Testing the Relationship Between Preferences for Infant-Directed Speech and Vocabulary Development: A Multi-Lab Study

Soderstrom M, Rocha-Hidalgo J, Munoz LE, Bochynska A, Werker JF, Skarabela B, Seidl A, Searle A, Ryjova Y, Rennels JL, Potter C, Paulus M, Ota M, Noble C, Nave C, Mayor J, Martin A, Machon LC, Lew-Williams C, Ko ES, Kim H, Kartushina N, Kammermeier M, Jessop A, Hay JF, Hannon EE, Hamlin JK, Havron N, Gonzalez-Gomez N, Gonzalez-Barrero AM, Gampe A, Fritzsche T, Frank MC, Floccia C, Durrant S, Delle Luche C, Davies C, Cashon C, Byers-Heinlein K, Black AK, Bergmann C, Anderson L, AlShakhori MK, Al-Hoorie AH, & Sin Mei Tsui A

Key Words: infant-directed speech, vocabulary development

Abstract

From early in life, infants show a preference for infant-directed speech (IDS) over adult-directed speech (ADS), and exposure to IDS has been correlated with different language outcome measures such as vocabulary. The present multi-laboratory study explores this issue by investigating whether there is a link between early preference for IDS and later vocabulary size. Infants' preference for IDS was tested as part of the ManyBabies1 project, and follow-up CDI data were collected from a subsample of this dataset at 18 and 24 months. A total of 341 (18 months) and 327 (24 months) infants were tested across 21 laboratories. In neither preregistered analyses with North American and UK English, nor exploratory analyses with a larger sample did we find evidence for a relation between IDS preference and later vocabulary (Bayes Factor analysis was inconclusive). We discuss the implications of this finding in light of recent work suggesting that IDS preference measured in the laboratory has low test-retest reliability.

Testing the Relationship Between Preferences for Infant-Directed Speech and Vocabulary Development: A Multi-Lab Study

Infant-directed speech (IDS) is a type of register characterized by higher pitch, exaggerated prosody, simplified structure, longer pauses, and slower speech rate than adult-directed speech (ADS), among other distinctive features (e.g., Fernald et al., 1989; Kuhl et al., 1997). There are now over 50 years of research supporting the idea that IDS plays an important role in language acquisition (e.g., Karzon, 1985; Kemler Nelson et al., 1989; Shneidman & Goldin-Meadow, 2012; Snow & Ferguson, 1977; Trainor & Desjardins, 2002; Weisleder & Fernald, 2013). Within this body of research, numerous laboratory studies have demonstrated the benefits of various characteristics of IDS for language acquisition (e.g., Graf Estes & Hurley, 2013; Kempe et al., 2003; Ma et al., 2011; Mintz, 2003; Thiessen et al., 2005). For instance, under controlled laboratory conditions, it has been reported that IDS, relative to ADS, facilitates word segmentation (Thiessen et al., 2005), word recognition (Singh et al., 2009), and word learning (Graf Estes & Hurley, 2013).

Studies outside of the laboratory have also found correlations between caregivers' use of IDS and child language outcomes, including vocabulary acquisition (e.g., Ramírez-Esparzal, 2014; Shneidman et al., 2013; Shneidman & Goldin-Meadow, 2012; Weisleder & Fernald, 2013) and speech perception abilities (e.g., Trainor & Desjardins, 2002). A meta-analysis also found evidence that speech conforming more to the prosodic characteristics of IDS correlates with infants' attention and with lexical development (Spinelli et al., 2017).

In tandem with this body of literature showing that *exposure* to IDS promotes language learning, there is a parallel body of literature showing that young infants *prefer* listening to IDS

over ADS. The basic finding - that young infants prefer IDS - has been replicated across infants of different ages and language backgrounds (Cooper & Aslin, 1994; Cooper et al., 1997; Fernald, 1985; Hayash et al., 2001; Kitamura & Lam, 2009; Newman & Hussain, 2006; Pegg et al., 1992; Santesso et al., 2007; Singh et al, 2002; Werker & McLeod, 1989), and is supported by a meta-analysis (Zettersten et al., in prep.). However, a number of studies also report that the IDS preference (Newman & Hussain, 2006; Robertson et al., 2013) and benefits of IDS during word learning may begin to decrease with age (Ma et al., 2011). The goal of the present study is thus to systematically examine whether or not a *preference* for IDS (rather than exposure to IDS) predicts later language outcomes using a large, multi-lab sample of linguistically diverse infants.

To our knowledge, only one study at the time of writing addressed this question directly. That study found that individual preferences for IDS over ADS between 6 and 12 months of age predict expressive language outcomes at 18 months, at least in typically developing infants (siblings of children with autism did not show this association; Droucker et al., 2013). The current study built on this finding with a larger and more diverse sample, which allowed us to more accurately measure the predictive effect of IDS preference and, with sufficient power, to examine additional questions about how other factors like age of testing influence this relationship.

A link between preference for IDS and language development may indicate a simple causal relationship: infants who have a greater preference for IDS fare better because their attention is drawn to the signal that is best matched to their learning needs. A wide variety of studies and theoretical papers suggest that IDS provides particularly rich language-learning opportunities, highlighting, for example, syntactic (e.g., Mintz, 2003; Soderstrom et al., 2008), morphological (e.g., Kempe et al., 2003), phonetic (e.g., Kuhl et al., 1997; Trainor & Desjardins,

2002; Werker et al., 2007, but e.g. Martin et al., 2015 for counterevidence), timbral (Piazza et al., 2017), and prosodic properties (e.g., Kemler Nelson et al., 1989). Other findings suggest that IDS might serve as a cue to infants that a speaker is a potential teacher or social partner who will provide learning opportunities (Begus et al., 2016). For example, infants prefer to look at a person who previously used IDS than a person who used ADS (Schachner & Hannon, 2011), and a speaker's prior use of IDS was critical for eliciting infants' subsequent gaze-following towards an object (Senju & Csibra, 2008).

It could therefore be argued that children who are able to preferentially focus on IDS as compared to ADS effectively enhance their exposure to the most appropriate type of input for language learning. However, more complex relationships might also be at play, as infants' preferences may be driven by their experiences with IDS. For example, exposure to IDS may enhance a pre-existing but small early preference for IDS over ADS (as supported by findings that even newborn infants prefer IDS; e.g., Cooper & Aslin, 1990). As familiarity with this speech register increases, the infant develops greater interest in IDS, which leads to more attention to this kind of input. Given the diversity of individual experiences in exposure to IDS (e.g., Weisleder & Fernald, 2013), a relation between IDS preference and language outcome may thus be intimately connected with experience. Indeed, in children with hearing loss, preferences for IDS are more closely tied to the child's hearing age than to their chronological age (Robertson et al., 2013). Additionally, the quality and quantity of IDS may be correlated with other beneficial aspects of infants' environments, including social factors such as attachment. Although it may be difficult to disentangle the relative influence of infants' underlying preferences for, versus experience with, IDS, determining whether IDS preference is indeed associated with language development presents an important starting point.

In order to directly address this question, the current project leveraged the unique opportunity afforded by the ManyBabies 1 project. In the ManyBabies1 project (The ManyBabies Consortium, 2020), 67 laboratories contributed data from 2329 infant participants on their relative preference for samples of North American English IDS and ADS, in 13 languages between 3 and 15 months of age. The current proposal builds on the primary ManyBabies1 project by assessing language development in the same participants whose IDS preferences were tested as infants, at two later age points. Evidence for the facilitative effects of IDS is most robust in lexical learning, thus we elected to measure our participants' vocabulary size, using parental report data (Communicative Development Inventories - CDI; Fenson et al., 2007) collected at two time points commonly used for measuring productive vocabulary: 18 and 24 months. A productive measure was used for greater comparability across the 18 and 24 month ages. The 18-24 month age range is a time of considerable diversity in toddlers' vocabulary size, characterized by a rapid rate of growth (Frank et al., 2017b; Ganger & Brent, 2004; McMurray, 2007). Data from toddlers at both 18 and 24 months will therefore allow us to not only establish whether there is a relation between IDS preference and vocabulary acquisition generally, but to also elucidate whether the magnitude of this relation varies or remains constant throughout early language development. In total, 21 labs participated in this follow-up study, yielding a sample of N = 341 infants at 18 months, and N= 327 at 24 months who contributed data on their preference for IDS and their later vocabulary size. This sample is much larger than those that can typically be gathered by a single laboratory, which will allow more statistical power to measure the potential relation between IDS preference and vocabulary size.

