Experience with research paradigms relates to infants' direction of preference.

Chiara Santolin[1], Gonzalo Garcia-Castro[1], Martin Zettersten[2], Nuria Sebastian-Galles[1], & Jenny Saffran[2]

[1] Center for Brain and Cognition, Universitat Pompeu Fabra

[2] Waisman Center & Department of Psychology, University of Wisconsin-Madison

Preprint submitted to peer-review on February 17th, 2020. Resubmitted addressing reviewers' comments on September 4th, 2020.

Abstract

Interpreting and predicting direction of preference in infant research has been a thorny issue for decades. Several factors have been proposed to account for familiarity versus novelty preferences, including age, length of exposure, and task complexity. The current study explores an additional dimension: experience with the experimental paradigm. We re-analyzed the data from 4 experiments on artificial grammar learning in 12-month-old infants run using the Head-turn Preference Procedure (HPP). Participants in these studies varied substantially in their number of laboratory visits. Results show that the number of HPP studies is related to direction of preference: infants with limited experience with the HPP setting were more likely to show familiarity preferences than infants who had amassed more experience with this paradigm. This evidence has important implications for the interpretation of experimental results: experience with a given method or, more broadly, with the lab environment, may affect infants' patterns of preferences.

Keywords: preferential looking, familiarity preference, novelty preference, head-turn preference procedure, linear mixed-effects model

Word count: 2307

Experience with research paradigms relates to infants' direction of preference.

Introduction

The importance of changes in preferential looking has been recognized since at least the 1960s, when Fantz (1964) showed that infants preferentially attend to novel visual stimuli. Subsequent studies extended this evidence to domains including auditory perception and cognition, revealing differences in direction of preference. Rather than representing a binary distinction, direction of preference can be construed as a continuum from more familiar to more novel (e.g., Thiessen et al., 2005). The infant's position along this continuum seems to be determined by a variety of factors related to the task and/or age (e.g., Houston-Price & Nakai, 2004; Aslin, 2007; Hunter & Ames, 1988). However, it is frequently the case that the observed direction of preference does not conform with expectations based on these dimensions; the infancy literature is rife with examples of counterintuitive patterns of preference (e.g., Fiser & Aslin, 2001; Bosch & Sebastián-Gallés, 2001; Dawson & Gerken, 2009; DePaolis, Keren-Portnoy, & Vihman, 2016; Johnson et al., 2009; Jusczyk & Aslin, 1995; Sebastián-Gallés & Bosch, 2009; Thiessen, 2012).

One frequently-overlooked factor is that infants do not arrive at the lab as naïve participants. Like adults, they bring significant prior experience that may influence their performance in lab tasks. Researchers attempt to override or sidestep those experiences by using novel stimuli (e.g., unfamiliar languages, shapes or sounds), or by integrating those experiences into their experimental designs (e.g., monolingual vs. bilingual infants; see Sebastian-Galles & Santolin, 2020 for a recent review). But there may also be forms of experience that go unidentified by researchers. One such factor is that many infants participate in multiple (putatively unrelated) experiments over the course of weeks or months. This common practice in infant research reflects the challenges of advancing a field of investigation that is based on a limited and hard-to-recruit population. Researchers are typically very careful to avoid stimulus contagion across unrelated studies, but it is possible that prior lab experience impacts infants' performance. The purpose of this article is to explore the effect of experience with experimental paradigms on direction of preference in learning tasks.

