Manybabies1 Test-Retest Supplementary Materials

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 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 of 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 with a time between sessions above 31 days were also included in the analyses (n=2).
- 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 analysis script for 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.), and the language background of the participants (S2.3.).

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 the 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 test session 1 (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 number of days between test sessions moderated the relationship between IDS preference at test session 1 and 2. Neither the main effect of time between sessions, β=-0.01, SE=0.03, p=.684, nor the interaction term, β=-0.01, SE=0.02, p=.465, showed significant effects.

S2.2. Participant Age.

S2.3. Language Background. NAE-learning infants showed greater IDS preferences than their non-NAE counterparts in MB1. We therefore also assessed if test-retest reliability interacted with children's language background. A multilevel analysis with Lab as random intercept, predicting the IDS preference in Session 2 based on the IDS preference in Session 1, NAE and the interaction of these two variables, revealed no interaction, β=0.29, SE=0.18, p=.115 (see Figure S1).

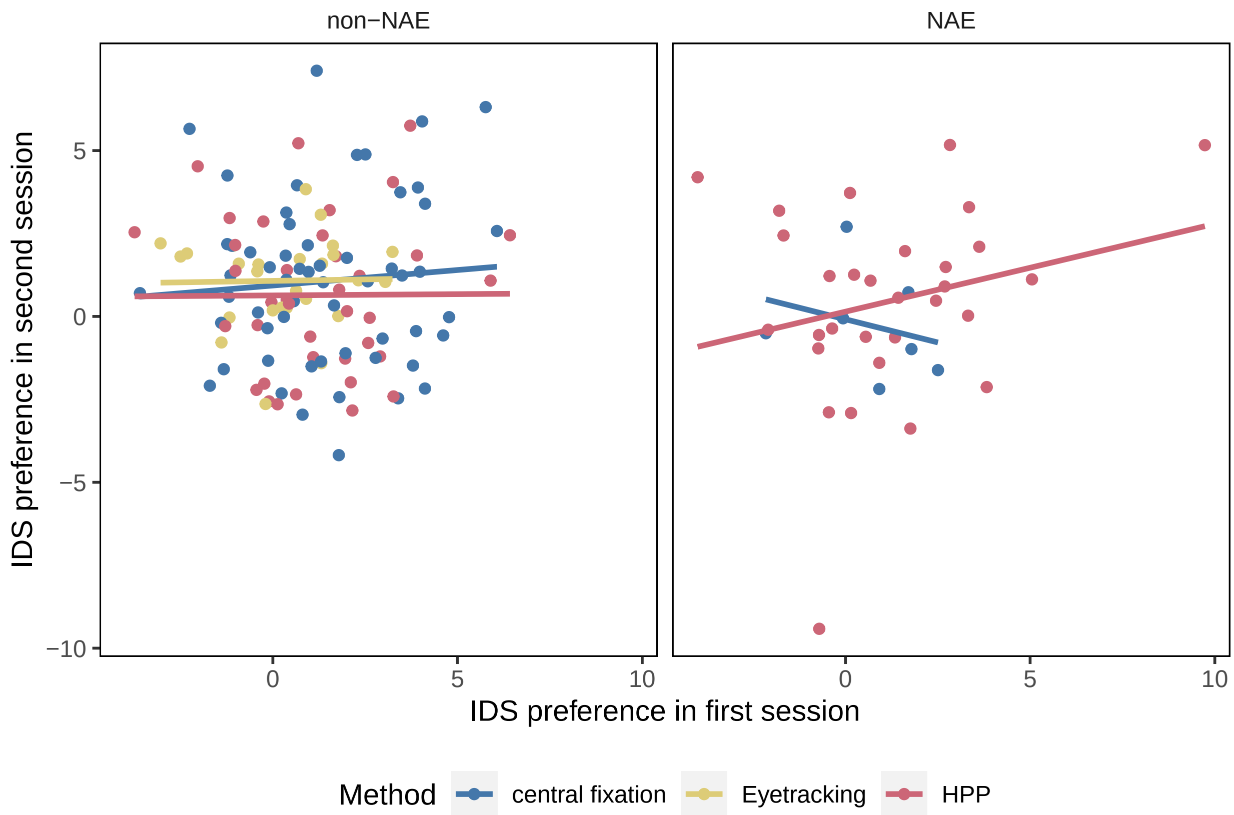


Figure 1. Infants' preference in Session 1 and Session 2 with individual data points and regression lines color-coded by method (central fixation, eye-tracking, or HPP). Results are plotted separately for North American English-learning infants (right panel) and infants learning other languages and dialects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