In addition to querying the overall strength of this relation and how it changes over development, the characteristics of the sample allow us to address whether the relation between

IDS preference and vocabulary is influenced by the child's age at the time of data collection (or the chronological distance between collection of the two measures). Recent findings suggest that the importance of IDS may decrease over development. In one study, the amount of speech with IDS-like characteristics diminished from 24 to 33 months, and speech with *less* IDS-like characteristics was associated with greater vocabulary at 33 months, but not speech with more IDS-like characteristics (Ramírez-Esparza et al., 2017). Thus, we might expect a smaller/less reliable relation with vocabulary development when IDS preference is measured at older ages.

Finally, the ManyBabies1 sample is linguistically diverse, with participating labs testing infants learning 13 different languages, often including multiple language varieties (e.g., American, Canadian, British, and Australian English). This diversity can, to some extent, begin to address the impact of the over-representation of North American English in child language research. Numerous studies point to IDS as a cross-linguistic phenomenon (Blount, 1972, 1984; Blount & Padgug, 1976, 1977; Englund & Behne, 2006; Farran et al., 2016; Fernald & Morikawa, 1993; Fernald & Simon, 1984; Fernald et al., 1989; Grieser & Kuhl, 1988; Katz et al., 1996; Kitamura et al.,, 2001; Morikawa et al., 1988; Niwano & Sugai, 2002; Newman, 2003; Papoušek et al., 1987; Shute & Wheldall, 1995; Zeidner, 1983). Infants' preference for IDS is also a crosslinguistic (and even cross-modal, Masataka, 1996) phenomenon (Cooper & Aslin, 1994; Cooper et al., 1997; Fernald, 1985; Hayashi et al., 2001; Kitamura & Lam, 2009; Newman & Hussain, 2006; Pegg et al., 1992; Santesso et al., 2007; Singh et al.,2002; Werker & McLeod, 1989). However, there is ample evidence that North American IDS is particularly extreme in its characteristics (e.g., Fernald et al., 1989). Indeed, prosodic differences in IDS registers have been implicated as a source of difference in lexical learning between infants exposed to North American versus British English (Floccia et al., 2016). In the ManyBabies 1 project, slightly less

than half of participating labs were from North America. We therefore used the cross-linguistic diversity of this sample to investigate the relation between specific linguistic experience, IDS preference (at least to North American IDS), and eventual vocabulary outcomes. This latter analysis is necessarily tentative in nature because linguistic experience/community is confounded with the measurement of vocabulary, given that many linguistic communities have their own language- or dialect-specific version of the CDI to measure vocabulary size, and because all language communities were tested on their "IDS preference" using North American English stimuli. Ultimately, we were able to recruit sizeable samples of infants learning North American English, British English, and other languages, which allowed us to test our hypotheses with groups outside of North American English contexts.

In sum, the unique opportunity of the ManyBabies1 project, which gathered data on infants' preference for IDS from 2,329 infants, allowed us the opportunity to follow up with N = 467 (N = 341 at 18 months and N = 327 at 24 months) of these infants by assessing their productive vocabulary size. These data allowed us to ask the following three research questions:

1. To what extent does infants' preference for IDS as measured in a laboratory setting predict their vocabulary at 18 and 24 months?
2. Does the relation between IDS preference and vocabulary size change over development?

3. Are there systematic differences in the strength of this relation across the language communities in our sample (exploratory)?

Method

Participants

Participants were a subsample of the primary ManyBabies1 dataset, based on participating laboratories' interest and ability to collect the follow-up CDI data from their participants. Only monolingual infants (90% or more exposure to the primary language based on parental report or via a detailed questionnaire depending on the laboratory) were included. A total of 21 laboratories (9 North American, 4 UK, 2 German, 1 New Zealand English, 1 Dutch, 1 Korean, 1 French, 1 Norwegian, 1 Swiss German) collected follow-up data, with a minimum sample of 10 infants per laboratory. In addition, three other laboratories initially signed up but withdrew due to: lack of sufficient participant interest combined with many of the participants not maintaining monolingual status (1), and author miscommunication (2). We asked that laboratories not impose any additional eligibility restrictions on their CDI collection beyond those of the primary study. The final sample consisted of 467 infants (228 North American English, 76 UK English, 163 other languages/dialects) who contributed data at 18 (N = 341) and/or 24 months (N= 327). In total, the final sample consisted of 668 CDI contributions (333 North American English, 92 UK English, and 243 other languages/dialects).

Additional participants who were part of the initial sample of CDI measures were excluded for the following reasons: CDI data from 88 infants were collected when infants were outside of the target age range, 123 infants did not complete at least one pair of IDS and ADS trials to provide an IDS score, and 2 infants were excluded as the participating laboratories

reported unusable and/or incomplete CDI data. See Table 1 for more information about the demographics and distribution of the participants.

Data

The data used in our analysis came from two sources. First, we made use of the ManyBabies1 primary dataset, which can be found at https://github.com/manybabies/mb1-analysis-public. Details about the creation of this dataset can be found in the ManyBabies1 published study (The ManyBabies Consortium, 2020), and information about the conceptual and methodological relevance of the ManyBabies Project can be found in Frank et al. (2017a). In brief, the ManyBabies1 dataset contains basic participant information (e.g., age in days, gender), and looking time data comparing interest to IDS and ADS. Three looking time paradigms were used across the laboratories: Headturn Preference Procedure, Central Fixation, and Eyetracking. Only data from the participants from laboratories contributing to the CDI follow-up, and for whom CDI data were collected were included in the current analysis. Note that in the ManyBabies study, all infants, regardless of their language background, were tested on the same set of stimuli recorded in North American English. Stimuli were recorded from mothers speaking to their infant (aged 4-8 months; i.e., IDS) and separately to an experimenter (i.e., ADS) about a set of objects. Clips from these recordings were then subjected to a selection process based on naïve rater ratings regarding the extent to which they sounded infant-directed vs. adult-directed, as well as other characteristics that were controlled for (e.g. naturalness, noisiness). Selected clips were also controlled for other characteristics such as object labels and speaker identity. These clips were then combined to create 16 total test trials of 18 s each. The visual stimulus used for the Central Fixation, Eyetracking and some of the labs using Headturn Preference

Procedure was a colourful checkerboard pattern. For the other Headturn Preference Procedure labs, lights or other visual displays were used.

Second, we collected four different types of CDI data. In the North American Primary Sample, we collected data from North American participants using webcdi, a web-based version of the MacArthur-Bates Communicative Development Inventories (Fenson et al., 2007). For this sample, data were collected at 18 months (16 - 20 months) and 24 months (22 - 26 months), and standardized scores were used in the analysis. In cases where standardized scores were unavailable, raw scores were used, and the age ranges were therefore narrowed to 17.5 - 18.5 and 23.5 - 24.5 months. In the UK Primary Sample, data were collected using an online version of the Oxford CDI (Hamilton et al., 2000). In the "Other Language/Dialects Primary Sample" we collected data from non-North American and non-UK language communities. For this sample, each laboratory selected the CDI that best matched their language community. Lastly, we allowed for the contribution of "samples of convenience" because some laboratories had specific policies already in place regarding the collection of CDI data for their participants at the time of testing (i.e., concurrently with the experimental test of IDS preference) or at other times than those specified in the primary collection protocol. Given the diversity of these latter two samples and insufficient details at the time of pre-registration, we planned to conduct exploratory analyses on these data. We conducted specific exploratory analyses for the "Other Language/Dialects Primary Sample", but there was insufficient data collected from samples of convenience to conduct analyses. Instructions provided to contributing laboratories can be found here: https://osf.io/t9mk5.

Analysis and Results

Data Preparation

See https://osf.io/7z4u6 for the original Stage 1 registered analysis. Full details of our analysis pipeline can be found at https://github.com/manybabies/mb1-cdi-followup. A cloud-based "Docker" image to facilitate result reproducibility is also available at https://github.com/manybabies/mb1-cdi-followup/blob/master/Docker%20instructions.md**.** Each participating laboratory provided a spreadsheet with laboratory and participant ID codes, and summary vocabulary scores at 18 and/or 24 months for each infant. By-item data were also collected but are not included in the analyses. For the North American sample, standard percentile scores using the Fenson et al. (2007) norms were used for the preregistered analysis. For any other language communities, we planned to use raw scores (i.e., number of vocabulary words) if no standardized scoring was available. We used this approach for the UK preregistered analysis. However, ultimately we were able to obtain percentile scores for all of the samples as described below in the exploratory analyses. Laboratory and participant ID codes were used to match each infant with their mean looking times (mean preference to IDS and ADS samples) in the ManyBabies1 dataset, as well as gender, age-in-days at test, and testing protocol (Headturn Preference, Central Fixation, or Eyetracking). A new variable "standardized mean preference for IDS" was created. To calculate this variable, looking time to each ADS trial was subtracted from its paired IDS trial to create a raw difference score. As a result we excluded any trial pairs in which there were missing data based on the ManyBabies1 criteria for trial inclusion. A mean difference score was then calculated for each child. This mean difference score was then divided by the mean looking time across both ADS and IDS trials for each infant, to control for differences in looking time due to methodological and age-related factors. This score could vary from 1.6 (indicating a complete IDS preference) to -1.6 (indicating a complete ADS preference. In the analyses that follow, we make the assumption that the "standardized mean preference for

IDS" can be used in a continuous fashion to represent the *degree* of preference to IDS for a given infant. We acknowledge that there are limitations to this assumption.