In an influential model of preferential behavior in infants, Hunter and Ames (1988) hypothesized three central factors to affect the strength and direction of preference: age, familiarization duration, and task complexity. In a given task, younger infants tend to prefer familiar stimuli whereas older infants are more likely to prefer novel stimuli (e.g., Colombo & Bundy, 1983; though see Bergmann & Cristia, 2016, for a meta-analysis suggesting that age does not predict shifts in preference). A shorter exposure to familiar stimuli prior to testing also leads infants to subsequently prefer the familiar items (for reviews, see Rose, Feldman, & Jankowski, 2004). Task complexity refers to the stage of stimulus processing. For example, in a visual recognition task, 4-month-old infants preferred familiar objects before subsequently showing a strong preference for the novel object (Roder, Bushneil, & Sasseville, 2000). Task complexity can also refer to the complexity of the stimuli. For example, sequential stimuli put greater strain on memory resources than materials in which all components are simultaneously available (e.g., Ferguson, Franconeri, & Waxman, 2018). A related dimension is the similarity between stimuli used during familiarization and test: when there is a close perceptual match, infants are more likely to show a novelty preference (e.g., Hunter & Ames, 1988; Thiessen & Saffran, 2003). The combination of these factors informs predictions concerning direction of preference in systematic ways. For example, Thiessen, Hill, and Saffran (2005) manipulated length of exposure and observed a flip from familiarity to novelty preference after doubling the amount of familiarization received by infants. Similarly, Ferguson et al. (2018) manipulated sequential vs. spatial presentation of visual patterns, and observed stronger novelty effects with (a) increasing age and (b) spatial presentation.

The idea behind the current paper emerged from a puzzling pattern of results in a replication of a published study focused on non-linguistic artificial grammar learning in 12-month-olds (Santolin & Saffran, 2019). We observed a flip in preference from novelty to familiarity between the original study and its replication (Santolin et al., 2019), despite the use of identical stimuli and procedures. While there were some differences between the studies (most notably, in the location in which the studies were run), one main factor stood out to us: many of the infants in the study that elicited a novelty preference had participated in prior studies using the Head-turn Preference Procedure (HPP), whereas most of the infants in the study that elicited a familiarity preference were first-time HPP participants. We reasoned that the more familiarity infants had with the lab apparatus and task demands, the more likely they would be to learn rapidly, leading to a novelty preference. To investigate this question, we conducted exploratory analyses combining the data from these two experiments with the data from two other published artificial grammar learning tasks with similar design that included 12-month-olds who ranged in the number of lab visits (Saffran et al., 2008, Exp. 1 Language P; and Saffran & Wilson, 2003, Exp. 2). Our hypothesis was that the amount of infants' experience with HPP would affect direction of preference.

Methods

A brief description of the four experiments included in this analysis, and our rationale for selecting them, is provided in the Supplementary Information (SI), Section 1 (see Fig. 1 for a summary of the results). Infants were aged between 11-13 months in all studies. A fully reproducible repository hosting data and analyses is available at https://osf.io/g95ub/.

We modeled results of all infants ($N = 102$) who completed the four studies. Number of HPP visits varied from one to six (including the current visit). We fit a linear mixed-effects model including *Looking Time* as the response variable, and *Test Item* (Familiar vs. Novel), *HPP* (number of experiments completed by infants) and their