S3: Meta-analysis of test-retest reliability

We used the metafor package (Viechtbauer, 2010) to fit a mixed-effects meta-analytic model on z-transformed correlations 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.

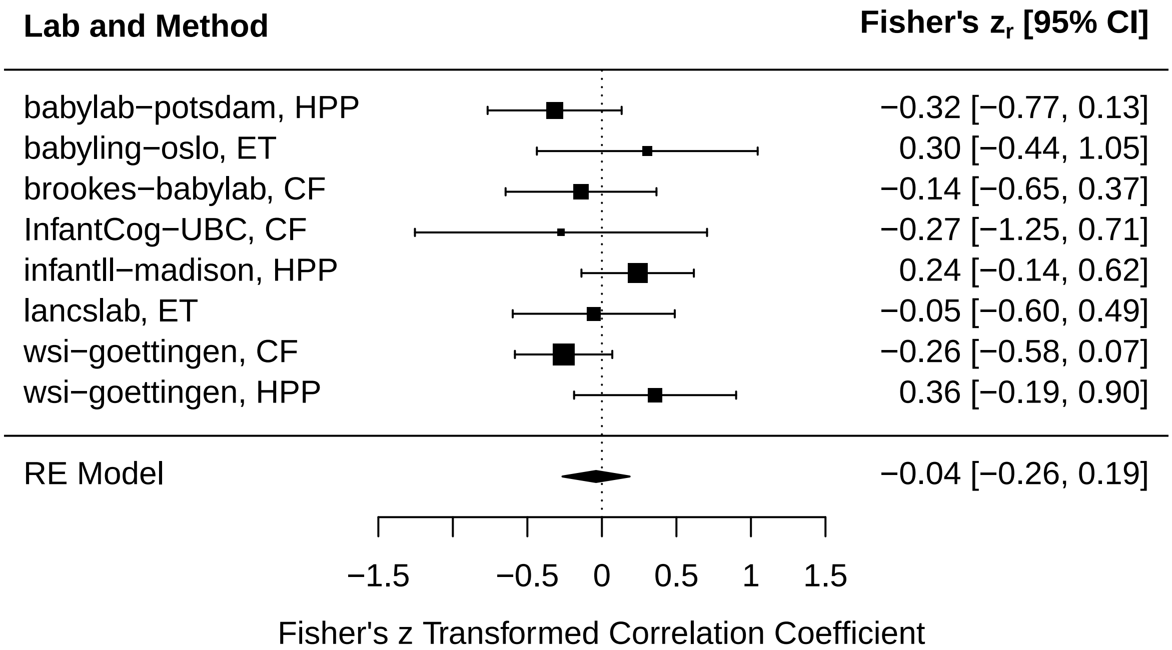


Figure 2. Forest plot of test-retest reliability effect sizes. Each row represents Fisher's z transformed correlation coefficient and 95% CI for a given lab and method (HPP = head-turn preference procedure; ET = eye-tracking; CF = central fixation). The black diamond represents the overall estimated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

S4: Alternative Dependent Variables

S4.1. Log-transformed looking times.

S4.2. Proportion novelty preference.

S5: 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, indicating that infants were not more likely than chance to maintain their preference from test session 1 to test session

2 (exact binomial test; $p = 0.05$). Of the 158 total infants, 44.90% of infants showed a consistent infant-directed speech preference and 13.30% showed a consistent adult-directed speech preference. 23.40% of infants switched from an infant-directed speech preference at test session 1 to an adult-directed speech preference at test session 2 and 18.40% switched from an adult-directed speech preference to an infant-directed speech 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. 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 simply track whether participants are approximately above or below the mean looking time difference for test session 1 and test session 2, and the diagnostic elbow plot shows little evidence of a qualitative improvement as the number of clusters is increased.

