Limited evidence of test-retest reliability in infant-directed speech preference in a large preregistered infant experiment

Melanie S. Schreiner[1,2], Martin Zettersten[3,4], Christina Bergmann[5], Michael C. Frank[6], Tom Fritzsche[7], Nayeli Gonzalez-Gomez[8], Kiley Hamlin[9], Natalia Kartushina[10], Danielle J. Kellier[11], Nivedita Mani[1,2], Julien Mayor[10], Jenny Saffran[3], Mohinish Shukla[12], Priya Silverstein[13, 14], Melanie Soderstrom[15], & Matthias Lippold[1,2]

[1] University of Goettingen

[2] Leibniz Science Campus PrimateCognition

[3] University of Wisconsin-Madison

[4] Princeton University

[5] Max Planck Insitute for Psycholinguistics

[6] Stanford University

[7] University of Potsdam

[8] Oxford Brookes University

[9] University of British Columbia

[10] University of Oslo

[11] University of Pennsylvania

[12] Università di Padova

[13] Institute for Globally Distributed Open Research

[14] Ashland University

[15] University of Manitoba

Author Note

Acknowledgements. This work was supported in part by a Leibniz ScienceCampus Primate Cognition seed fund awarded to MSc and ML, a grant from the Research Council of Norway (project number 301625) and its Centres of Excellence funding scheme (project number 223265) awarded to NK, an ERC Grant (agreement number 773202 – ERC 2017, "BabyRhythm") awarded to MSh, a ManyBabies SSHRC Partnership Development Grant awarded to MSo, and a grant from the NSF awarded to MZ (NSF DGE-1747503).

Conflict of Interest Statement. The authors declare that there are no conflicts of interest for this work.

Data Availability Statement. All code for reproducing the paper is available at https://github.com/msschreiner/MB1T. Data and materials are available on OSF (https://osf.io/zeqka/?view_only=e027502f4e7f49408cfb2cba38f7b506).

CRediT author statement. Outside of the position of the first, the second, and the last author, authorship position was determined by sorting authors' last names in alphabetical order. An overview of authorship contributions following the CRediT taxonomy can be viewed here: https://docs.google.com/spreadsheets/d/1jDvb0xL1U6YbXrpPZ1UyfyQ7yYK9aXo002UaArqy35U/edit?usp=sharing.

Correspondence concerning this article should be addressed to Melanie S. Schreiner, Gosslerstr. 14, 37073 Göttingen. E-mail: melanie.schreiner@psych.uni-goettingen.de

Abstract

Test-retest reliability — establishing that measurements remain consistent across multiple testing sessions — is critical to measuring, understanding, and predicting individual differences in infant language development. However, previous attempts to establish measurement reliability in infant speech perception tasks are limited, and reliability of frequently-used infant measures is largely unknown. The current study investigated the test-retest reliability of infants' preference for infant-directed speech over adult-directed speech in a large sample (N=158) in the context of the ManyBabies1 collaborative research project (Frank et al., 2017; ManyBabies Consortium, 2020). Labs were asked to bring in participating infants for a second appointment retesting infants on their preference for infant-directed speech. This approach allowed us to estimate test-retest reliability across three different methods used to investigate preferential listening in infancy: the head-turn preference procedure, central fixation, and eye-tracking. Overall, we found no consistent evidence of test-retest reliability in measures of infants' speech preference (overall r = .09, 95% CI [-.06,.25]). While increasing the number of trials that infants needed to contribute for inclusion in the analysis revealed a numeric growth in test-retest reliability, it also considerably reduced the study's effective sample size. Therefore, future research on infant development should take into account that not all experimental measures may be appropriate for assessing individual differences between infants.

Keywords: language acquisition; speech perception; infant-directed speech; adult-directed speech; test-retest reliability

Word count: 5383

Limited evidence of test-retest reliability in infant-directed speech preference in a large preregistered infant experiment

Obtaining a quantitative measure of infants' cognitive abilities is an extraordinarily difficult endeavor. The most frequent way to assess what infants know or prefer is to track overt behavior. However, measuring overt behavior at early ages presents many challenges: participants' attention span is short, they do not follow instructions, their mood can change instantly, and their behavior is often inconsistent. Therefore, most measurements are noisy and the typical sample size of an infant study is small (around 20 infants per group), resulting in low power (Oakes, 2017). In addition, there is individual and environmental variation that may add even more noise to the data (e.g., Johnson & Zamuner, 2010). Despite these demanding conditions, reliable and robust methods for assessing infants' behavior are critical to understanding development.

In order to address these challenges, the ManyBabies collaborative research consortium was formed to conduct large-scale, conceptual, consensus-based replications of seminal findings to identify sources of variability and establish best practices for experimental studies in infancy (Frank et al., 2017). The first ManyBabies collaborative research project (hereafter, MB1, ManyBabies Consortium, 2020) explored the reproducibility of the well-studied phenomenon that infants prefer infant-directed speech (hereafter, IDS) over adult-directed speech (hereafter, ADS, Cooper & Aslin, 1990). Across many different cultures, infants are commonly addressed in IDS, which typically is characterized by higher pitch, greater pitch range, and shorter utterances, compared to the language used between interacting adults (Fernald et al., 1989). A large body of behavioral studies finds that infants show increased looking times when hearing IDS compared to ADS stimuli across ages and methods (Cooper & Aslin, 1990; see Dunst, Gorman, & Hamby, 2012 for a meta-analysis). This attentional enhancement is also documented in neurophysiological studies showing increased neural activation during IDS compared to

ADS exposure (Naoi et al., 2012; Zangl & Mills, 2007). IDS has also been identified as facilitating early word learning. In particular, infants' word segmentation abilities (Floccia et al., 2016; Schreiner & Mani, 2017; Singh, Nestor, Parikh, & Yull, 2009; Thiessen, Hill, & Saffran, 2005) and their learning of word-object associations (Graf Estes & Hurley, 2013; Ma, Golinkoff, Houston, & Hirsh-Pasek, 2011) are enhanced in the context of IDS. In sum, several lines of evidence suggest that IDS is beneficial for early language development.