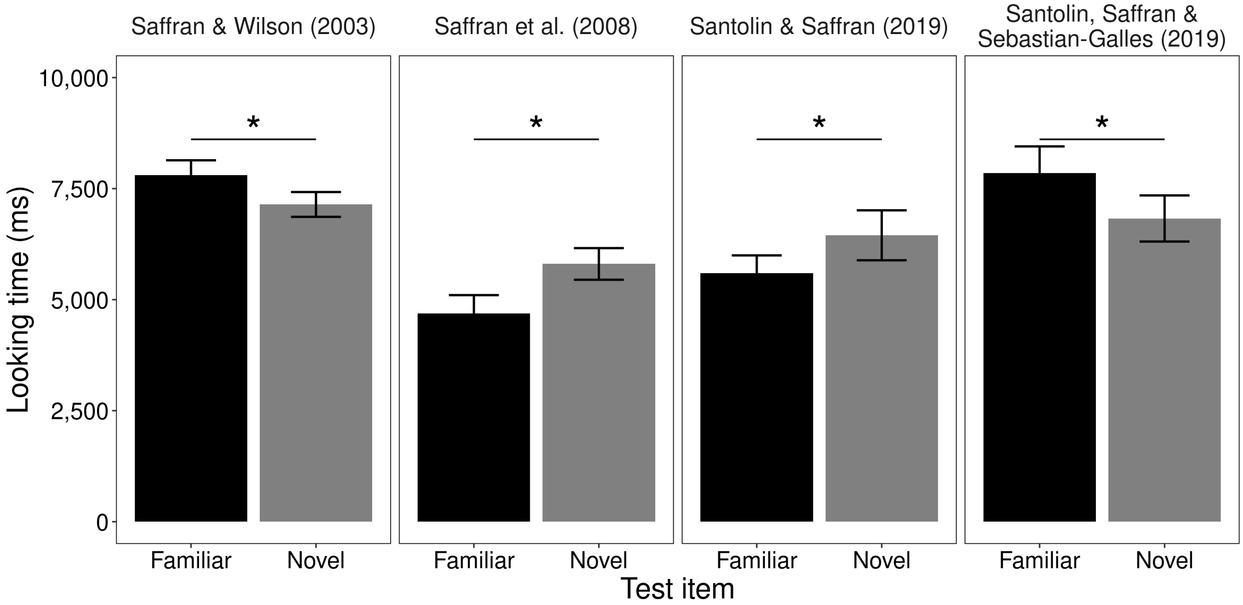

Figure 1. Looking time for familiar and novel test stimuli of the original studies. Stimuli vary based on the experiment. Error bars indicate the standard error of the mean.

interaction as fixed effects. We also included by-participant and by-study random intercepts (4 levels: Santolin & Saffran, 2019; Saffran et al., 2008; Saffran & Wilson, 2003; Santolin et al., 2019). The *HPP* predictor was coded as a continuous variable indicating each infant's total number of HPP experiments. *Test Item* was centered on familiar test items (Familiar = 0; Novel = 1). Since the experiments differ at distinct levels (e.g., different stimuli, lab location), the model accounted for cross-participant and cross-study differences in looking time. Degrees of freedom were approximated using the Kenward-Rogers approach (e.g., Judd, Westfall, & Kenny, 2012), which can result in non-integer values. See SI, Section 3, for additional details.

We predicted a *Test Item* (familiar vs. novel) by number of *HPP* studies interaction, indicating that the duration of infants' looking towards familiar versus novel items would depend on infants' HPP experience. An interaction could result from at least three different patterns of results: an increase in looking time for novel items, a decrease in looking time for familiar items, or both, as a result of additional HPP experience.

Results

The interaction was statistically significant, $F(1{,}100.00) = 11.99$, $p = .001$, suggesting that the effect of Test Items on looking time differences was affected by the number of HPP experiments infants had participated in (Table 1, Fig. 2). In line with our predictions, the size of the difference between looking times on familiar and novel test items changed as a function of number of HPP visits.

The main effect of the HPP predictor was also significant, $F(1{,}133.12) = 4.80$, $p =$ .030, indicating that the Test Item by HPP interaction is mainly driven by a significant decrease in looking time to familiar items as the number of HPP visits increases. There was no evidence that a greater number of HPP visits was accompanied by longer looking to novel items, $F(1{,}133.12) = 0.27$, $p = .606$.

The number of infants in our dataset who had participated in many HPP studies were very small; in particular, the five and six HPP visits groups each included only a single infant. We thus reanalyzed the data to ensure that the pattern of results was not driven by the small number of infants who had visited the lab far more times than most; these participants may not be representative of our samples more generally. The pattern of results was unchanged, indicating that the interaction effect was not driven exclusively by participants with an unusually high number of visits (HPP 1-5: $F(1{,}99.00) = 10.29$, $p =$ .002; HPP 1-4: $F(1{,}98.00) = 10.42$, $p = .002$; HPP 1-3: $F(1{,}92.00) = 4.56$, $p = .035$). Notably, the interaction is significant even with the subset of infants who participated in 1-2 HPP experiments only $F(1{,}78.00) = 4.05$, $p = .048$; see SI, Section 4, for details).

In addition, we conducted the main analysis on the older datasets of Saffran and Wilson (2003) and Saffran et al. (2008) alone, and found a similar significant interaction between test item (novel vs. familiar) and number of HPP visits ($F(1{,}50.00) = 11.00$, $p =$ .002); see SI, Section 5 for details).

Table 1

Summary of the results of the linear mixed-effects model.

