1   Manybabies1 Test-Retest Supplementary Materials

2

3

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), 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).

S2. Secondary analyses investigating possible moderating variables

S2.1. Descriptives and power

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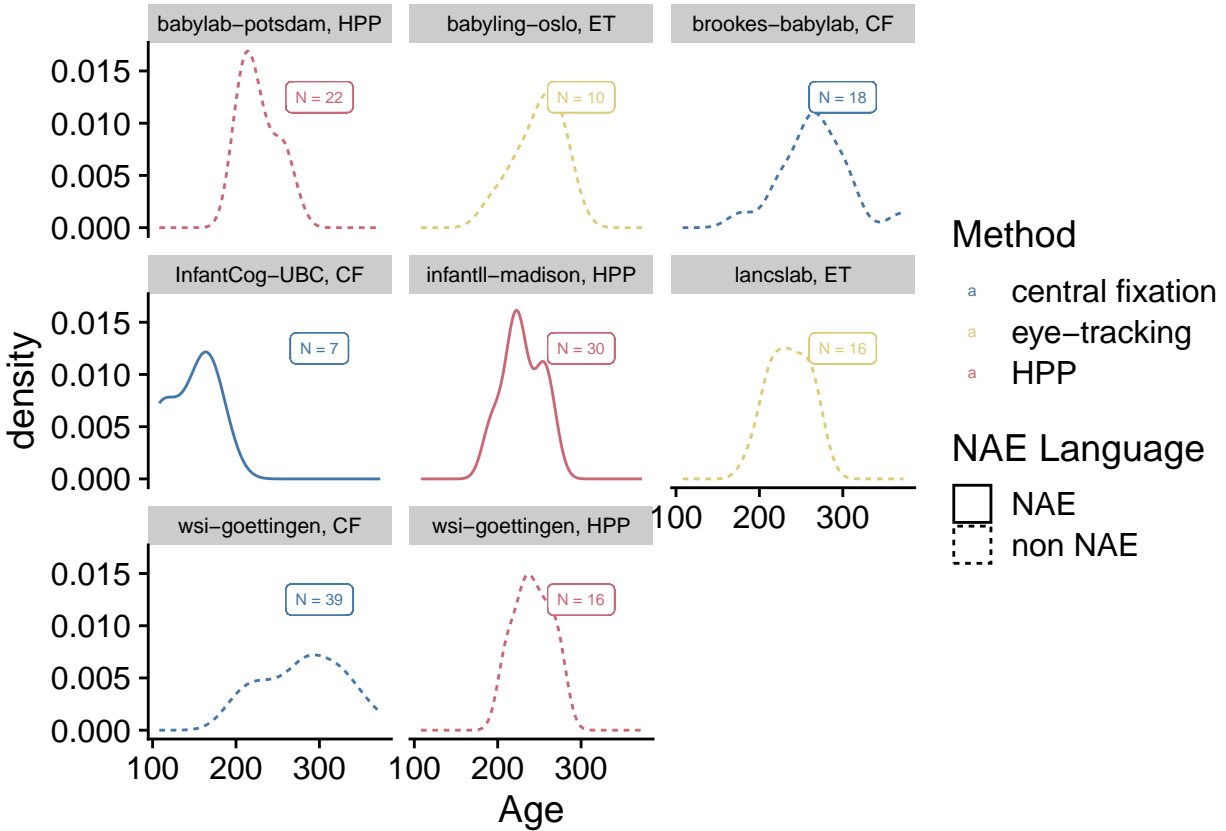
To highlight the distributions of the key moderators of interest, we include an additional plot representing the distribution of infant age among the 7 participating labs, split by method and language background (Figure 1).

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.

100 Given this sample size, differences between the two samples would have to be substantial in

101 order to have reasonable power to detect a difference: assuming $r = 0$ for one sample, we

102 would only reach 80% power to detect a difference if $r \sim 0.5$ for the other sample. We had

103 slightly more power to detect differences for method, where we had $n = 68$ HPP

104 observations and $n = 90$ non-HPP observations. For example, again assuming $r = 0$ for

105 one sample, we would reach 80% power to detect differences once $r \sim 0.43$ for the second

106 sample. Given the limited power to detect all but large effect sizes in our moderation

107 analyses, we planned to treat any significant results from the moderator analyses with

108 caution.

S2.2. Time between test sessions

S2.2.1. Reliability moderated by time between test sessions.   The number

111 of days between the first and second testing session varied widely across participants

112 (mean: 10 days; range: 1 - 49 days). We therefore tested for the possibility that the time

113 between sessions might have an impact on test-retest reliability. We fit a linear

114 mixed-effects model predicting IDS preference in Session 2 from IDS preference in Session 1

115 (mean-centered), number of days between testing sessions (mean-centered), and their

116 interaction, including a by-lab random intercept and random slope for IDS preference in

117 Session 1. A more complex random effects structure including additional random slopes for

118 number of days between test sessions and its interaction with IDS preference in Session 1

119 did not converge. We found no evidence that the number of days between test sessions

120 moderated the relationship between IDS preference in Session 1 and 2. Neither the main

121 effect of time between sessions, β=-0.01, SE=0.03, $t(148.70)$=-0.41, p=.684, nor the

122 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

125 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).

There were two key results. 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. IDS preference moderated by method. In ManyBabies1, infants who participated in the headturn preference procedure 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, $\chi^2=1.11$, $p=.575$.

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, $\chi^2=3.85$, $p=.146$.

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, $\chi^2=6.44$, $p=.040$. This effect appears to be mainly driven by older infants showing some evidence of test-retest reliability for the headturn preference procedure, $r = 0.45$, $p = 0.02$ (see Figure 2B). However, we believe

176 these tentative findings should be treated with caution, due to the small size of our infant

177 sample once binned by multiple moderating factors.

