, TF, RLAF, AG, JG, NGG, AG, LEH, JKH, EEH, NH, JH, MH, BH, DMH, LHH, MI, SI, IJ, KVJ, MJ, SPJ, CJ, DK, NK, TKP, KK, ESK, JEK, HEK, AARK, FK, JL, RJL, ML, CL, CL, UL, LL, SGL, RAL, VMC, NM, CM, AM, MM, VM, JM, KM, CM, YM, BM, KMN, CN, MAN, NMO, AJO, MO, RP, SPE, MP, CP, LP, CP, HR, SR, JLR, GDR, KCR, CR, DR, YR, JS, AS, SS, AS, GS, MSS, AS, EAS, LS, BS, GS, MS, AT, AT, LJT, SET, AST, ASMT, KT, KVH, YW, SW, SW, AW, DY, KZ, MZ, MS contributed to data collection. MCF, CB, AC, MK, JEK, ML, HR, ASMT, AW, MZ, MS contributed to data analysis. MCF, EB, CB, KBH, AC, RC, CF, JG, NGG, JKH, EEH, MK, CLW, RAL, TN, HR, JLR, MS contributed to the stage 1 manuscript. MCF, CB, KBH, AC, JG, JKH, MK, ML, CM, JLR, MS contributed to the stage 2 manuscript.

Conflicts of Interest

The authors declare 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 through the Association for Psychological Science from the Laura and John Arnold Foundation. Individual participating labs further acknowledge funding support from: the Natural Sciences and Engineering Research Council of Canada (12R81103, 2018-05823, and 402470-2011); a Social Sciences and Humanities Research Council of Canada Insight Grant (12R20580); the UK Economic and Social

Research Council (ES/L008955/1 and ES/N005635/1); Agence Nationale de la Recherche (ANR-17-EURE-0017); a European Research Council Synergy Grant, SOMICS (609819); the Alvin V., Jr. and Nancy C. Baird Professorship; the Korean National Research Fund (NRF-2016S1A2A2912606); the US National Institutes of Health (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.

Prior Versions

Our pre-registered protocol was posted prior to data collection at https://psyarxiv.com/s98ab/.

Disclosures

Preregistration

Our manuscript was reviewed prior to data collection; in addition, we registered our instructions and materials prior to data collection (https://osf.io/gf7vh/).

Data, materials, and online resources

All materials, data, and analytic code are available at https://osf.io/re95x/; the specific code and data required to render this document are available at 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 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 IRB protocols permitted, video recordings were also uploaded to Databrary, a central controlled-access database accessible to other researchers (Databrary, n.d.).

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi:10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development*, *89*(6), 1996–2009. doi:10.1111/cdev.13079

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365. doi:10.1038/nrn3475

Byers-Heinlein, K. (2015). Methods for studying infant bilingualism. In J. Schwieter (Ed.), *The Cambridge Handbook of Bilingual Processing* (p. 133—154). Cambridge, UK: Cambridge University Press.

Byers-Heinlein, K., Bergmann, C., Black, A., Carbajal, J. M., Fennell, C. T., Frank, M. C., … Tsui, A. S. M. (accepted pending data collection). A multi-lab study of bilingual infants: Exploring the preference for infant-directed speech. *Advances in Methods and Practices in Psychological Science.*

Cooper, R. P., & Aslin, R. N. (1990). Preference for infant-directed speech in the first month after birth. *Child Development*, *61*(5), 1584–1595.

doi:10.1111/j.1467-8624.1990.tb02885.x

Cooper, R. P., & Aslin, R. N. (1994). Developmental differences in infant attention to the spectral properties of infant-directed speech. *Child Development*, *65*(6), 1663–1677. doi:10.1111/j.1467-8624.1994.tb00841.x

Cooper, R. P., Abraham, J., Berman, S., & Staska, M. (1997). The development of infants' preference for motherese. *Infant Behavior and Development*, *20*(4), 477–488. doi:10.1016/S0163-6383(97)90037-0

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153. doi:10.1016/j.tics.2009.01.005

Csibra, G., Hernik, M., Mascaro, O., Tatone, D., & Lengyel, M. (2016). Statistical treatment of looking-time data. *Developmental Psychology*, *52*(4), 521–536. doi:10.1037/dev0000083

Cusack, R., & Carlyon, R. P. (2003). Perceptual asymetries in audition. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(3), 713. doi:10.1037/0096-1523.29.3.713

Databrary. (n.d.). *The databrary project: A video data library for developmental science.* Retrieved 2012, from http://databrary.org