	Coefficient	*SEM*	95% CI	*F*	Den. *df*	*p*
Intercept	7,679.0	673.287	[6389.7, 9294.6]	124.638	9.060	< .001
Test Item	-1,398.8	411.324	[-2204.9, -589.1]	11.565	100.000	.001
HPP	-539.7	238.688	[-999.9, -74.7]	4.800	133.117	.030
Test Item × *HPP*	667.1	192.645	[247.2, 1028.5]	11.992	100.000	.001

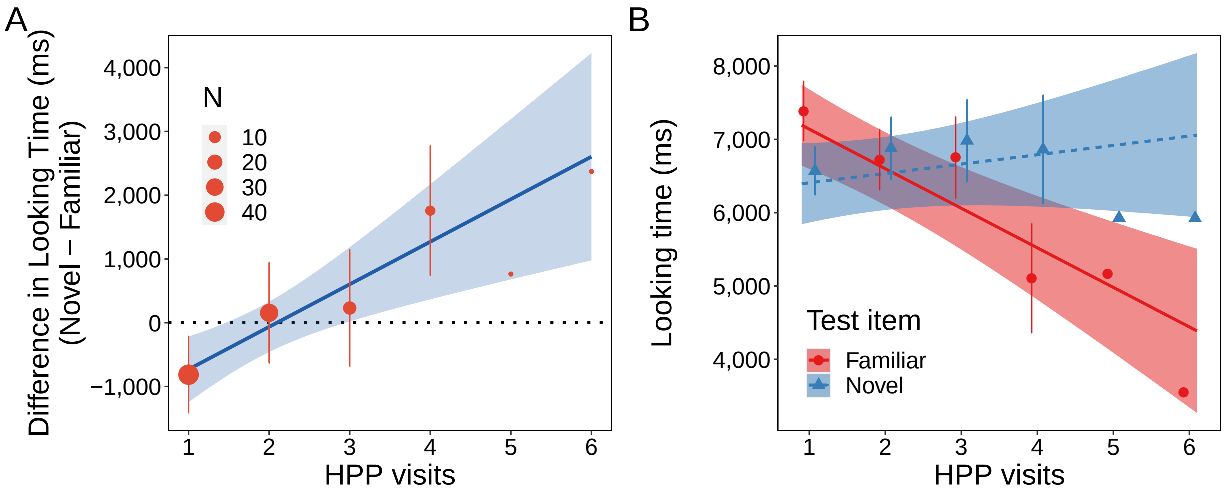

Figure 2. A: Difference in looking time between novel and familiar trials, as a function of HPP visits. Shaded bands indicate 95% CIs. Points represent group means, with error bars representing 95% CIs. B: Predicted looking time (in ms) for familiar and novel test items plotted against number of HPP visits. Shaded bands represent +1/-1 SEs. Points represent group means with +1/-1 SEs as error bars.

Discussion

Experience with the Head-turn Preference Procedure affects direction of preference, at least for the subset of studies examined in this article. The exploratory analyses included data from four experiments with 12-month-old infants performing artificial grammar learning tasks. Infants who had *not* previously experienced the HPP setting were more likely to show familiarity preferences than infants who had prior experience. One possible explanation for this finding relates to the structure of the HPP task. There are at least two types of information that must be simultaneously encoded during an infant's first HPP experiment: 1) visual-auditory contingency (i.e., sounds appear contingently on the infant looking at the screen), and 2) the experiment stimuli (e.g., word sequences, sound streams). When experiencing HPP for the first time, infants must both learn the structure of the HPP method and solve the learning problem itself (e.g., grammatical pattern learning). Such double-processing of information likely increases the task complexity, biasing results towards familiarity preferences. Infants who return to the lab for subsequent HPP experiments may be more able to focus on the learning problem, resulting in better learning as evidenced by novelty preferences.

