Methods (HPP3)

Participants

We retrieved data from 85 participants: 37 from Santolin et al. (2019), 24 tested in Wisconsin, 37 tested in the replication study in Barcelona (Santolin & Saffran, 2019), and 24 from Saffran & Wilson (2003), tested in Wisconsin. Two datapoints were available for each participant: one for the mean looking time in "familiar" trials, and one for the mean looking time in "novel" trials. A total of 170 datapoints were included in the analysis. Participants included in our analysis were those included in the final version of both studies.

Studies

The two studies conducted by Santolin and colleages compared looking times in novel and familiar trials, and revealed that infants displayed a preference toward one of the trial types. The direction of the preference, however, was different for the two studies.

Data analysis

We fit a model predicting looking time (LT) from the fixed effects of test item ($Item$, Familiar vs. Novel), the number of HeadTurn Preference Procedure experiments completed by infants (HPP), and their interaction ($Item \times HPP$) as fixed effects. *Familiar* trials were set as the baseline. Participant ($Participant$) and study ($Study$) were included as random effects. Following Barr, Levy, Scheepers, & Tily (2013), we fit a model with the maximal random effects structure that allowed the model to converge. The maximal random effects structure included by-participant and by-study random intercepts and by-participant and by-study random slopes of HPP. Due to lack of convergence, we subsequently pruned the random effects structure until convergence was achieved. The final model included by-participant and by-study random intercepts. The particular random effects structure chosen does not qualitatively impact the estimates and conclusions from the model.

This model accounts for cross-participants variability in overall looking time (i.e. some infants are long lookers, some are short lookers), and for cross-studies differences in overall looking time, and allows the effect of HPP to vary across studies. We specifiyied by-study random effects for two strong reasons. First, participants from different linguistic/cultural environments were included in both studies. This may have led to participants in one of the locations to looking longer in average than those from the other location. Second, in spite of their similarity both studies were not identical, which could also have led to differences in overall looking time.

Table 1: Estimates of the linear mixed-effects model and outcomes of the Kenward-Roger F-tests performed on fixed effects. 95% confidence intervals were bootstrapped.

Term	Coefficient	SEM	95% CI	F	Den. df	p
Intercept	8192.662	818.840	6590.87, 9902.66	87.890	5.683	0.000
Item	-1487.614	558.698	-2688.43, -376.81	7.090	83.000	0.009
HPP	-535.325	390.562	-1408.99, 227.97	1.592	100.365	0.210
Item * HPP	657.089	314.644	0.26, 1328.39	4.361	83.000	0.040

We found a statistically significant interaction term, $F(1, 83) = 4.36$, $p = 0.04$, 95% CI = [0.26, 1328.39], suggesting that the effect of trial type on looking time was influenced by the number of HPP experiments each participant participated in. The interaction shows that experience with a higher number of HPP experiments

