ManyBabies1 Test-Retest Supplementary Materials

Contents

S1. Notes on and deviations from the preregistration

S1.1. Deviations from the preregistration

Below, we have compiled a list of 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 ($n = 3$). We included these participants for two reasons. First, the general philosophy in ManyBabies studies has been to err on the side of being inclusive, as long as the data from a given participant adds valid information to the study in question. Secondly, time between test session varied continuously across participants and we planned to assess the impact of time between test on reliability. We expected that including these participants should (if anything) provide additional information (and statistical power) by extending the range of a continuous predictor variable (time between test sessions) in our moderator analyses.
- 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.), participant age (S2.2.), method (S2.3.), and the language background of the participants (S2.4.).
- While we fit all models described in the secondary analyses of the preregistration, including models investigating interactions between moderators, we interpret the more complex, three-way interaction models with caution. Our final sample size was smaller than we anticipated, which made our sample less well-powered to investigate more complex relationships between moderators. Moreover, the baseline model for these secondary interaction models was incorrectly specified in the preregistration (lower-order terms for the moderator were incorrectly removed in the planned baseline model), and we opt instead to report estimates using the more conventional method of comparing parameters of interest to models including all predictors except the main predictor of interest (e.g., estimating significance of three-way interaction terms by comparing the model fit to a model including only all lower-order predictors).
- In the by-lab meta-analysis of test-retest reliability (S3), we also separated the data by method, such that the data from one lab was split into data from the head-turn

preference procedure and the central fixation method. We separated the data by both lab and method because differences in IDS preference across methods were observed in MB1 and because this approach was more consistent with the analytic approach in the rest of the manuscript. This decision does not qualitatively affect the conclusions of the meta-analysis, namely that there was no consistent evidence of test-retest reliability.

S1.2. Additional notes

While the original idea was to retest infants that contributed data to the original study, some labs had already finished data collection for MB1 but nevertheless agreed to collect a new set of data for the MB1 test-retest spin off project. In addition, one lab already started data collection prior to the preregistration, however, this data had not been inspected or analyzed prior to the preregistration. We here present a detailed list of our collected data in relation to the original MB1 study and the preregistration (see Table 1).

Table 1

Additional notes on data collection status of each lab in relation to preregistration and MB1.

Lab	Method	Collection prior to preregistration	MB1 as Session 1
babylab-potsdam	HPP	No	No
babyling-oslo	eye-tracking	No	No
brookes-babylab	central fixation	No	No
InfantCog-UBC	central fixation	No	Yes
infantll-madison	HPP	No	No
lancslab	eye-tracking	No	No
wsi-goettingen	central fixation	Yes (n=14)	Yes
wsi-goettingen	HPP	No	No

S2. Secondary analyses investigating possible moderating variables

S2.1. Descriptives and power

S2.1.1. Additional descriptive information. To highlight the distributions of the key moderators of interest, we include an additional plot representing the distribution of infant age among the 7 participating labs, split by method and language background (Figure 1).

S2.1.2. A note on (post-hoc) power. Our final sample size (N = 158) — although quite large for typical infant looking time studies — had limited power to detect moderation effects. As a heuristic for approximate post-hoc power, we can consider the power to detect differences between correlations for our final sample. For the moderator of language background, we had n = 37 participants with a North American English language background and n =121 participants with non-North American English backgrounds.

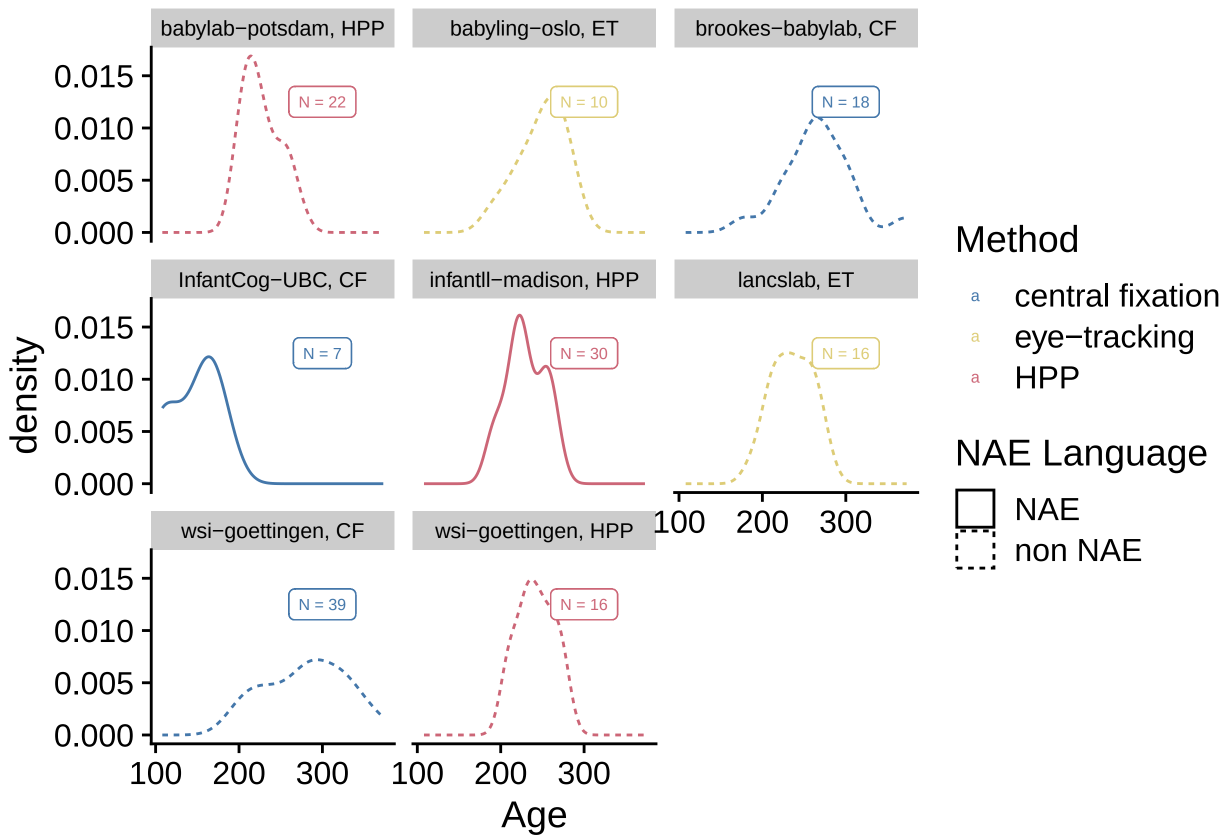

Figure 1. Distribution of participant age for each lab and method. Method is highlighted by color and language status is indicated by line type (solid = North American English; dashed = non-North American English).