Power Analyses

Although there is no equivalent study on which to conduct a power analysis, the one study that has looked directly at the relation between preference for IDS and vocabulary size found a correlation of r = .504 for children without siblings with autism (Droucker et al., 2013). For the current study, we conducted a prospective power analysis and set the smallest effect of interest to a more modest r = .3, which would account for approximately 10% of the variance in the vocabulary size. A comparable level of effect size has been reported in studies investigating the relation between infants' segmentation skills at 7.5 months and their productive vocabulary size at 24 months (Singh et al., 2012). A power analysis using the *pwr* package (Champely, 2015) in the *R* programming language (R Core Team, 2017) showed that a minimum of 84 infants would be necessary to detect a main effect of this size with a power threshold of 80% and alpha at 5%[1]. Although our final sample was larger than 84 for all three of our main samples, there may be a reduction in power due to interactions involving the IDS preference and the greater model complexity. To address this concern, we pre-registered that we would conduct power checks during the analysis phase (see original Stage 1 registered report . Unfortunately, this was not possible due to singularity issues. Removing significant effects based on models that have singularity issues would not produce reliable results and thus we excluded this power analysis from the Stage 2 registered report. Instead, we conducted a sensitivity analysis (see Supplement 1).

Confirmatory Statistical Models

[1] pwr.r.test(r=0.3, sig.level = 0.05, power=0.8)

Some necessary deviations from our original analytic plan were implemented. Please see Supplement 2 for details.

For our primary confirmatory analyses, we applied a series of mixed-effects models (one for the North American primary sample, one for the UK primary sample, and one combined across the two samples) using the *lme4* package (Bates et al., 2015) in the latest version of the *R* programming language (4.2.3) available at the time of completing the analysis (R Core Team, 2017). The dependent factor of the models was the productive vocabulary score of each child in the CDI data (for the North American sample this is the standardized score, whereas raw scores were used for the UK sample). The models included the following predictors as fixed effects (Note that the grand mean is interpreted at the reference levels for all binary variables and at the average levels for all continuous variables in the model):

ids_pref: Standardized mean preference for IDS (described above) as a centered continuous variable. The intercept of this factor represents the CDI value for the grand mean of ids_pref.

test_age: Age (in months) at time of IDS preference testing, entered as a centered continuous variable. The intercept of this factor represents the CDI value for the grand mean of test_age.

cdi_age: Age (in months) at which the CDI measure was taken, entered as a centered continuous variable. The intercept of this factor represents the CDI value for the grand mean of cdi_age.

gender: Male/female as an effect coded fixed factor to test for effects of gender. The intercept of this factor represents a hypothetical CDI value where gender is neither male nor female.

protocol: The testing protocol used to assess IDS preference (3 levels: central fixation, eye-tracking, and head-turn preference), entered as a deviation coded factor. The intercept of this factor represents the difference between CDI values for the mean of each level (e.g., central fixation) and the grand mean of all levels.

In order to keep the models to a manageable level of complexity, we restricted the interaction terms to those that could be motivated theoretically. The main factor of interest, the effect of IDS preference, may be conditioned by age at time of IDS testing, age when CDI was taken, or the testing protocol used to test IDS preference. The model therefore included two-way interactions between *ids_pref* and *test_age*, *ids_pref* and *cdi_age*, and *ids_pref* and *protocol, as well as main effects of ids_pref, test_age and cdi_age*. The factor *gender* was included to address known gender differences in vocabulary size, and therefore entered only as a main effect. In addition, *lab* was entered as a random factor in order to control for variance between the participating laboratories. This is particularly important given the allowed variation in methodology across laboratories in the original ManyBabies study, even within a protocol. The resulting starting model for each of the NAE and UK primary samples had the following structure with 6 fixed effects along with their random intercepts and slopes:

CDI vocabulary ~ ids_pref:test_age + ids_pref:cdi_age + ids_pref:protocol + ids_pref + cdi_age + gender + test_age + protocol + (1 + ids_pref:test_age + ids_pref:cdi_age + ids_pref:protocol + ids_pref + cdi_age + gender | lab)

As noted above, we preregistered two criteria that this mixed-effects model needed to meet and for which the model was simplified if necessary. First, the model had to reach

convergence. To achieve convergence, we iteratively simplified the random effects structure of the model by sequentially removing random slopes for lab, starting with the highest order interaction terms that explained the least amount of random variance (Barr et al., 2013). The final pruned model for the NAE was:

CDI_vocabulary ~ ids_pref:test_age + ids_pref:cdi_age + ids _pref:protocol + ids_pref + cdi_age + gender + protocol + test_age + (1 | lab) + (1| participant)

And the final pruned model for the UK model was:

CDI_vocabulary ~ ids_pref:test_age + ids_pref:cdi_age + ids _pref:protocol + ids_pref + cdi_age + gender + protocol + test_age + (1| participant)

In both finalized models, we controlled for the random effects (i.e., random intercept) of the participants to handle repeated measurements because some participants have completed CDI twice. The lmerTest R package (Kuznetsova, Brockhoff, & Christensen, 2017) was used to run the model using Type III error Sum of Square for consistency with the original ManyBabies 1 study.

Our second criterion involved a power calculation which we were unable to complete and was therefore not implemented (see above).

In addition, our pre-registered analysis plan included a Kappa test on the possibility of collinearity between *test_age* and *cdi_age*. A 'c' number higher than 10 from the Kappa test would result in residualizing *test_age* against *cdi_age* to allow the use of both in our models. We carried out the Kappa test and found a value of 3.26, suggesting that we did not violate the multicollinearity assumption and can include both age at CDI test and age at IDS test in the same planned mixed-effect models.

Separate NAE and UK models

A summary of the NAE and UK models can be found in Table 2. In the NAE model, one of the predictors in the model was statistically significant: child's CDI age. In the UK model, there were two significant effects: the main effect of age at CDI test and the main effect of age at time of IDS preference testing. Neither the main effect of ids_pref (Research Question #1) nor the interaction(s) of ids_pref with age (Research Question #2) were significant. In addition, we ran a preregistered Bayesian analysis to probe the strength of the evidence in favor of the null effects for our research questions (see Supplement 2). Bayes factors ranged between .87 and 1.04, which did not reach our established threshold (.33) of support for the null hypothesis. These were calculated using the 'brms' packaged in r (version 2.18.0; Bürkner, 2017).

Combined NAE and UK model

For the third research question, we ran a third analysis parallel to the first two, but that combined across the North American and UK samples, and included a new variable:

Dialect: NA/UK as an effect coded fixed factor

For this analysis, only North American infants in the more restricted 17.5-18.5 and 23.5-24.5 age ranges were included, and proportional scores (raw score divided by the total number of items in the CDI) were used, to ensure greater comparability across the samples. The initial model fitted to the data had the following structure:

CDI vocabulary ~ ids_pref:test_age + ids_pref:cdi_age + ids_pref:protocol + ids_pref:dialect + ids_pref + cdi_age + gender + dialect + protocol + test_age + (1 +

ids_pref:test_age + ids_pref:cdi_age + ids_pref:protocol +ids_pref:dialect + ids_pref + cdi_age + gender + dialect | lab)

After pruning, the final model was:

CDI vocabulary ~ ids_pref:test_age + ids_pref:cdi_age + ids_pref:protocol + ids_pref:dialect + ids_pref + cdi_age + gender + dialect + protocol + test_age + (1 | lab) + (1 | participant)

Just like the NAE and UK model, we controlled for the random effects (i.e., random intercept) of the participants because of repeated measurements of CDI (see Table 2 for summary).