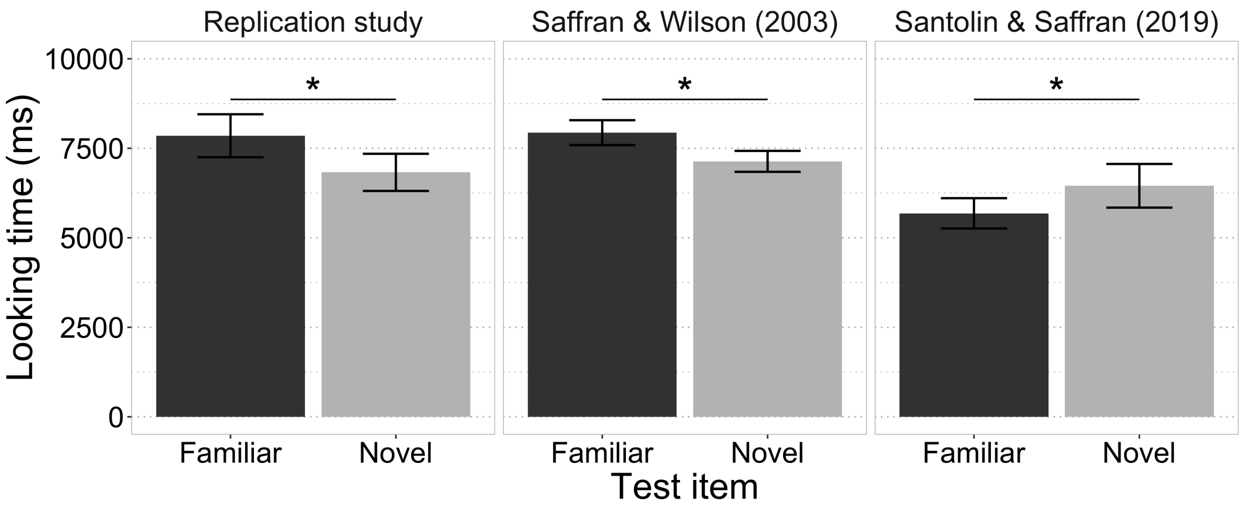


Figure 1: Looking time (ms) in familiar and novel trials split by Study and Location. Error bars indicate the standard error of the mean, respectively. Grey lines indicate participant level mean looking times.

is associated with a stronger novelty preference.

Appendices

Appendix 1: Linear Mixed Model

Following Barr et al. (2013) guidelines, we initially specified a maximal random structure, including random intercepts for *Participant* and *Study*, and random slopes for *Item*, *HPP*, and *Item* × *HPP*, by *Participant* and *Study*. Due to lack of convergence, the random structure was simplified until the model converged successfully and fit was no longer singular. The final model included only by-participant and by-study random intercepts. The code and formula of the resulting mode are presented below:

```
LookingTime ~ Item * HPP + (1 | Participant) + (1 | Study)
```

$$
\begin{aligned}
LT_{ips} = & \\
& \beta_0 + Participant_{0p} + Study_{0s} + \\
& \beta_1 \times HPP_i + \\
& \beta_2 \times Item_i + \\
& \beta_3 \times (HPP \times Itemi) + \\
& e_{ips}, e_{ips} \sim N(0, \sigma^2), Participant_{0p} \sim N(0, \tau_{00}^2), Study_{0p} \sim N(0, \quad \omega_{00}^2)
\end{aligned}
$$

Where:

- LT_{ips} is the looking time in trial i, in participant p from study s
- β_0 is the fixed intercept (the grand mean of all looking times from all trials)
- $Participant_{0p}$ is the by-participant random intercept (the overall looking time of participant p, assumed to have been sampled from a normal distribution with mean 0 and variance τ_{00}^2
- $Study_{0s}$ is the by-study random intercept (the overall looking time in study s, assumed to have been sampled from a normal distribution with mean 0 and variance ω_{00}^2.
- β_1 is the coefficient of the fixed effect of the HPP predictor