Within MB1, 67 labs contributed data from 2,329 infants showing that babies generally prefer to listen to IDS over ADS. Nevertheless, the overall effect size of $d = 0.35$ was smaller than a previously reported meta-analytic effect size of $d = 0.67$ (Dunst et al., 2012). The results revealed several additional factors that influenced the effect size. First, older infants showed a larger preference of IDS over ADS. Second, the stimulus language was linked to IDS preference, with North American English learning infants showing a larger IDS preference than infants learning other languages. Third, comparing the different methods employed, the head-turn preference procedure yielded the highest effect size, while the central fixation paradigm and eye-tracking methods revealed smaller effects. Finally, exploratory analyses assessed the effect of different inclusion criteria. Across methods, using stricter inclusion criteria led to an increase in effect sizes despite the larger proportion of excluded participants (see also Byers-Heinlein, Bergmann, & Savalei, 2022).

However, there is a difference between a result being reliable in a large sample of infants and the measurement of an individual infant being reliable. In studies tracking individual differences, the measured behavior during an experimental setting is often used to predict a cognitive function or specific skill later in life. Individual differences research of this kind often has substantial implications for theoretical and applied work. For example, research showing that infants' behavior in speech perception tasks can be linked to later language development (see Cristia, Seidl, Junge, Soderstrom, & Hagoort, 2014 for a meta-analysis) has the potential to identify infants at risk for later language delays or disorders. A necessary precondition for this link to be observable is that individual differences between infants can be measured with high reliability at these earlier stages, in order to ensure that measured inter-individual variation mainly reflects differences in children's abilities rather than measurement error. How reliable are the measures used in infancy research?

Previous attempts to address the reliability of measurements have typically been limited to adult populations (Hedge, Powell, & Sumner, 2018; Oliveira, Hayiou-Thomas, & Henderson, 2023), or have been conducted with small sample sizes (e.g., Colombo, Mitchell, & Horowitz, 1988; Houston, Horn, Qi, Ting, & Gao, 2007). For example, Houston et al. (2007) tested 10 9-month-old infants' speech discrimination in a visual habituation procedure in two test sessions 1-3 days apart and found a large correlation (r = .7). These data were subsequently included in a much larger systematic investigation of test-retest reliability in infant speech perception (Cristia, Seidl, Singh, & Houston, 2016). Cristia et al. (2016) analyzed 13 different experiments assessing test-retest reliability in infant speech perception tasks, with the retest session occurring 0-18 days after the first session. The experiments were conducted at three different labs with different implementations of the individual studies. Hence, it was only after completed data collection that the data was pooled together by the different labs revealing potential confounds. Nevertheless, the results showed that reliability was extremely variable across the different experiments and labs and low overall (meta-analytic r = .07). In sum, there remains limited evidence and considerable uncertainty about the test-retest reliability of infant looking time measures.

Against this background, the current study investigated test-retest reliability of infants' performance in a speech preference task. Within MB1, a multi-lab collaboration, we examined whether infants' preferential listening behavior to IDS and ADS is reliable across two different test sessions. By collecting data from multiple labs, we were able to conduct a preregistered, large-scale analysis of test-retest reliability within a standardized looking-time task that yields reliable condition effects in infants. In addition to assessing overall test-retest reliability, we also planned to investigate the influence of several potential moderators on the reliability of IDS preference: the experimental method, infants' age and linguistic background, and the time between test sessions.

One main moderator analysis of interest was whether there were any differences in test-retest reliability between three widely used methods: central fixation (CF), eye-tracking (ET), and the head-turn preference procedure (HPP). Exploring differences in CF, ET, and HPP, Junge et al. (2020) provided experimental and meta-analytic evidence in favor of using HPP in speech segmentation tasks. Similarly, the MB1 project reported an increase in the effect size for HPP compared to CF and ET (ManyBabies Consortium, 2020). HPP requires gross motor movements relative to other methods, such as CF and ET paradigms, for which subtle eye movements towards a monitor located in front of the child are sufficient. One possible explanation for the stronger effects with HPP may be a higher sensitivity to the contingency of the presentation of auditory stimuli and infants' head turns away from the typical forward-facing position. While these findings suggest that HPP may be a more sensitive index of infant preference, they do not necessarily imply higher reliability for individual infants' performance using HPP. For example, Marimon and Höhle (2022) found no evidence for the test-retest reliability of infants' prosodic preferences (as measured by a difference score) using the HPP method across three testing sessions, each 7-8 days apart on average. It remains an open question whether the same measures that produce larger effect sizes at the group-level also have higher test-retest reliability for individual infants (Byers-Heinlein et al., 2022). In the current study, we therefore assessed whether HPP yields higher test-retest reliability compared to CF and ET in looking-time measures of IDS preference.

In our second set of moderator analyses, we aimed to address whether characteristics of the infant, specifically their age and linguistic background, were associated with differences in test-retest reliability. In MB1, older infants yielded larger effect sizes. Given that older infants have had more linguistic experience, we predicted that their preference may be more stable than that of younger infants. Infants also varied with respect to their linguistic background. All infants were tested using a North American English (hereafter, NAE) stimulus set, which was either their native or non-native language. We predicted that infants for whom NAE was not their native language — and who therefore had little or no experience with NAE — would demonstrate more variable and less reliable looking behavior than English-learning infants. In our final planned moderator analysis, we assessed whether time between test and retest influenced the reliability of the preference measure by investigating whether test-retest reliability decreased for participants with longer durations between their first and second test session.

Method

Preregistration

We preregistered the current study on the Open Science Framework (https://osf.io/v5f8t). Section S1 in the Supplementary Materials contains additional notes on the preregistration decisions and any deviations from the preregistered analytic plan.

Data Collection

A call was issued to all labs participating in the original MB1 study on January 29th, 2018 (ManyBabies Consortium, 2020). The collection of retest session data was initially set to end on May 31st, 2018, one month after the end date of the original MB1 project. Due to the fact that the original MB1 project extended the time frame for data collection and the late start of data collection for the MB1 test-retest study, we also allowed participating labs to continue data collection past the scheduled end date.

Participants

Contributing labs were asked to recruit monolingual participants between the ages of 6 to 12 months. If participating labs could not commit to test either of these age groups, they were also allowed to recruit participants from the youngest age group of 3- to 6-month-olds and/or the oldest age group of 12- to 15-month-olds. Labs were asked to contribute half (n=16) or full samples (n=32), however, a lab's data was included in the study regardless of the number of included infants. The study was approved by each lab's respective ethics committee and parental consent was obtained for each infant prior to participation in the study.