Dunst, C., Gorman, E., & Hamby, D. (2012). Preference for infant-directed speech in preverbal young children. *Center for Early Literacy Learning*, *5*(1), 1–13. Retrieved from http://www.earlyliteracylearning.org/cellreviews/cellreviews_v5_n1.pdf

Englund, K., & Behne, D. (2006). Changes in infant directed speech in the first six months. *Infant and Child Development: An International Journal of Research and Practice*, *15*(2), 139–160. doi:10.1002/icd.445

Farran, L. K., Lee, C.-C., Yoo, H., & Oller, D. K. (2016). Cross-cultural register differences in infant-directed speech: An initial study. *PloS One*, *11*(3), e0151518. doi:10.1371/journal.pone.0151518

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development*, *8*(2), 181–195. doi:10.1016/S0163-6383(85)80005-9

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: Is the melody the message? *Child Development*, *60*(6), 1497–1510. doi:10.2307/1130938

Fernald, A., & McRoberts, G. W. (1996). Prosodic bootstrapping: A critical analysis of the argument and the evidence. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 365–388). Mahwah, NJ: Lawrence Erlbaum Associates.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., Boysson-Bardies, B. de, & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477–501.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., … Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. doi:10.1111/infa.12182

Gervain, J., & Mehler, J. (2010). Speech perception and language acquisition in the first year of life. *Annual Review of Psychology*, *61*, 191–218.

Graf Estes, K., & Hurley, K. (2013). Infant-directed prosody helps infants map sounds to meanings. *Infancy*, *18*(5), 797–824. doi:10.1111/infa.12006

Grieser, D. L., & Kuhl, P. K. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese. *Developmental Psychology*, *24*(1), 14–20. doi:10.1037/0012-1649.24.1.14

Hayashi, A., Tamekawa, Y., & Kiritani, S. (2001). Developmental change in auditory preferences for speech stimuli in japanese infants. *Journal of Speech, Language, and Hearing Research*, *44*(6), 1189–1200. doi:10.1044/1092-4388(2001/092)

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal*, *327*(7414), 557.

Hirsh-Pasek, K., Nelson, D. G. K., Jusczyk, P. W., Cassidy, K. W., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition*, *26*(3), 269–286. doi:10.1016/S0010-0277(87)80002-1

Huedo-Medina, T., Sanchez-Meca, J., Marin-Martinez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or i2 index? CHIP documents. 2006; paper 19. *Psychological Methods*, *11*(2), 193–206.

Johnson, E. K. (2016). Constructing a proto-lexicon: An integrative view of infant language development. *Annual Review of Linguistics*, *2*, 391–412.

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*, 601–625. doi:10.1146/annurev-psych-122414-033702

Kaplan, P. S., Jung, P. C., Ryther, J. S., & Zarlengo-Strouse, P. (1996). Infant-directed versus adult-directed speech as signals for faces. *Developmental Psychology*, *32*(5), 880–891. doi:10.1037/0012-1649.32.5.880

Karzon, R. G. (1985). Discrimination of polysyllabic sequences by one-to four-month-old infants. *Journal of Experimental Child Psychology*, *39*(2), 326–342. doi:10.1016/0022-0965(85)90044-X

Katz, G. S., Cohn, J. F., & Moore, C. A. (1996). A combination of vocal f0 dynamic and summary features discriminates between three pragmatic categories of infant-directed speech. *Child Development*, *67*(1), 205–217. doi:10.1111/j.1467-8624.1996.tb01729.x

Kawahara, H., & Morise, M. (2011). Technical foundations of tandem-straight, a speech analysis, modification and synthesis framework. *Sadhana*, *36*(5), 713–727. doi:10.1007/s12046-011-0043-3

Kitamura, C., & Burnham, D. (2003). Pitch and communicative intent in mother's speech: Adjustments for age and sex in the first year. *Infancy*, *4*(1), 85–110. doi:10.1207/S15327078IN0401_5

Kitamura, C., & Lam, C. (2009). Age-specific preferences for infant-directed affective intent. *Infancy*, *14*(1), 77–100. doi:10.1080/15250000802569777

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology*, *45*, 142–152. doi:10.1027/1864-9335/a000178

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. doi:10.18637/jss.v082.i13

Ma, W., Golinkoff, R. M., Houston, D. M., & Hirsh-Pasek, K. (2011). Word learning in infant-and adult-directed speech. *Language Learning and Development*, *7*(3), 185–201.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542. doi:10.1177/1745691612460688

Martin, A., Igarashi, Y., Jincho, N., & Mazuka, R. (2016). Utterances in infant-directed speech are shorter, not slower. *Cognition*, *156*, 52–59.