As with the individual models, we found a significant main effect of age at the time when CDI was collected. We also found a significant main effect of dialect, with UK infants showing a higher vocabulary proportional score than North American infants. Finally, we found a significant main effect of gender in which females have a higher proportion of vocabulary than males. However, as with the individual models, we did not find any significant effect of the IDS preference nor any significant interaction between IDS preference and other factors, including dialect (Research Question #3).

The calculated Bayes factor for the main effect of IDS preference was .75. For the IDS preference and age of IDS test interaction it was .35, and for the interaction between dialect and IDS preference it was .37. These did not quite reach our established threshold (.33) of support for the null hypothesis.

Exploratory Analysis

Next we conducted an analysis including all of the data across all 21 laboratories to test our hypotheses with a larger and more diverse sample. At the time of preregistration, it was unknown if we would have enough non-English laboratories to conduct additional analyses, so the following analysis was not registered and should be considered exploratory. In addition to the NAE and UK English-speaking laboratories, we included data from German (including Swiss German), Dutch, French, Norwegian, and Korean-speaking laboratories, as well as an additional English sample from New Zealand.

For this exploratory analysis, we took a different approach from our confirmatory analysis, in two primary ways. First, we generated normed data for all of our datasets using a process described in Frank et al. (2017b), rather than relying on proportional scores. This approach better accounts for differences across the instruments and languages and is more sensitive to age effects within a sample (see below). Second, we used a beta regression model to better capture the structure of the percentile scores, which are not normally distributed and bounded between 0 and 1.

To create the normed data that could be compared across instruments, data for German, Dutch, French, Norwegian, Korean, British and North American English vocabulary score norms were retrieved from WordBank (Frank et al., 2017b; retrieved April 14th, 2022).[2] The vocabulary score norms from the countries' participating labs were divided by age for each CDI instrument. Norming data for all instruments, except Norwegian (for which this process had already been conducted for a prior study), were collected using the child's age in months, as this is how they are reported in the WordBank system. Given the rapid expansion of vocabulary during the period from 18-24 months, we wanted our norms to capture a more granular level of

[2] The New Zealand sample used the North American instrument and was therefore normed on the North American data.

analysis at the level of individual days. Therefore, a quantile regression was performed for each country's norming data. This was done using 1 percent quantiles with the R function "gcrq" from the package "quantregGrowth" (Muggeo, 2023) and followed the procedure introduced in Kartushina et al., 2022). This process resulted in rankings from 1 to 99 for each infant age, in days, thereby controlling for vocabulary size differences attributable to infants' sex, age and language. Each of these rankings interpolated data from months to days by dividing each month by the average length of a month in days (30.457 days). Raw scores from our participants were then compared to the raw score derived from the norms to the closest age in days. The column containing a CDI score with the closest value to our participant's score was assigned as the participant's percentile ranking. Participants outside the age range used for the CDI were removed from further analysis. To adjust for the fact that reporting age in months in the norms would have been centered on the middle of the month, 15 days were added to the reported age of each child when comparing to the norms. For the Norwegian data, no adjustment was used since the data were originally collected in days. IDS preference and test_age were also z-transformed to more easily interpret the estimates.

daily_percentile: The percentile vocabulary score normed to each language at the specific age in days for each child.

z.IDS_pref: Standardized mean preference for IDS (described above) as a centered continuous variable. The intercept of this factor represents the CDI value for the grand mean of ids_pref.

z.age_months: Age (in months) at time of IDS preference testing, entered as a centered continuous variable. The intercept of this factor represents the CDI value for the grand mean of test_age.

CDI.agerange: Age (in months) at which the CDI measure was taken, entered as a factor variable. The intercept of this factor represents the CDI value at 18 months old.

gender: Male/female as an effect coded fixed factor to test for effects of gender. The intercept of this factor represents a hypothetical CDI value where gender is neither male nor female.

method: The testing protocol used to assess IDS preference (3 levels: central fixation, eye-tracking, and head-turn preference), entered as a deviation coded factor. The intercept of this factor represents the difference between CDI values for the mean of each level (e.g., central fixation) and the grand mean of all levels.

nae: TRUE/FALSE coded fixed factor to test for effects of North-American English. The intercept of this factor represents FALSE.

For the model analysis, a beta regression was conducted using the function glmmTMB in the package of the same name (1.1.2; Brooks et al., 2017). To evaluate model assumption of multicollinearity the function "vif" was used (R package 'car'; version 3.0-12; Fox et al, 2019). A "full-null" model paradigm was used to account for the large number of viable possible models that could be used given our data (Forstmeier & Schielzeth, 2011). Overdispersion was also investigated by checking that the dispersion parameter for the full model was not above 1.

Full Model:

daily_percentile ~ z.IDS_pref * z.age_months + z.IDS_pref * CDI.agerange + z.IDS_pref * method + z.IDS_pref * nae + gender + (1 |labid) + (1 | subid_unique)

Null Model:

daily_percentile ~ z.age_months + CDI.agerange + method + nae + gender + (1 | labid) + (1 | subid_unique)

The full-null model comparison was performed with the function “anova” (R package “lmtest”; version 0.9.40; Zeileis and Hothorn 2002).

Across all the labs 668 datapoints were collected. The final sample consisted of a total sample of 625 data points with 447 unique infants from 21 labs using 3 different methods. 43 observations were removed from the initial sample due to incomplete data, specifically, missing a computed percentile because their age range was outside of the age for the CDI from their respective assessments.

There was no evidence of collinearity between any of the predictors (maximum VIF was 1.17). No evidence of overdispersion was found (dispersion parameter = 0.5366). There was no evidence from the likelihood ratio test of the full-null model comparison: $\chi 2 = 6.612$, df = 6, $p = 0.358$ that the interactions of IDS preference and the other factors are associated with the daily percentile vocabulary scores. Thus, we performed another model comparison without IDS preference in any interactions as seen below.

Full Model (No Interactions):

daily_percentile ~ z.IDS_pref + CDI.agerange + method + nae + gender + (1 |labid) + (1 | subid_unique)

Null Model (No Interactions):

daily_percentile ~ CDI.agerange + method + nae + gender + (1 | labid) + (1 | subid_unique)

There was no evidence of overdispersion in this new full model without interactions (dispersion parameter = 0.532). There was still no evidence that including IDS preference in the model statistically significantly increased the model fit the full-null model comparison $\chi 2$ = 1.441, df = 1, p = 0.230. See **Table 3** for the model estimates for the full and null models without interactions.

In sum, our analyses do not support the hypotheses that preference for IDS as measured in the ManyBabies preference task is associated with later vocabulary (measured via CDI parental report), even with this expanded dataset and more granular level of analysis with respect to percentile scoring. We also did not find support for an interaction effect with age at testing (neither for the preference test nor the age of CDI collection), or method. We did find significant main effects in two of our nuisance variables – gender and language grouping.

Discussion

Our primary goal in this study was to test for a possible relation between infants' preference for IDS (as measured in a large scale IDS preference study) and later vocabulary knowledge (as measured by CDI parental reports) at 18 and 24 months. Secondarily, we were interested in knowing whether any such relation might change over development or based on the infant's linguistic experience. Across both our preregistered and exploratory analyses, we found no evidence for a relation between IDS preference and later vocabulary. However, Bayesian analyses did not reach our pre-established threshold to support the null hypothesis, so we cannot take this null finding as direct evidence that such a relation does not exist. Furthermore, the effect of age and language community on this relation was not significant.

Before exploring the implications of these null findings, a brief note on some unanticipated findings with "nuisance" variables is in order. We found significant effects of the age at which the CDI was collected, across several models. This result was expected (and would have been troubling if not found) for the UK-only model, which used raw scores - it is simply capturing that infants' vocabulary grows from 18 to 24 months. However, we also found this effect in the North American English (NAE)-only model, which would not be predicted given these were percentile scores, which should not show a systematic increase with age, due to norming. It's important to note that these scores were collected using the new web-CDI version of the North American CDI during an initial pilot phase (DeMayo et al., 2021). One possible explanation is that the percentile scores used were based on older normed data collected via a different approach, which may not have fully accounted for age effects in more recent web-based samples. However, we cannot be certain why this effect emerged. The effect of age that we found, although significant, was not large in this model and is unlikely to have an impact on the conclusion of our main research question.