Our final sample consisted of 158 monolingual infants from 7 different labs (Table 1). In order to be included in the study, infants needed a minimum of 90% first language exposure, to be born full term with no known developmental disorders, and normal hearing and vision. We excluded 22 additional participants (see Data exclusion section for details). The mean age of infants included in the study was 245 days (approximately 8.06 months; range: 108 – 373 days).

Materials

Visual stimuli. The visual stimuli and instructions were identical to MB1. For the CF paradigm and ET, labs used a multicolored static checkerboard as the fixation stimulus as well as a multicolored moving circle with a ringing sound as an attention-getter between trials. For the HPP method, labs used their standard procedure, as in MB1.

Auditory stimuli. Our study was faced with a critical design choice: what stimuli to use to assess test-retest reliability. One constraint on our study was that, since it was a follow-on to MB1, any stimulus we used would always be presented after the MB1 stimuli. One option would be simply to bring back infants and have them hear exactly the same stimulus materials. A weakness of this design would be the potential for stimulus familiarity effects, however, since infants would have heard the materials before. Further complicating matters, infants might show a preference for or against a familiar stimulus depending on their age (Hunter & Ames, 1988). The ideal solution then would be to create a brand new stimulus set with the same characteristics. Unfortunately, because of the process of how MB1 stimuli were created, we did not have enough normed raw recordings available to make brand new stimulus items that conformed to the same standards as the MB1 stimuli. We therefore chose an intermediate path: we reversed the ordering of MB1 stimuli. A second set of naturalistic IDS and ADS recordings of mothers either talking to their infant or to an experimenter was created for the retest session by reversing the order of clips within each sequence of the original study. This resulted in eight reordered sequences of natural IDS and eight reordered sequences of natural ADS with a length of 18 seconds each. Average looking times in MB1 were always lower than 9s per trial, even for the youngest children on the earliest trials (the group who looked the longest on average), so most children in MB1 did not hear the second half of most trials. Thus, by reversing the order, we had a perfectly matched stimulus set that was relatively unfamiliar to most infants. The disadvantage of this design was that infants who looked longer might be more likely to hear a familiar clip heard in the previous session. If infants then showed a familiarity preference — an assumption which might not be true — the end result could be to inflate our estimates of test-retest reliability slightly, since longer lookers would on average look longer at retest due to their familiarity preference. We view this risk as relatively low, but do note that it is a limitation of our design. In addition to the 16 reversed-order IDS and ADS speech stimuli, we used the identical training stimuli of piano music from MB1.

Procedure. Infants were retested using the identical procedure as during the first testing day: CF, HPP, or ET. Participating labs were asked to schedule test and retest sessions 7 days apart with a minimum number of 1 day and a maximum number of 31 days. However, infants whose time between test and retest exceeded 31 days were still included in the analyses ($n = 3$). The mean number of days between test and retest was 10 (range: 1 - 49).

A total of 18 trials, including two training, eight IDS, and eight ADS trials, were presented in one of four pseudo-randomized orders. Trial length was either infant-controlled or fixed depending on the lab's standard procedure: a trial stopped either if the infant looked away for 2 seconds or after the total trial duration of 18 seconds. All Labs using CF or HPP had infant-controlled trial length whereas labs using ET had fixed trial length. The online coding experimenter and the parent listened to music masked with the stimuli of the study via noise-cancelling headphones. If the experimenter was in an adjacent room separate from the testing location, listening to masking music was optional for the experimenter.

Data exclusion. In total, 22 participants were excluded from the analysis. 4 participants were excluded for being preterm (defined as a gestation time of less than 37 weeks). 6 participants were excluded due to session errors involving an experimenter error (e.g., inaccurate coding or presentation of retest stimuli on the first test session). Individual trials were excluded if they were marked as trial errors (5.45% of remaining trials), i.e., if the infant was reported as fussy, an experimental or equipment error occurred, or there was parental interference during the task (e.g., if the parent spoke with the infant during the trial). Trials were also excluded if the minimum looking time of 2 s was not met (12.60% of the remaining trials). If a participant was unable to contribute at least one IDS and one ADS trial for either test or retest after trial-level exclusions, all data of that participant was excluded from the test-retest analyses (12 additional participants).

Results

IDS preference

First, we conducted confirmatory analyses examining infants' preference for IDS in both sessions. Two-samples t-tests comparing the difference in average looking time between IDS and ADS to zero revealed that infants showed a preference of IDS over ADS in Session 1, $t(157) = 6.47$, $p < .001$, and Session 2, $t(157) = 4.19$, $p < .001$, replicating the main finding from MB1 (Table 2; see Supplementary Materials S5 for robustness analyses

Table 1

Statistics of the included labs. N refers to the number of infants included in the final analysis.

Lab	Method	Language	Mean age (days)	N
InfantCog-UBC	central fixation	English	147	7
babylab-potsdam	HPP	German	227	22
babyling-oslo	eye-tracking	Norwegian	249	10
brookes-babylab	central fixation	English	267	18
infantll-madison	HPP	English	230	30
lancslab	eye-tracking	English	236	16
wsi-goettingen	HPP	German	242	16
wsi-goettingen	central fixation	German	280	39

using alternative dependent measures). 68.35% of infants in Session 1 and 63.29% of infants in Session 2 showed a preference for IDS. In order to test whether there was a difference in the strength of the preference effect across sessions, we fit a linear mixed-effects model predicting infants' average difference in looking time between IDS and ADS from test session (1 vs. 2), including by-lab and by-participant random intercepts. There was no significant difference in the magnitude of infants' preference between the two sessions, β=-0.30, SE=0.24, p=.208.

Reliability

We assessed test-retest reliability in two planned, confirmatory analyses. First, we fit a linear mixed-effects model predicting IDS preference in Session 2 from IDS preference in Session 1, including a by-lab random intercept. The results revealed no significant relationship between IDS preference in Session 1 and 2 (Table 3). Second, we calculated the

Table 2

Average looking times (in seconds) for each session and condition.

Trial type	Session 1 Mean	Session 1 *SD*	Session 2 Mean	Session 2 *SD*
ADS	7.71	2.77	6.96	2.92
IDS	8.76	2.84	7.75	2.75

Pearson correlation coefficient. While a simple correlation coefficient might overestimate the test-retest reliability in our sample because it does not control for the differences between different labs and methods (HPP, CF, and ET), we felt it was important to also conduct a Pearson correlation as it is commonly used to assess reliability. The size of the correlation coefficient was not statistically different from zero and the estimate was small, $r = .09$, 95% CI $[-.06, .25]$, $t(156) = 1.19$, $p = .237$. Moreover, no significant correlations emerged in each sample considered separately (Figure 1; see Supplementary Materials S3 for a meta-analytic approach). 41.77% of the infants reversed their direction of preference for IDS versus ADS from the test to the retest session (see Supplementary Materials S7 for additional analyses of infants' patterns of preferential looking across sessions).