Given this sample size, differences between the two samples would have to be substantial in order to have reasonable power to detect a difference: assuming $r = 0$ for one sample, we would only reach 80% power to detect a difference if $r \sim 0.5$ for the other sample. We had slightly more power to detect differences for method, where we had $n = 68$ HPP observations and $n = 90$ non-HPP observations. For example, again assuming $r = 0$ for one sample, we would reach 80% power to detect differences once $r \sim 0.43$ for the second sample. Given the limited power to detect all but large effect sizes in our moderation analyses, we planned to treat any significant results from the moderator analyses with caution.

S2.2. Time between test sessions

S2.2.1. Reliability moderated by 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.

S2.2.2. Change in preferential looking moderated by time between test sessions.

In addition to assessing the influence of moderators on test-retest reliability, we also tested whether the difference in magnitude of the IDS preference between Session 1 and Session 2 depended on moderators of interest. To investigate the influence of time between test sessions, we fit a linear mixed-effects model predicting average IDS preference from Session (centered; Session 1 vs. Session 2), days between test sessions (mean-centered), and their interaction. We included by-lab and by-participant random intercepts (more complex random effects structures did not converge due to singular fits). We found no evidence that the change in preferential looking to IDS between Session 1 and Session 2 was moderated by days between test sessions, β=-0.02, *SE*=0.04, $t(156)$=-0.48, p=.634.

S2.3. Participant age

S2.3.1. Reliability moderated by 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 (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.

S2.3.2. Change in preferential looking moderated by participant age. To investigate the potential of moderators to influence the overall magnitude of the IDS effect between Session 1 and 2, we fit a linear mixed-effects model predicting average IDS preference from Session (centered; Session 1 vs. Session 2), participant age (mean-centered), and their interaction. We included by-lab and by-participant random intercepts (more complex random effects structures did not converge due to singular fits). We found no evidence that the change in preferential looking to IDS between Session 1 and Session 2 was moderated by participant age, β=0.00, *SE*=0.00, $t(157.50)$=-0.56, p=.577.

S2.4. Method

S2.4.1. Differences in IDS preference across method. In MB1, infants who participated in the HPP showed a significantly larger magnitude of IDS preference, compared to central fixation and eye-tracking methods. Therefore, in the current study, we also explored whether the magnitude of IDS preference differed as a function of method. We fit a linear mixed-effects model predicting IDS preference from Session and Method (dummy-coded, with central fixation as the reference level), including by-lab and by-participant random intercepts. We found no significant difference in IDS preference across methods, χ^2=1.11, p=.575.

S2.4.2. Reliability moderated by method. We tested whether method (eye-tracking vs. central fixation vs. headturn preference procedure) moderated test-retest reliability by fitting a linear mixed-effects model predicting IDS preference in Session 2 from IDS preference in Session 1 (mean-centered), Method (dummy-coded, with central fixation as the reference level) and their interaction. The model included a by-lab random intercept and a by-lab random slope for IDS preference in Session 1 (models with more complex random effects structure including by-lab random effects for Method did not converge). We found no evidence that Method influenced test-retest reliability as indicated by the interaction between IDS preference in Session 1 and age, χ^2=3.85, p=.146.

S2.4.3. Reliability and its interaction with both method and age. In a more complex linear mixed-effects model (preregistered as part of our planned secondary analyses) including the interaction between IDS preference in Session 1 (mean-centered), Method (dummy-coded, with central fixation as the reference level), participant age (mean-centered), and all lower order interactions, we find evidence for an interaction between method and age in predicting reliability, χ^2=6.44, p=.040. This effect appears to be mainly driven by older infants showing some evidence of test-retest reliability for the HPP, $r = 0.45$, $p = 0.02$ (see Figure 3B). However, we believe these tentative findings should be treated with caution, due to the small size of our infant sample once binned by multiple moderating factors.

S2.4.4. Change in preferential looking moderated by age and method. We fit a linear mixed-effects model predicting average IDS preference from the three-way interaction of Session (centered; Session 1 vs. Session 2), participant age (mean-centered), Method (dummy-coded, with central fixation as the reference level), and all lower order predictors. We included a by-participant random intercept (more complex random effects structures did not converge due to singular fits). We found no evidence that the change in preferential looking to IDS between Session 1 and Session 2 was moderated by participant age and Method, β=-0.01, SE=0.02, t(155.40)=-0.58, p=.562.

S2.5. Language background

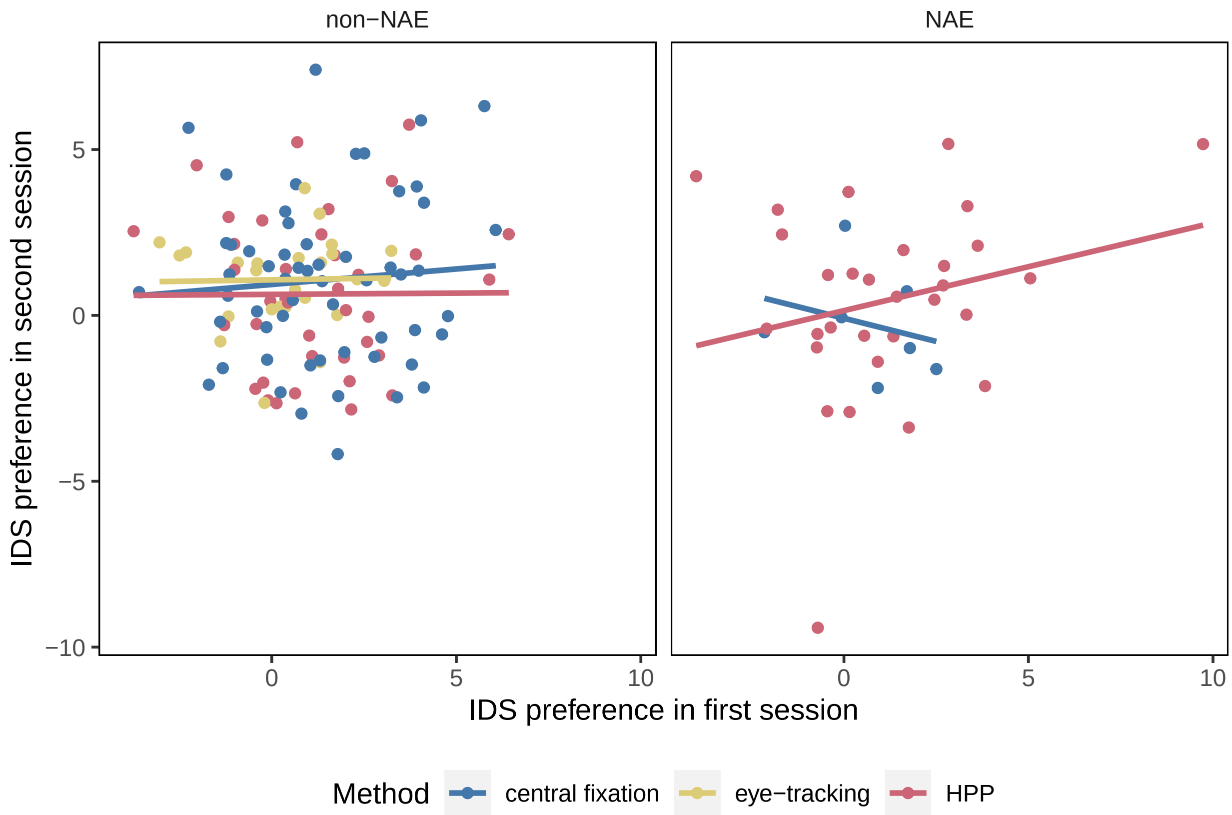

Figure 2. Infants' preference in Session 1 and Session 2 with individual data points and regression lines color-coded by method (CF, ET, or HPP). Results are plotted separately for North American English-learning infants (right panel) and infants learning other languages and dialects (left panel).