Maurer, D., & Werker, J. F. (2014). Perceptual narrowing during infancy: A comparison of language and faces. *Developmental Psychobiology*, *56*(2), 154–178. doi:10.1002/dev.21177

McRoberts, G. W., McDonough, C., & Lakusta, L. (2009). The role of verbal repetition in the development of infant speech preferences from 4 to 14 months of age. *Infancy*, *14*(2), 162–194. doi:10.1080/15250000802707062

Mills-Smith, L., Spangler, D. P., Panneton, R., & Fritz, M. S. (2015). A missed opportunity for clarity: Problems in the reporting of effect size estimates in infant developmental science. *Infancy*, *20*(4), 416–432. doi:10.1111/infa.12078

Nelson, D. G. K., Jusczyk, P. W., Mandel, D. R., Myers, J., Turk, A., & Gerken, L. (1995). The head-turn preference procedure for testing auditory perception. *Infant Behavior and Development*, *18*(1), 111–116. doi:10.1016/0163-6383(95)90012-8

Newman, R. S. (2003). Prosodic differences in mothers' speech to toddlers in quiet and noisy environments. *Applied Psycholinguistics*, *24*(4), 539–560. doi:doi:10.1017/S0142716403000274

Newman, R. S., & Hussain, I. (2006). Changes in preference for infant-directed speech in low and moderate noise by 4.5-to 13-month-olds. *Infancy*, *10*(1), 61–76.

doi:doi:10.1207/s15327078in1001_4

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*(9), 1105–1107. doi:10.1038/nn.2886

Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy*, *22*(4), 436–469.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. doi:10.1126/science.aac4716

Pegg, J. E., Werker, J. F., & McLeod, P. J. (1992). Preference for infant-directed over adult-directed speech: Evidence from 7-week-old infants. *Infant Behavior and Development*, *15*(3), 325–345. doi:10.1016/0163-6383(92)80003-D

Santesso, D. L., Schmidt, L. A., & Trainor, L. J. (2007). Frontal brain electrical activity (eeg) and heart rate in response to affective infant-directed (id) speech in 9-month-old infants. *Brain and Cognition*, *65*(1), 14–21. doi:10.1016/j.bandc.2007.02.008

Schachner, A., & Hannon, E. E. (2011). Infant-directed speech drives social preferences in 5-month-old infants. *Developmental Psychology*, *47*(1), 19–25. doi:10.1037/a0020740

Segal, J., & Newman, R. S. (2015). Infant preferences for structural and prosodic properties of infant-directed speech in the second year of life. *Infancy*, *20*(3), 339–351.

Shute, H. B. (1987). Vocal pitch in motherese. *Educational Psychology*, *7*(3), 187–205. doi:10.1080/0144341870070303

Singh, L., Morgan, J. L., & Best, C. T. (2002). Infants' listening preferences: Baby talk or happy talk? *Infancy*, *3*(3), 365–394. doi:10.1207/S15327078IN0303_5

Snow, C. E. (1977). The development of conversation between mothers and babies. *Journal of Child Language*, *4*(1), 1–22. doi:10.1017/S0305000900000453

Soderstrom, M., Blossom, M., Foygel, R., & Morgan, J. L. (2008). Acoustical cues and grammatical units in speech to two preverbal infants. *Journal of Child Language*, *35*(4), 869–902. doi:10.1017/S0305000908008763

Soderstrom, M., Werker, J., Tsui, A., Skarabela, B., Seidl, A., Searle, A., & Anderson, L. (accepted pending data collection). Testing the relationship between preferences for infant-directed speech and vocabulary development: A multi-lab study. *Journal of Child Language.*

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, *7*(1), 53–71. doi:10.1207/s15327078in0701_5

Trainor, L. J., & Desjardins, R. N. (2002). Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels. *Psychonomic Bulletin & Review*, *9*(2), 335–340. doi:10.3758/BF03196290

Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *Journal of Statistical Software*, *36*(3). doi:10.18637/jss.v036.i03

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

Werker, J. F., & McLeod, P. J. (1989). Infant preference for both male and female infant-directed talk: A developmental study of attentional and affective responsiveness. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *43*(2), 230. doi:10.1037/h0084224

Werker, J. F., Pegg, J. E., & McLeod, P. J. (1994). A cross-language investigation of infant preference for infant-directed communication. *Infant Behavior and Development*, *17*(3), 323–333. doi:10.1016/0163-6383(94)90012-4

Zangl, R., & Mills, D. L. (2007). Increased brain activity to infant-directed speech in 6-and 13-month-old infants. *Infancy*, *11*(1), 31–62. doi:10.1207/s15327078in1101_2