Limited evidence of test-retest reliability in infant-directed speech preference in a large pre-registered infant sample

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 (hereafter, IDS) over adult-directed speech (hereafter, ADS) in a large sample (N=158) in the context of the ManyBabies1 collaborative research project (hereafter, MB1; Frank et al., 2017; ManyBabies Consortium, 2020). Labs of the original MB1 study were asked to bring in participating infants for a second appointment retesting infants on their IDS preference. This approach allows 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 find 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: 3998

Limited evidence of test-retest reliability in infant-directed speech preference in a large pre-registered infant sample

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, 2021).

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. However, 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), or have been conducted with small sample sizes (e.g., Houston, Horn, Qi, Ting, & Gao, 2007). For example, Colombo, Mitchell, and Horowitz (1988) used a paired-comparison task, in which infants were familiarized with a stimulus and presented with the familiarized and a novel stimulus side-by-side at test. Results indicated that infants' novelty preference was extremely variable from task to task. Assessing infants' performance from one week to another revealed that infants' attention measures were moderately reliable. However, reliability seemed to increase with the number of tasks infants completed in the younger age group, suggesting that reliability is influenced by the number of assessments. In addition, infants' performance from 4 to 7 months was longitudinally stable but somewhat smaller than week-to-week reliability. Cristia, Seidl, Singh, and Houston (2016) also retested infant populations by independently conducting 12 different experiments on infant speech perception 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).

Against this background, the current study investigates test-retest reliability of infants' performance in a speech preference task. Within MB1, a multi-lab collaboration, we examine whether infants' preferential listening behavior to IDS and ADS is reliable across two different test sessions. We also investigate the influence of various moderators on the reliability of IDS preference (e.g., time between test and retest; infants' language background).

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. 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 study. 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.

The current study also explores whether there are 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) provide experimental and meta-analytic evidence in favor of using the

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 test-retest reliability when testing infants' prosodic preferences using the HPP method. 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, Bergmann, et al., 2021). Therefore, assessing the test-retest reliability of the different preference measures is crucial, so that researchers can make informed decisions about the appropriate methods for their particular research question. Critically, only measures with high test-retest reliability should be used for studies of individual differences.

Method

Preregistration

Prior to the start of data collection, we preregistered the current study on the Open Science Framework (https://osf.io/v5f8t; see S1 in the Supplementary Materials for details).

Data Collection

A call was issued to all labs participating in the original MB1 study on January 24th, 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 re-recruit their monolingual participants between the ages of 6 to 12 months who had already participated in the MB1 project. If participating labs had not committed to testing either of these age groups, they were also allowed to re-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 11 participants due to session errors and 11 participants who did not have at least one valid trial per condition (IDS and ADS) at their first or second session. The mean age of infants included in the study was 245 days (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.

Speech stimuli. We used the identical training stimuli of piano music from MB1. 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 new sequences of natural IDS and eight new sequences of natural ADS with a length of 18 seconds each.

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. 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. A child was excluded if they had a session error, i.e., an experimenter error (e.g., inaccurate coding, or presentation of retest stimuli on the first test session) or equipment failure (visual stimuli continued to play after the end of a trial).

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

Trials were excluded if they were marked as trial errors, 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. If a participant was unable to contribute at least one IDS and one ADS trial for either test or retest, all data of that participant was excluded from the test-retest analyses.

Results

IDS preference

First, we examined 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

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	7.72	2.77	6.96	2.92
IDS	8.76	2.85	7.75	2.75

Session 2, $t(157) = 4.19$, $p < .001$, replicating the main finding from MB1 (Table 2). 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 ways. 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 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

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

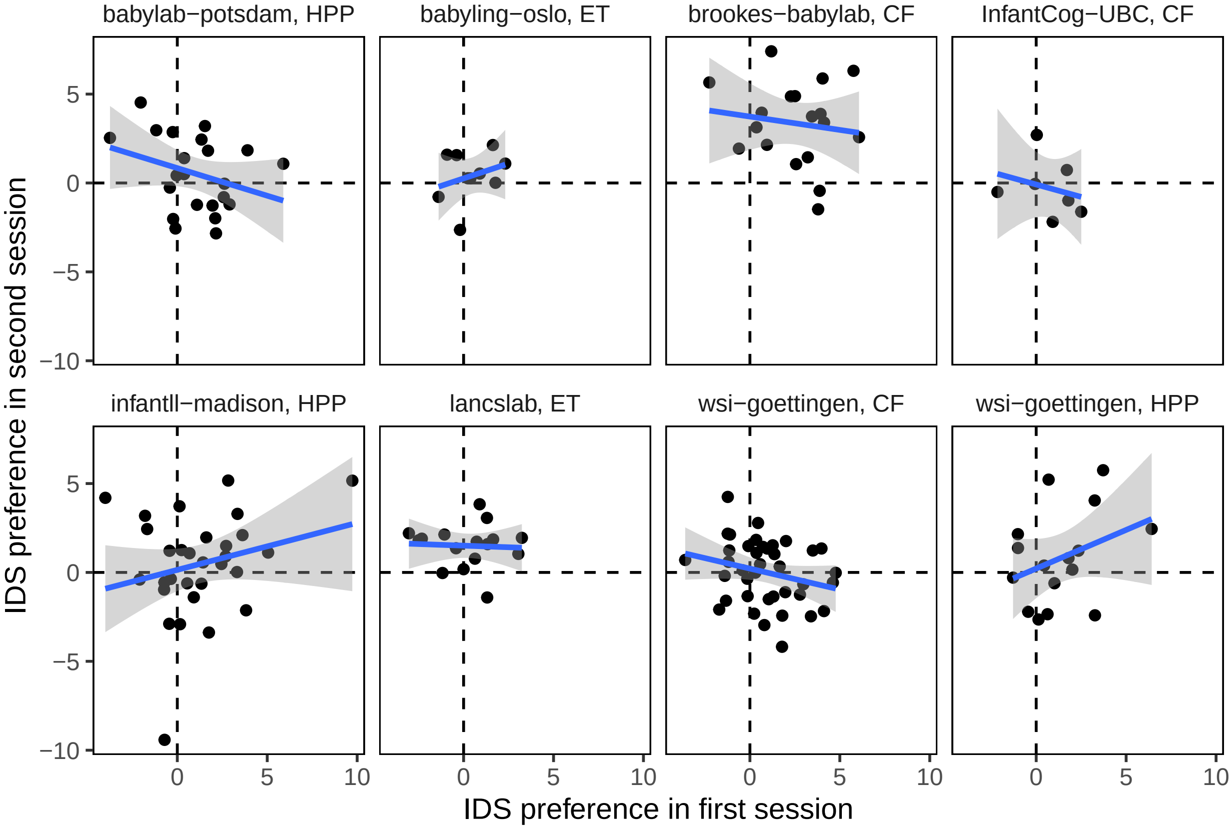

Figure 1. Correlation between IDS Preference in Session 1 and Session 2 in each lab and method. Dots indicate individual participants. Error bands represent 95 percent confidence intervals. The dashed line indicates no preference (i.e., a value of zero) for the first and second session, respectively.