To investigate the test-retest reliability of each specific method, we computed Pearson correlation coefficients and the same mixed-effects model described above for HPP, CF, and ET separately (Table 4) in additional exploratory analyses. None of the three methods showed evidence of test-retest reliability. Neither the Pearson correlation coefficients nor the coefficients of the multilevel analysis were significant, all p-values > 0.12. In planned secondary analyses, we found that experimental method, time between test sessions, participant age, and language background did not moderate the relationship between IDS preference in Session 1 and Session 2 (see Supplementary Materials S2). Taken together, we find no significant evidence of test-retest reliability across our preregistered analyses.

Table 3

Coefficient estimates from a linear mixed effects model predicting IDS preference in Session 2.

	Estimate	SE	t	p
Intercept	0.87	0.46	1.92	0.10
IDS Preference Session 1	0.04	0.09	0.41	0.68

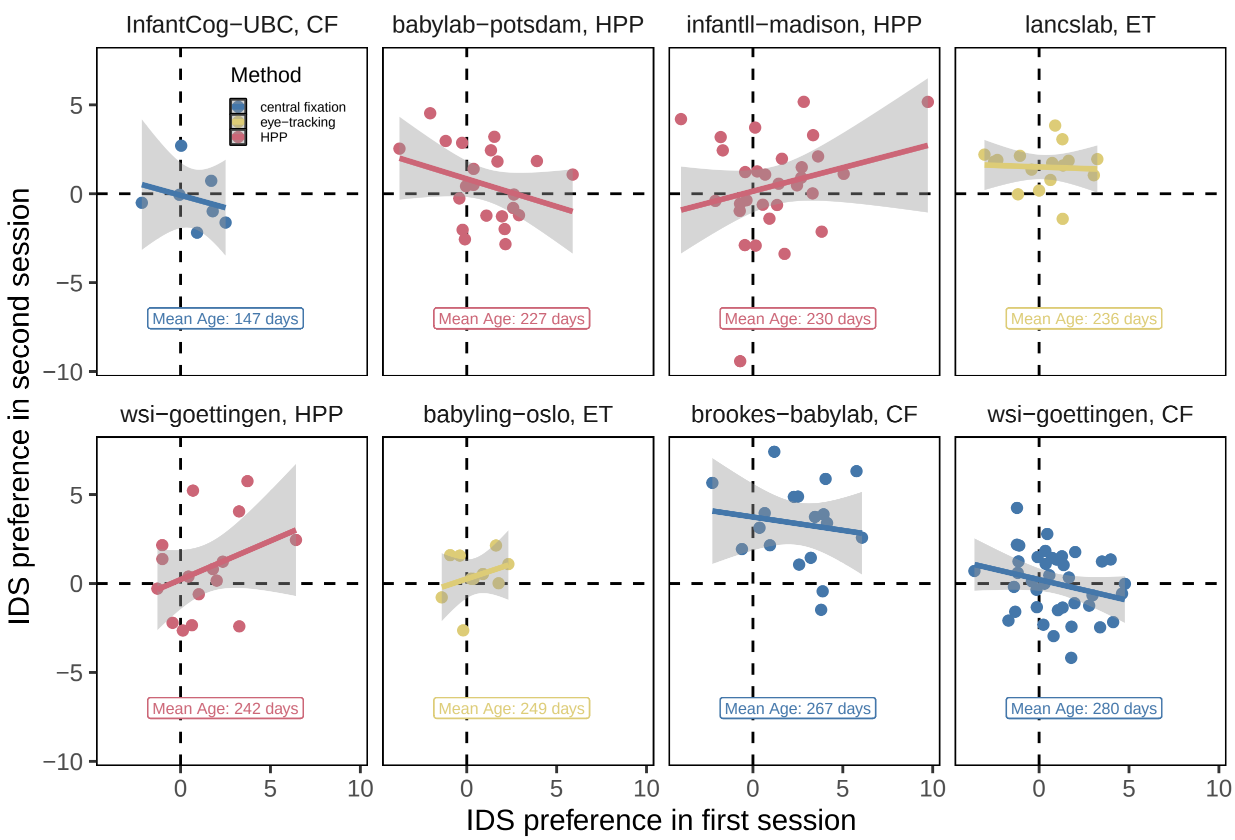

Figure 1. Correlation between IDS Preference in Session 1 and Session 2 in each lab and method. Dots indicate individual participants. Error bands represent 95 percent confidence intervals. The dashed line indicates no preference (i.e., a value of zero) for the first and second session, respectively.

Table 4

Coefficient estimates from a linear mixed effects model predicting IDS preference in Session 2 and Pearson correlation coefficient for each method separately.

Method	*beta*	*SE*	*p*	Pearson *r*
central fixation	-0.20	0.12	0.12	0.08
HPP	0.15	0.14	0.28	0.13
eye-tracking	0.03	0.16	0.84	0.02

Exploratory analyses with different inclusion criteria

To this point, all analyses were performed using the inclusion criteria from MB1, which required only that infants contribute at least one trial per condition for inclusion (i.e., one IDS and one ADS trial). However, more stringent inclusion criteria yielded larger effect sizes in MB1. We therefore conducted exploratory analyses assessing test-retest reliability after applying progressively stricter inclusion criteria, requiring two, four, six, and eight valid trials per condition. Applying stricter criteria — and thereby increasing the number of test trials — increased reliability numerically from $r = 0.07$ to $r = 0.34$ (Figure 2). In part due to a decrease in sample size, only one of these correlations was statistically significant (when requiring six trial pairs): two valid trial pairs, $t(152) = 0.90$, $p = .367$; four valid trial pairs, $t(143) = 1.03$, $p = .306$; six valid trial pairs, $t(98) = 2.23$, $p = .028$; eight valid trial pairs — all trials in both sessions — $t(22) = 1.68$, $p = .108$. The analyses provide tentative evidence that stricter inclusion criteria may lead to higher test-retest reliability, but at the cost of substantial decreases in sample size. In Supplementary Materials S4 and S6, we provide additional exploratory analyses, including moderator analyses using a more restricted sample (S4) and a more fine-grained analysis of the increase in reliability with stricter inclusion criteria (S6). In particular, in S6.2, we conduct simulations demonstrating that, holding participant groups constant, test-retest reliability rises systematically as more trials are resampled within each participant, indicating that increasing the number of observations per participant is critical to improving reliable measurement.