It is important to notice that this effect may not just be limited to experiencing the HPP setting *per se*, but may also be influenced by the laboratory visit itself. When infants visit the lab for the first time, they face an unusual situation: a new environment with unfamiliar people, testing rooms with a peculiar design (e.g., monochrome walls with big screens), and novel sounds and images (e.g., blinking lights). This is a significant amount of information for a young infant to process at once. In contrast, as infants come back to the lab for subsequent studies, the location, testing room and research staff may become more familiar, reducing the information load (see Rovee-Collier, 1997, for effects of consistent training and testing contexts on reminding infants of details of prior experiences). In the current study, the number of laboratory visits was significantly correlated with the number of HPP visits, $r(100 = .92$, $p < .001$, 95% CI = [.88, .94]), therefore the current analyses cannot discern which type of previous experience (HPP procedure and/or lab setting) is responsible for the observed results.

Our findings have important implications for the interpretation of directions of preference in future studies. Prior experience with a lab or research paradigm could account for distinct, and sometimes counterintuitive, patterns of preference. We encourage researchers to track number of visits as part of their lab's workflow, and to consider this form of prior experience when preregistering analytic plans and interpreting results. Doing so may be particularly informative when unpredicted directions of preferences emerge, as in the replication that spawned the current set of analyses. Recording the type of task implemented with HPP might also be informative. Accumulating experience with different tasks (e.g., those measuring spontaneous preferences versus those measuring learning over the course of an experiment) might have a different effect on the results than having experienced only tasks including a learning phase.

It is also possible that apparent null effects may be driven by variability in the number of lab visits; infants with more lab experience may show novelty preferences while infants with less lab experience may exhibit familiarity preferences, leading to an overall lack of preference across the sample. Effects of prior research experience are less likely to be evident in studies with large effect sizes, where there is less intra-infant variability. In addition, apparent age differences may conceivably be the result not of age per se, but of the number of prior studies, since older infants are likely to have participated in more experiments than younger infants, on average. By tracking infants' study participation, it becomes possible to examine these potential effects, which may be especially apparent in tasks that yield relatively small effects (as most infant studies do).

A related hypothesis suggests that less-common directions of preference for studies addressing a given topic (e.g., rule learning) likely represent sign errors (a sampling error in which the estimated effect has the wrong sign, e.g. a novelty preference is incorrectly estimated to be a familiarity preference; see also Gelman & Carlin, 2014) as opposed to true infant preferences (Bergmann, Rabagliati, & Tsuji, 2019; Rabagliati, Ferguson, & Lew-Williams, 2019). While this may be the case, it is also possible that some discrepancies in preferential looking are related to factors like those investigated in the current study: prior experience with the testing environment. For this reason, unexpected directions of preference may actually be meaningful and informative about the state of infant learners in specific studies.

These results also suggest extensions of models of the factors inducing different patterns of preference (e.g., Hunter & Ames, 1988). The current results suggest that the dimension of task complexity could be expanded beyond the specific task content (e.g., how complex are the stimuli presented) to include *infants' familiarity with the paradigm.* Our findings, in fact, suggest that the learning outcome of a given task is constrained by how much task experience infants have accumulated through prior lab visits. Therefore, the amount of novel information infants must process in parallel during a study increases the task demands, and the likelihood of showing a familiarity preference. This may well include the novelty of the experimental paradigm. Ongoing efforts in the infant research community to facilitate large-scale replications of studies (e.g., The ManyBabies Consortium, 2020) provide a unique opportunity to determine whether experience with different paradigms influences preferential behavior. Expanding our findings to other paradigms (e.g., infant-controlled preferential looking procedures, visual-world paradigms) would continue to advance our understanding of how task/laboratory experience modulates infants' performance. These efforts, in turn, will bring us closer to connecting our research paradigms with the pressing questions about infant behavior that we hope to answer.

References

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48–53. https://doi.org/10.1111/j.1467-7687.2007.00563.x

Bergmann, C., & Cristia, A. (2016). Development of infants' segmentation of words from native speech: A meta-analytic approach. *Developmental Science*, *19*(6), 901–917. https://doi.org/10.1111/desc.12341

Bergmann, C., Rabagliati, H., & Tsuji, S. (2019). What's in a looking time preference? https://doi.org/10.31234/osf.io/6u453

Bosch, L., & Sebastián-Gallés, N. (2001). Evidence of Early Language Discrimination Abilities in Infants From Bilingual Environments. *Infancy*, *2*(1), 29–49. https://doi.org/10.1207/S15327078IN0201_3