S6: Relationship between the number of trials infants contribute in each session

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 their 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

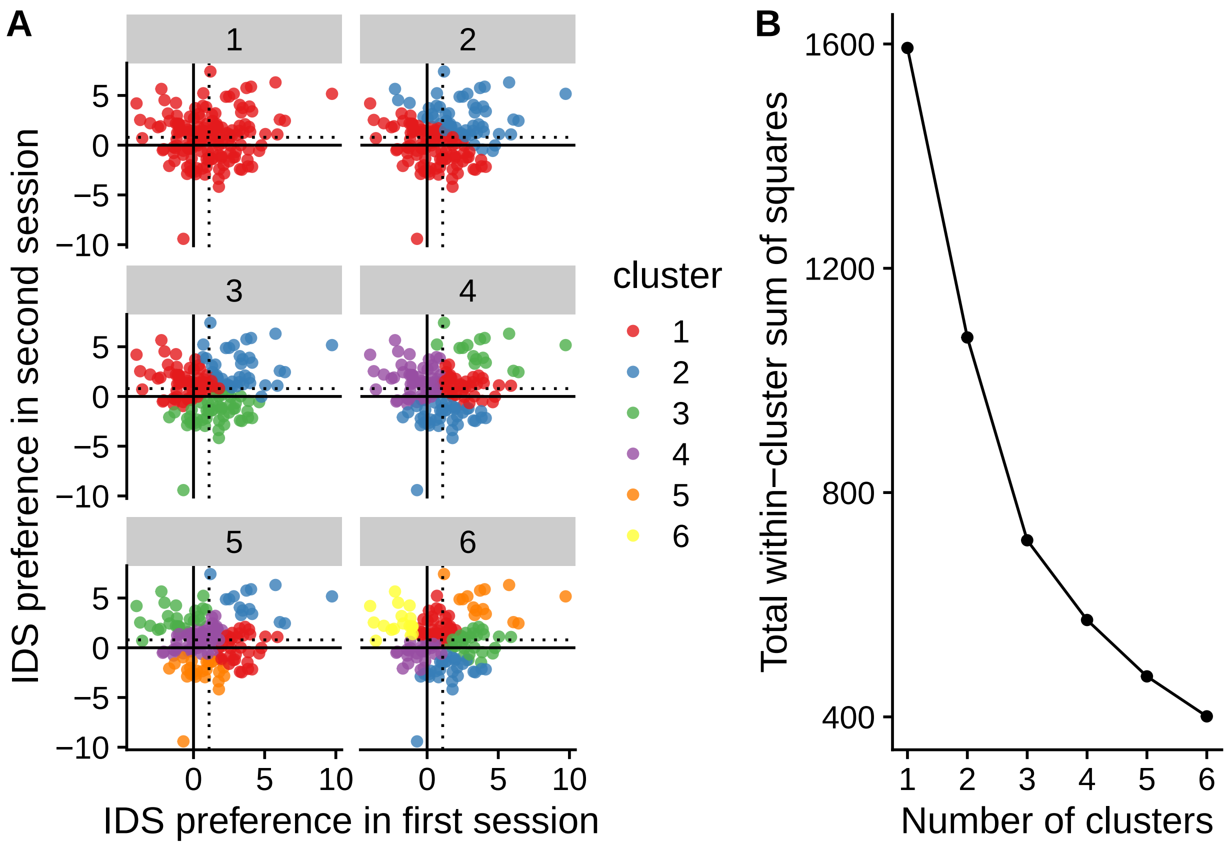


Figure 3. (A) Results from the k-means clustering analysis of IDS preference in session 1 and 2 for different numbers of k and (B) the corresponding elbow plot of the total within-cluster sum of squares. In (A), points represent indvidual participants' magnitude of looking time difference at test sessions 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.