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
HPP	0.15	0.14	0.28	0.13
ET	0.03	0.16	0.84	0.02
CF	-0.20	0.12	0.12	0.08

meta-analytic approach). 41.77% of the infants reversed their direction of preference for IDS versus ADS from the test to the retest session.

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). 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 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.

Results 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 (see Supplementary Materials S5 for additional analyses).

General Discussion

The current study investigated the test-retest reliability of infants' preference for IDS over ADS. We retested the IDS preference of infants participating in the original MB1 project to assess the extent to which their pattern of preference would remain consistent across multiple testing sessions. 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. In other words, we were unable to detect 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). 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. Therefore, infant preference measures may be a good approach for capturing group-level phenomena, but may be less appropriate for examining individual differences in development.

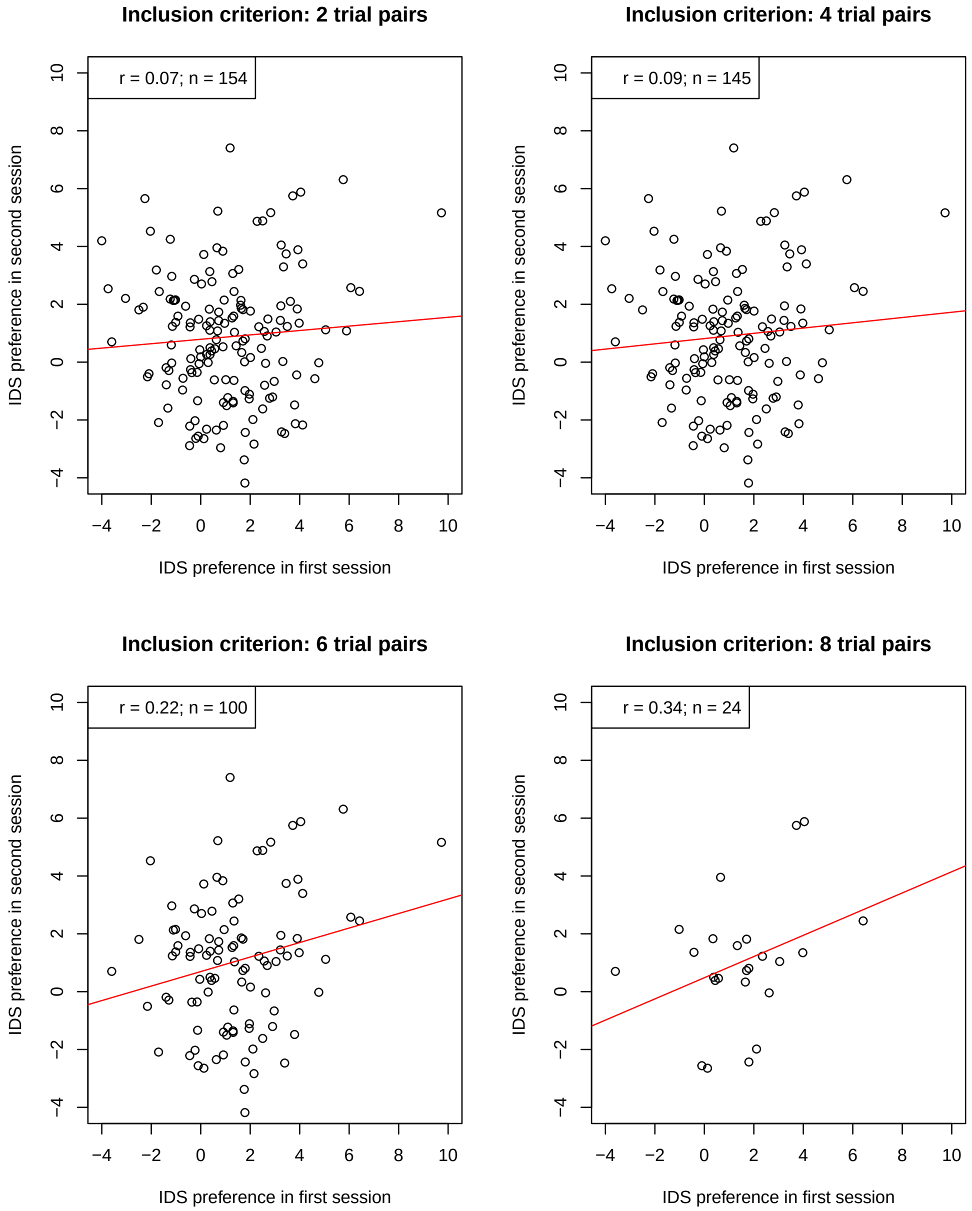

Figure 2. IDS preferences of both sessions plotted against each other for each inclusion criterion. n indicates the number of included infants, r is the Pearson correlation coefficient as the indicator for reliability.

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 restrictions in 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 may not be random, and so having high drop-out rates can further limit the generalizability of a study. In fact, the number of trials individual infants contributed was highly correlated between test sessions in the current study (see Supplementary Materials S6). 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).

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) are often inconsistent and difficult to predict (Bergmann & Cristia, 2016). 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.

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. In addition to a delayed call, the extra effort of having to schedule a second lab visit for each participant and the fact that 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), might have contributed to a low participation rate. A higher sample size and a larger 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.

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. (2021). Six solutions for more reliable infant research. *Infant and Child Development*, 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

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.

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.

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

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

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

Manybabies1 Test-Retest Supplementary Materials

Contents

S1. Notes on and deviations from the preregistration

Below, we have compiled a list of notes on and deviations from the preregistered methods and analyses available at https://osf.io/v5f8t.

- All infants with usable data for both test and retest session were included in the analyses, regardless of the number of total infants a lab was able to contribute after exclusion. This decision is consistent with past decisions in ManyBabies projects to be as inclusive about data inclusion as possible (ManyBabies Consortium, 2020).
- A small number of infants whose time between sessions exceeded 31 days were still included in the analyses ($n = 3$).
- Consistent with analytic decisions in ManyBabies 1 (ManyBabies Consortium, 2020), total looking times were truncated at 18 seconds (the maximum trial time) in the small number of cases where recorded looking times were slightly greater than 18s (presumably due to small measurement error in recording infant looking times).
- In assessing differences in IDS preference between test and retest sessions, we preregistered an additional linear mixed-effects model including a by-lab random slope for session. This model yielded qualitatively equivalent results (see R markdown of the main manuscript). However, the model resulted in a singular fit, suggesting that the model specification may be overly complex and that its estimates should be interpreted with caution. We therefore focused only on the first preregistered model (including only by-lab and by-participant random intercepts) in reporting the analyses in the main manuscript.
- In assessing the reliability of IDS using a linear mixed-effects model predicting IDS preference in session 2 from IDS preference in session 1, we also assessed the robustness of the results by fitting a second preregistered model with more complex random effects structure, including a by-lab random slope for IDS preference in session 1. This model is included in the main R markdown script and yields

qualitatively equivalent results to the model reported in the manuscript that includes a by-lab random intercept only.