There was also a significant main effect of the age of IDS preference testing in the model for the UK sample (but not the NAE sample) in predicting CDI scores. The reason for this finding is unknown, but it may have been an artifact of non-random assignment of infant age of testing in the CDI data collection - i.e. it is possible that the age window we imposed for follow-up CDI data collection might have unintentionally given rise to cohort effects within the IDS preference sampling. Finally, a significant main effect of the "language zone" in the combined model (i.e., UK vs. NAE) suggests that the proportional measure used as the outcome measure in that model did not fully calibrate between the two languages' instruments. A similar main effect of language zone (NAE vs. others) emerged in our exploratory analysis which used percentile

scores and an alternative analytic approach, suggesting that even with the percentile-based approach, we were not fully able to calibrate across the instruments. These findings reinforce the challenges for work that combines and/or compares across vocabulary instruments within a single analysis. However, despite these challenges, the failure to find a relation between IDS preference and vocabulary was consistent across the exploratory and confirmatory analyses. We also found an effect of gender in our analyses, with males scoring on average lower than females, as is common in the literature (e.g., Eriksson et al., 2012; Frota et al., 2016; Nylund et al., 2021; Sansavini et al., 2010; Schults & Tulviste, 2016).

Reliability of individual differences in preference studies

One possible lens with which to understand our null findings is to raise the question of whether infant preference measures of the type used in our study actually capture meaningful individual variation. At a much broader level, this question raises the often underappreciated distinction between "differential" research approaches, which emphasize individual differences and are thus optimized to maximize between-subjects variance, and "experimental" approaches which emphasize group-level differences between (experimentally manipulated) conditions and are thus optimized to maximize within-subjects variance (Draheim et al., 2019). It is possible that a group of infants would show a robust preference for one stimulus type over another, without it being the case that individual variation in the size of that preference is meaningful. More concretely, although it is possible that larger differences in the measured looking toward IDS over ADS for a given infant capture real differences in that infant's underlying preference for IDS relative to another infant who showed longer looking toward ADS, it is also possible that differences in performance are simply capturing attentional differences in the task, or transient effects of distraction or mood on the day of testing.

A way to probe this analytically is to examine the extent to which infant preference as measured in the laboratory is stable across repeated testing. A separate follow-on study (Schreiner et al., under review) to the original ManyBabies study did just this with a subset of the sample used in our analysis. Specifically, a total of 158 infants across 7 laboratories were brought back for a second day of testing about one week after the first test (range = 1–31 days). Although an IDS preference at the group level was also found in the retest session, replicating the group effect of preference for IDS, they found a lack of consistent evidence for test-retest reliability in measures of infants' speech preference at the individual level.

A second analytic approach to examining the reliability of the IDS preference task is to examine its internal consistency - the extent to which infants show a consistent preference for IDS vs. ADS across trials within the same test session. Byers-Heinlein et al. (2022) undertook such an analysis and found that the internal consistency measured via the intraclass correlation coefficient across the 8 trial pairs was .14. Note that values below .5 indicate poor reliability (Koo & Li, 2016), so again this analysis indicates that this task is not a reliable measure of individual preference.

The goal in the current study was to investigate the correlation between IDS preference in this task and CDI, and our ability to do so crucially depended on the reliability of both measures, as well as the sample size. The CDI is optimized to measure individual differences and test-retest reliability of the CDI is quite good, estimated to be between .86 and .90 (Dale et al., 1989; Jahn-Samilo et al., 2001; Simonsen et al., 2014). Combined with the estimated reliability of the IDS preference task as well as the sample sizes of our groups, a sensitivity analysis revealed that our design had 80% power to detect a true correlation between IDS preference and CDI of .46 or greater for the NAE sample, and .89 or greater for the UK sample. Thus, given the possibly low

reliability of the IDS preference task, even with the large samples we were able to collect, our study would have been underpowered to detect more moderately-sized correlations. A much larger sample, or ideally a more reliable measure of infants' individual IDS preferences, would in the future be more revealing of whether a relation between attention to IDS and vocabulary development exists.

Implications for Theory

The above commentary raises concerns about the extent to which we were able to capture individual variation in IDS preference sufficiently well to detect an effect, and it is worth noting that our Bayes analysis did not permit us to claim direct evidence in favour of the null hypothesis. Indeed, our Bayes Factor for the key factor of interest was close to 1, indicating close to equal support for the null hypothesis and for an effect of IDS preference. Moreover, our findings are in contradiction to those of Droucker et al. 2013, who found a significant relationship between preference for IDS and CDI scores at 18 months. There are some methodological differences between that study and ours, including the sample size and population tested, their use of the Words and Gestures form rather than the Words and Sentences form, and the details of how preference for IDS was measured (e.g. number of trials, specific nature of the IDS and ADS stimuli). But it is not possible at this point to know whether a methodological difference led to the different findings or simple statistical variation. Therefore, we must nonetheless consider the implications of the possibility that our findings are a true null result - i.e., that there is no relation between an infant’s underlying individual preference for IDS and their later language development.

This finding needs to be contextualized within, on the one hand, experimental evidence for the benefits of IDS in infant language processing tasks (e.g., Thiessen et al., 2005; Ma et al.,

2011), widespread cross-cultural/cross-linguistic IDS usage (Hilton et al., 2022), and correlational findings of a relation between caregiver usage of IDS and infant language development within Western contexts (e.g. Weisleder & Fernald, 2013), and on the other, large cross-cultural differences in the usage of IDS that do not appear to be reflected in cultural differences in language acquisition milestones (e.g. Cristia et al., 2019; Casillas et al., 2020, 2021). In other words, there is compelling evidence that IDS can be important for language development, but not necessarily for all cultures.

One possibility, therefore, is that preference for IDS, rather than capturing a construct of relevance for language outcomes, is capturing individual differences in infant experience with IDS in a way that influences the extent to which IDS matters in the development of an individual child. Possible evidence in favour of this idea comes from the stronger preference for IDS found in the original ManyBabies 1 study for North American English-learning infants compared to infants learning other languages. This finding could be driven by differential experience with IDS across different languages, since North American English IDS is often considered to be a relatively extreme version of IDS (e.g. Byers-Heinlein et al., 2021), which could lead to systematic differences in the importance of IDS in the language development process, as infants' perceptual systems tune to the ambient language experience. (However, there is an important confound in that study, in that infants were all tested with North American English IDS.) Alternatively, IDS might be similarly important to all infants regardless of cross-cultural/cross-linguistic differences in experience, but in a passive way, such that individual differences in preference are simply irrelevant to the IDS effect. This latter alternative is possible, but would contradict mainstream theories that a crucial role for IDS is in drawing the infant's attention to the speech signal (e.g. Soderstrom, 2007). Finally, it is possible that the gap between the time

when IDS preference was tested and the vocabulary size was reported was too large to reveal a reliable relation between IDS and language development, as it might be more robust when measured concurrently. The more extreme possibility, that IDS plays no role at all in supporting language development, seems even more unlikely given the wealth of evidence (at least in Western contexts) to the contrary.

Limitations and Future Directions

The primary limitation of this study is what we have already discussed in detail, which is the low test-retest reliability of the IDS preference measure for capturing individual differences. In addition and relatedly, our sample, although larger than many infant preference studies, may still have been underpowered to detect a relation between our primary variables of interest. Similarly, while our study was more geographically diverse than many of its type, our sample was one of convenience and not representative of the populations from which it was sampled, let alone representative of global diversity.

Our findings, together with those of Schreiner et al. (under review), point to the importance of further considering the relation between group-based effects like the preference for IDS and individual variation within those effects. This approach is important both for general theory (understanding the role that IDS plays in language development) and due to the potential for perceptual measures of this type to be used in the study of special developmental populations and intervention (e.g. Droucker et al., 2013).

Conclusion

Across 467 infants from 21 labs and several analytic approaches, our findings provide little support for a relation between preference for infant-directed speech as measured by laboratory perceptual tests, and later vocabulary measured by parental reports. A lack of test-retest

reliability suggests that we may not be sufficiently capturing individual variation in infant preference to robustly detect relations between IDS preferences and vocabulary, and points to the importance of differentiating between group effects and individual differences in interpreting infant preference data. Future research should strive to improve the reliability of preference measures and expand the sample diversity. By exploring the relation between group effects and individual differences, we can gain a deeper understanding of the complex interplay between IDS, infant preferences, and language development in diverse populations.

Table 1. Participant demographics by laboratory. HPP = Headturn Preference Procedure. CF = Central Fixation. ET = Eye Tracking.