S2.5.1. Reliability moderated by 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 2).

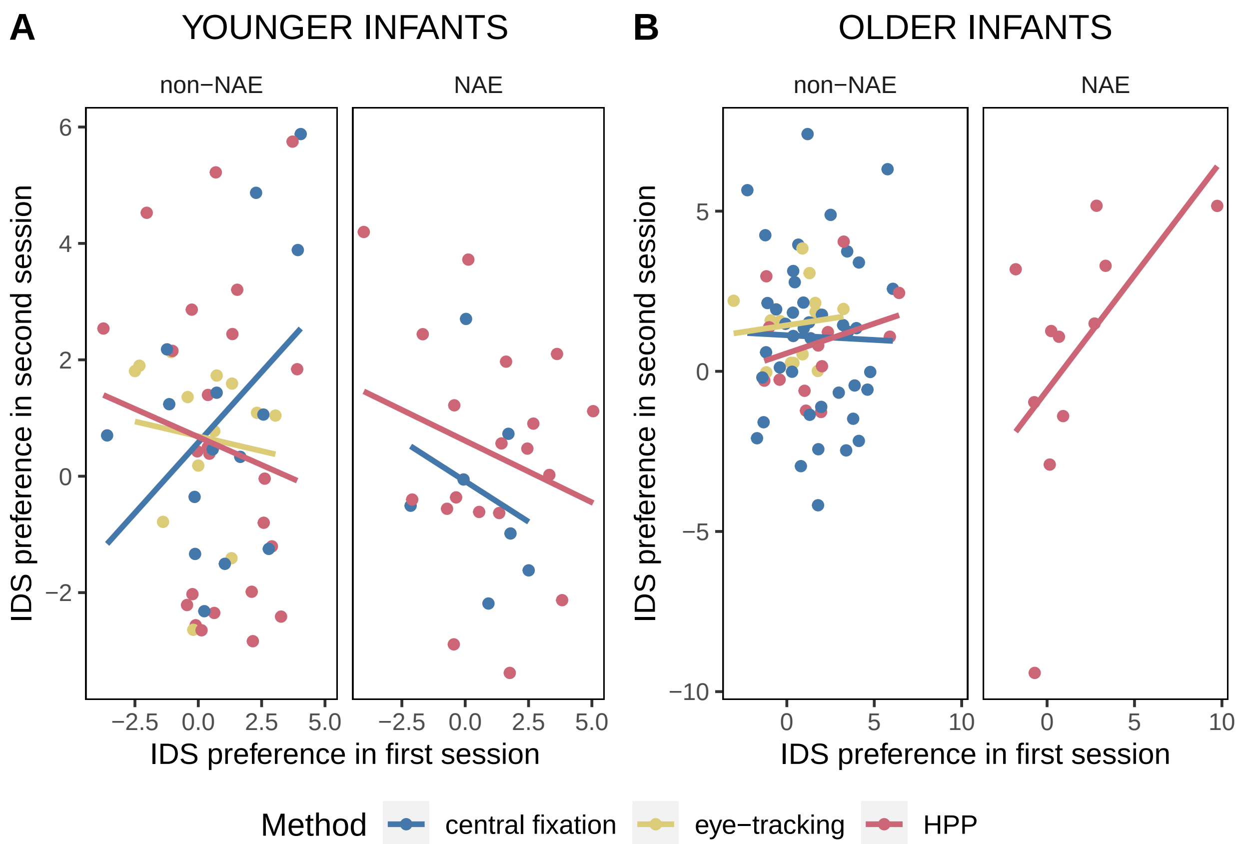

Figure 3. Infants' preference in Session 1 and Session 2 with individual data points and regression lines color-coded by method for (A) younger and (B) older infants (median-split). Results are plotted separately for North American English-learning infants and infants learning other languages and dialects.

S2.5.2. Reliability and its interaction between language background and age. We also fit a preregistered linear mixed-effects model predicting IDS preference in Session 2 from the three-way interaction between IDS preference in Session 1 (mean-centered), NAE (centered), participant age (mean-centered), and all lower order interactions. We find evidence for an interaction between language background and age in predicting reliability, β=0.01, *SE*=0.00, $t(63.70)$=2.43, p=.018. Figure 3 illustrates that this interaction was driven by a small set of older infants (all from a single lab and participating in the HPP method) showing a somewhat more reliable relationship between

Session 1 and Session 2 looking. Note that the mixed-effects analyses use Age as a continuous predictor — age is median-split in Figure 3 to ease visualization. Given the small number of infants driving the three-way interaction and the confounded nature of this sample (with method and lab), we do not draw strong conclusions from the existence of this three-way interaction, but report it here to spur future investigations into how age and experience interacts with test-retest reliability.

S2.5.3. Change in preferential looking moderated by age and language background. We fit a linear mixed-effects model predicting average IDS preference from the three-way interaction of Session (centered; Session 1 vs. Session 2), participant age (mean-centered), NAE (centered), and all lower order predictors. We included by-lab and by-participant random intercepts and by-lab random slope for Session (more complex random effects structures did not converge due to singular fits). We found no evidence that the change in preferential looking to IDS between Session 1 and Session 2 was moderated by participant age and language background, β=0.01, SE=0.02, t(114.60)=0.95, p=.347.

S3. Meta-analysis of test-retest reliability

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 4.