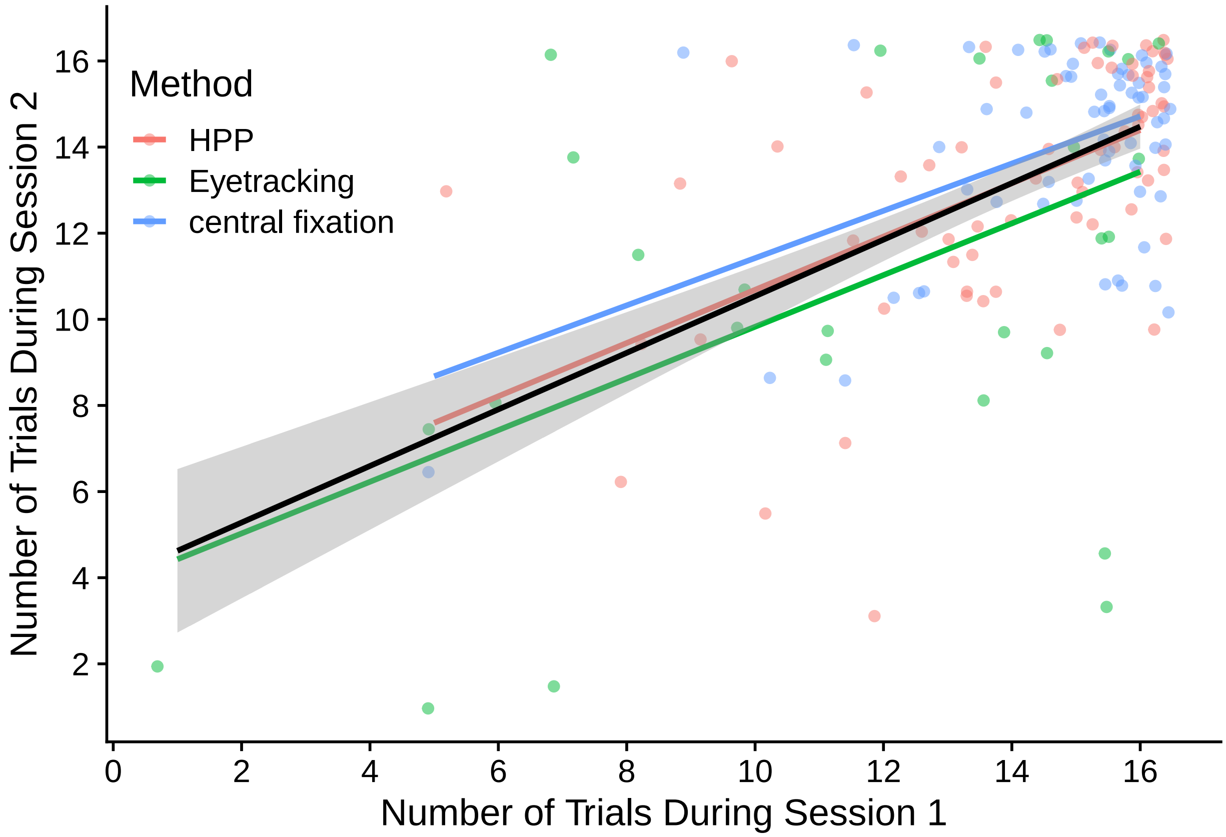


Figure 4. Correlation between the number of trials contributed in session 1 and session 2. Each data point represents one infant. Colored lines represent linear fits for each method.

the first and the second session $r = .58$, 95% CI $[.47, .68]$, $t(159) = 9.05$, $p < .001$ (see Figure 1). 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).

S7: Correlations in average looking times between sessions

S7.1: Relations between overall looking time in session 1 and 2. There is a strong relationship between average overall looking in the first test session and the second test session, even after controlling for number of trials in the first and second session.