Laboratory (Protocol)	**Language (CDI form)**	**MB1 Age Range**	**N 18 Month CDI (N Males)**	**N 24 Month CDI (N Males)**	**Country**
bllumanitoba (HPP)	Canadian English (webcdi)	147 - 446 days	39 (21)	32 (16)	Canada
unlvmusiclab (CF)	North American English (webcdi)	100 - 179 days	12 (4)	10 (3)	United States
princetonbabylab (HPP)	North American English (webcdi)	101 - 448 days	18 (9)	16 (9)	United States
infantcogUBC (ET)	Canadian English (webcdi)	189 - 262 days	17 (9)	9 (5)	Canada
lllliv (ET)	British English (Oxford CDI)	213 - 451 days	14 (8)	9 (2)	United Kingdom
babylabbrookes (CF)	British English (Oxford CDI)	92 - 445 days	17 (8)	14 (4)	United Kingdom
purdueinfantspeech	North	273 - 446 days	27 (11)	23 (13)	United States

(HPP)	American English (webcdi)				
weescienceedinburgh (CF)	British English (Oxford CDI)	184 - 250 days	22 (10)	4 (3)	United Kingdom
bcrlunlv (CF)	North American English (webcdi)	372 - 455 days	2 (1)	15 (8)	United States
infantlanglabutk (HPP)	North American English (webcdi)	189 - 452 days	29 (16)	43 (22)	United States
babylablmu (ET)	German (FRAKIS)	276 - 442 days	11 (4)	28 (13)	Germany
lcduleeds (ET)	British English (Oxford CDI)	368 - 446 days	4 (2)	8 (3)	United Kingdom
babylabvuw (CF)	New Zealand English (webcdi)	107 - 278 days	10 (6)	11 (6)	New Zealand
babylabnijmegen (HPP)	Dutch (NCDI)	206 - 358 days	23 (9)	27 (12)	Netherlands
chosunbaby (HPP)	Korean (K-CDI)	183 - 451 days	21 (11)	19 (8)	Korea

lscppsl (ET)	French (French short form)	369 - 442 days	3 (2)	5 (2)	France
infantstudiesubc (HPP)	Canadian English (webcdi)	94 - 231 days	9 (4)	16 (9)	Canada
babylingoslo (ET)	Norwegian (Norwegian CDI)	184 - 270 days	17 (9)	16 (8)	Norway
infantcoglablouisville (CF)	North American English (webcdi)	281 - 330 days	11 (5)	5 (1)	United States
weltentdeckerzurich (ET)	Swiss German (FROSCH)	367 - 384 days	8 (6)	0	Switzerland
babylabpotsdam (HPP, SS)	German (FRAKIS)	249 - 364 days	27 (13)	17 (9)	Germany

Table 2. Coefficient Estimates from the preregistered mixed-effects models (North American primary sample, UK primary sample, and the combination of the two samples)

	NAE Full Model (Standard Error)	UK Full Model (Standard Error)	NAE+UK Full Model (Standard Error)
(Intercept)	31.269***	182.366***	0.324***
	(6.471)	(10.307)	(0.019)
CDI.z_age_months (Months)	1.157***	33.925***	0.052***
	(0.341)	(2.345)	(0.003)
gender Male (Female as reference)	-1.407	10.142	0.029*
	(1.888)	(10.156)	(0.011)
z_age_months	-0.104	-8.461*	-0.004
	(0.629)	(3.340)	(0.004)
method Eye Tracking (Central Fixation as reference)	6.537	-12.223	-0.014
	(6.580)	(12.511)	(0.025)
method Head Turn Preference (Central Fixation as reference)	-15.000		0.001

	(11.927)		(0.039)
z.IDS_pref	-33.917	25.907	0.008
	(24.793)	(27.203)	(0.040)
Method (Central Fixation vs. Eye Tracking) * z.IDS_pref	26.253	-25.499	-0.057
	(25.436)	(29.465)	(0.053)
Method (Central Fixation vs. Head Turn Preference) x z.IDS_pref	-56.683		0.043
	(49.501)		(0.088)
CDI_age : z.IDS_pref	0.555	10.598	0.010
	(1.047)	(7.059)	(0.008)
Test_age : z.IDS_pref	2.182	-8.352	-0.005
	(2.083)	(8.978)	(0.011)
language_zone UK (NAE as reference)			0.097***
			(0.025)
language_zone : z.IDS_pref			0.029

			(0.058)
SD (Intercept subid_unique)	23.546	69.528	0.144
SD (Observations)	14.803	46.997	0.142
SD (Intercept labid)	5.900		0.039
Num.Obs.	307	92	397
R2 Marg.	0.033	0.580	0.399
R2 Cond.	0.738	0.868	0.715
AIC	2823.2	1023.5	-65.6
BIC	2875.4	1051.3	-1.9
RMSE	9.99	28.69	0.11

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Table 3. Coefficient estimates from exploratory mixed effects analyses (no interactions model, null model and interactions model)

	No Interactions Model	Null Model	[a]Interactions Model

(Intercept)	0.303	0.304	0.283
	(0.186)	(0.187)	(0.191)
z.IDS_pref	0.072		0.044
	(0.06)		(0.128)
z.age_months	-0.075	-0.067	-0.072
	(0.07)	(0.069)	(0.07)
CDI.agerange 24 months (18 months as reference level)	-0.096	-0.095	-0.089
	(0.073)	(0.074)	(0.074)
Method hpp (eye-tracking as reference level)	0.345	0.350	0.373
	(0.239)	(0.241)	(0.251)
method singlescreen (eye-tracking as reference level)	0.448	0.441	0.442
	(0.240)	(0.241)	(0.249)
NAE TRUE (FALSE as reference level)	-0.918***	-0.921***	-0.94***

	(0.185)	(0.187)	(0.195)
Gender Male (Female as reference level)	-0.435***	-0.438***	-0.432***
	(0.119)	(0.119)	(0.119)
z.IDS_pref : z.age_days			0.012
			(0.071)
z.IDS_pref : CDI.agerange			0.048
			(0.074)
z.IDS_pref : method hpp			0.240
			(0.180
z.IDS_pref : method singlescreen			0.036
			(0.190)
z.IDS_pref : nae TRUE			-0.261
			(0.150)
SD (Intercept subid_unique)	1.051	1.054	1.039

SD (Intercept labid)	0.222	0.226	0.250
Num.Obs.	625	625	625
R2 Marg.	0.176	0.171	0.189
R2 Cond.	0.990	0.990	0.990
AIC	-131.8	-132.4	-131.8
BIC	-83	-88	-83
RMSE	0.110	0.110	0.110

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

[a] The estimates from this model cannot be interpreted since the interaction effects did not improve model fit over the null model. The main effects are also uninterpretable due to the insignificant interaction effects (Engqvist 2005; Lorah 2020).

Figure 1

Sample distribution by age, laboratory, and language type.