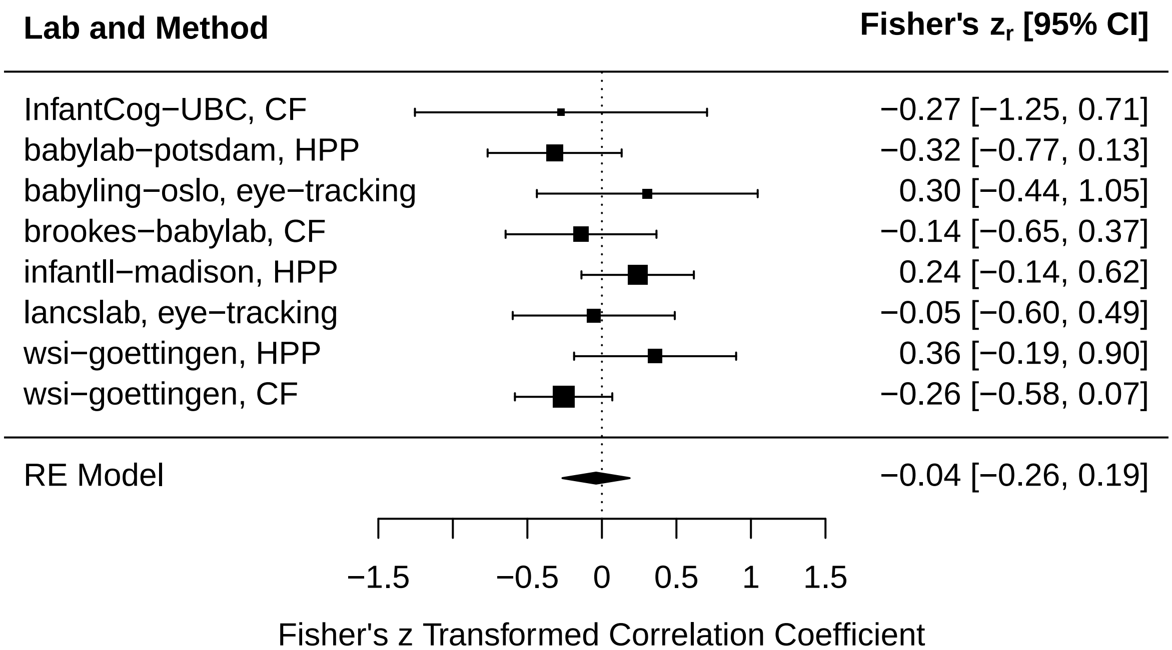

Figure 4. Forest plot of test-retest reliability effect sizes. Each row represents Fisher's z transformed correlation coefficient and 95% CI for a given lab and method (HPP = head-turn preference procedure; ET = eye-tracking; CF = central fixation). The black diamond represents the overall estimated effect size from the mixed-effects meta-analytic model.

S4. Analyses including a more restricted sample

Given that we found that restricting the sample to participants contributing at least 6 ADS and IDS trials in both sessions, we conducted the central analyses with this more restricted infant sample.

S4.1. Descriptives and IDS preference for the restricted sample

The participants in the restricted sample — contributing at least 6 IDS and ADS trials for both sessions — were distributed across the contributing labs, methods, and

Table 2

Statistics of the included labs for the restricted sample (min 6 trials contributed per session). N refers to the number of infants included in the analysis.

Lab	Method	Language	Mean age (days)	N
InfantCog-UBC	central fixation	English	136	5
babylab-potsdam	HPP	German	224	18
babyling-oslo	eye-tracking	Norwegian	250	1
brookes-babylab	central fixation	English	254	15
infantll-madison	HPP	English	233	12
lancslab	eye-tracking	English	235	10
wsi-goettingen	HPP	German	240	13
wsi-goettingen	central fixation	German	281	26

language backgrounds (Table 2). There was no difference in average age between the main sample and the restricted sample, $t(204.57) = -0.33$, $p = .744$. There was a robust preference for IDS in both Session 1, $t(99) = 6.67$, $p < .001$, and Session 2, $t(99) = 4.42$, $p < .001$. We observed no difference in IDS preference between the two sessions, β=-0.34, SE=0.28, p=.225.

Interestingly, while there was a significant simple correlation between IDS preference in session 1 and session 2, $r = .22$, 95% CI [.02, .40], $t(98) = 2.23$, $p = .028$, we found that IDS preference in Session 1 did not significantly predict IDS preference in Session 2 in a linear mixed-effects model including a by-lab random intercept, β=0.12, SE=0.11, p=.255.

S4.2. Moderator analyses including a more restricted sample

S4.2.1. Time between test sessions. As in the analyses with the full dataset, 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=-0.03, SE=0.03, t(95.80)=-0.96, p=.342, nor the interaction term, β=-0.01, SE=0.03, t(93.60)=-0.22, p=.828, showed significant effects.

S4.2.2. Participant age

To investigate the possibility that age moderated test-retest reliability in the restricted sample, we fit a linear mixed-effects model predicting IDS preference in Session 2 from IDS preference in Session 1 (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(43.20)=-0.69, p=.494.

S4.2.3. Method

We tested whether method (ET vs. CF vs. HPP) moderated test-retest reliability by fitting a linear mixed-effects model predicting IDS preference in Session 2 from IDS preference in Session 1 (mean-centered), Method (dummy-coded, with CF as the reference level), 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 Method influenced test-retest reliability as indicated by the interaction between IDS preference in Session 1 and age, χ^2=3.85, p=.146. There was no significant relationship between IDS preference for Session 1 and Session 2 for each method considered separately, CF: β=-0.06,

SE=0.16, p=.704; HPP: β=0.26, SE=0.17, p=.139; eye-tracking: β=-0.04, SE=0.26, p=.866.

S4.2.4. Language background

As in the main sample, 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=0.31, SE=0.24, $t(95.10)$=1.29, p=.199.

Table 3

Correlations between alternative dependent measures

	1	2	M	SD
1. Diff	-		1.21	2.22
2. Prop	.96***	-	0.54	0.07
3. Diff_log_lt	.95***	.96***	0.16	0.30

Note. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

S5. 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.

S5.1. Correlations between alternative dependent variables

First, we consider the correlations between the three dependent measures we considered for IDS preference: (a) a simple difference score between average IDS and ADS looking times (main manuscript), (b) a difference score between average log-transformed looking times, and (c) the proportion-based preference measure. As expected, the correlations between the alternative dependent measures were very high, all r-values >= 0.95 (Table 3).

Table 4

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

S5.2. 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 4). 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 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 5).