178   **S2.4.4. Change in preferential looking moderated by age and method.**   We

179 fit a linear mixed-effects model predicting average IDS preference from the three-way

180 interaction of Session (centered; Session 1 vs. Session 2), participant age (mean-centered),

181 Method (dummy-coded, with central fixation as the reference level) and all lower order

182 predictors. We included a by-participant random intercept (more complex random effects

183 structures did not converge due to singular fits). We found no evidence that the change in

184 preferential looking to IDS between Session 1 and Session 2 was moderated by participant

185 age and Method, β=-0.01, SE=0.02, $t(155.40)$=-0.58, p=.562.

186 **S2.5. Language background**

187   **S2.5.1. Reliability moderated by language background.**   NAE-learning

188 infants showed greater IDS preferences than their non-NAE counterparts in MB1. We

189 therefore also assessed whether test-retest reliability interacted with children's language

190 background. A linear mixed-effects model predicting IDS preference in Session 2 based on

191 IDS preference in Session 1 (mean-centered), NAE (centered), and their interaction,

192 including Lab as a random intercept, revealed no interaction, β=0.29, SE=0.18,

193 $t(151.30)$=1.59, p=.115 (Figure 1).

194   **S2.5.2. Reliability and its interaction between language background and**

195 **age.**   We also fit a preregistered linear mixed-effects model predicting IDS preference in

196 Session 2 from the three-way interaction between IDS preference in Session 1

197 (mean-centered), NAE (centered), participant age (mean-centered), and all lower order

198 interactions. We find evidence for an interaction between language background and age in

199 predicting reliability, β=0.01, SE=0.00, $t(63.70)$=2.43, p=.018. Figure 2 illustrates that

200 this interaction was driven by a small set of older infants (all from a single lab and

201 participating in the HPP method) showing a somewhat more reliable relationship between

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 (left panel).

202 Session 1 and Session 2 looking. Note that the mixed-effects analyses use Age as a

203 continuous predictor — age is median-split in Figure 2 to ease visualization. Given the

204 small number of infants driving the three-way interaction and the confounded nature of

205 this sample (with method and lab), we do not draw strong conclusions from the existence

206 of this three-way interaction, but report it here to spur future investigations into how age

207 and experience interacts with test-retest reliability.

208 **S2.5.3. Change in preferential looking moderated by age and language**

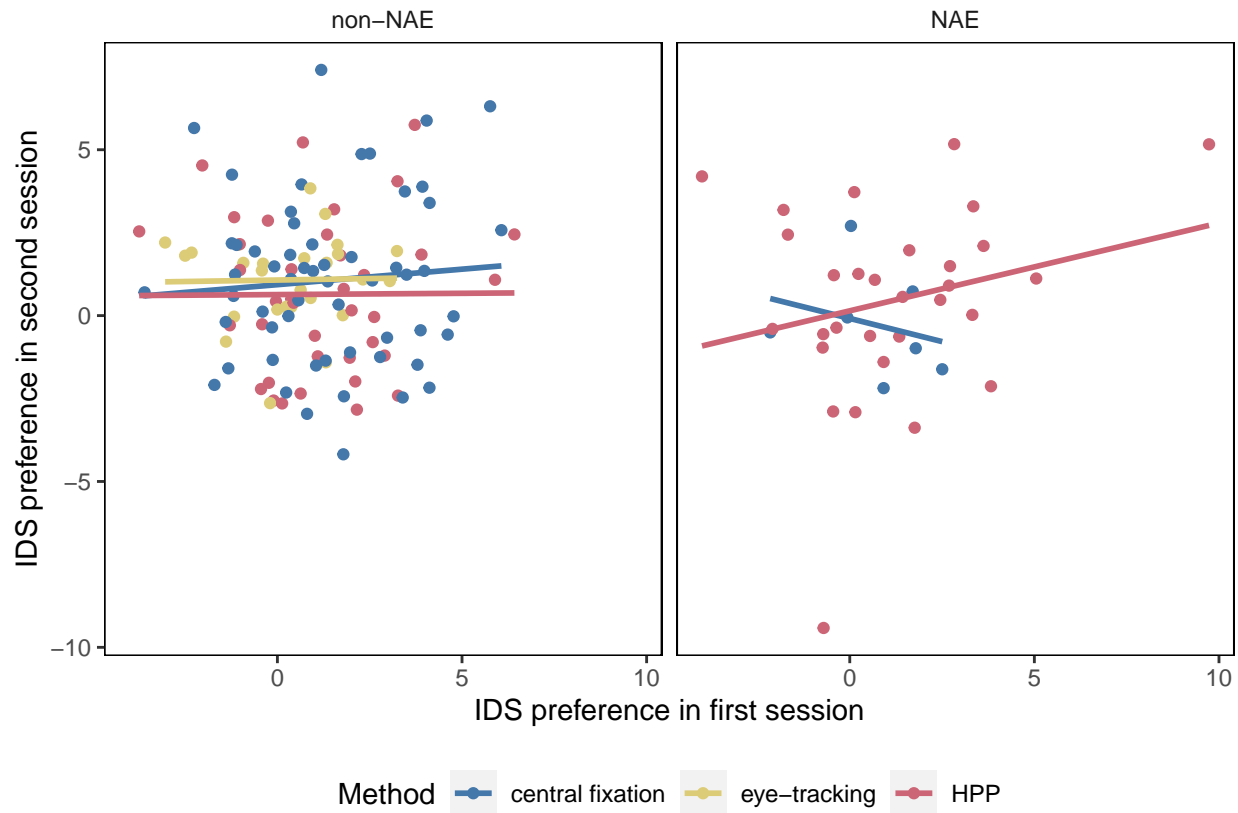209 **background.** We fit a linear mixed-effects model predicting average IDS preference from
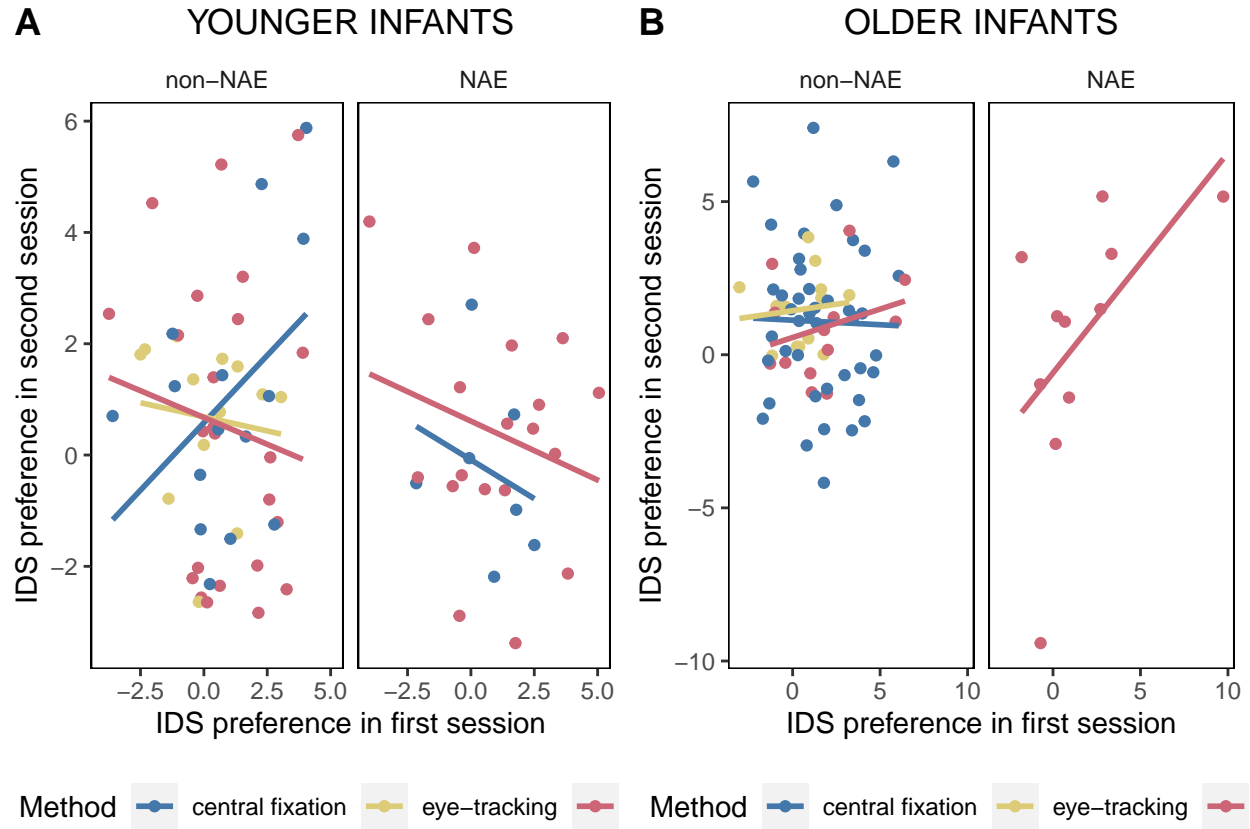
Figure 2. Infants' preference in Session 1 and Session 2 with individual data points and regression lines color-coded by method for (A) younger and (B) older infants (median-split). Results are plotted separately for North American English-learning infants and infants learning other languages and dialects