Correlations between sessions for number of trials contributed and overall looking time

In exploratory analyses, we also investigated whether there were stable individual differences in (a) the number of trials an infant contributed across the two test sessions and (b) infants' overall looking times.

Number of trials contributed. We found a strong positive correlation between number of trials contributed during the first and the second session, $r = .58$, 95% CI $[.47, .67]$, $t(160) = 9.00$, $p < .001$ (Figure 3A). In other words, if infants contributed a higher number of trials in one session, compared to other infants, they were likely to contribute a higher number of trials in their next session. This finding is consistent with the hypothesis that how attentive infants are throughout an experiment (and hence how many trials they contribute) is a stable individual difference, at least for some infant looking time tasks.

Overall looking times. To what extent are participants looking times between the two sessions related? To test this question, we investigated whether participants' overall looking times — irrespective of condition — were correlated between the first and second session. There was a robust correlation between average looking time in Session 1 and Session 2: infants with longer looking times during their first session also tended to look longer during their second session, $r = .45$, 95% CI $[.31, .57]$, $t(156) = 6.28$, $p < .001$ (Figure 3B). This relationship held even after controlling for number of trials, $b = 0.42$, 95% CI $[0.27, 0.58]$, $t(154) = 5.52$, $p < .001$, and participants' average age, $b = 0.44$, 95% CI $[0.30, 0.59]$, $t(155) = 6.16$, $p < .001$, across the two test sessions in linear regression

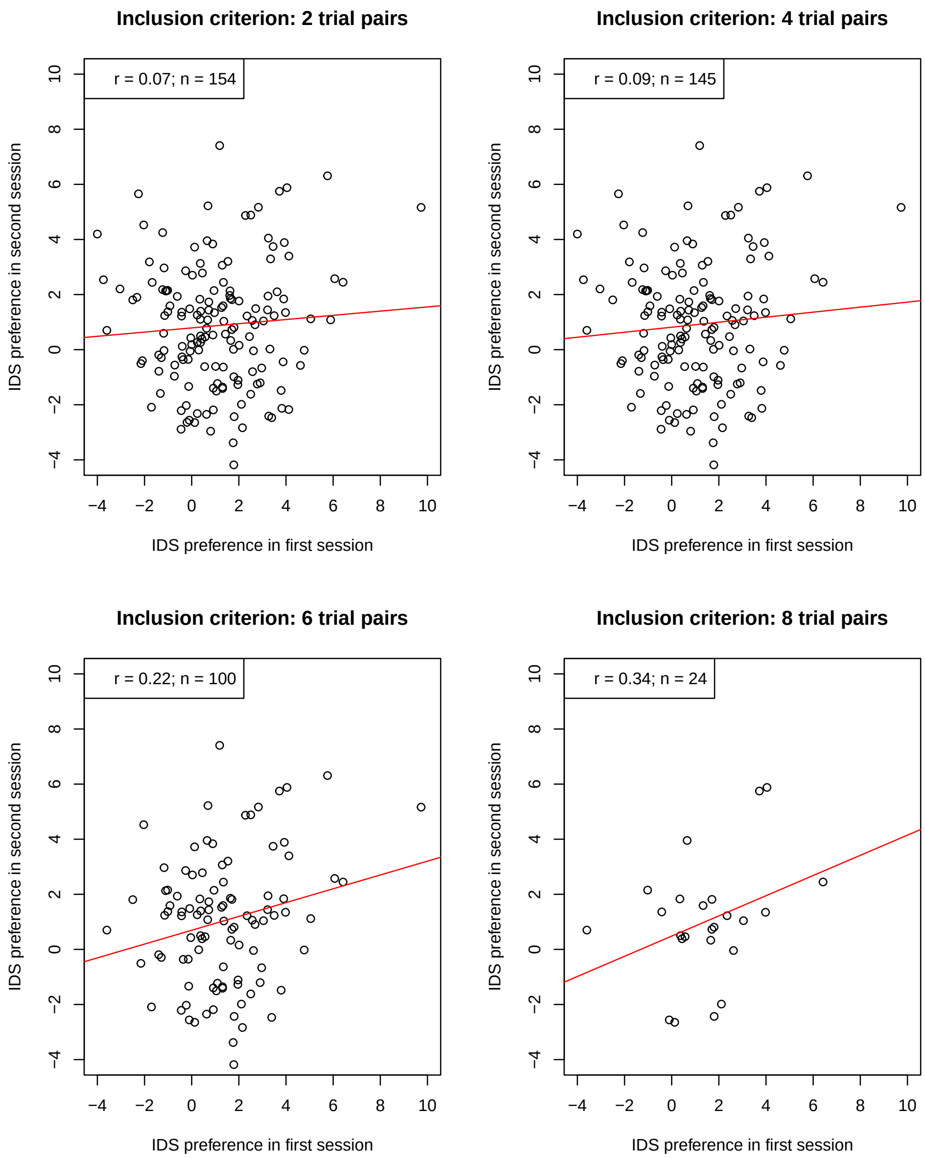

Figure 2. IDS preferences of both sessions plotted against each other for each inclusion criterion. n indicates the number of included infants, r is the Pearson correlation coefficient as the indicator for reliability.

models. Finally, we found similar correlations in average looking time to IDS stimuli in Session 1 and 2, $r = .38$, 95% CI [.24, .51], $t(156) = 5.19$, $p < .001$, and ADS stimuli in Session 1 and 2, $r = .40$, 95% CI [.26, .53], $t(156) = 5.49$, $p < .001$ (Figure 3C; see Supplementary Materials S8 and S9 for further details, including an investigation of item-level correlations).