- We report a series of secondary planned analyses in the Supplementary Materials exploring potential moderating variables of time between test sessions (S2.1), the language background of the participants (S2.2.), and participant age (S2.3.).
- We did not fit all models (in particular, the models investigating interactions between moderators) described in the secondary analyses of the preregistration, because our final sample size was smaller than we anticipated, which made it less feasible to investigate more complex relationships between moderators.

S2. Secondary analyses investigating possible moderating variables

S2.1. Time between test sessions

The number of days between the first and second testing session varied widely across participants (mean: 10 days; range: 1 - 49 days). We therefore tested for the possibility that the time between sessions might have an impact on test-retest reliability. We fit a linear mixed-effects model predicting IDS preference in Session 2 from IDS preference in Session 1 (mean-centered), number of days between testing sessions (mean-centered), and their interaction, including a by-lab random intercept and random slope for IDS preference in Session 1. A more complex random effects structure including additional random slopes for number of days between test sessions and its interaction with IDS preference in Session 1 did not converge. We found no evidence that the number of days between test sessions moderated the relationship between IDS preference in Session 1 and 2. Neither the main effect of time between sessions, β=-0.01, *SE*=0.03, t(148.70)=-0.41, p=.684, nor the interaction term, β=-0.01, *SE*=0.02, t(149.10)=-0.73, p=.465, showed significant effects.

S2.2. Language background

NAE-learning infants showed greater IDS preferences than their non-NAE counterparts in MB1. We therefore also assessed whether test-retest reliability interacted with children's language background. A linear mixed-effects model predicting IDS preference in Session 2 based on IDS preference in Session 1 (mean-centered), NAE (centered), and their interaction, including Lab as a random intercept, revealed no interaction, β=0.29, *SE*=0.18, t(151.30)=1.59, p=.115 (Figure 1).

S2.3. Participant age

To investigate the possibility that age moderated test-retest reliability, we fit a linear mixed-effects model predicting IDS preference in Session 2 from IDS preference in Session 1

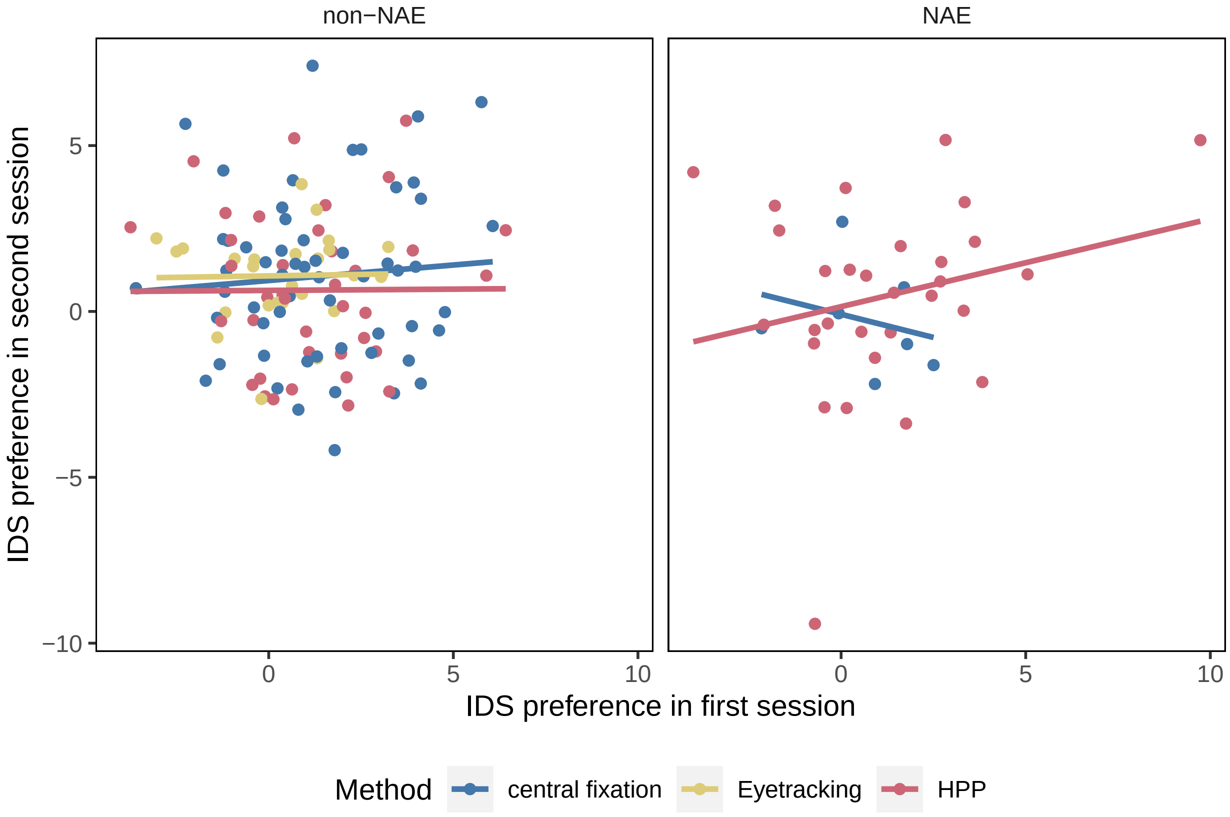

Figure 1. Infants' preference in Session 1 and Session 2 with individual data points and regression lines color-coded by method (CF, ET, or HPP). Results are plotted separately for North American English-learning infants (right panel) and infants learning other languages and dialects (right panel).

(mean-centered), participant age (mean-centered) and their interaction. The model included a by-lab random intercept and a by-lab random slope for IDS preference in Session 1. We found no evidence that age influenced test-retest reliability as indicated by the interaction between IDS preference in Session 1 and age, β=0.00, SE=0.00, t(76.60)=-0.85, p=.398.

