Supplementary Information

S1. Experiments included in the linear mixed-effects model.

The selected experiments each consist of an artificial grammar learning task with 12-month-old infants. These experiments are characterized by variability in the number of infants' prior HPP visits[1]. They include all studies run in the two senior authors' labs that included (a) 12-month-old participants; (b) HPP; (c) artificial grammar learning (linguistic or non-linguistic); (d) 2 - to - 5 minutes of exposure; (e) an a priori hypothesis that infants would show learning; (f) visit numbers recorded at the time of testing. The studies are thus as well matched as is possible given the retrospective nature of this analysis.

Saffran & Wilson (2003) demonstrated that 12-month-old infants can compute multiple regularities from a finite-state grammar. Infants were able to first segment words from running speech based on transitional probabilities, then detect permissible orderings of the segmented words. Test items consisted of grammatical and ungrammatical sentences that could only be discriminated based on word-level information (transitional probabilities between syllables were not informative about the "grammaticality" of test items). Infants showed a significant familiarity preference: $F(1.00, 38.00)= 5.37$, $p<.05$.

Saffran et al. (2008) demonstrated that infants could detect simple phrases (i.e., clusters of nonsense words grouped together based on statistical regularities) from artificial grammars. In Exp. 1, infants in the Predictive Language condition were familiarized with a grammar including predictive (statistical) dependencies between words. The test items consisted of familiar sentences vs. novel sentences violating the grammar. Infants showed a significant novelty preference: $t(11.00)= 2.52$, $p<.05$.

Santolin & Saffran (2019) is a conceptual replication of Saffran et al. (2008) using non-linguistic sounds (e.g., computer alert sounds) to implement the grammars. Infants exposed to the Predictive language showed a significant novelty preference: $t(26.00)=2.45$, $p=.021$, $d=0.47$.

We replicated the Predictive Language condition of Santolin & Saffran (2019) at the University Pompeu Fabra, Barcelona (Santolin et al., 2019), using identical stimuli and procedures. We found significant discrimination of the test stimuli but observed the opposite direction of preference: Infants listened longer to familiar than novel strings: $t(23.00)=2.30$, $p=.030$, $d=0.47$. All results are shown in Fig. 1 of the main manuscript.

S2. Participants information

We retrieved data from 102 12-month-old infants who had participated in a range of 1-6 studies. Three of the studies were run in Madison, WI (University of Wisconsin-Madison): Saffran & Wilson, 2003 (Exp. 2): $N=40$, mean age: 11.5 months; Saffran et al., 2008 (Exp. 1, Condition P-Language): $N=12$), mean age: 12.8 months; Santolin & Saffran, 2019 (Condition 1); $N=26$, mean age: 12.9 months. One study was run in Barcelona, Spain (Universitat Pompeu Fabra): Santolin, Saffran & Sebastian-Galles, 2019: $N=24$, mean age: 13 months. Two data points (average looking time for familiar and novel test items) were available for each participant. Participants included in the current analysis are those included in the final version of the studies. In additional analyses, we included only the infants who participated in 1- to- 4 HPP studies overall ($n=100$).

S3. Linear mixed-effects model

We fit a model predicting looking time (LT) from the fixed effects of test item ($TestItem$, Familiar vs. Novel), the number of Head-turn Preference Procedure experiments completed by infants (HPP), and their interaction ($Item \times HPP$) as fixed effects. Participant and study (4 levels: Santolin & Saffran, 2019, @santolin2019a,

[1] At the time of publication of Saffran & Wilson (2003), the first author noted that there appeared to be an association between the number of prior studies completed by the infants and the direction of preference. The analysis was included in the original manuscript submission but was removed from later revisions based on the manuscript reviews.

@saffran2003 as a third study, and @saffran2008 as a fourth study) were included as random effects. Following Barr, Levy, Scheepers, & Tily (2013), we fit a model with the maximal random effects structure including random intercepts by-participant and by-study, and random slopes of HPP by-participant and by-study. However, due to lack of convergence, we pruned the random effects structure until convergence was achieved (e.g., Brauer & Curtin, 2018). The final model included by-participant and by-study random intercepts only. The particular random effects structure chosen does not qualitatively impact the estimates and conclusions from the model. This model accounts for cross-participant variability in overall looking time (as some infants look longer than others), and for cross-studies differences in overall looking time. The model was fit using the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015) from the R environment (R Core Team, 2018). We used the `Anova` function from the `car` R package (Fox & Weisberg, 2019) to perform F-tests on fixed effects using Kenward-Roger's approximation to degrees of freedom (e.g., Judd, Westfall, & Kenny, 2012).

S4. Results sub-setting data to participants with 1-4 HPP visits

Consistent with the results of the entire dataset (reported in the manuscript), we found a statistically significant interaction of Test item on the number of HPP visits ($F(1, 1, 1, 1, 98.00)=10.43$, $p= 1$), indicating that the original effect was not driven by infants with a high number of visits only.

Table 1: Summary of the results of the linear mixed-effects model (subset to infants with 1-4 visits). Degrees of freedom were approximated using the Kenward-Rogers approach, thus sometimes result in non-integers.

	Coefficient	SEM	95%CI	F	Den. df	p
Intercept	7,611.06	719.37	6188.42, 9275.57	107.35	10.46	$< .001$
Test Item	-1,491.11	452.08	-2348.39, -578.5	10.88	98.00	.001
HPP	-500.79	278.70	-1070.22, 98.18	3.04	131.92	.083
Test Item * HPP	726.22	224.92	294.24, 1144.99	10.43	98.00	.002

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

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389–411. https://doi.org/10.1037/met0000159

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/

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*(1), 54–69. https://doi.org/10.1037/a0028347

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Saffran, J., Hauser, M., Seibel, R., Kapfhamer, J., Tsao, F., & Cushman, F. (2008). Grammatical pattern learning by human infants and cotton-top tamarin monkeys. *Cognition*, *107*(2), 479–500. https://doi.org/10.1016/j.cognition.2007.10.010

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.