S5.3. 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

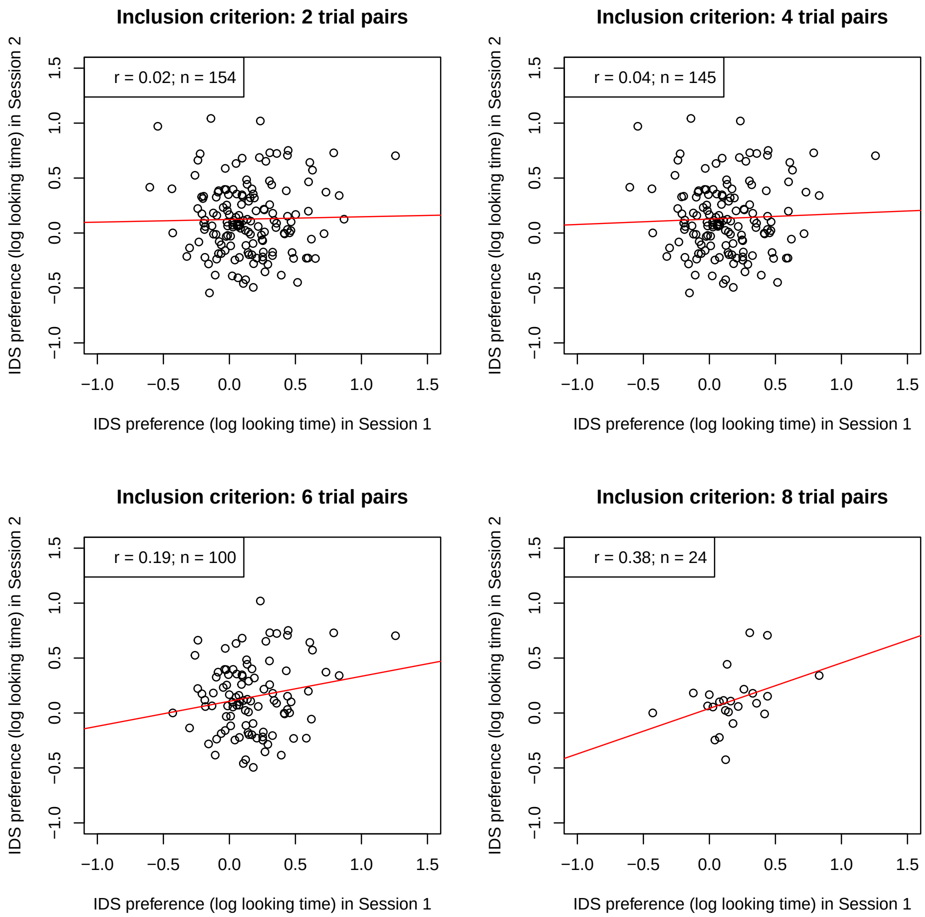

Figure 5. IDS preferences (based on average log-looking times) of both sessions plotted against each other for each inclusion criterion. n indicates the number of included infants, r is the Pearson correlation coefficient as the indicator for reliability.

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 5). 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 6).

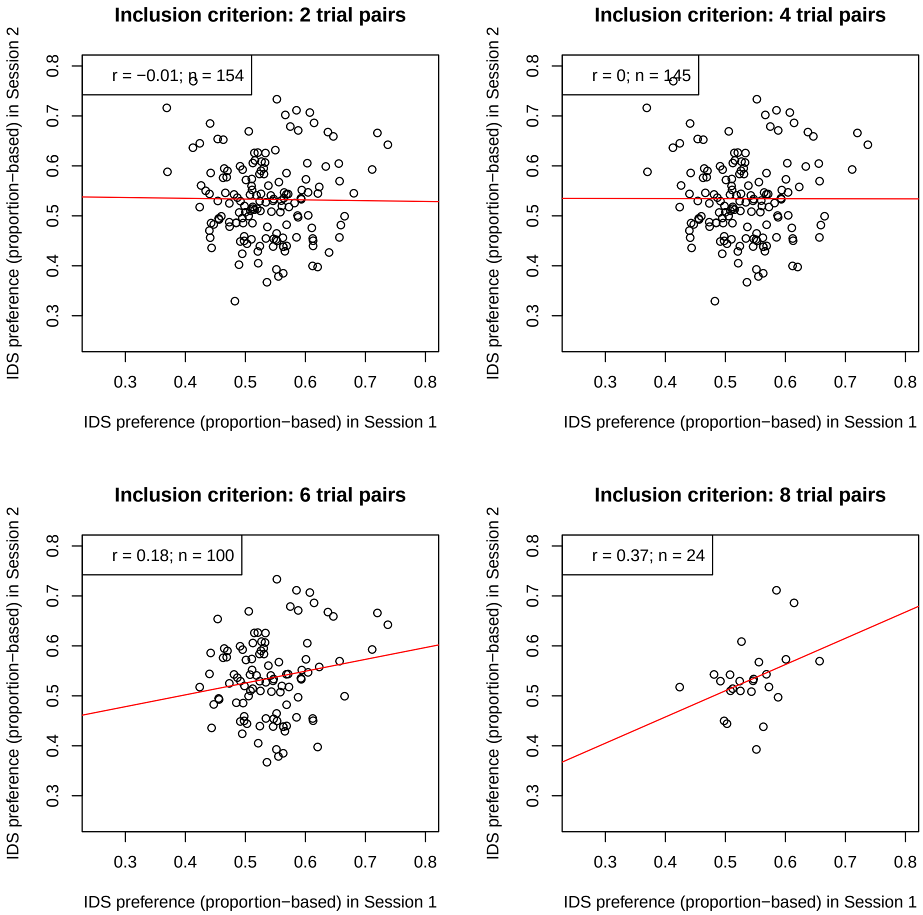

Figure 6. IDS preferences (based on proportion IDS looking) of both sessions plotted against each other for each inclusion criterion. n indicates the number of included infants, r is the Pearson correlation coefficient as the indicator for reliability.

Table 5

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

S6. 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 7 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.