210 the three-way interaction of Session (centered; Session 1 vs. Session 2), participant age

211 (mean-centered), NAE (centered), and all lower order predictors. We included by-lab and

212 by-participant random intercepts and by-lab random slope for Session (more complex

213 random effects structures did not converge due to singular fits). We found no evidence that

214 the change in preferential looking to IDS between Session 1 and Session 2 was moderated

215 by participant age and language background, $\beta=0.01$, $SE=0.02$, $t(114.60)=0.95$, $p=.347$.

Lab and Method   **Fisher's z_r [95% CI]**

InfantCog–UBC, CF	−0.27 [−1.25, 0.71]
babylab–potsdam, HPP	−0.32 [−0.77, 0.13]
babyling–oslo, eye–tracking	0.30 [−0.44, 1.05]
brookes–babylab, CF	−0.14 [−0.65, 0.37]
infantll–madison, HPP	0.24 [−0.14, 0.62]
lancslab, eye–tracking	−0.05 [−0.60, 0.49]
wsi–goettingen, HPP	0.36 [−0.19, 0.90]
wsi–goettingen, CF	−0.26 [−0.58, 0.07]
RE Model	−0.04 [−0.26, 0.19]

−1.5   −0.5   0   0.5   1   1.5

Fisher's z Transformed Correlation Coefficient
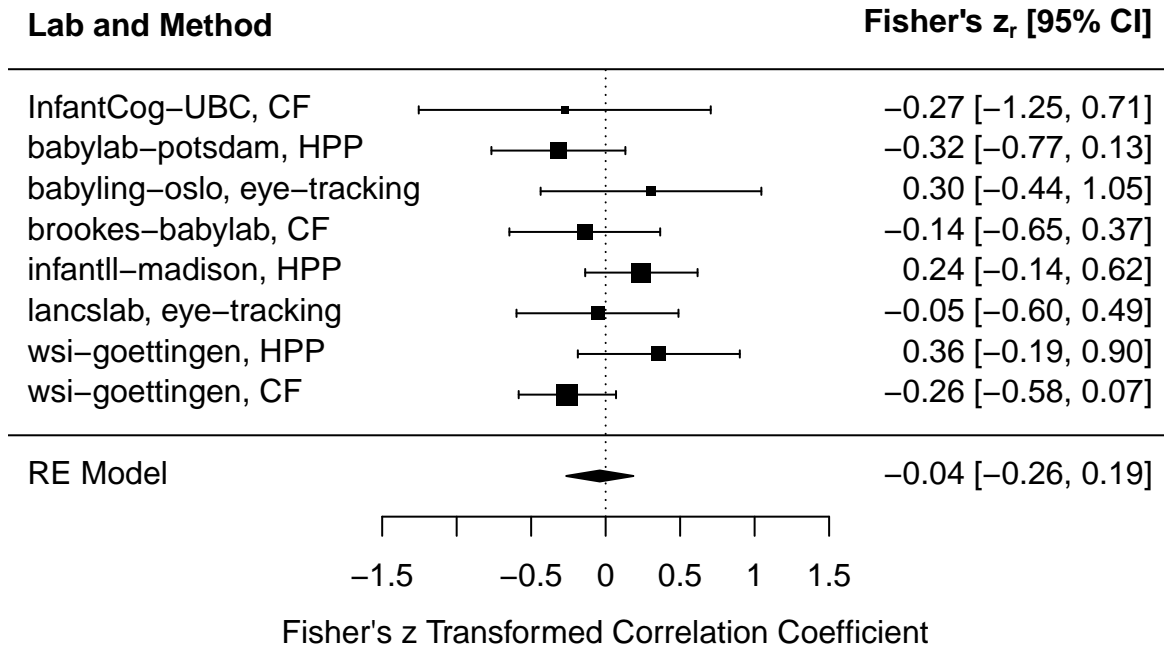
Figure 3. Forest plot of test-retest reliability effect sizes. Each row represents Fisher's z transformed correlation coefficient and 95% CI for a given lab and method (HPP = head-turn preference procedure; ET = eye-tracking; CF = central fixation). The black diamond represents the overall estimated effect size from the mixed-effects meta-analytic model.