S3. Meta-analysis of test-retest reliability

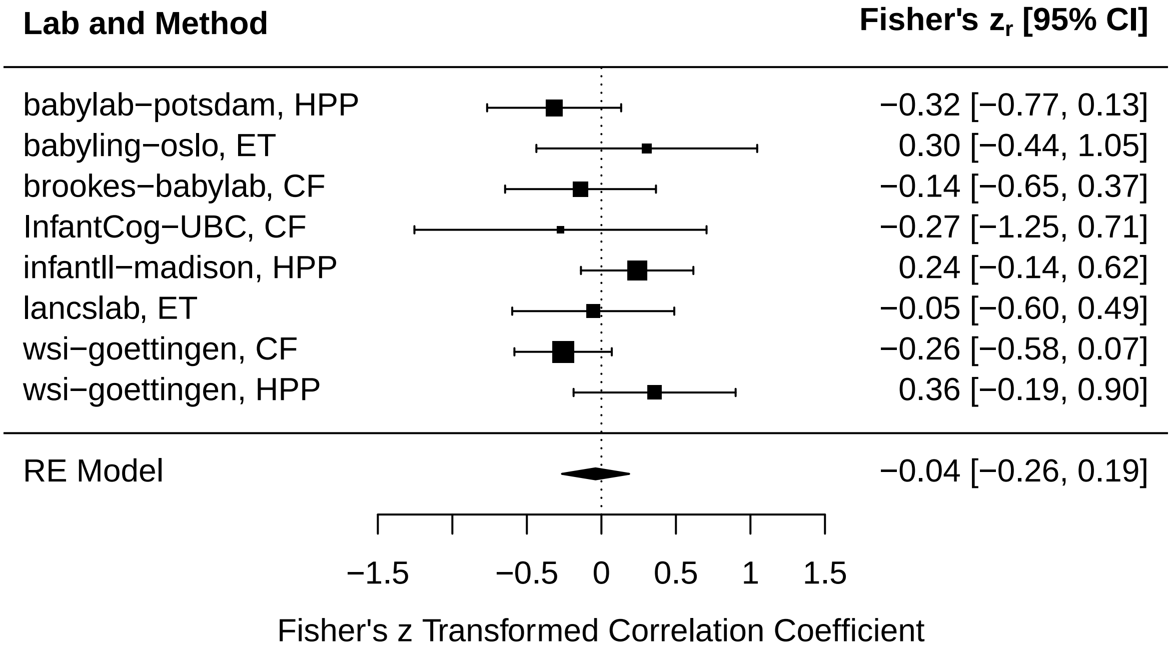

Figure 2. Forest plot of test-retest reliability effect sizes. Each row represents Fisher's z transformed correlation coefficient and 95% CI for a given lab and method (HPP = head-turn preference procedure; ET = eye-tracking; CF = central fixation). The black diamond represents the overall estimated effect size from the mixed-effects meta-analytic model.

In addition to the methods for assessing test-retest reliability reported in the main manuscript, we also investigated test-retest reliability across labs using a meta-analytic approach. We used the metafor package (Viechtbauer, 2010) to fit a mixed-effects meta-analytic model on z-transformed correlations for each combination of lab and method using sample size weighting. The model included random intercepts for lab and method. The overall effect size estimate was not significantly different from zero, b = -0.04, 95% CI = [-0.26, 0.19], p = 0.73. A forest plot of the effect sizes for each lab and method is shown in Figure 2.

Table 1

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

	Estimate	SE	t	p
Intercept	0.14	0.07	2.05	0.09
Log LT IDS Preference Session 1	-0.06	0.09	-0.68	0.50

S4. Alternative dependent variables

To check the robustness of our results, we also investigated whether we obtained similar results with other possible dependent measures: average log-transformed looking times and a proportion-based preference measure. For each alternative dependent variable, we conducted the main analyses of test-retest reliability reported in the manuscript: the overall Pearson correlation, the test-retest linear mixed-effects model, and an inspection of applying stricter inclusion criteria for number of trials contributed.

S4.1. Log-transformed looking times

In these analyses, we calculated IDS preference by first log-transforming looking times for each trial, computing the average log-transformed looking time for IDS and ADS for each participant, and calculating the difference between average IDS and ADS log-transformed looking times. 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. As in the analyses using average raw looking times, the results revealed no significant relationship between IDS preference in Session 1 and 2 (Table 1). The Pearson correlation coefficient was also not statistically significant, $r = .03$, 95% CI $[-.12, .19]$, $t(156) = 0.43$, $p = .670$. Applying successively stricter inclusion criteria — by requiring a

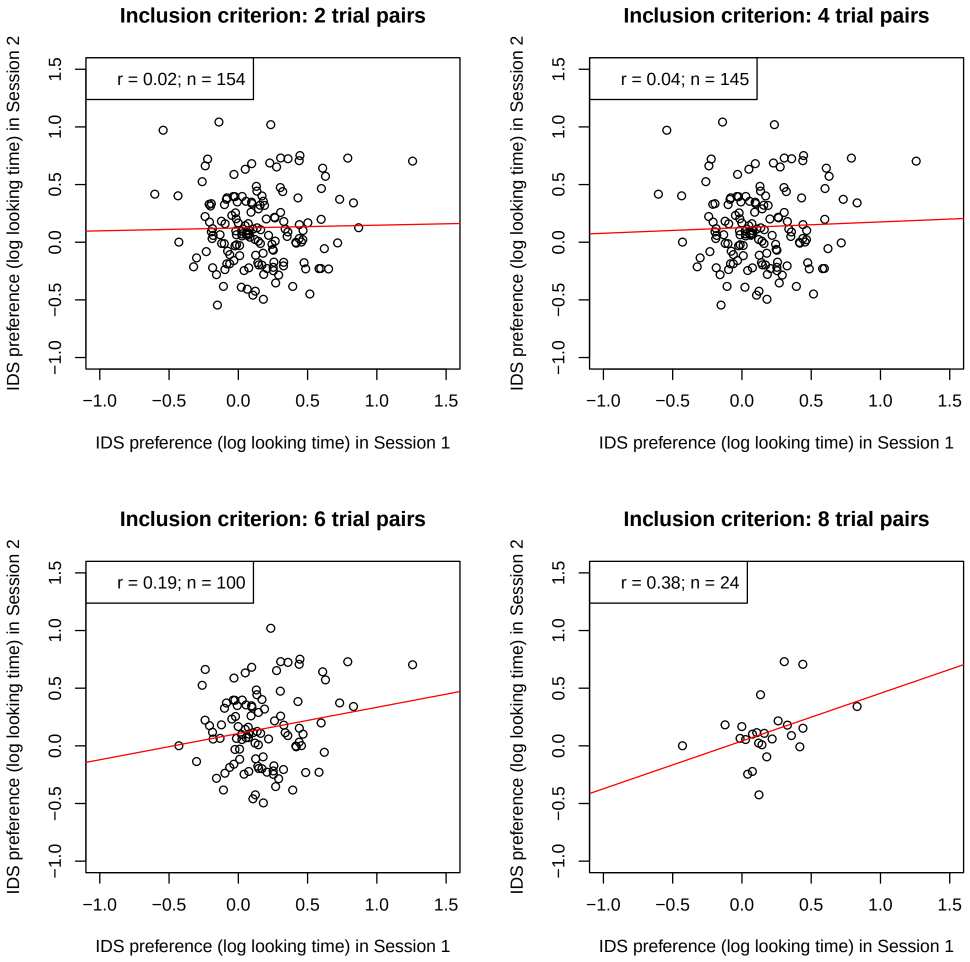

Figure 3. IDS preferences (based on average log-looking times) of both sessions plotted against each other for each inclusion criterion. n indicates the number of included infants, r is the Pearson correlation coefficient as the indicator for reliability.

higher number of valid trials per condition in each session — showed a similar pattern to the main manuscript, such that correlations increased somewhat with stricter inclusion criteria, but substantially reduced the sample size at the same time (Figure 3).