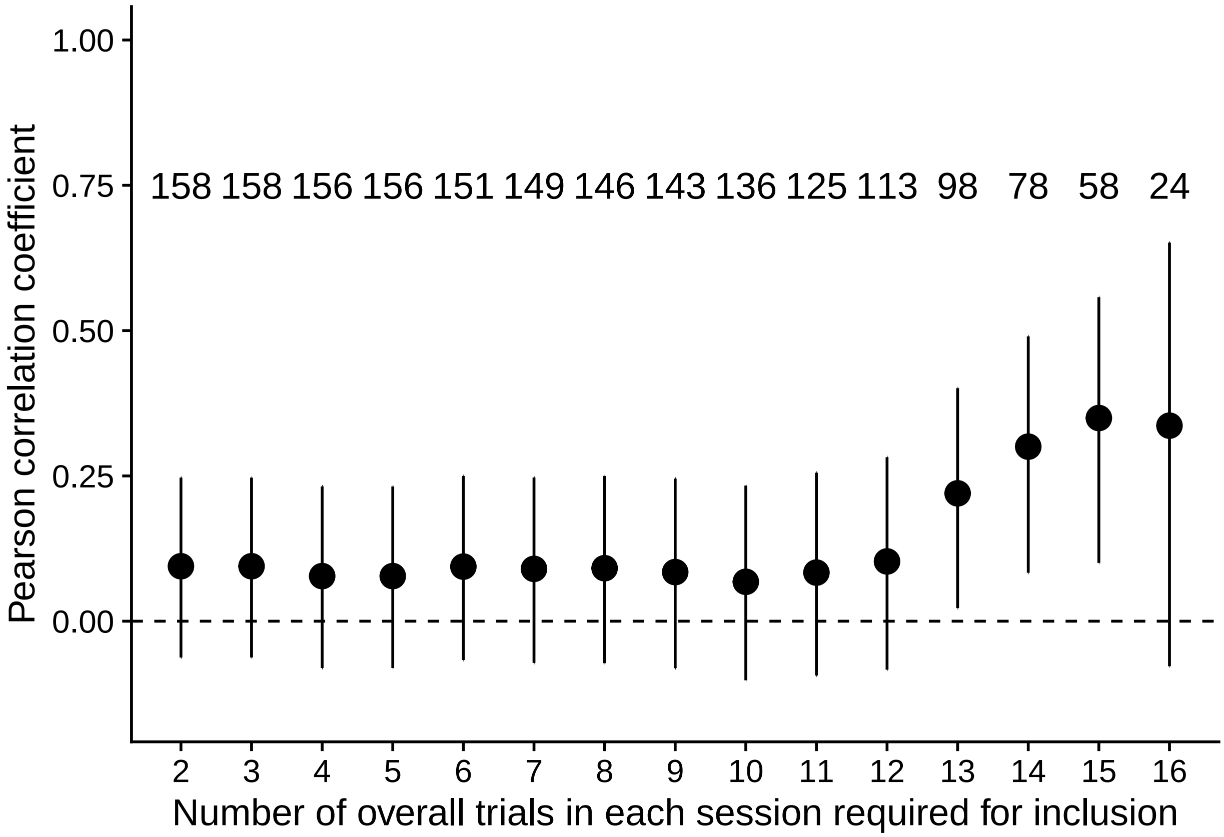

Figure 7. Pearson correlation coefficient with increasingly strict trial-level inclusion criteria. The x-axis depicts the required number of overall valid trials in both Session 1 and Session 2. Dots represent corresponding correlation coefficients, with 95 percent CIs. The sample size is shown above each dot.

S7. 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 Session 1 and 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 Session 1 to an ADS preference at

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 Session 1 and Session 2 (Figure 8A). 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 8B).

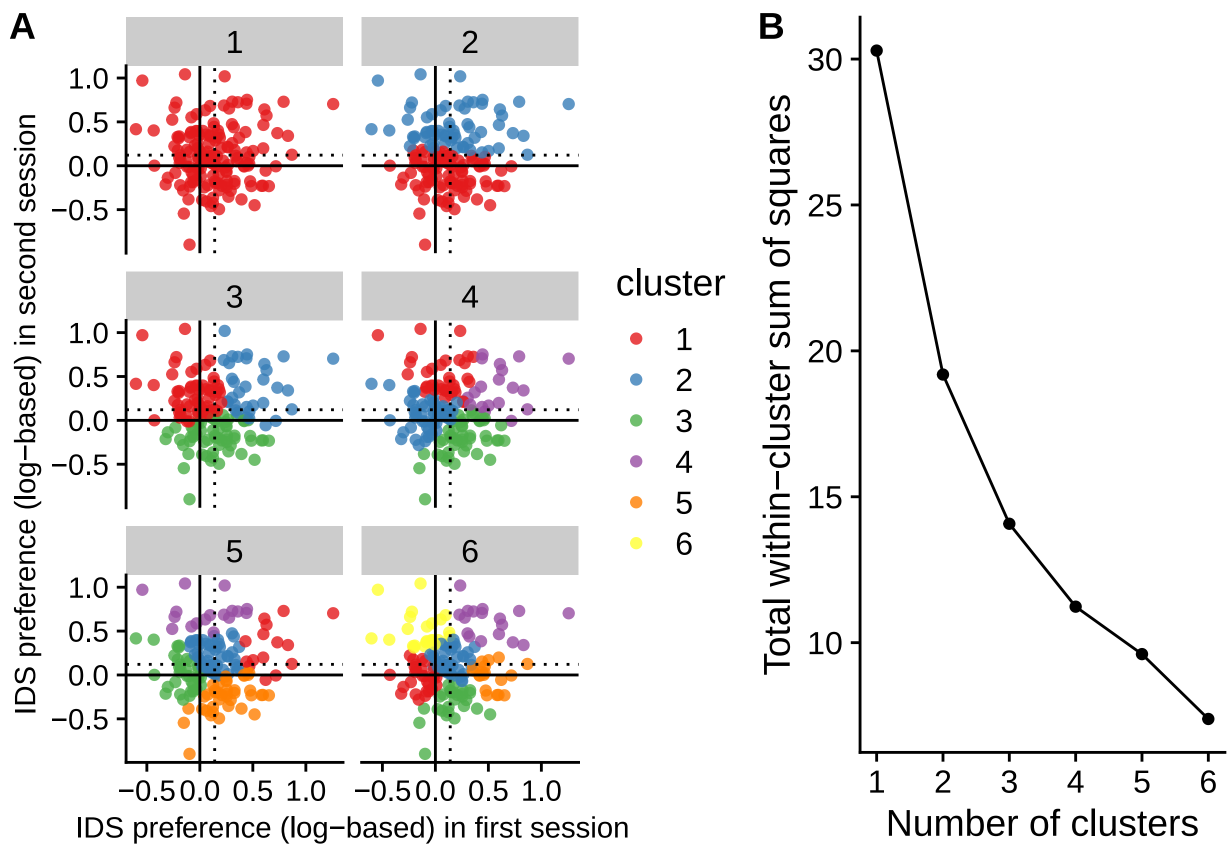

Figure 8. (A) Results from the k-means clustering analysis of IDS preference (based on average log looking times) in Session 1 and 2 for different numbers of k, and (B) the corresponding elbow plot of the total within-cluster sum of squares. In (A), points represent indvidual participants' magnitude of looking time difference at 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 Session 1 and 2, respectively. Colors indicate clusters from the k-means clustering for different values of k.

S8. Correlations in average looking times between sessions

As reported in the main manuscript, we found that infants' average looking time was correlated between their two testing sessions. We also 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 $[.26, .53]$, $t(156) = 5.49$, $p < .001$. 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.

We next 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 9). Similar results hold if looking times are log-transformed.