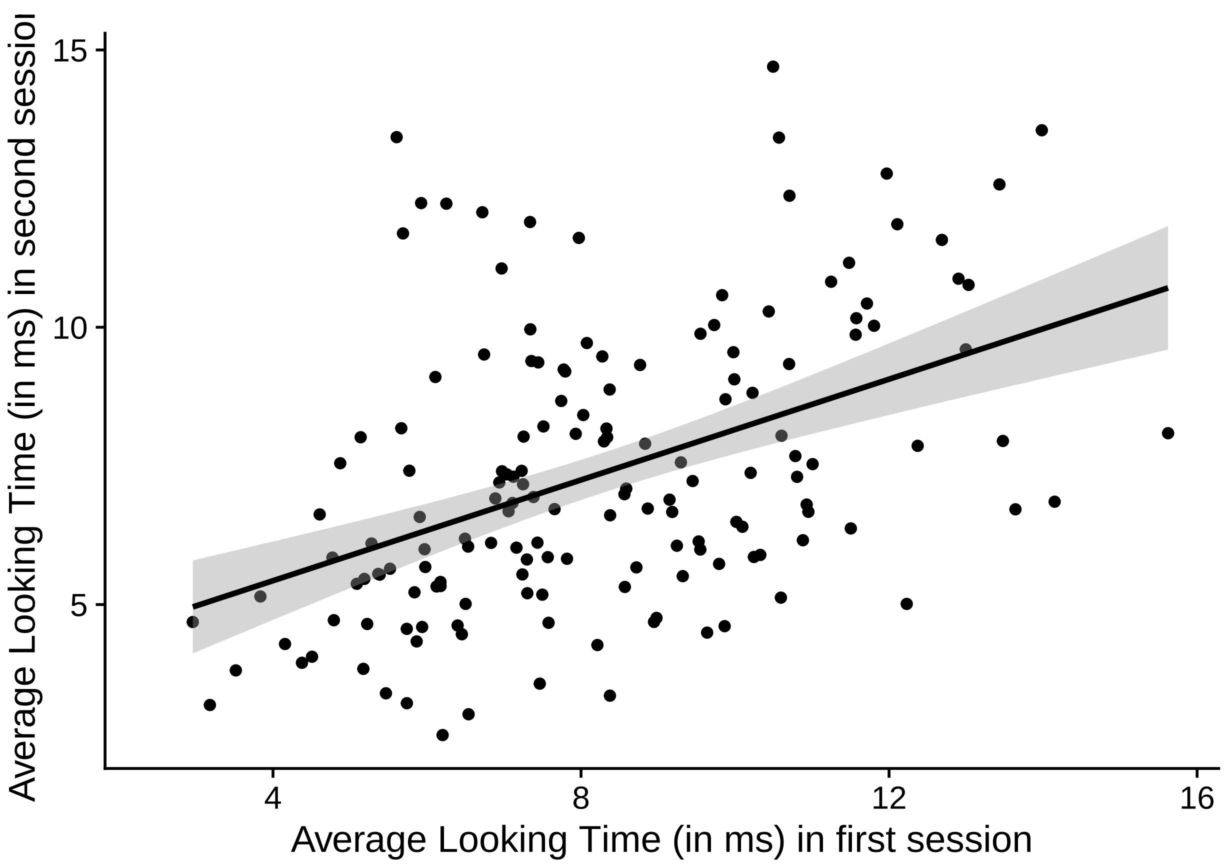
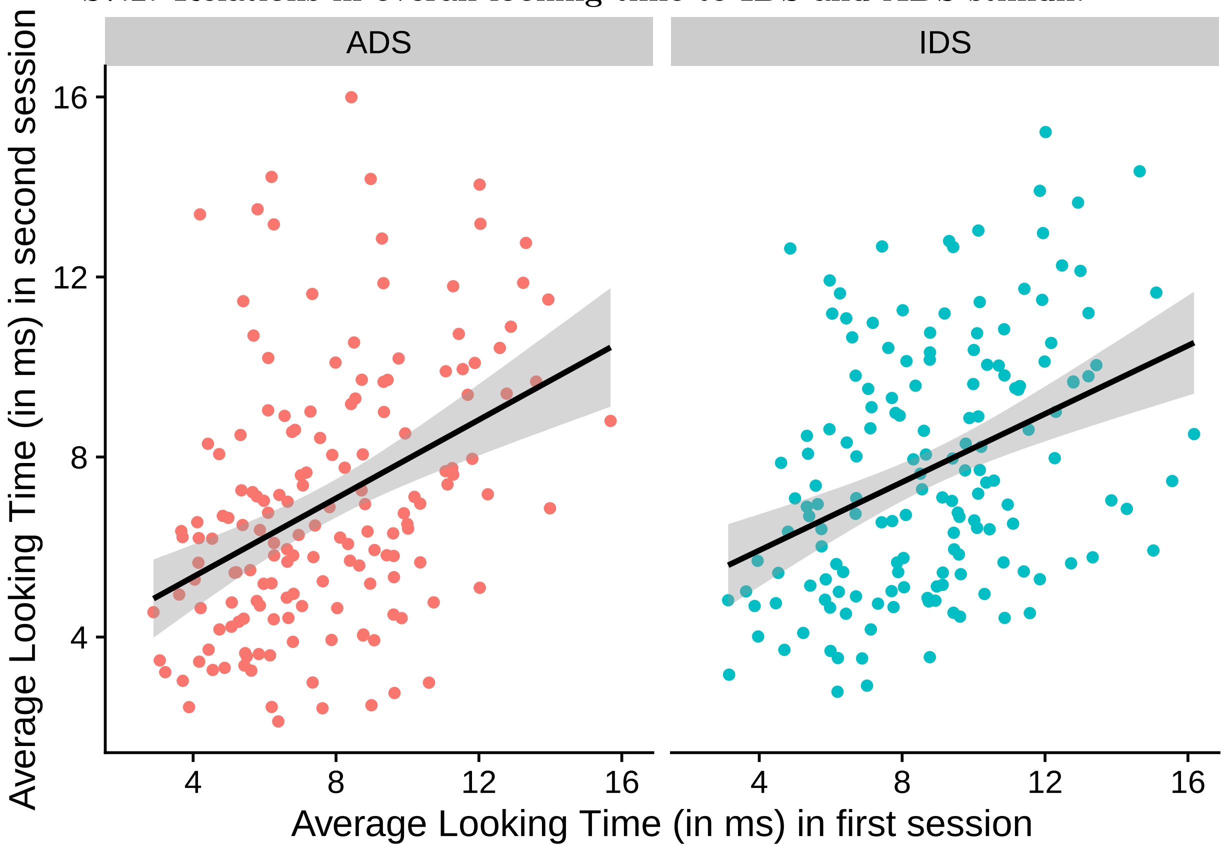