S4.2. Proportion looking to IDS

Next, we calculated a proportion-based IDS preference measure by computing the average proportion (raw) looking time to IDS relative to total (raw) looking time to IDS and ADS for each subject (i.e., IDS looking time / (ADS looking time + IDS looking time)). We fit a linear mixed-effects model predicting proportion-based IDS preference in Session 2 from proportion-based IDS preference in Session 1, including a by-lab random intercept. As in the analyses using other measures of IDS preference, the results revealed no significant relationship between IDS preference in Session 1 and 2 (Table 2). The Pearson correlation coefficient based on proportional IDS looking was also not statistically significant, $r = .01$, 95% CI $[-.15, .16]$, $t(156) = 0.09$, $p = .927$. Stricter inclusion criteria increased the correlation somewhat, as in previous analyses (Figure 4).

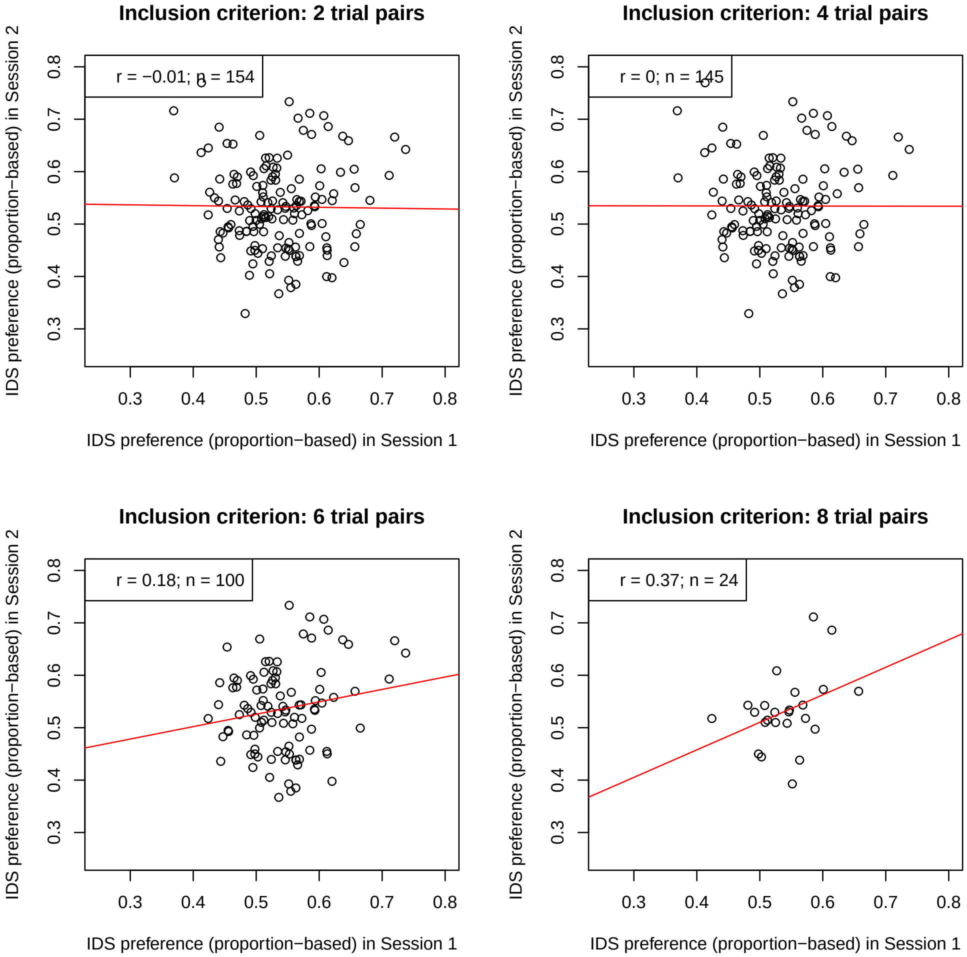

Figure 4. IDS preferences (based on proportion IDS looking) of both sessions plotted against each other for each inclusion criterion. n indicates the number of included infants, r is the Pearson correlation coefficient as the indicator for reliability.

Table 2

Coefficient estimates from a linear mixed-effects model predicting IDS preference (based on proportion IDS looking) in Session 2.

	Estimate	SE	t	p
Intercept	0.59	0.05	10.70	0.00
IDS Preference (proportion measure) Session 1	-0.10	0.10	-1.01	0.31

S5. Sensitivity of test-retest reliability to trial number inclusion criteria

To conduct a more fine-grained analysis of how stricter trial inclusion criteria affect test-retest reliability, we computed correlations while gradually increasing the number of total valid trials required for inclusion. For this analysis, we required a minimum of one IDS and one ADS trial and gradually increased the number of total valid trials required in both sessions (irrespective of IDS and ADS condition) from 2 to 16 (the maximum number of total trials). Figure 5 depicts the Pearson correlation coefficients for increasingly stricter requirements for the overall trial numbers of a given participant in both sessions. Correlations only increase and reach conventional levels of significance once the number of total required trials for both sessions is greater than 12.