General Discussion

The current study investigated the test-retest reliability of infants' preference for IDS over ADS. As a spin-off of the original MB1 project, we tested the IDS preference of infants in two separate test sessions to assess the extent to which their pattern of preference would remain consistent. While we replicated the original effect of infants' speech preference for IDS over ADS for both the test and retest session on the group-level, we found that infants' speech preference measures showed no evidence of test-retest reliability across all preregistered analyses. In other words, we were unable to find confirmatory evidence of stable individual differences in infants' preference for IDS. This finding is consistent with past research suggesting low test-retest reliability in other infant paradigms (Cristia et al., 2016; Marimon & Höhle, 2022). Given that most experimental procedures conducted in infant research are interested in the comparison of groups, individual differences between participants within a specific condition are usually minimized by the experimental procedure while differences between conditions are maximized. As a consequence, infant preference measures may be a good approach for capturing group-level phenomena, but may be less appropriate for examining individual differences in development.

We also found no robust evidence that several hypothesized moderators influenced test-retest reliability. While previous research suggests that HPP may be a more sensitive index of infant preference and speech segmentation abilities (Junge et al., 2020; ManyBabies Consortium, 2020), we found no evidence for higher test-retest reliability for HPP relative to ET and CF. Similarly, our planned analyses found no evidence that the

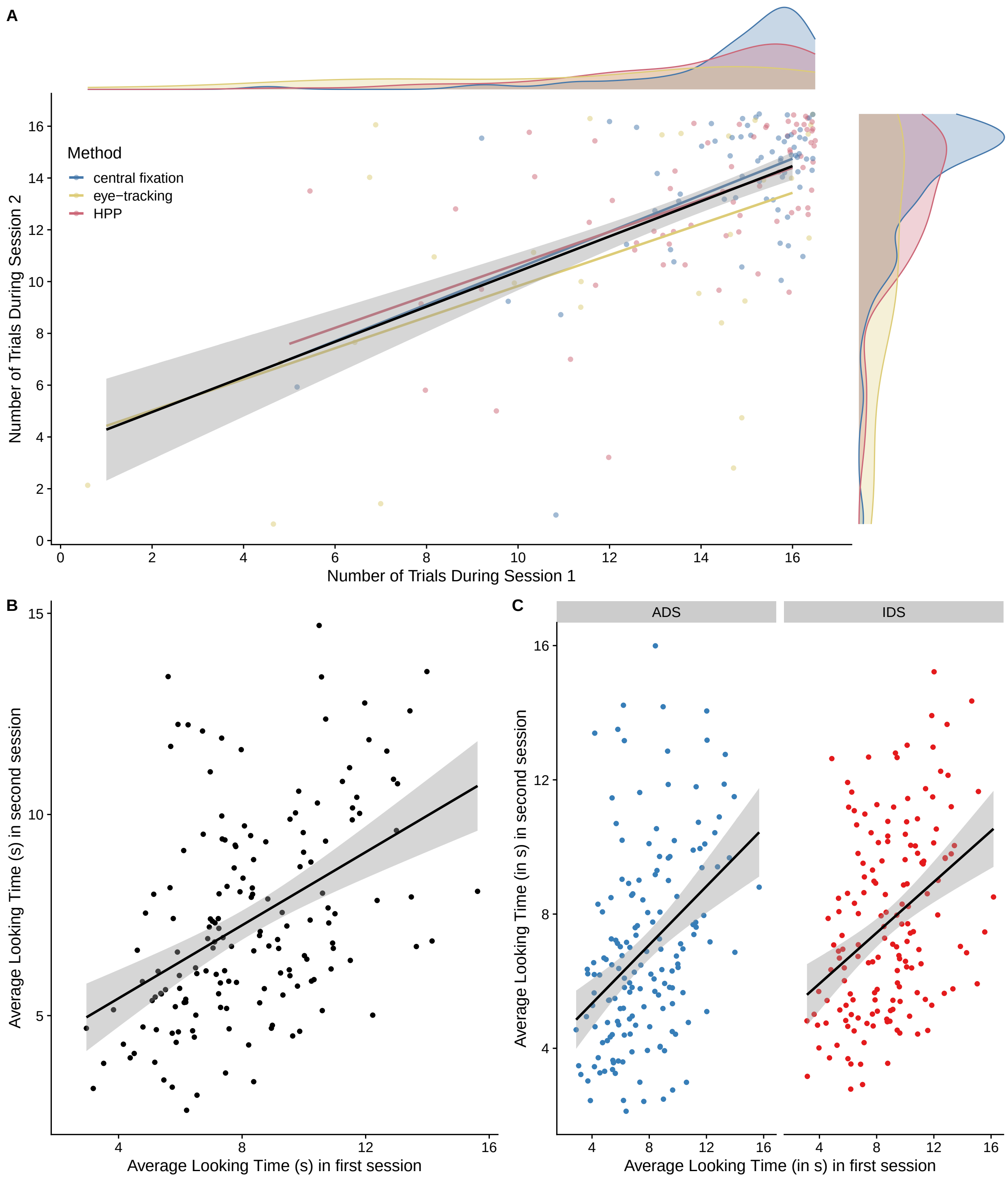

Figure 3. (A) Correlation between the number of trials contributed in Session 1 and Session 2. Each data point represents one infant. Colored lines represent linear fits for each method. (B) Overall correlations in average looking time (in s) between Session 1 and 2. (C) Correlations in average looking time (in s) between sessions, split by IDS/ADS condition.

test-retest reliability of IDS preference varied as a function of the time between test sessions, infants' language background, or infants' age. The absence of evidence for moderating effects should be treated with caution, given that fewer labs contributed to the study than we had anticipated by the time of the preregistration, resulting in limited power to detect interaction effects (see S2.1 for more detail on power considerations). However, these analyses also suggest that the lack of test-retest reliability was not be due to variation across any of our hypothesized moderators.

Consistent with general psychometric theory (e.g., DeBolt, Rhemtulla, & Oakes, 2020), stricter inclusion criteria — and consequently a larger number of included test trials per participant — tended to increase the magnitude of the correlation between test sessions. However, this association was based on exploratory analyses and was in part only observed descriptively, and hence should be interpreted with caution. A similar effect on the group-level was found in the MB1 project, where a stricter inclusion criterion led to bigger effect sizes (ManyBabies Consortium, 2020). As in MB1, higher reliability through strict exclusions came at a high cost. In particular, with the strictest criterion, only a small portion of the original sample size (24 out of 158 infants) could be included in the final sample. In other words, applying stricter criteria leads to a higher drop-out rate and can dramatically reduce the sample size. In the case of studies in the field of developmental science, where there are many practical constraints on collecting large samples of infants (e.g., birth rate in the area, restricted lab capacities, budget restrictions), a strict drop-out criterion may often be difficult to implement. Note that studies in developmental science already have above-average drop-out rates (Miller, 2017). In addition, drop out is likely not random: the number of trials individual infants contributed was highly correlated between test sessions in the current study. Therefore, implementing strict inclusion criteria that result in high drop-out rates can further limit the generalizability of a study. Particularly in the context of turning individual differences measures into diagnostic tools, high drop-out rates have an additional limitation of not being broadly usable.