In MB1, the ordering of stimuli was counterbalanced, but some stimuli still appeared earlier in the experiment than others. For example, the IDS1 and ADS1 speech stimuli appeared on trials 1,2,5, or 6, while the IDS8 and ADS8 speech stimuli always occurred on the final two trials (trial number 15 or 16). This means that the interpretation of the correlations between individual speech stimuli must also take into account these stimuli tend to be occurring in earlier or later portions of the experiment (when infants are more or less attentive and show longer looking times in general). To further investigate the impact of trial number on by-item correlations in looking time, we fit an interaction model testing whether the magnitude of the item-level correlation depended on the trial number for a given session. We fit a linear mixed-effects model predicting item-level looking time in Session 2 from the interaction between item-level looking time in Session 1 and trial number in Session 1 (trial numbers across sessions are almost always identical). The model included random intercepts for participant, item, and lab, as well as random slopes for item-level looking time and trial number in Session 1 for participant and lab. We indeed found that the magnitude of the item-level correlations in looking time between sessions depended on trial number, $\hat{\beta} = -0.01$, 95% CI $[-0.02, 0.00]$, $t(1, 200.31) = -2.53$, $p = .012$, with the strength of the relation between sessions declining as trial number increased. While trial number was a strong predictor of Session 2 looking time, $\hat{\beta} = -0.28$, 95% CI $[-0.36, -0.20]$, $t(8.67) = -6.85$, $p < .001$, item-level looking in Session 1 only marginally predicted Session 2 looking when controlling for trial number, $\hat{\beta} = 0.10$, 95% CI $[0.01, 0.20]$, $t(6.47) = 2.12$, $p = .075$. Variation in item-level correlations is therefore at least partially due to the ordering of the stimuli in the experiment, rather than a sole function of differences between the stimuli *per se*.

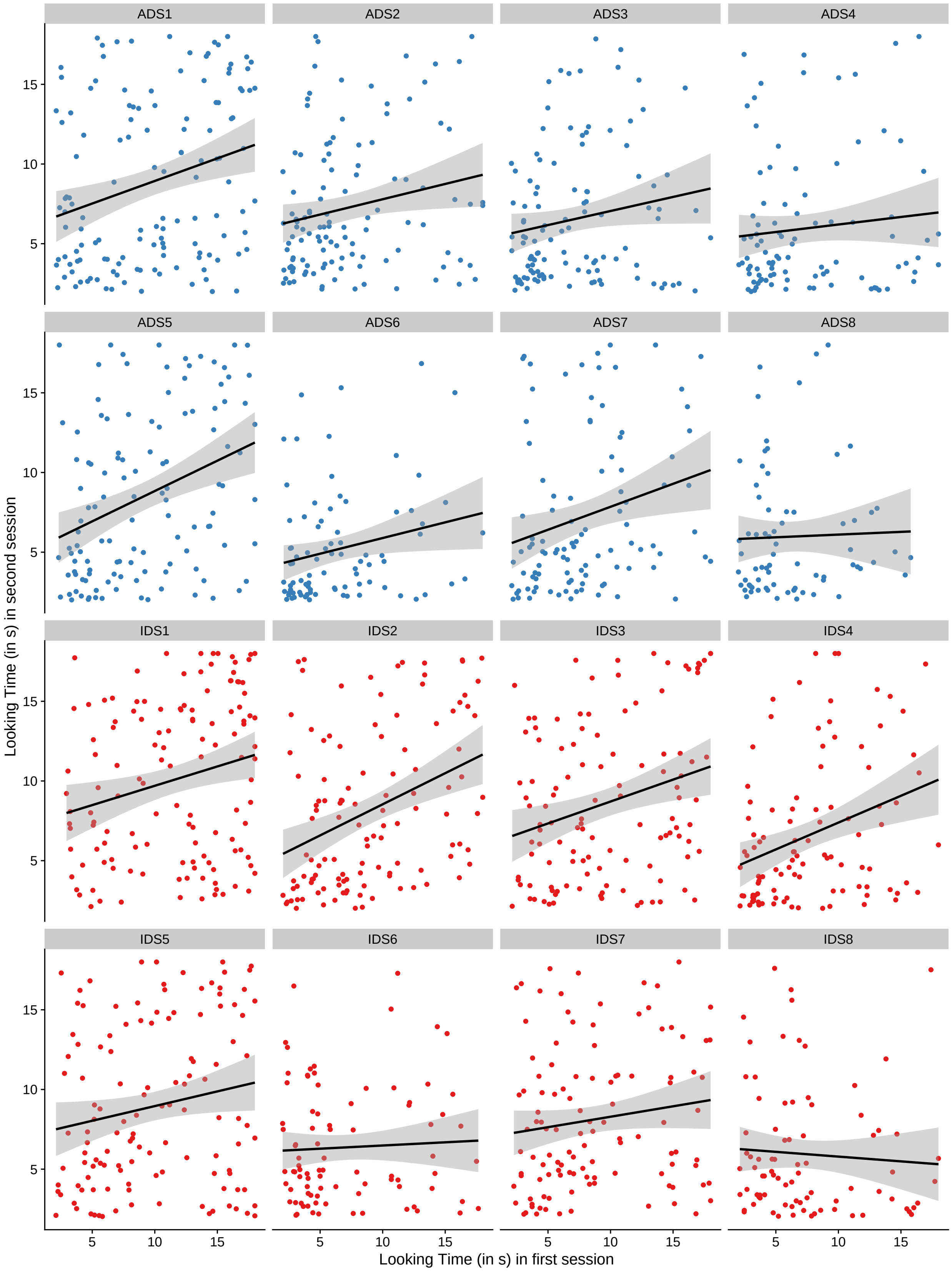

Figure 9. Correlations in average looking time (in s) between Session 1 and 2 (A) overall, (B) by condition, and (C) by item.

Table 6

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 6).

S10. Overall looking times and test-retest IDS preference

S10.1. Correlations between overall looking time and IDS preference

We also investigated whether preferential looking behavior varied as a function of infants' tendencies to look for longer or shorter periods of time on average across all stimuli. We found no evidence for a correlation between infants' average looking time and the magnitude of IDS preference (Figure 10), either in Session 1 ($r = .05$, 95% CI $[-.11, .20]$, $t(156) = 0.61$, $p = .543$) or in Session 2 ($r = -.06$, 95% CI $[-.21, .10]$, $t(156) = -0.70$, $p = .483$).