Colombo, J., & Bundy, R. S. (1983). Infant response to auditory familiarity and novelty. *Infant Behavior & Development*, *6*(3), 305–311. https://doi.org/10.1016/S0163-6383(83)80039-3

Dawson, C., & Gerken, L. (2009). From Domain-Generality to Domain-Sensitivity: 4-Month-Olds Learn an Abstract Repetition Rule in Music That 7-Month-Olds Do Not. *Cognition*, *111*(3), 378–382. https://doi.org/10.1016/j.cognition.2009.02.010

DePaolis, R. A., Keren-Portnoy, T., & Vihman, M. (2016). Making sense of infant familiarity and novelty responses to words at lexical onset. *Frontiers in Psychology*, *7*, 715. https://doi.org/10.3389/fpsyg.2016.00715

Fantz, R. L. (1964). Visual Experience in Infants: Decreased Attention to Familiar Patterns Relative to Novel Ones. *Science*, *146*(3644), 668–670. https://doi.org/10.1126/science.146.3644.668

Ferguson, B., Franconeri, S. L., & Waxman, S. R. (2018). Very young infants learn abstract rules in the visual modality. *PloS One*, *13*(1), e0190185. https://doi.org/10.1371/journal.pone.0190185

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*(6), 499–504. https://doi.org/10.1111/1467-9280.00392

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/https://doi.org/10.1177/1745691614551642

Houston-Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development*, *13*(4), 341–348. https://doi.org/10.1002/icd.364

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. In *Advances in infancy research, Vol. 5.* (pp. 69–95). Westport, CT, US: Ablex Publishing.

Johnson, S. P., Fernandes, K. J., Frank, M. C., Kirkham, N., Marcus, G., Rabagliati, H., & Slemmer, J. A. (2009). Abstract Rule Learning for Visual Sequences in 8- and 11-Month-Olds. *Infancy : The Official Journal of the International Society on Infant Studies*, *14*(1), 2–18. https://doi.org/10.1080/15250000802569611

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

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*(1), 1–23. https://doi.org/10.1006/cogp.1995.1010

Rabagliati, H., Ferguson, B., & Lew-Williams, C. (2019). The profile of abstract rule learning in infancy: Meta-analytic and experimental evidence. *Developmental Science*, (1), e12704. https://doi.org/10.1111/desc.12704

Roder, B. J., Bushneil, E. W., & Sasseville, A. M. (2000). Infants' Preferences for Familiarity and Novelty During the Course of Visual Processing. *Infancy*, *1*(4), 491–507. https://doi.org/10.1207/S15327078IN0104_9

Rose, S. A., Feldman, J. F., & Jankowski, J. J. (2004). Infant visual recognition memory. *Developmental Review*, *24*(1), 74–100. https://doi.org/10.1016/j.dr.2003.09.004

Rovee-Collier, C. (1997). Dissociations in infant memory: Rethinking the development of implicit and explicit memory. *Psychological Review*, *104*(3), 467.

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.

Sebastian-Galles, N., & Santolin, C. (2020). Bilingual acquisition: The early steps. *Annual Review of Developmental Psychology (in Press)*.

Sebastián-Gallés, N., & Bosch, L. (2009). Developmental shift in the discrimination of vowel contrasts in bilingual infants: Is the distributional account all there is to it? *Developmental Science*, *12*(6), 874–887. https://doi.org/10.1111/j.1467-7687.2009.00829.x

The 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. https://doi.org/10.1177/2515245919900809

Thiessen, E. D. (2012). Effects of inter- and intra-modal redundancy on infants' rule learning. *Language Learning and Development*, *8*(3), 197–214. https://doi.org/10.1080/15475441.2011.583610

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-Directed Speech Facilitates Word Segmentation. *Infancy*, *7*(1), 53–71. https://doi.org/10.1207/s15327078in0701_5

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, *39*(4), 706–716. https://doi.org/10.1037/0012-1649.39.4.706

Appendix

S1: Experiments included in the linear mixed-effects model

The selected experiments 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) 11- to 13-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: $F(1, 38) = 5.37$, $p < .05$.