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

Table 1

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

²²⁴ in Figure 3.

S4. Analyses including a more restricted sample

²²⁶ Given that we found that restricting the sample to participants contributing at least ²²⁷ 6 ADS and IDS trials in both sessions, we conducted the central analyses with this more ²²⁸ restricted infant sample.

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

²³⁰ The participants in the restricted sample — contributing at least 6 IDS and ADS ²³¹ trials for both sessions — were distributed across the contributing labs, methods, and ²³² language backgrounds (Table 1). There was no difference in average age between the main ²³³ sample and the restricted sample ($t(204.57) = -0.33$, $p = .744$). There was a robust

234 preference for infant-directed speech in both session 1 ($t(99) = 6.67$, $p < .001$) and session 2

235 ($t(99) = 4.42$, $p < .001$). We observed no difference in IDS preference between the two

236 sessions, β=-0.34, SE=0.28, p=.225.

237 Interestingly, while there was a significant simple correlation between IDS preference

238 in session 1 and session 2 ($r = .22$, 95% CI [.02, .40], $t(98) = 2.23$, $p = .028$), we found that

239 IDS preference in session 1 did not significantly predict IDS preference in session 2 in a

240 linear mixed-effects model including a by-lab random intercept, β=0.12, SE=0.11, p=.255.

241 **S4.2. Moderator analyses including a more restricted sample**

242 **S4.2.1. Time between test sessions.** As in the analyses with the full dataset, we

243 found no evidence that the number of days between test sessions moderated the

244 relationship between IDS preference in Session 1 and 2. Neither the main effect of time

245 between sessions, β=-0.03, SE=0.03, $t(95.80)$=-0.96, p=.342, nor the interaction term,

246 β=-0.01, SE=0.03, $t(93.60)$=-0.22, p=.828, showed significant effects.

247 **S4.2.2. Participant age**

248 To investigate the possibility that age moderated test-retest reliability in the

249 restricted sample, we fit a linear mixed-effects model predicting IDS preference in Session 2

250 from IDS preference in Session 1 (mean-centered), participant age (mean-centered) and

251 their interaction. The model included a by-lab random intercept and a by-lab random

252 slope for IDS preference in Session 1. We found no evidence that age influenced test-retest

253 reliability as indicated by the interaction between IDS preference in Session 1 and age,

254 β=0.00, SE=0.00, $t(43.20)$=-0.69, p=.494.

S4.2.3. Method

²⁵⁶   We tested whether method (eye-tracking vs. central fixation vs. headturn preference

²⁵⁷ procedure) moderated test-retest reliability by fitting a linear mixed-effects model

²⁵⁸ predicting IDS preference in Session 2 from IDS preference in Session 1 (mean-centered),

²⁵⁹ Method (dummy-coded, with central fixation as the reference level) and their interaction.

²⁶⁰ The model included a by-lab random intercept and a by-lab random slope for IDS

²⁶¹ preference in Session 1. We found no evidence that Method influenced test-retest reliability

²⁶² as indicated by the interaction between IDS preference in Session 1 and age, χ^2=3.85,

²⁶³ p=.146. There was no significant relationship between IDS preference for session 1 and

²⁶⁴ session 2 for each method considered separately (central fixation:β=-0.06, SE=0.16,

²⁶⁵ 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

²⁶⁷   As in the main sample, a linear mixed-effects model predicting IDS preference in

²⁶⁸ Session 2 based on IDS preference in Session 1 (mean-centered), NAE (centered), and their

²⁶⁹ interaction, including Lab as a random intercept, revealed no interaction, β=0.31,

²⁷⁰ SE=0.24, $t(95.10)$=1.29, p=.199.

Table 2

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 was very high (all rs > 0.95; Table 2).

Table 3

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 3). 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 3).