An alternative approach to increasing the number of valid trials is to increase the number of experimental trials. This approach seeks to increase the likelihood that participants will contribute sufficient trials (after trial-level exclusions) to allow for precise individual-level estimates (DeBolt et al., 2020; see also Silverstein, Feng, Westermann, Parise, & Twomey, 2021). While this approach is promising, it may not always be feasible, because the attention span of a typical infant participant is limited. Therefore, prolonging the experimental procedure to maximize the absolute number of trials is often challenging in practice. Other avenues for obtaining higher numbers of valid trials may include changes in the procedure (e.g., Egger, Rowland, & Bergmann, 2020) or implementing multi-day test sessions (Fernald & Marchman, 2012), which could become easier with the adoption of online testing methods (Lo, Hermes, Kartushina, Mayor, & Mani, 2023; Scott & Schulz, 2017; Weaver, Zettersten, & Saffran, 2022).

Even under best-case scenarios focused on infants contributing larger numbers of trials, reliability remained low. For example, when restricting the sample to infants contributing at least six trials in each condition in both sessions, we obtained a correlation of $r = 0.22$ and an intra-class correlation coefficient of $\alpha = 0.36$. As Byers-Heinlein et al. (2022) outline, low measurement reliability severely restricts power for detecting relationships between measures. Using the same approach as Byers-Heinlein et al. (2022), we estimate that over 682 infants would be needed to have at least 80% power to observe a true correlation of $r = .3$ between two measurements, assuming an intra-class correlation coefficient as large as that observed in our restricted sample ($\alpha = 0.36$). Even a very large true correlation of $r = .7$ would require a sample size of over 120 infants. In other words, even under optimistic estimates of reliability based on strict inclusion criteria, the low reliability of IDS preference measures would severely limit the feasibility of individual difference and longitudinal research using current methods.

As our results are only based on the phenomenon of IDS preference (albeit, with three widely used methods: HPP, CF, ET), it is essential to further assess the underlying reliability of preferential looking measures within other areas of speech perception (Marimon & Höhle, 2022). While most infants prefer IDS over ADS (Dunst et al., 2012), patterns of preferential looking in other tasks (e.g., speech segmentation, artificial grammar learning) are often inconsistent and can vary based on factors such as infants' experience with the testing paradigm (Bergmann & Cristia, 2016; Santolin, Garcia-Castro, Zettersten, Sebastian-Galles, & Saffran, 2021). These inconsistencies in looking behavior are especially important to consider in the context of relating a direction of preference to later language development, and can sometimes lead to seemingly contradictory findings. That is, both familiarity and novelty responses have been suggested to be predictive of infants' later linguistic abilities (DePaolis, Vihman, & Keren-Portnoy, 2014; Newman, Ratner, Jusczyk, Jusczyk, & Dow, 2006; Newman, Rowe, & Ratner, 2016). In light of our findings, researchers conducting longitudinal studies with experimental data from young infants predicting future outcomes should be cautious, as there may be large intra-individual variability affecting preference measurement.

While we observed limited evidence for test-retest reliability using preference measures, we observed robust correlations for average looking times between Session 1 and 2, both overall and for IDS and ADS stimuli considered separately (see also Supplementary Materials S8 for an investigation of item-level correlations). This finding is consistent with past results in infant looking time studies finding robust correlations in average looking times across multiple sessions (Marimon & Höhle, 2022). It also raises an apparent puzzle: why are overall looking times for ADS and IDS stimuli correlated while difference scores are not? One explanation is that infants have stable individual differences in how long they look to stimuli, but little or no stable individual differences in their preference for one stimulus type over another. This only partially explains the current pattern of results, however, because IDS looking time in Session 1 predicted IDS looking time in Session 2 even when controlling for ADS looking time, and vice versa (see Supplementary Materials S8). In other words, the condition-specific looking time correlations are not fully explained by overall looking behavior. Another long-established explanation is that difference scores tend to have poor measurement reliability, because difference scores combine error from individual measurements into a composite score, increasing the ratio of error relative to the variance between participants (Hedge et al., 2018; Lord, 1956). Given the limitations of difference scores (and composite scores in general), one goal for future research will be to assess the use of trial-by-trial model-based approaches for estimating reliability (Haines et al., 2020; Rouder & Haaf, 2019). One key to building more reliable models of infant looking time will be to understand and quantify the main sources of variation in looking behavior within and across participants. In S11 and S12 of the Supplementary Materials, we report exploratory analyses identifying additional predictors of trial-to-trial variation in looking time and decomposing the variance explained by focal within- and between-participant predictors of interest.

Limitations

While we had an above-average sample size for a study in infant research, we were unable to approach the number of participants collected within the original MB1 study. Several factors likely contributed to the lower participation rate. The call to participate in the test-retest spin-off was delayed relative to the MB1 call, scheduling a second lab visit for each participant involved a significant additional effort for participating labs, and there were already other collaborative studies taking place simultaneously (MB1B, Byers-Heinlein, Tsui, Bergmann, et al., 2021; MB1G, Byers-Heinlein, Tsui, Van Renswoude, et al., 2021). A larger sample size and a greater number of participating labs from different countries would have enabled us to conduct a more highly-powered test of differences in test-retest reliability across different methods, language backgrounds, and participant age.

A further limitation concerns the stimuli. While the order of the audio recording clips presented to infants within a given trial differed between the first and second session, the exact same stimulus material as in MB1 was used in both sessions. In particular, all children heard the exact same voices in Session 1 and in Session 2. From a practical point of view, this was the most straightforward solution for coordinating the experiment within the larger MB1 project. However, familiarity effects might have influenced infants' looking behavior. Infants with longer looking times in their first session might have had more opportunity to recognize familiar audio clips in their second session. For infants with short looking times, familiar audio clips would only occur towards the end of second-session trials, thus offering infants less opportunity to recognize voices from their first session. Therefore, inconsistent familiarity with the stimulus material in the second session across infants might have artificially lowered test-retest reliability. However, in supplementary analyses, we found that test-retest reliability was not significantly moderated by infants' overall looking time during their first testing session (see Supplementary Materials S10). Moreover, infants' looking times generally declined in Session 2 compared to Session 1 (consistent with past work, e.g., Marimon & Höhle, 2022), limiting opportunities for infants to encounter previously experienced stimulus material.