Table 1

Average Looking during session 1 predicted from average looking at session 2, controlling for trial number for each session.

Predictor	b	95% CI	t	df	p
Intercept	2.55	[0.38, 4.73]	2.32	154	.022
Mean lt 1	0.42	[0.27, 0.58]	5.52	154	$<$ .001
N 1	-0.08	[-0.24, 0.08]	-0.96	154	.338
N 2	0.18	[0.04, 0.32]	2.52	154	.013

S7.2: Relations in overall looking time to IDS and ADS stimuli.



##

```
## Call:
## lm(formula = LT_Retest_IDS ~ LT_Test_IDS + LT_Test_ADS, data = agg_by_subj_condition_
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2721 -1.7567 -0.2799  1.4822  6.4805
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.9749     0.6902   5.759 4.41e-08 ***
## LT_Test_IDS   0.2123     0.1008   2.105   0.0369 *
## LT_Test_ADS   0.2467     0.1044   2.362   0.0194 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.52 on 155 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.1771, Adjusted R-squared:  0.1665
## F-statistic: 16.68 on 2 and 155 DF,  p-value: 2.751e-07
```

```
##
## Call:
## lm(formula = LT_Retest_ADS ~ LT_Test_IDS + LT_Test_ADS, data = agg_by_subj_condition_
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.556 -1.771 -0.489  1.254  8.901
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.2374     0.7356   4.401    2e-05 ***
## LT_Test_IDS    0.1103     0.1075   1.026  0.30641
## LT_Test_ADS    0.3563     0.1113   3.201  0.00166 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.686 on 155 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.1677, Adjusted R-squared:  0.157
## F-statistic: 15.62 on 2 and 155 DF,  p-value: 6.619e-07
```

S7.3: Relations for specific ADS and IDS stimuli between the first and second session.