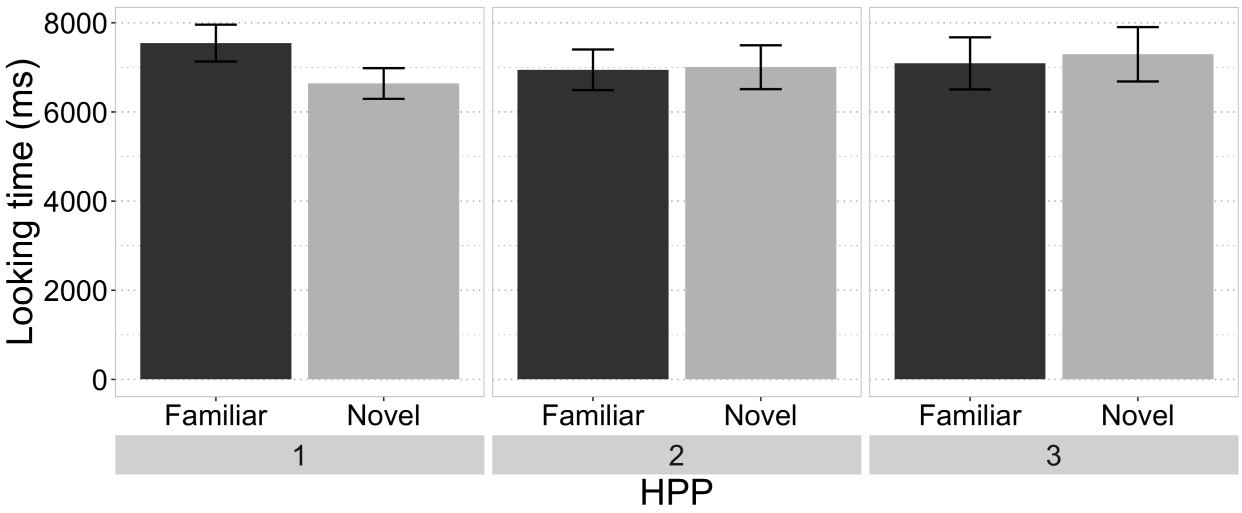


Figure 2: Looking times of familiar and novel trials, split by number of HeadTurn Preference Procedure experiments. Black points and error bars represent the group-level mean looking time and SEM, respectively. Grey lines represent participant-level mean looking time.

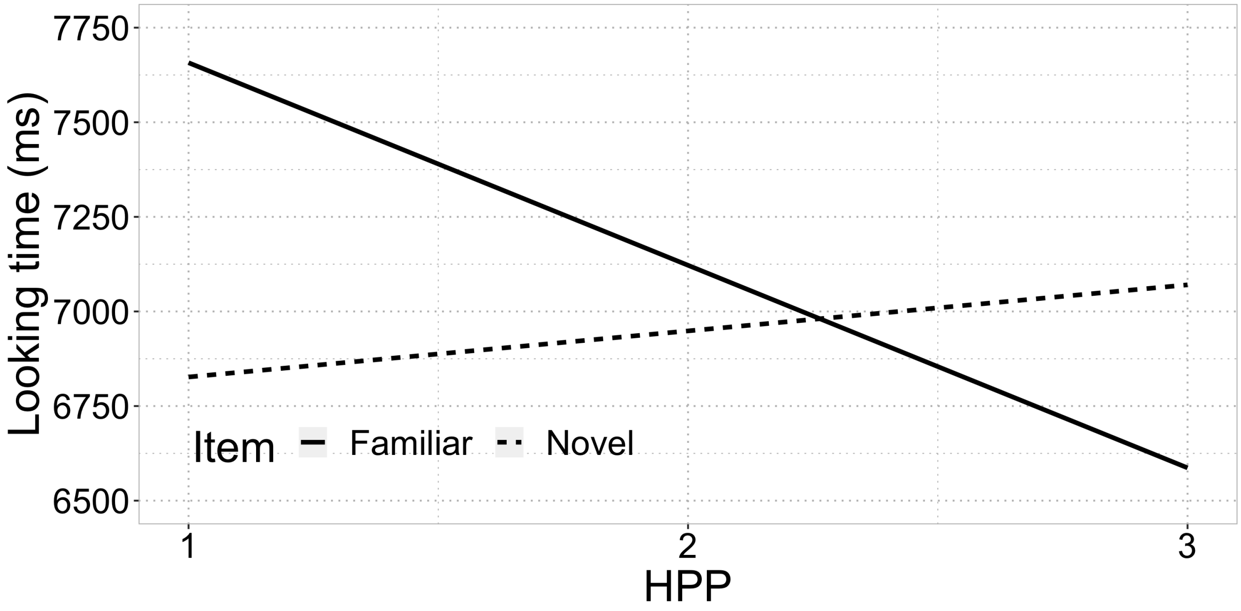


Figure 3: Predicted looking times plotted against HPP, split by test item.