Conclusion

Following the MB1 protocol, the current study could not detect test-retest reliability in measures of infants' preference for IDS over ADS. Subsequent analyses provided tentative evidence that stricter criteria for the inclusion of participants may enhance test-retest reliability at the cost of high drop-out rates. Developmental studies relying on stable individual differences between their participants need to consider the underlying reliability of their measures, and we recommend a broader assessment of test-retest reliability in infant research.

References

Bergmann, C., & Cristia, A. (2016). Development of infants' segmentation of words from native speech: A meta-analytic approach. *Developmental Science*, *19*(6), 901–917.

Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2022). Six solutions for more reliable infant research. *Infant and Child Development*, *31*(5), e2296.

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.

Byers-Heinlein, K., Tsui, R. K.-Y., Van Renswoude, D., Black, A. K., Barr, R., Brown, A., ... Singh, L. (2021). The development of gaze following in monolingual and bilingual infants: A multi-laboratory study. *Infancy*, *26*(1), 4–38.

Colombo, J., Mitchell, D. W., & Horowitz, F. D. (1988). Infant visual attention in the paired-comparison paradigm: Test-retest and attention-performance relations. *Child Development*, 1198–1210.

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

Cristia, A., Seidl, A., Junge, C., Soderstrom, M., & Hagoort, P. (2014). Predicting individual variation in language from infant speech perception measures. *Child Development*, *85*(4), 1330–1345.

Cristia, A., Seidl, A., Singh, L., & Houston, D. (2016). Test–retest reliability in infant speech perception tasks. *Infancy*, *21*(5), 648–667.

DeBolt, M. C., Rhemtulla, M., & Oakes, L. M. (2020). Robust data and power in infant research: A case study of the effect of number of infants and number of trials in visual preference procedures. *Infancy*, *25*(4), 393–419.

DePaolis, R. A., Vihman, M. M., & Keren-Portnoy, T. (2014). When do infants begin recognizing familiar words in sentences? *Journal of Child Language*, *41*(1), 226–239.

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

Egger, J., Rowland, C. F., & Bergmann, C. (2020). Improving the robustness of infant lexical processing speed measures. *Behavior Research Methods*, *52*(5), 2188–2201.

Fernald, A., & Marchman, V. A. (2012). Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. *Child Development*, *83*(1), 203–222.

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

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.

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

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

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., . . . Turner, B. (2020). *Theoretically Informed Generative Models Can Advance the Psychological and Brain Sciences: Lessons from the Reliability Paradox* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/xr7y3

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186.

Houston, D. M., Horn, D. L., Qi, R., Ting, J. Y., & Gao, S. (2007). Assessing speech discrimination in individual infants. *Infancy*, *12*(2), 119–145.

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research*, *5*, 69–95.

Johnson, E., & Zamuner, T. (2010). Using infant and toddler testing methods in language acquisition research. In E. Blom & S. Unsworth (Eds.), *Experimental methods in language acquisition research* (pp. 73–93). Amsterdam: John Benjamins Publishing Company.

Junge, C., Everaert, E., Porto, L., Fikkert, P., Klerk, M. de, Keij, B., & Benders, T. (2020). Contrasting behavioral looking procedures: A case study on infant speech segmentation. *Infant Behavior and Development*, *60*, 101448.

Lo, C. H., Hermes, J., Kartushina, N., Mayor, J., & Mani, N. (2023). E-Babylab: An open-source browser-based tool for unmoderated online developmental studies. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02200-7

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement*, *16*, 421–437. https://doi.org/10.1177/001316445601600401

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.

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.

Marimon, M., & Höhle, B. (2022). Testing prosodic development with the headturn preference procedure: A test-retest reliability study. *Infant and Child Development*, e2362.

Miller, S. A. (2017). *Developmental research methods*. Sage publications.

Naoi, N., Minagawa-Kawai, Y., Kobayashi, A., Takeuchi, K., Nakamura, K., Yamamoto, J., & Shozo, K. (2012). Cerebral responses to infant-directed speech and the effect of talker familiarity. *Neuroimage*, *59*(2), 1735–1744.

Newman, R., Ratner, N. B., Jusczyk, A. M., Jusczyk, P. W., & Dow, K. A. (2006). Infants' early ability to segment the conversational speech signal predicts later language development: A retrospective analysis. *Developmental Psychology*, *42*(4), 643.

Newman, R., Rowe, M. L., & Ratner, N. B. (2016). Input and uptake at 7 months predicts toddler vocabulary: The role of child-directed speech and infant processing skills in language development. *Journal of Child Language*, *43*(5), 1158–1173.

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

Oliveira, C. M., Hayiou-Thomas, M. E., & Henderson, L. M. (2023). The reliability of the serial reaction time task: Meta-analysis of test–retest correlations. *Royal Society Open Science*, *10*(7), 221542. https://doi.org/10.1098/rsos.221542

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*(2), 452–467. https://doi.org/10.3758/s13423-018-1558-y

Santolin, C., Garcia-Castro, G., Zettersten, M., Sebastian-Galles, N., & Saffran, J.

R. (2021). Experience with research paradigms relates to infants' direction of preference. *Infancy*, *26*(1), 39–46. https://doi.org/10.1111/infa.12372

Schreiner, M. S., & Mani, N. (2017). Listen up! Developmental differences in the impact of IDS on speech segmentation. *Cognition*, *160*, 98–102.

Scott, K. M., & Schulz, L. E. (2017). Lookit (part 1): A new online platform for developmental research. *Open Mind*, *1*(1), 4–14. https://doi.org/doi:10.1162/opmi_a_00002

Silverstein, P., Feng, J., Westermann, G., Parise, E., & Twomey, K. E. (2021). Infants learn to follow gaze in stages: Evidence confirming a robotic prediction. *Open Mind*, 1–15.

Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009). Influences of infant-directed speech on early word recognition. *Infancy*, *14*(6), 654–666.

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

Weaver, H., Zettersten, M., & Saffran, J. (2022). *Becoming word meaning experts: Infants' processing of familiar words in the context of typical and atypical exemplars.* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/njh38

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. https://doi.org/10.1207/s15327078in1101_2