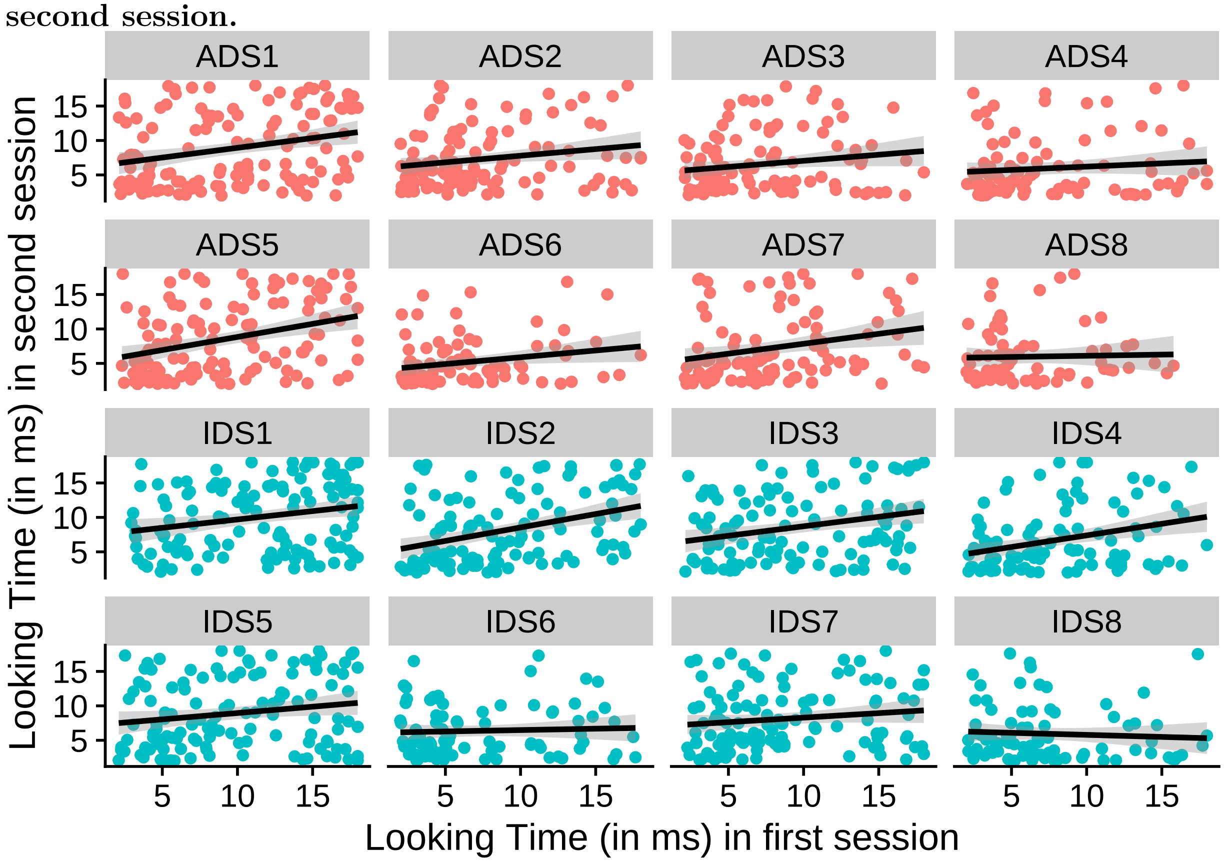


Table 2

Mixed-effects model results predicting looking time during session 1 from looking time at session 2 at the stimulus level.