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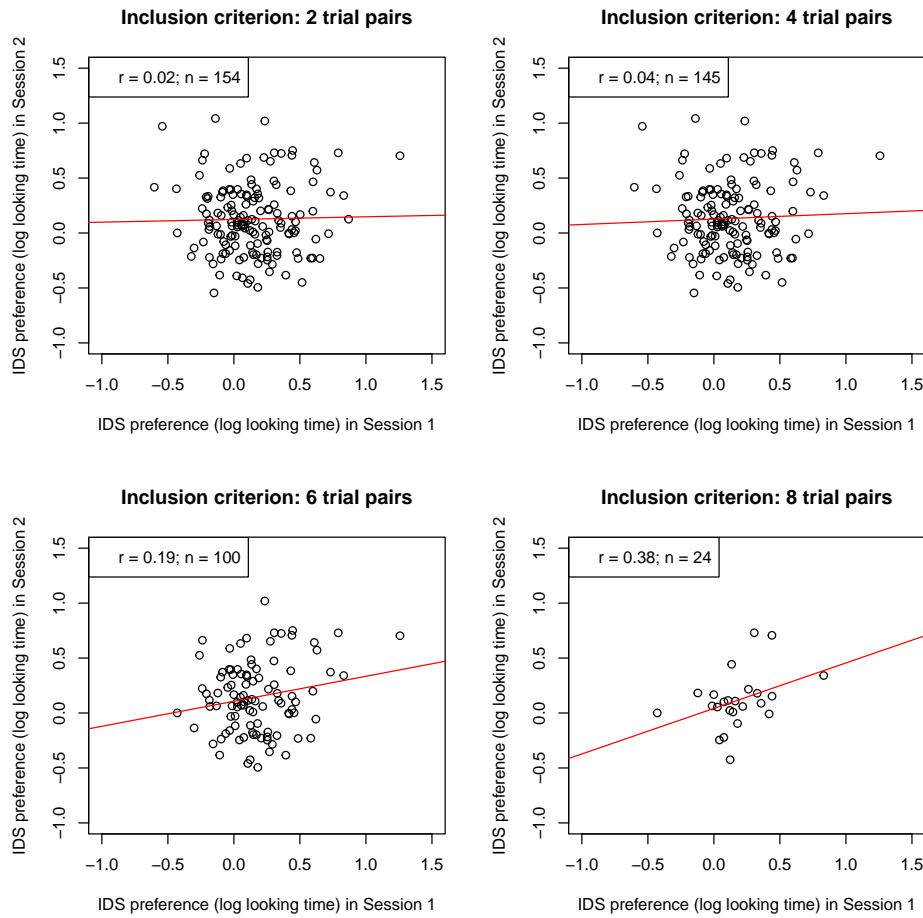
Figure 4. IDS preferences (based on average log-looking times) of both sessions plotted against each other for each inclusion criterion. n indicates the number of included infants, r is the Pearson correlation coefficient as the indicator for reliability.

302 time)). We fit a linear mixed-effects model predicting proportion-based IDS preference in

303 Session 2 from proportion-based IDS preference in Session 1, including a by-lab random

304 intercept. As in the analyses using other measures of IDS preference, the results revealed

305 no significant relationship between IDS preference in Session 1 and 2 (Table 4). The

306 Pearson correlation coefficient based on proportional IDS looking was also not statistically

307 significant, $r = .01$, 95% CI $[-.15, .16]$, $t(156) = 0.09$, $p = .927$. Stricter inclusion criteria

308 increased the correlation somewhat, as in previous analyses (Figure 4).
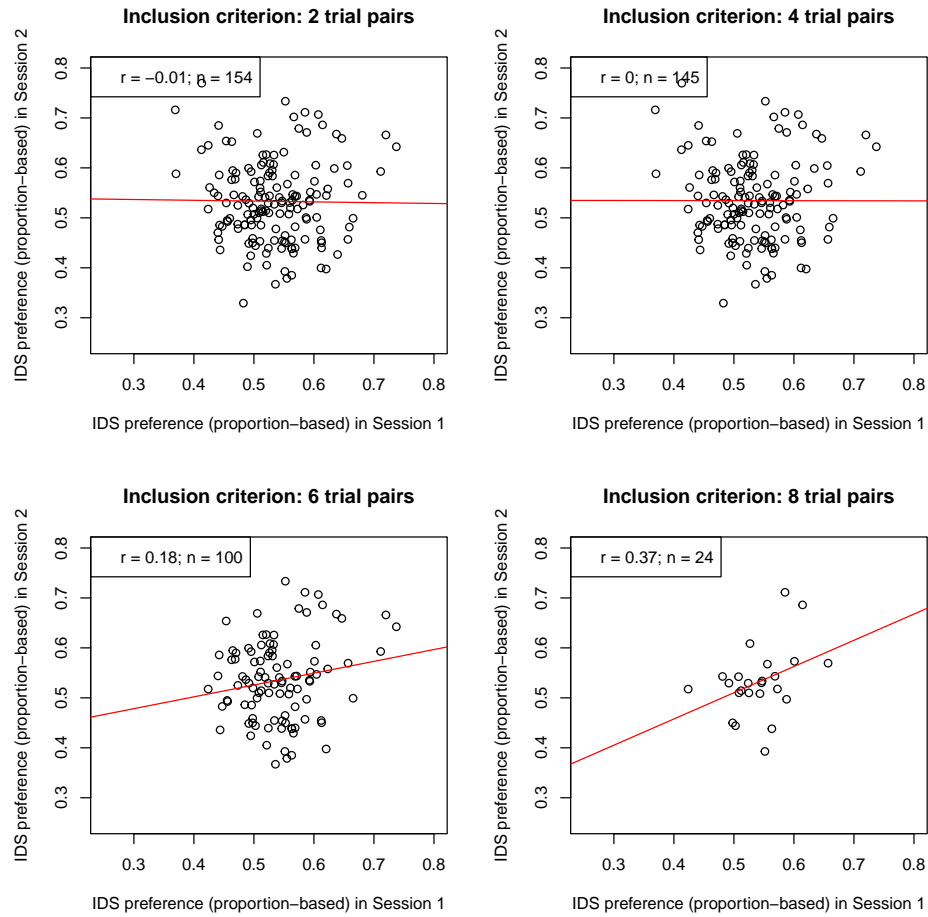
Figure 5. IDS preferences (based on proportion IDS looking) of both sessions plotted against each other for each inclusion criterion. n indicates the number of included infants, r is the Pearson correlation coefficient as the indicator for reliability.