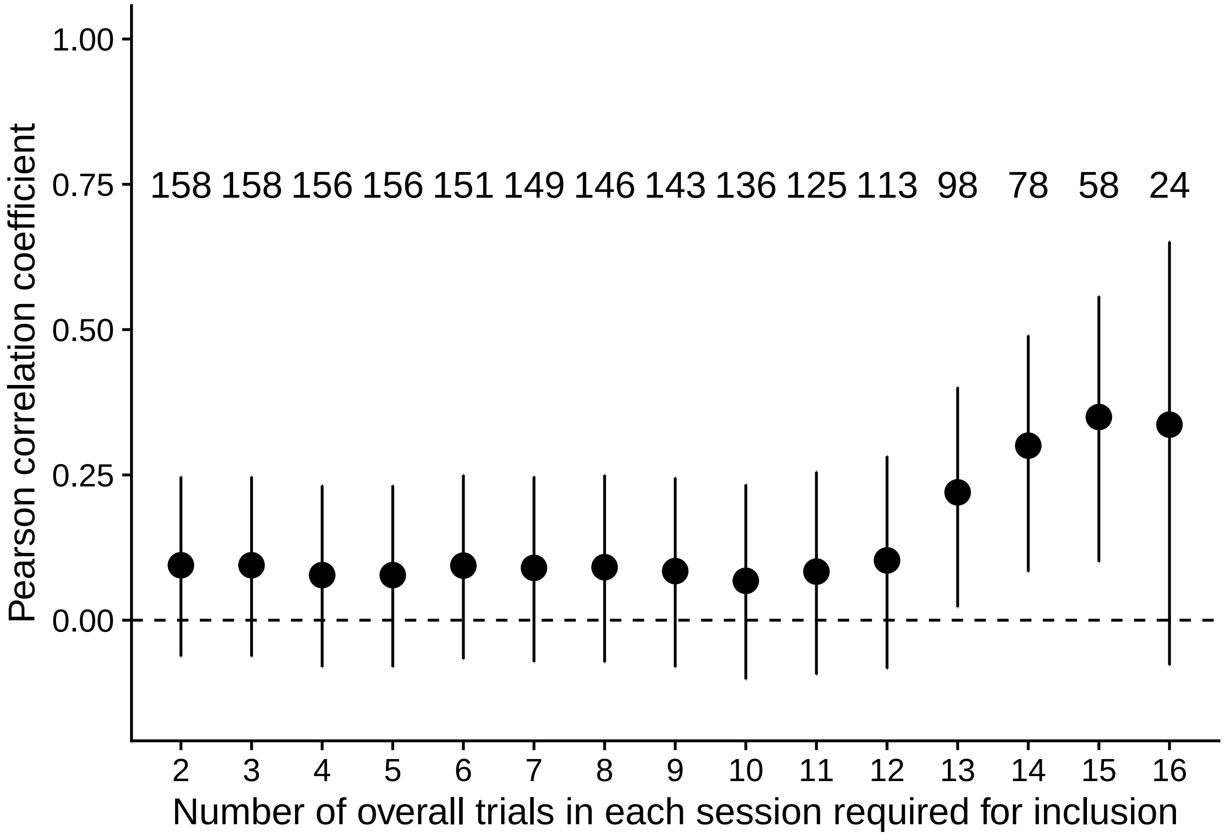

Figure 5. Pearson correlation coefficient with increasingly strict trial-level inclusion criteria. The x-axis depicts the required number of overall valid trials in both session 1 and session 2. Dots represent corresponding correlation coefficients, with 95 percent CIs. The sample size is shown above each dot.

S6. Patterns of preference across sessions

We also conducted analyses to explore whether there were any patterns of preference reversal across test sessions. While there was no strong correlation in the magnitude of IDS preference between test session 1 and test session 2, here we asked whether infants consistently expressed the same preference across test sessions. Overall, 58.20% of the infants had a consistent preference from test to retest session. Of the 158 total infants, 44.90% of infants showed a consistent IDS preference and 13.30% showed a consistent ADS preference. 23.40% of infants switched from an IDS preference at test session 1 to an ADS preference at test session 2 and 18.40% switched from an ADS preference to an IDS preference.

Next, we explored whether we could detect any systematic clustering of infants with distinct patterns of preference across the test and retest session. We took a bottom-up approach and conducted a *k*-means clustering of the test-retest difference data (here using log-transformed looking time data). We found little evidence of distinct clusters emerging from these groupings: the clusterings ranging from k=2 (2 clusters) to k=4 (4 clusters) appear to mainly track whether participants are approximately above or below the mean looking time difference for test session 1 and test session 2 (Figure 6A). The diagnostic elbow plot shows little evidence of a qualitative improvement as the number of clusters is increased, which suggests little evidence for a distinctive set of clusters of participants who showed similar patterns of looking across the test and retest sessions (Figure 6B).