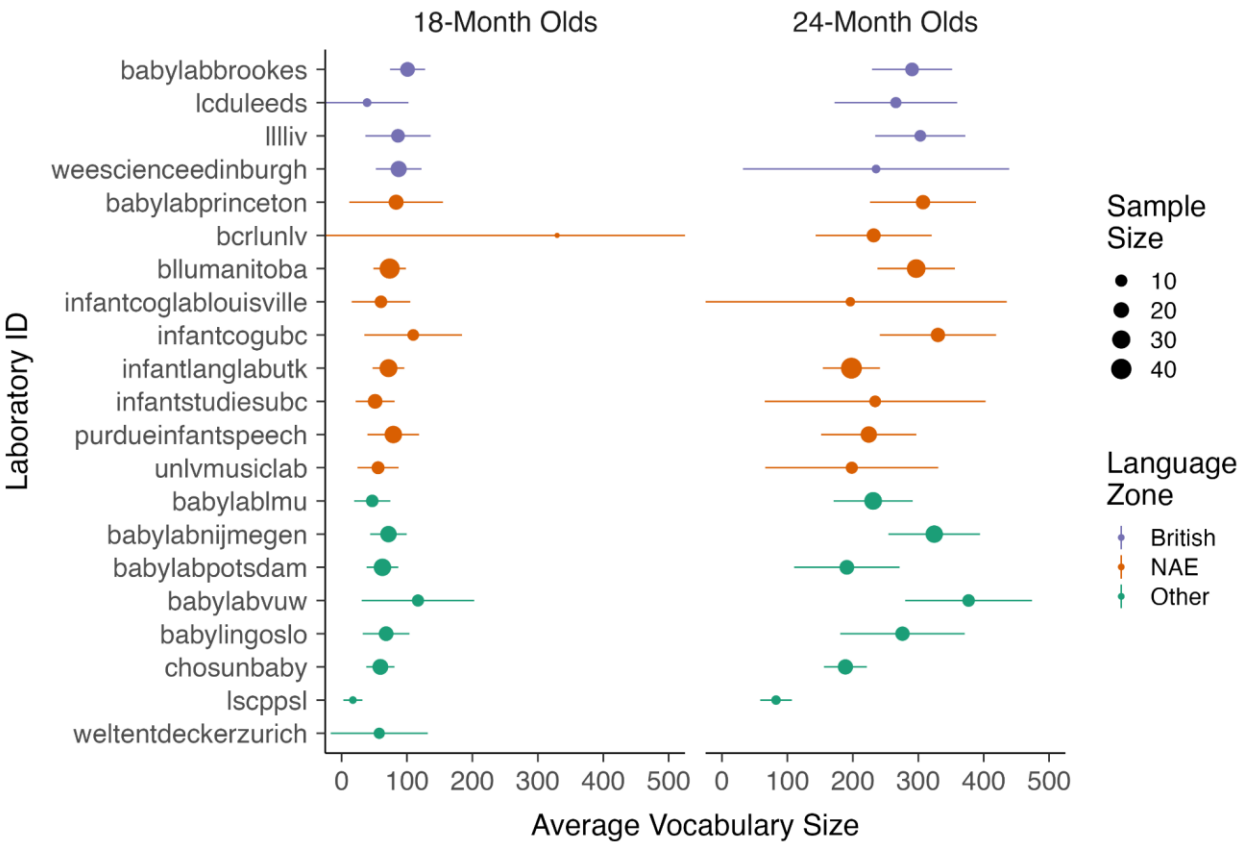
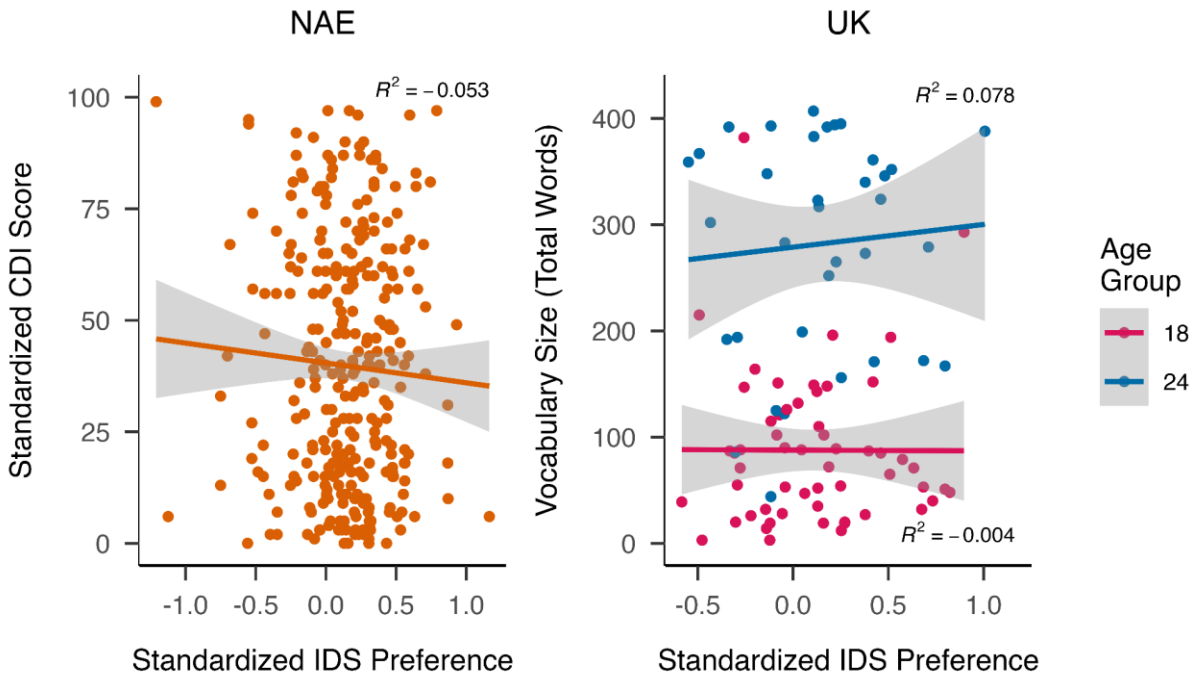


Figure 2

Scatterplots of Correlations between Vocabulary (either as standardized CDI score or Total Vocabulary size) and IDS Preference



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 lme4. *Journal of Statistical Software, 67*(1), 1-48.

Begus, K., Gliga, T., & Southgate, V. (2016). Infants' preferences for native speakers are associated with an expectation of information. *Proceedings of the National Academy of Sciences, 113*, 12397-12402.

Blount, B. G. (1972). Parental speech and language acquisition: Some Luo and Samoan examples. *Anthropological Linguistics*, 119-130.

Blount, B. G. (1984). Mother-infant interaction: Features and functions of parental speech in English and Spanish. *The development of oral and written language in social contexts, 13*, 3-29.

Blount, B. G., & Padgug, E. J. (1976). *Mother and father speech: Distribution of parental speech features in English and Spanish. Papers and Reports on Child Language Development*. ERIC Clearinghouse.

Blount, B. G., & Padgug, E. J. (1977). Prosodic, paralinguistic, and interactional features in parent-child speech: English and Spanish. *Journal of Child Language, 4*(1), 67-86.

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., ... & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. The R journal, 9(2), 378-400.

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software, 80(1),* 1-28.<doi:10.18637/jss.v080.i01>

Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2022). Six solutions for more reliable infant research. *Infant and Child Development*, *31*(5), e2296. https://doi.org/10.1002/icd.2296

Byers-Heinlein, K., Tsui, A. S. M., Bergmann, C., Black, A. K., Brown, A., Carbajal, M. J., ... & Wermelinger, S. (2021). A multilab study of bilingual infants: Exploring the preference for infant-directed speech. Advances in Methods and Practices in Psychological Science, 4(1), 2515245920974622.

Casillas, M., Brown, P., & Levinson, S. C. (2020). Early language experience in a Tseltal Mayan village. Child Development, 91(5), 1819-1835.

Casillas, M., Brown, P., & Levinson, S. C. (2021). Early language experience in a Papuan community. Journal of Child Language, 48(4), 792-814.

Champely, S. (2015). pwr: Basic Functions for Power Analysis. R package version 1.1-3. http://CRAN.R-project.org/package=pwr

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.

Cristia, A., Dupoux, E., Gurven, M., & Stieglitz, J. (2019). Child-directed speech is infrequent in a forager-farmer population: A time allocation study. Child development, 90(3), 759-773.

Dale, P. S., Bates, E., Reznick, J. S., & Morisset, C. (1989). The validity of a parent report instrument of child language at twenty months. Journal of Child Language, 16(2), 239-249.

DeMayo, B., Kellier, D., Braginsky, M., Bergmann, C., Hendriks, C., Rowland, C. F., ... & Marchman, V. (2021). Web-CDI: A system for online administration of the MacArthur-Bates Communicative Development Inventories. Language Development Research.

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. Psychological Bulletin, 145(5), 508.

Droucker, D., Curtin, S., & Vouloumanos, A. (2013). Linking infant-directed speech and face preferences to language outcomes in infants at risk for Autism Spectrum Disorder. *Journal of Speech, Language, and Hearing Research, 56*, 567-576.

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

Eriksson, M., Marschik, P. B., Tulviste, T., Almgren, M., Pereira, M. P., Wehberg, S., Marjanovič, U. L., Gayraud, F., Kovacevic, M., & Gallego, C. (2012). Differences between girls and boys in emerging language skills: Evidence from 10 language communities. British Journal of Developmental Psychology, 30(2), 326–343. https://doi-org.ezproxy.library.unlv.edu/10.1111/j.2044-835X.2011.02042.x

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.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories: User's guide and technical manual (2nd ed.).* Baltimore, MD: Brookes.

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

Fernald, A., & Morikawa, H. (1993). Common themes and cultural variations in Japanese and American mothers' speech to infants. *Child Development, 64*(3), 637-656.

Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology, 20*(1), 104.

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*(3), 477-501.

Floccia, C., Keren-Portnoy, T., DePaolis, R., Duffy, H., Delle Luche, C., Durrant, S., ... & Vihman, M. (2016). British English infants segment words only with exaggerated infant-directed speech stimuli. *Cognition*, *148*, 1-9.

Forstmeier, W., & Schielzeth, H. (2011). Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. Behavioral ecology and sociobiology, 65, 47-55.