Table 4

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 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.
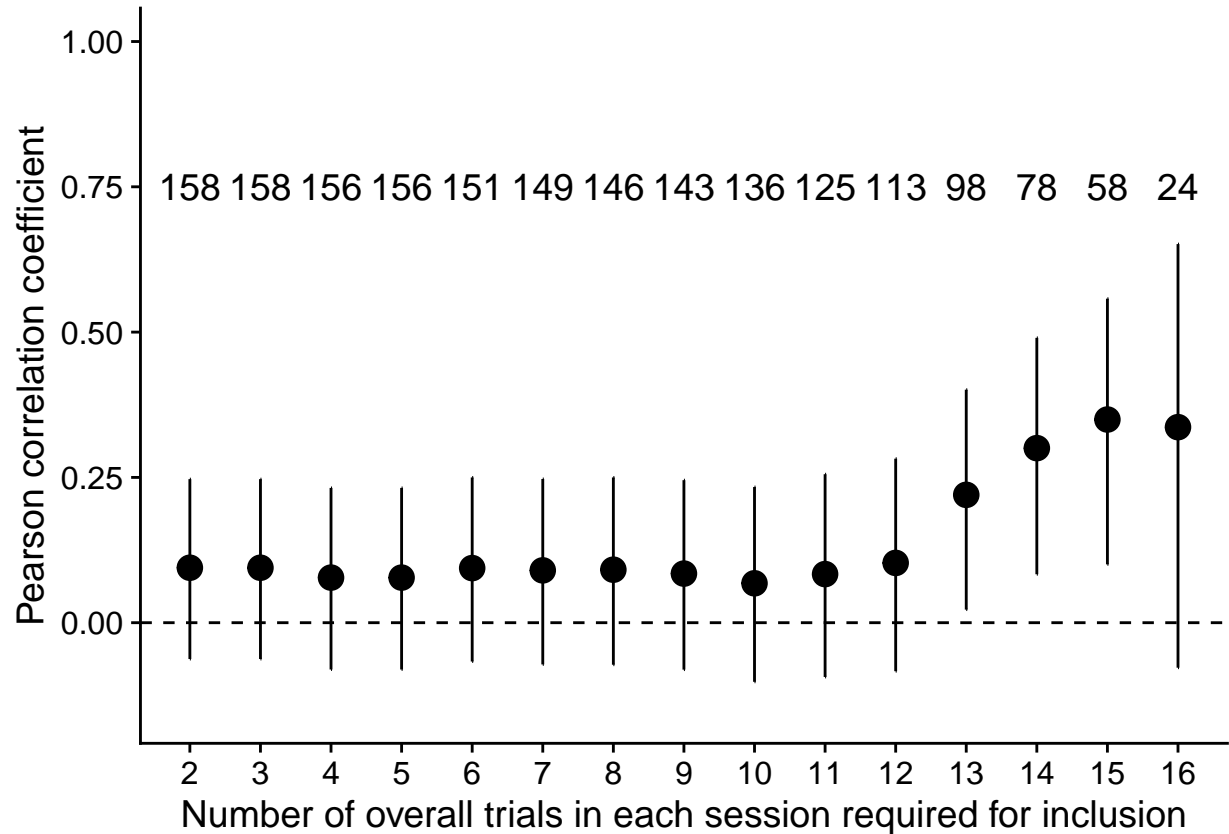
Figure 6. Pearson correlation coefficient with increasingly strict trial-level inclusion criteria. The x-axis depicts the required number of overall valid trials in both session 1 and session 2. Dots represent corresponding correlation coefficients, with 95 percent CIs. The sample size is shown above 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 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).
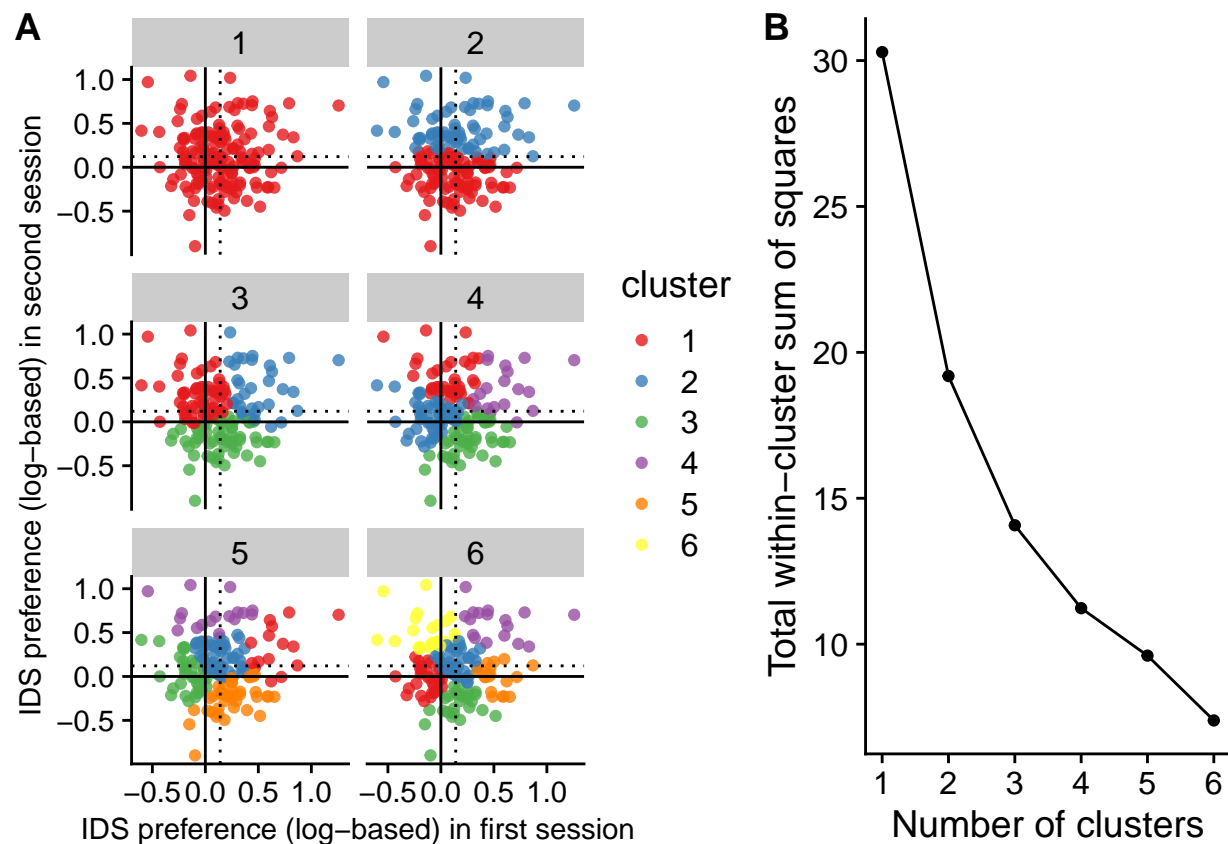
Figure 7. (A) Results from the k-means clustering analysis of IDS preference (based on average log looking times) in session 1 and 2 for different numbers of k and (B) the corresponding elbow plot of the total within-cluster sum of squares. In (A), points represent indvidual participants' magnitude of 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.