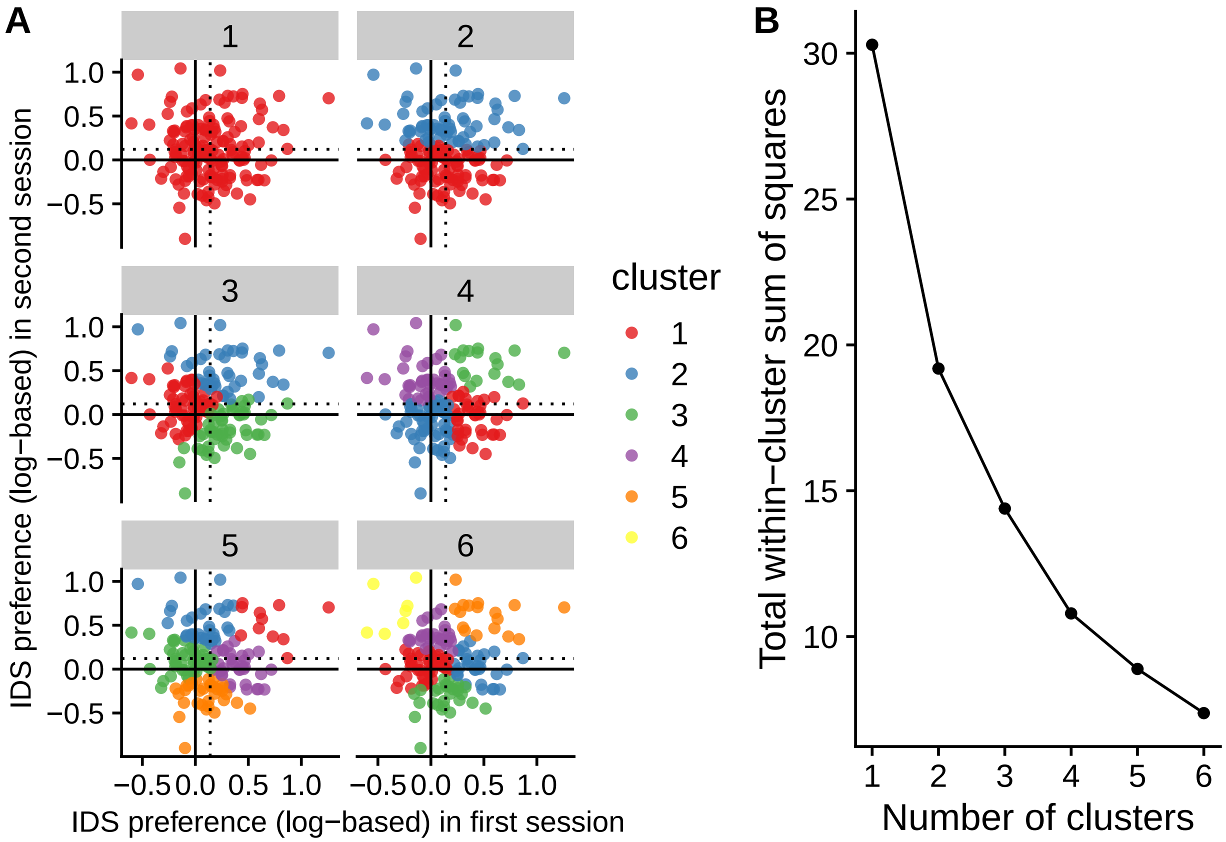

Figure 6. (A) Results from the k-means clustering analysis of IDS preference (based on average log looking times) in session 1 and 2 for different numbers of k and (B) the corresponding elbow plot of the total within-cluster sum of squares. In (A), points represent indvidual participants' magnitude of looking time difference at test sessions 1 (x-axis) and 2 (y-axis). The solid line indicates no preference for IDS vs. ADS, the dotted lines indicate mean IDS preference at test session 1 and 2, respectively. Colors indicate clusters from the k-means clustering for different values of k.

S7. Relation between number of contributed trials in each session

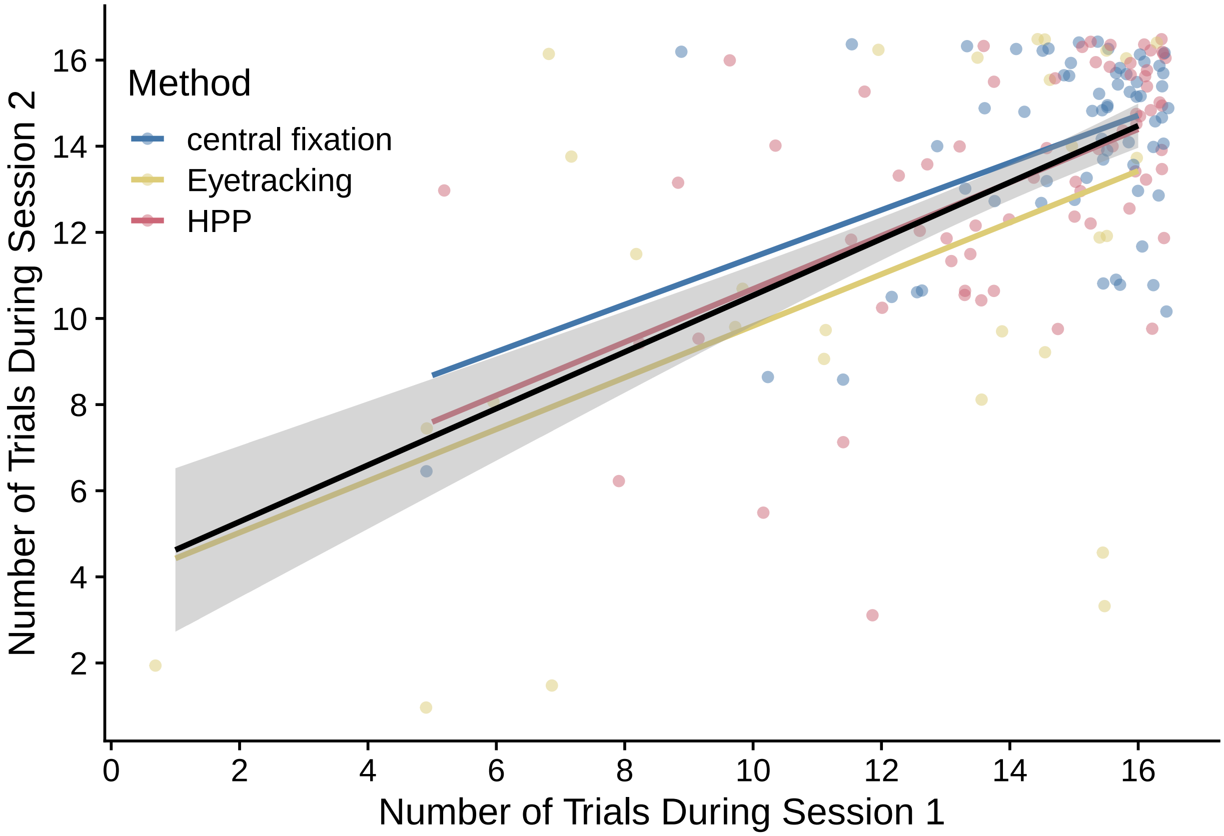

Figure 7. Correlation between the number of trials contributed in Session 1 and Session 2. Each data point represents one infant. Colored lines represent linear fits for each method.

Are there stable individual differences in how likely an infant is to contribute a high number of trials? To answer this question, we conducted an exploratory analysis investigating whether there is a relationship between the number of trials an infant contributed in Session 1 and Session 2. Do infants who contribute a higher number of trials during their first testing session also tend to contribute more trials during their second testing session? A positive correlation between trial numbers during the first and second session would indicate that there is some stability in a given infants' likelihood of remaining attentive throughout the experiment. On the other hand, the absence of a correlation would indicate that the number of trials a given infant contributes is not predictive of how many trials they might contribute during their next session.

We found a strong positive correlation between number of trials contributed during the first and the second session $r = .58$, 95% CI $[.47, .68]$, $t(159) = 9.05$, $p < .001$ (Figure 7). This result suggests that if infants contribute a higher number of trials in one session, compared to other infants, they are 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. Researchers should therefore be mindful of the fact that decisions about including or excluding infants based on trials contributed may selectively sample a specific sub-set of the infant population they are studying (Byers-Heinlein, Bergmann, & Savalei, 2021; DeBolt, Rhemtulla, & Oakes, 2020).

S8. Correlations in average looking times between sessions

To what extent are participants looking times between the two sessions related? To test this question, we first 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 $[.31, .57]$, $t(156) = 6.28$, $p < .001$. This relationship held even after controlling for number of trials in the first and second session, suggesting that the relation between average looking in Session 1 and 2 could not be entirely explained by the correlation in the number of trials contributed between the two sessions (S7), $b = 0.42$, 95% CI $[0.27, 0.58]$, $t(154) = 5.52$, $p < .001$ (Figure 8A). The result is also similar when controlling for participants' average age across the two test sessions, $b = 0.44$, 95% CI $[0.30, 0.59]$, $t(155) = 6.16$, $p < .001$.