Term	$\hat{\beta}$	95% CI	t	df	p
Intercept	6.04	[4.99, 7.08]	11.35	6.88	$< .001$
LT Test	0.13	[0.05, 0.20]	3.46	25.38	.002

S8: By-item-pair preference scores across sessions

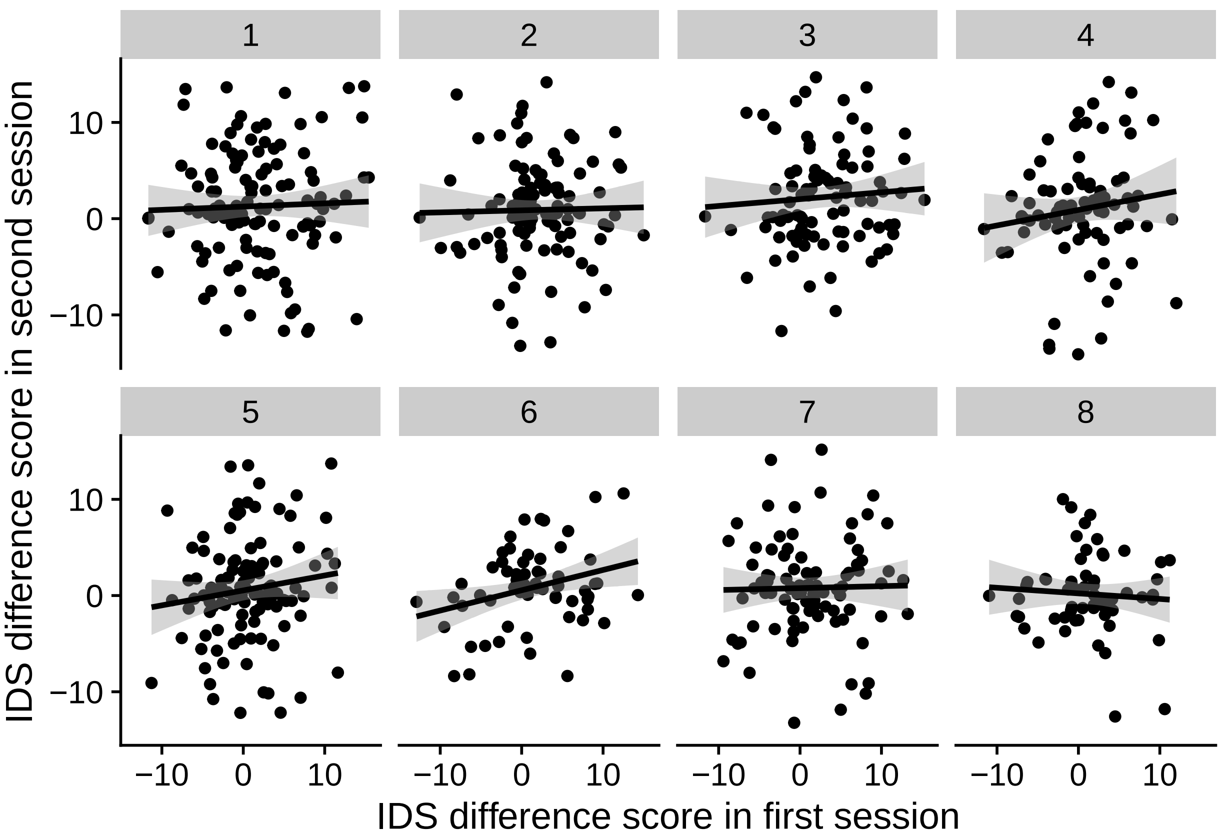


Table 3

Mixed-effects model results predicting IDS preference during session 1 from IDS preference at session 2 at the stimulus level.

Term	$\hat{\beta}$	95% CI	t	df	p
Intercept	0.87	[0.45, 1.30]	4.04	122.79	$< .001$
Diff 1	0.10	[-0.02, 0.22]	1.63	6.31	.151

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