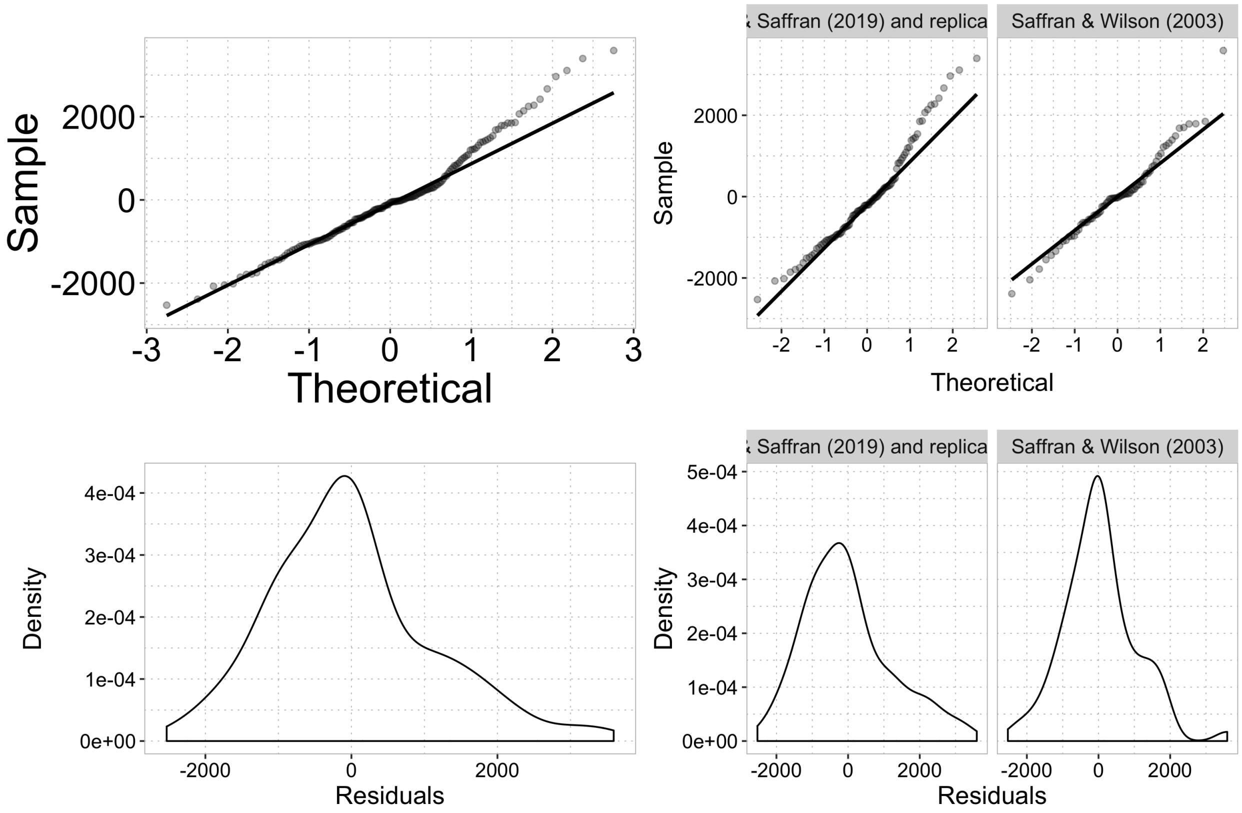
- β_2 is the coefficient of the fixed effect of the $Item$ predictor
- β_3 is the coefficient of the fixed effect of the $Item \times HPP$ interaction
- e_{ips} is the error of the model in trial i, of participant p from study s, assumed to be normally distributed with mean 0 and variance σ^2