Next, we explored the extent to which average looking times for IDS and ADS stimuli were related. First, 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 8B). To test whether these correlations were specific to looking times for IDS or ADS stimuli alone, we fit linear regression models predicting average looking to IDS (or ADS) stimuli in Session 2 from average looking to IDS and ADS stimuli in Session 1. We found that average looking to IDS stimuli in Session 2 could be predicted from average looking to IDS stimuli in Session 1, even after controlling for average looking to ADS stimuli in Session 1, $b = 0.21$, 95% CI $[0.01, 0.41]$, $t(155) = 2.11$, $p = .037$. Conversely, average looking to ADS stimuli in Session 2 could be predicted from average looking to ADS stimuli in Session 1, even after controlling for average looking to IDS stimuli in Session 1, $b = 0.36$, 95% CI $[0.14, 0.58]$, $t(155) = 3.20$, $p = .002$. These results suggest that the condition-specific correlations in average looking time cannot be fully explained by the fact that infants' overall looking times between sessions are correlated.

Finally, we inspected item-level correlations between the two test sessions. Specifically, we investigated the relation between items composed of the same recording clips in Session 1 and Session 2 (but with a reversed order of clips between the two sessions). We fit a linear mixed-effects model predicting item-level looking time in Session 2 from item-level looking time in Session 1, including random intercepts for participant, item, and lab, as well as a random slope for item-level looking time in Session 1 for participant and lab. Item-level looking in Session 2 was related to item-level looking in Session 1, $\hat{\beta} = 0.17$, 95% CI $[0.07, 0.27]$, $t(5.52) = 3.38$, $p = .017$ (Figure 8C). Similar results hold if looking times are log-transformed

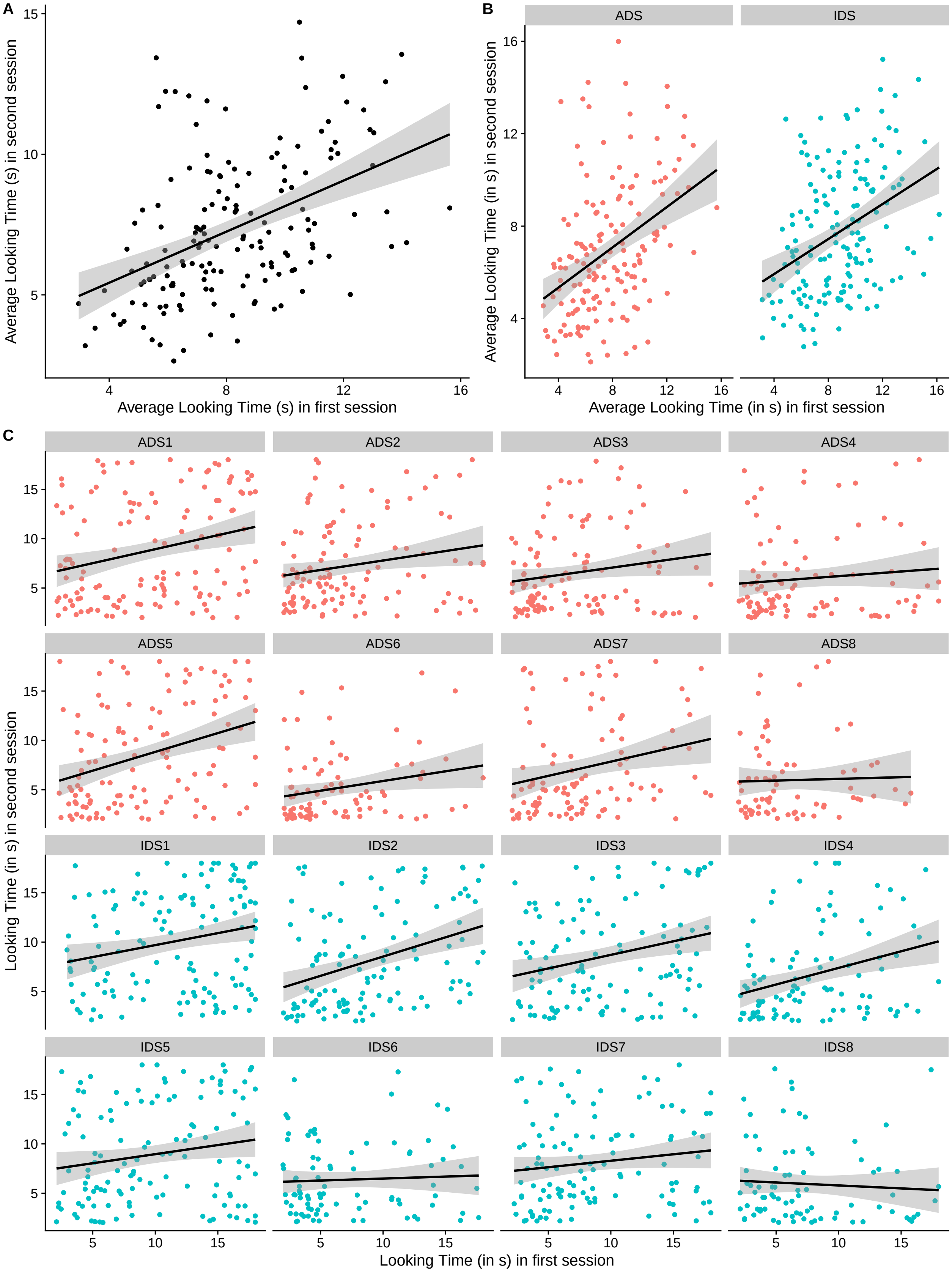

Figure 8. Correlations in average looking time (in s) between Session 1 and 2 (A) overall, (B) by condition, and (C) by item.

Table 3

Linear mixed-effects model results predicting IDS preference in Session 2 from IDS preference in Session 1 at the stimulus level.

Term	$\hat{\beta}$	95% CI	*t*	*df*	*p*
Intercept	1.02	[0.14, 1.90]	2.27	6.55	.060
Diff 1	0.07	[-0.01, 0.14]	1.79	718.46	.074

S9. By-item-pair preference scores across sessions

Finally, we inspected on a more fine-grained item level whether IDS preference in Session 1 was related to IDS preference in Session 2. To do so, we exploited the fact the specific IDS and ADS stimuli were paired together in test orders in both sessions, such that one IDS stimulus (e.g., IDS1) always occurred adjacently to a specific ADS stimulus (e.g., ADS1). We therefore computed stimulus-specific IDS preference scores by calculating the difference in raw looking time for each of the eight IDS-ADS stimulus pairs for each participant (whenever both trials in a given pair were available). We then fit a linear mixed-effects model predicting stimulus-specific IDS preference in Session 2 from stimulus-specific IDS preference in Session 1, including by-participant and by-lab random intercepts (models with more complex random effects structure, including by-item random effects, failed to converge). There was a marginal, but non-significant relation in stimulus-specific IDS preference between the two test sessions (Table 3).

References

Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2021). Six solutions for more reliable infant research. *Infant and Child Development*, e2296.

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.

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.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. Retrieved from https://doi.org/10.18637/jss.v036.i03