Fox, J., Weisberg, S., Price, B., Adler, D., Bates, D., Baud-Bovy, G., & Bolker, B. (2019). car: Companion to Applied Regression. R package version 3.0-2. Website https://CRAN. R-project. org/package= car [accessed 17 March 2020].

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. (2017a). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*, 421-435.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017b). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677-694.

Frota, S., Butler, J., Correia, S., Severino, C., Vicente, S., & Vigário, M. (2016). Infant communicative development assessed with the European Portuguese MacArthur–Bates Communicative Development Inventories short forms. First Language, 36(5), 525–545. https://doi-org.ezproxy.library.unlv.edu/10.1177/0142723716648867

Ganger, J., & Brent, M. R. (2004). Reexamining the vocabulary spurt. *Developmental Psychology*, *40*(4), 621.

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.

Hamilton, A., Plunkett, K., & Schafer, G. (2000). Infant vocabulary development assessed with a British communicative development inventory. *Journal of Child Language*, *27*(3), 689-705.

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.

Hilton, C. B., Moser, C. J., Bertolo, M., Lee-Rubin, H., Amir, D., Bainbridge, C. M., ... & Mehr, S. A. (2022). Acoustic regularities in infant-directed speech and song across cultures. Nature Human Behaviour, 6(11), 1545-1556.

Jahn-Samilo, J., Goodman, J., Bates, E., & Sweet, M. (2001). Vocabulary learning in children from 8 to 30 months of age: A comparison of parental report and laboratory measures. Retrieved from https://crl.ucsd.edu/bates/papers/pdf/from-meiti/39-Jahn%20et%20al.2000.pdf

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

Kartushina, N., Mani, N., Aktan-Erciyes, A., Alaslani, K., Aldrich, N. J., Almohammadi, A., ... & Mayor, J. (2022). COVID-19 first lockdown as a window into language acquisition: associations between caregiver-child activities and vocabulary gains. Language Development Research, 2(1).

Katz, G. S., Cohn, J. F., & Moore, C. A. (1996). A combination of vocal f_0 dynamic and summary features discriminates between three pragmatic categories of infant-directed speech. *Child Development*, *67*(1), 205-217.

Kemler Nelson, D. G. K., Hirsh-Pasek, K., Jusczyk, P. W., & Cassidy, K. W. (1989). How the prosodic cues in motherese might assist language learning. Journal of Child Language, 16(1), 55-68.

Kempe, V., Brooks, P. J., Mironova, N., & Fedorova, O. (2003). Diminutivization supports gender acquisition in Russian children. *Journal of Child Language*, *30*(2), 471-485.

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

Kitamura, C., Thanavishuth, C., Burnham, D., & Luksaneeyanawin, S. (2001). Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language. *Infant Behavior and Development*, *24*(4), 372-392.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155–163. https://doi.org/10.1016/j.jcm.2017.10.001

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U. & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*(5326), 684-686.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. Journal of statistical software, 82, 1-26.

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.

Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., & Cristia, A. (2015). Mothers speak less clearly to infants than to adults: A comprehensive test of the hyperarticulation hypothesis. *Psychological science*, *26*(3), 341-347.

Masataka, N. (1996). Perception of motherese in a signed language by 6-month-old deaf infants. Developmental Psychology, 32(5), 874.

McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science*, *317*(5838), 631-631.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91-117.

Morikawa, H., Shand, N., & Kosawa, Y. (1988). Maternal speech to prelingual infants in Japan and the United States: Relationships among functions, forms and referents. *Journal of Child Language, 15*(2), 237-256.

Muggeo, V. M., (2023). Package 'quantregGrowth'. Statistics in Medicine, 25, 1369-1382.

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

Niwano, K., & Sugai, K. (2002). Acoustic determinants eliciting Japanese infants' vocal response to maternal speech. *Psychological Reports, 90*(1), 83-90.

Nylund, A., Ursin, P. A., Korpilahti, P., & Rautakoski, P. (2021). Vocabulary Growth in Lexical Categories Between Ages 13 and 24 Months as a Function of the Child's Sex, Child, and Family Factors. Frontiers in Communication, 6, 709045.

Papoušek, M., Papoušek, H., & Haekel, M. (1987). Didactic adjustments in fathers' and mothers' speech to their 3-month-old infants. *Journal of Psycholinguistic Research, 16*(5), 491-516.

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.

Piazza, E. A., Iordan, M. C., & Lew-Williams, C. (2017). Mothers consistently alter their unique vocal fingerprints when communicating with infants. *Current Biology, 27,* 3162-3167.

R Core Team. (2017). R: A Language and Environment for Statistical Computing. https://www.R-project.org/

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2014). Look who's talking: Speech style and social context in language input to infants are linked to concurrent and future speech development. *Developmental Science*, *17*(6), 880-891.

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2017). Look who's talking NOW! parentese speech, social context, and language development across time. *Frontiers in Psychology, 8,* 1008.

Robertson, S., von Hapsburg, D. & Hay, J. F. (2013). The effect of hearing loss on the perception of infant- and adult-directed speech. *Journal of Speech, Language, and Hearing Research, 56,* 1108-1119.

Sansavini, A., Bello, A., Guarini, A., Savini, S., Stefanini, S., & Caselli, M. C. (2010). Early development of gestures, object-related-actions, word comprehension and word production, and their relationships in Italian infants: A longitudinal study. Gesture, 10(1), 52–85. https://doi-org.ezproxy.library.unlv.edu/10.1075/gest.10.1.04san

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

Schreiner, M. S., Zettersten, M., Bergmann, C., Frank, M. C., Fritzsche, T., Gonzalez-Gomez, N., ... & Lippold, M. (under review). Limited evidence of test-retest reliability in infant-directed speech preference in a large pre-registered infant sample.

Schults, A., & Tulviste, T. (2016). Composition of Estonian infants' expressive lexicon according to the adaptation of CDI/Words and Gestures. First Language, 36(5), 485–504. https://doi-org.ezproxy.library.unlv.edu/10.1177/0142723716648864

Senju, A. & Csibra, G. (2008). Gaze following in human infants depends on communicative signals. *Current Biology, 18,* 668-671.

Shneidman, L. A., & Goldin-Meadow, S. (2012). Language input and acquisition in a Mayan village: How important is directed speech?. *Developmental Science*, *15*(5), 659-673.

Shneidman, L. A., Arroyo, M. E., Levine, S. C., & Goldin-Meadow, S. (2013). What counts as effective input for word learning?. *Journal of Child Language*, *40*(3), 672-686.

Shute, B., & Wheldall, K. (1995). The incidence of raised average pitch and increased pitch variability of British motherese speech and the influence maternal characteristics and discourse form. *First Language, 15,* 35–55.

Simonsen, H. G., Kristoffersen, K. E., Bleses, D., Wehberg, S., & Jørgensen, R. N. (2014). The Norwegian Communicative Development Inventories: Reliability, main developmental trends and gender differences. First Language, 34(1), 3-23.

Singh, L., Morgan, J. L., & Best, C. T. (2002). Infants' listening preferences: Baby talk or happy talk?. *Infancy*, *3*(3), 365-394.Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009). Influences of infant-directed speech on early word recognition. Infancy, 14(6), 654-666.Singh, L., Steven Reznick, J., & Xuehua, L. (2012). Infant word segmentation and childhood vocabulary development: a longitudinal analysis. Developmental science, 15(4), 482-495.

Snow, C. E. & Ferguson, C. A. (eds.) (1977). *Talking to Children*. Cambridge University Press.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. Developmental Review, 27(4), 501-532.

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.

Spinelli, M., Fasolo, M., & Mesman, J. (2017). Does prosody make the difference? A meta-analysis on relations between prosodic aspects of infant-directed speech and infant outcomes. *Developmental Review*, *44*, 1-18.

The ManyBabies 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.

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.

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.

Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., & Amano, S. (2007). Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition*, *103*(1), 147-162.

Zeidner, M. (1983). 'Kitchie-koo'in modern Hebrew: the sociology of Hebrew baby talk. *International Journal of the Sociology of Language*, *1983*(41), 93-114.

Zeileis A, Hothorn T (2002). Diagnostic Checking in Regression Relationships. R News, 2(3), 7–10.

Zettersten, M., Cox, C., Bergmann, C., Soderstrom, M., Tsui, A.S.T., Mayor, J., Lundwall, R.A., Lewis, M., Kosie, J.E., Kartushina, N., Fusaroli, R., Frank, M.C., Byers-Heinlein, K., Black, A.K., & Mathur, M.B. (in prep). Evidence for infant-directed speech preference is consistent across large-scale, multi-site replication and meta-analysis.