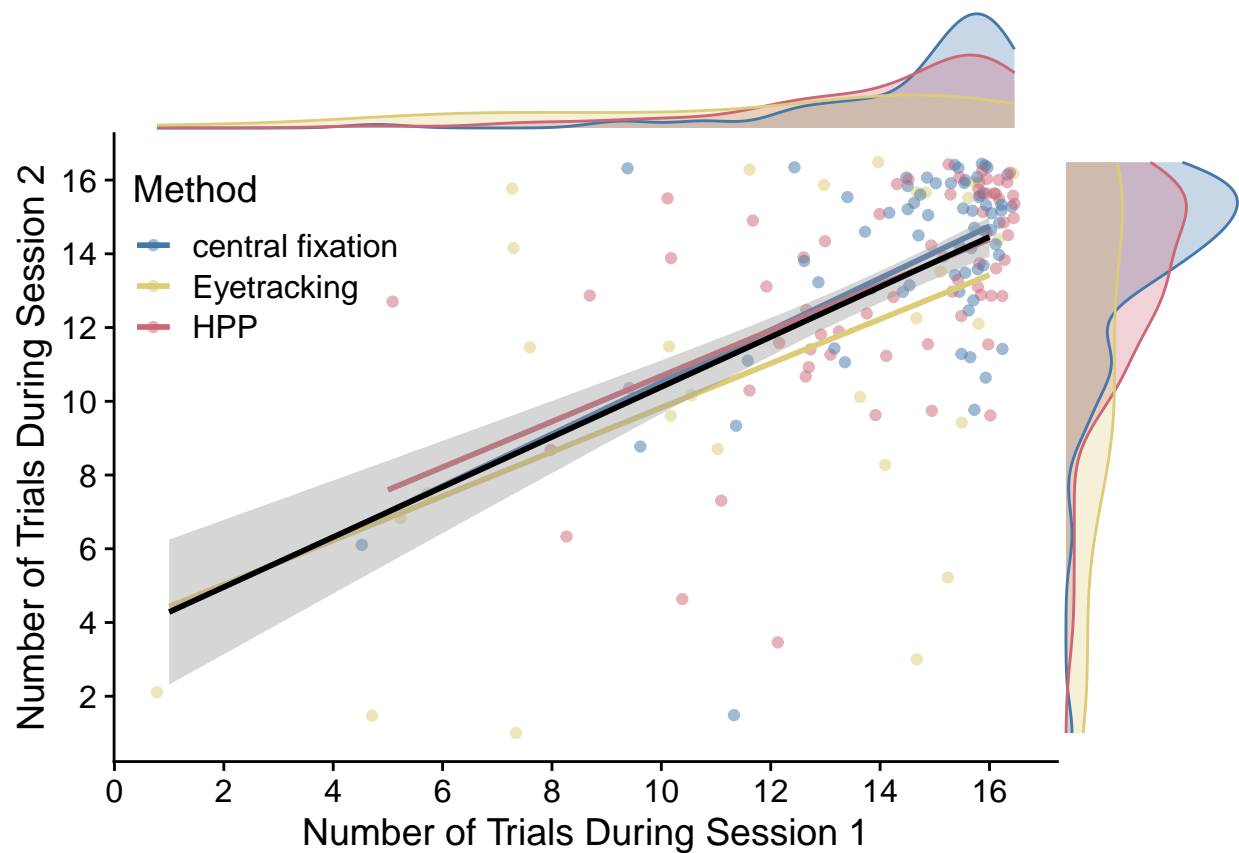**S8. Relation between number of contributed trials in each session**

Figure 8. 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

349 predictive of how many trials they might contribute during their next session.

350   We found a strong positive correlation between number of trials contributed during
351 the first and the second session $r = .58$, 95% CI $[.47, .67]$, $t(160) = 9.00$, $p < .001$ (Figure
352 7). This result suggests that if infants contribute a higher number of trials in one session,
353 compared to other infants, they are likely to contribute a higher number of trials in their
354 next session. This finding is consistent with the hypothesis that how attentive infants are
355 throughout an experiment (and hence how many trials they contribute) is a stable
356 individual difference, at least for some infant looking time tasks. Researchers should
357 therefore be mindful of the fact that decisions about including or excluding infants based on
358 trials contributed may selectively sample a specific sub-set of the infant population they are
359 studying (Byers-Heinlein, Bergmann, & Savalei, 2021; DeBolt, Rhemtulla, & Oakes, 2020).

₃₆₀   **S9. Correlations in average looking times between sessions**

₃₆₁   To what extent are participants looking times between the two sessions related? To

₃₆₂ test this question, we first investigated whether participants' overall looking times —

₃₆₃ irrespective of condition — were correlated between the first and second session. There was

₃₆₄ a robust correlation between average looking time in Session 1 and Session 2: infants with

₃₆₅ longer looking times during their first session also tended to look longer during their second

₃₆₆ session, $r = .45$, 95% CI $[.31, .57]$, $t(156) = 6.28$, $p < .001$. This relationship held even after

₃₆₇ controlling for number of trials in the first and second session, suggesting that the relation

₃₆₈ between average looking in Session 1 and 2 could not be entirely explained by the

₃₆₉ correlation in the number of trials contributed between the two sessions (S7), $b = 0.42$, 95%

₃₇₀ CI $[0.27, 0.58]$, $t(154) = 5.52$, $p < .001$ (Figure 8A). The result is also similar when

₃₇₁ controlling for participants' average age across the two test sessions, $b = 0.44$, 95% CI

₃₇₂ $[0.30, 0.59]$, $t(155) = 6.16$, $p < .001$.