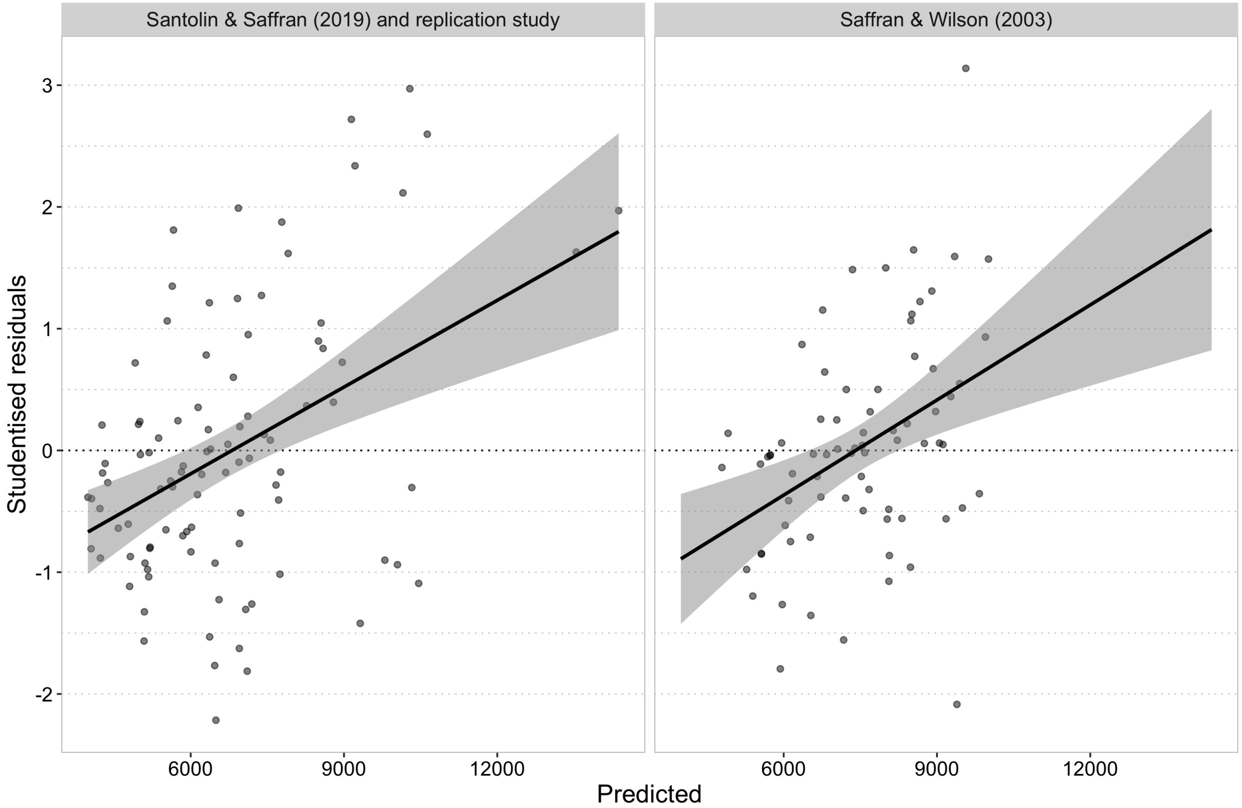
The model was fit using the `lme4` R package (Bates, Mächler, Bolker, & Walker, 2015). We used the `Anova` function from the `car` R package (Fox & Weisberg, 2019) to F-tests to fixed effects in the model using Kenward-Roger's (Kenward & Roger, 2009) approximation to degrees of freedom

Posterior predicted simulations (Gelman & Hill, 2006) indicated that the coefficients obtained by our model are plausible: The observed inter-quartile range (IQR) of looking times lied within the IQR obtained from 1000 simulated datasets generated from the model in more than 95% of the simulations ($p = 0.587$).

Appendix 2: Checking assumptions of the Linear Mixed Model

Residuals seem to approximate a normal distribution, though the distributions in the Santolin et al. (2019) and Santolin & Saffran (2019) studies seem to be somewhat skewed.





We observed little evidence of multicollinearity in the predictors we included in the model:

Term	VIF	Tolerance
Item	6.10	0.16
HPP	1.19	0.84
Item * HPP	6.29	0.16

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (Third). Thousand Oaks CA: Sage. Retrieved from https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Gelman, A., & Hill, J. (2006). *Data analysis using regression and Multilevel/Hierarchical models*. Cambridge University Press. Retrieved from https://books.google.es/books?id=c9xLKzZWoZ4C

Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics & Data Analysis*, *53*(7), 2583–2595. https://doi.org/10.1016/j.csda.2008.12.013

Saffran, J. R., & Wilson, D. P. (2003). From Syllables to Syntax: Multilevel Statistical Learning by 12-Month-Old Infants. *Infancy*, *4*(2), 273–284. https://doi.org/10.1207/S15327078IN0402_07

Santolin, C., & Saffran, J. R. (2019). Non-Linguistic Grammar Learning by 12-Month-Old Infants: Evidence for Constraints on Learning. *Journal of Cognition and Development*, *20*(3), 433–441. https://doi.org/10.1080/15248372.2019.1604525

Santolin, C., Saffran, J. R., & Sebastian-Galles, N. (2019). Non-linguistic artificial grammar learning in 12-month-old infants: A cross-lab replication study. In. Potsdam, Germany.