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

The ManyBabies Consortium

Abstract

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

Keywords: language acquisition; speech perception; infant-directed speech; reproducibility; experimental methods

Introduction

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 underpowered (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), 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 collect 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., 2016). 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. 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; Fernald et al., 1989; Farran, Lee, Yoo, & Oller, 2016; 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; 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 $d = .72$ – quite a large effect size for an experiment with infants (Lewis et al., 2016).

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 (Hirsh-Pasek et al., 1987; cf. Fernald & McRoberts, 1996). In addition, word segmentation (Thiessen, Hill, & Saffran, 2005) and word learning (Graf Estes & Hurley, 2013) 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).

Methods

Participation Details

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.

Materials

Visual stimuli. For labs using the Head-turn Preference Procedure (HPP), we asked labs to use their typical visual stimulus, which varied considerably across laboratories. Some labs

used flashing lights as the visual 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, etc.), 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: a ball, a shoe, a cup, a block, a train, 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 4 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 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 8 and 18 due to variability in random assignment. Affect and IDS ratings were made using low-pass filtered recordings (a 120 dB 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 for noise ratings, and was .85).

Ratings from the tasks were then used to produce 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 (> or < 4, depending on register, 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 8 IDS and 8 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 RMS amplitude-normalized to 70 dB 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.* Both familiar and unfamiliar objects were used during the recording session. 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 are more clips per stimulus in the IDS samples, there are also more utterances boundaries. This property is consistent with the literature on the natural characteristics of IDS.

- *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.

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 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.

Procedure

Onset of each trial. At the beginning of each trial, a centrally positioned visual stimulus (typically a laughing baby, 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 light or a similar video for HPP). The stimulus appeared to the left or right of the infant in HPP setups.

Trials. At the beginning of the session, there were 2 warm-up trials that familiarized infants with the general procedure. The auditory stimulus for warm-up trials was an 18-second clip of pleasant 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 the 16 test trials presenting the IDS and ADS auditory stimuli.

Minimum looking time. There was no minimum 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 2 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 2 IDS and 2 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

Minimizing caregiver bias. We created a custom blend of instrumental music and a pastiche of randomly timed and random amplitude stimulus materials (available as part of the study materials). This masking stimulus was played to the caregiver over noise-reducing 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.

Minimizing experimenter bias. Experimenters making online coding 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: gender, 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 siblings (gender and date of birth). A standard recommended participant questionnaire was distributed to participating labs, although labs were permitted to use their own forms as long as they gathered the necessary information.

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. N.B.: the first three criteria preemptively prevent participation (except in case of erroneously running the experiment with children outside of the inclusion guidelines).

- Full-term. We defined full term as gestation times greater than or equal to 37 weeks.
- No diagnosed developmental disorders. We excluded infants with parent-reported developmental disorders (e.g., chromosomal abnormalities, etc.) or diagnosed hearing impairments. Due to concerns about the accuracy of parent reports, we did not plan exclusions based on self-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. 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). In the exploratory analyses below, we

consider how exclusion decisions affected our effect size estimates.

- 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 parent/outside interference noted by participating labs

  (e.g., talking or pointing by parents, construction noise, sibling pounding on door)

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.

    Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects

    models using me4. *Journal of Statistical Software, 67*(1), 1-48.

    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), 365-376.

Byers-Heinlein, K. (2015). Methods for studying infant bilingualism. In J. W. Schwieter, (Ed.).

    Cambridge Handbook of Bilingual Processing. Cambridge, UK: Cambridge University

    Press, 133–154.

Cooper, R. P., & Aslin, R. N. (1990). Preference for infant-directed speech in the first month

    after birth. *Child Development, 61*(5), 1584-1595.

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.

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.

    Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4),

    148-153.

Csibra, G., Hernik, M., Mascaro, O., Tatone, D., & Lengyel, M. (2016). Statistical treatment of

    looking-time data. *Developmental Psychology*, 52(4), 521-536.

Cusack, R., & Carlyon, R. P. (2003). Perceptual asymmetries in audition. *Journal of Experimental Psychology: Human Perception and Performance, 29*(3), 713-725.

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.

Englund, K., & Behne, D. (2006). Changes in infant directed speech in the first six months. *Infant and Child Development*, *15*(2), 139-160.

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.

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development*, *8*(2), 181-195

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

Fernald, A., & McRoberts, G. (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*. New York and London: Psychology Press: pp 365-388.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(03), 477-501.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S., Yurovsky, D. (2016). A

collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. Preprint retrieved from https://osf.io/27b43/.

Graf Estes, K., & Hurley, K. (2013). Infant-directed prosody helps infants map sounds to meanings. *Infancy*, *18*(5), 797-824.

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.

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.

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.

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*, 1–17. Advance online publication.

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.

Karzon, R. G. (1985). Discrimination of polysyllabic sequences by one- and four-month-old infants. *Journal of Experimental Child Psychology*, *39*(2), 326-342.

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.

Kawahara, H. & Morise, M. (2011). Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework. *Sadhana, 36*(5)*,* 713-727.

Kemler Nelson, D. G., Jusczyk, P. W., Mandel, D. R., Myers, J., Turk, A., & Gerken, L. A. (1995). The Headturn Preference Procedure for testing auditory perception. *Infant Behavior and Development, 18*(1)*,* 111-116.

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

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

Klein, R., Ratliff, K., Vianello, M., Adams Jr, R., Bahník, S., Bernstein, M., …, & Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" Replication Project. *Social Psychology, 45*(3), 142-152.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-30.

Lewis, M. L., Braginsky, M., Tsuji, S., Bergmann, C., Piccinini, P., Cristia, A., Frank, M. C. (2016). A quantitative synthesis of early language acquisition using meta-analysis. Preprint retrieved from https://osf.io/htsjm/.

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.

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

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.

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

Newman, R. S. (2003). Prosodic differences in mothers' speech to toddlers in quiet and noisy environments. *Applied Psycholinguistics*, *24*(04), 539-560.

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.

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

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 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.

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.

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

Shute, H. B. (1987). Vocal pitch in motherese. *Educational Psychology*, *7*(3), 187-205.

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

Snow, C. E. (1977). The development of conversation between mothers and babies. *Journal of Child Language*, *4*(01), 1-22.

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*(04), 869-902.

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

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.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1-48.

Weisleder, A., & Fernald, A. (2013). Talking to children matters early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143-2152.

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), 230-246.

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.

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.