₃₇₃   Next, we explored the extent to which average looking times for IDS and ADS stimuli

₃₇₄ were related. First, we found similar correlations in average looking time to IDS stimuli in

₃₇₅ Session 1 and 2, $r = .38$, 95% CI $[.24, .51]$, $t(156) = 5.19$, $p < .001$, and ADS stimuli in

₃₇₆ Session 1 and 2, $r = .40$, 95% CI $[.26, .53]$, $t(156) = 5.49$, $p < .001$ (Figure 8B). To test

₃₇₇ whether these correlations were specific to looking times for IDS or ADS stimuli alone, we

₃₇₈ fit linear regression models predicting average looking to IDS (or ADS) stimuli in Session 2

₃₇₉ from average looking to IDS and ADS stimuli in Session 1. We found that average looking

₃₈₀ to IDS stimuli in Session 2 could be predicted from average looking to IDS stimuli in

₃₈₁ Session 1, even after controlling for average looking to ADS stimuli in Session 1, $b = 0.21$,

₃₈₂ 95% CI $[0.01, 0.41]$, $t(155) = 2.11$, $p = .037$. Conversely, average looking to ADS stimuli in

₃₈₃ Session 2 could be predicted from average looking to ADS stimuli in Session 1, even after

₃₈₄ controlling for average looking to IDS stimuli in Session 1, $b = 0.36$, 95% CI $[0.14, 0.58]$,

₃₈₅ $t(155) = 3.20$, $p = .002$. These results suggest that the condition-specific correlations in

386 average looking time cannot be fully explained by the fact that infants' overall looking

387 times between sessions are correlated.

388   Finally, we inspected item-level correlations between the two test sessions.

389 Specifically, we investigated the relation between items composed of the same recording

390 clips in Session 1 and Session 2 (but with a reversed order of clips between the two

391 sessions). We fit a linear mixed-effects model predicting item-level looking time in Session

392 2 from item-level looking time in Session 1, including random intercepts for participant,

393 item, and lab, as well as a random slope for item-level looking time in Session 1 for

394 participant and lab. Item-level looking in Session 2 was related to item-level looking in

395 Session 1, $\hat{\beta} = 0.17$, 95% CI $[0.07, 0.27]$, $t(5.52) = 3.38$, $p = .017$ (Figure 8C). Similar

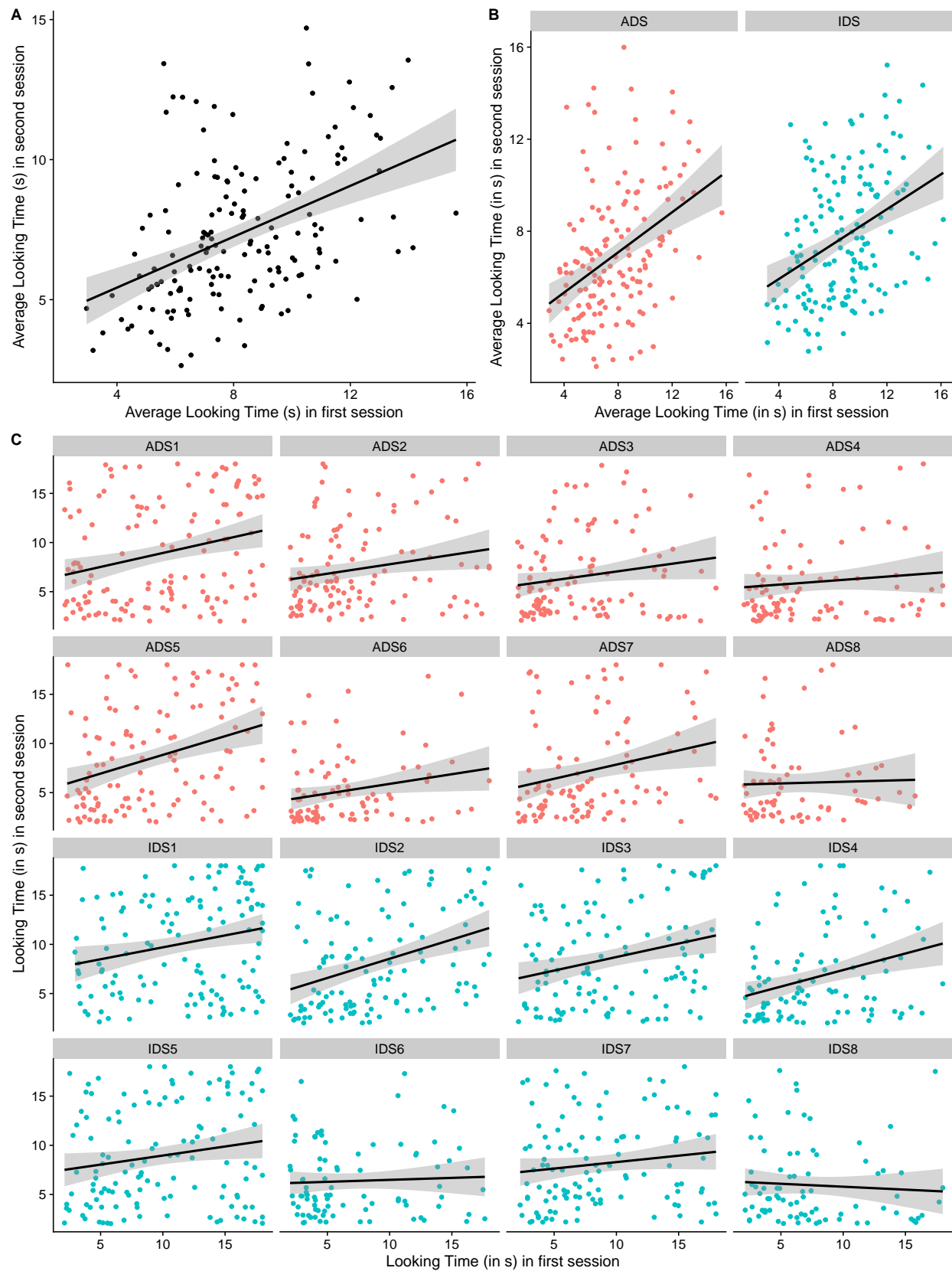396 results hold if looking times are log-transformed

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

Table 5

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

S10. 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 5).

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