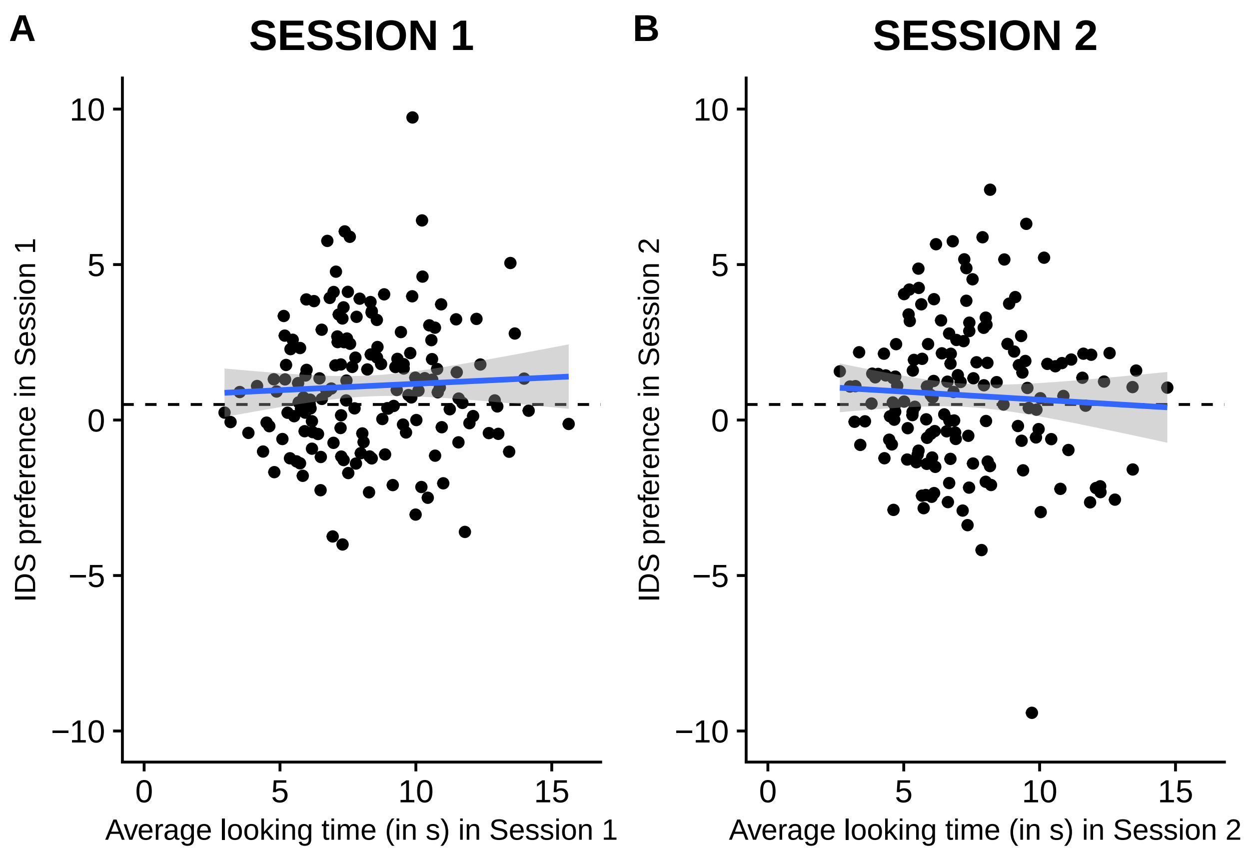

Figure 10. Correlations between average looking time (in s) and IDS preference in (A) Session 1 and (B) Session 2.

S10.2. Does average looking time moderate test-retest reliability?

Do longer lookers or shorter lookers show a tendency towards higher test-retest reliability? Next, we tested whether infants' tendency to look for longer or shorter periods (during Session 1) moderated test-retest reliability. We fit a linear mixed-effects model predicting IDS preference in Session 2 from the interaction of IDS preference in Session 1 and average looking time in Session 1. The model included a by-lab random intercept and a by-lab random slope for IDS preference (Session 1). Average looking time during Session 1 did not significantly moderate test-retest reliability, β=0.07, SE=0.04, t(121.50)=1.79, p=.076. The direction of this marginal, non-significant effect is consistent with a slight increase in test-retest reliability as average looking time increases (Figure 11). However, overall, we find no significant evidence for robust differences between long and short lookers.

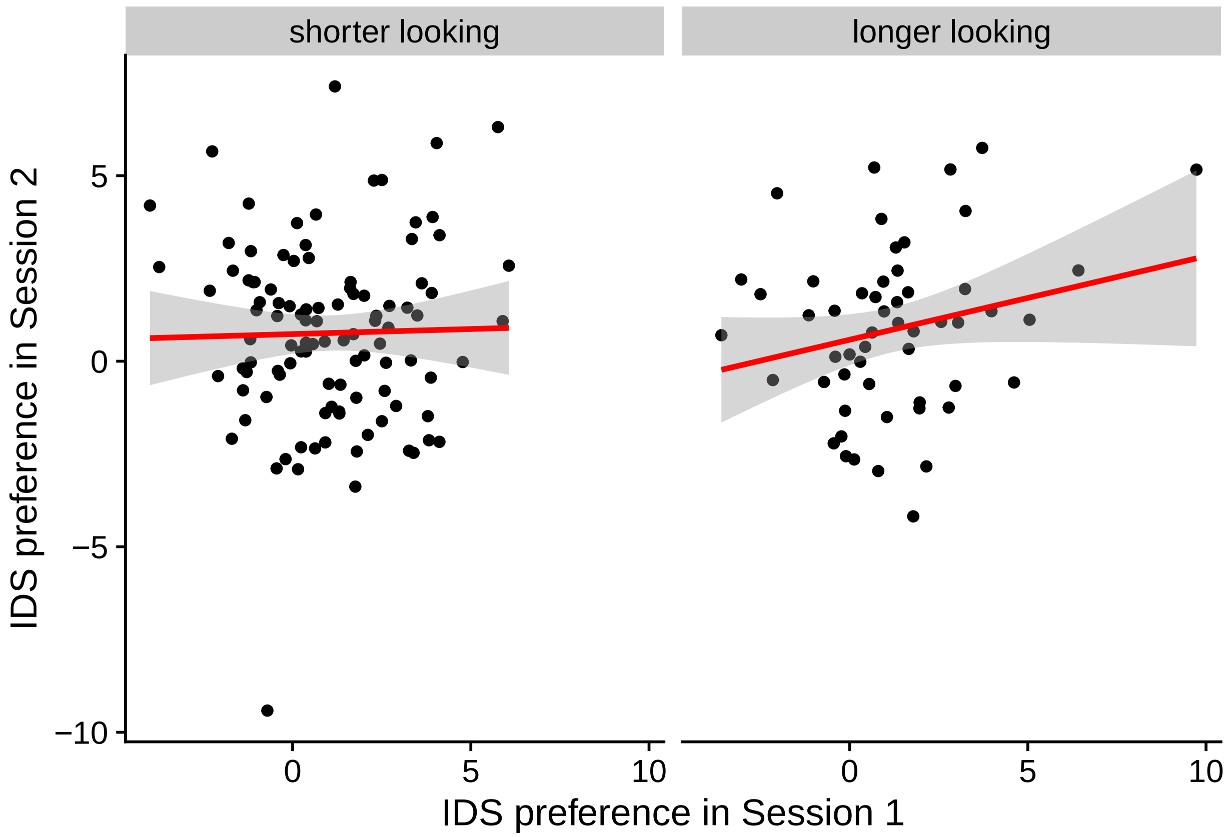

Figure 11. Correlations between average looking time (in s) and IDS preference in (A) Session 1 and (B) Session 2.

References

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