Saffran, Hauser, Seibel, Kapfhamer, Tsao, & Cushman (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

[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 reviewer suggestions.

sentences violating the grammar. Infants showed a significant novelty preference: $t(11) = 2.52$, $p < .05$.

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: $t(26) = 2.45$, $p = .021$, $d = 0.47$.

We replicated the Predictive Language condition of Santolin & Saffran (2019) at the University Pompeu Fabra, Barcelona (*Santolin, Saffran & Sebastian-Galles, 2019*, 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: $t(23) = 2.30$, $p = .030$, $d = 0.47$. All results are shown in Figure 1 of the main manuscript.

S2: Participants information

We retrieved data from 102 infants who had participated in a range of 1-6 HPP visits. Three of the experiments 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). All studies were conducted according to guidelines provided by the Declaration of Helsinki, with written informed consent obtained from a caregiver for each child before any assessment or data collection. Ethical approval was granted by the University of Wisconsin-Madison Social and Behavioral Sciences IRB for Saffran & Wilson (2003), Saffran et al. (2008), and Santolin & Saffran (2019), and by the Comitè Etic d'Investigació Clinica, Parc de Salut Mar Barcelona, for Santolin et al. (2019).

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.

S3: Linear mixed-effects model - additional information

We fit a model predicting looking time (LT) including $Item$ (Familiar vs. Novel), 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), Santolin, Saffran & Sebastian-Galles (2019), Saffran et al. (2008), Saffran & Wilson (2003)] 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. This model accounts for cross-participant variability in overall looking time (as some infants look longer than others), and for cross-study differences in overall looking time. The model was fit using the `lme4` R package (Bates, Kliegl, Vasishth, & Baayen, 2015). 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 of the degrees of freedom (e.g., Judd, Westfall, & Kenny, 2012).

S4: Results sub-setting data to participants with less than 6, 5, 4, and 3 HPP studies

Consistent with the results of the entire dataset, we found a statistically significant interaction of *Test Item* with the number of *HPP* visits when reducing the sample to the infants who participated in less than 6, 5, 4, and 3 HPP experiments. Below, a table reporting the output of the linear mixed-effects model fitted on the original and reduced samples.

Table A1

Summary of the results of the linear mixed-effects model performed on the reduced data.

Subset	**Term**	**Coefficient**	***SEM***	**95% CI**	**F**	**Den. *df***	***p***
Original	*Intercept*	7,679.0	673.3	[6389.7, 9294.6]	124.6	9.1	< .001
	Test Item	-1,398.8	411.3	[-2204.9, -589.1]	11.6	100.0	.001
	HPP	-539.7	238.7	[-999.9, -74.7]	4.8	133.1	.030
	Test Item × *HPP*	667.1	192.6	[247.2, 1028.5]	12.0	100.0	.001
HPP 1-5	*Intercept*	7,675.0	691.6	[6452.7, 9029.3]	118.3	10.1	< .001
	Test Item	-1,416.1	435.4	[-2237.7, -543.7]	10.6	99.0	.002
	HPP	-535.6	261.2	[-1081.1, -37.5]	4.0	133.6	.048
	Test Item × *HPP*	677.8	211.3	[241, 1081.9]	10.3	99.0	.002
HPP 1-4	*Intercept*	7,611.1	719.3	[6188.6, 9275.3]	107.4	10.5	< .001
	Test Item	-1,491.1	452.1	[-2348.4, -578.5]	10.9	98.0	.001
	HPP	-500.8	278.7	[-1070.2, 98.2]	3.0	131.9	.083
	Test Item × *HPP*	726.2	224.9	[294.2, 1145]	10.4	98.0	.002
HPP1-3	*Intercept*	7,470.0	794.6	[6007.1, 9172.6]	83.8	14.3	< .001
	Test Item	-1,349.9	532.3	[-2426.9, -267.6]	6.4	92.0	.013
	HPP	-395.8	366.6	[-1182.6, 316.1]	1.1	122.3	.299
	Test Item × *HPP*	627.9	294.1	[3, 1252.6]	4.6	92.0	.035
HPP 1-2	*Intercept*	7,301.9	1009.9	[5253.3, 9362.1]	48.9	23.9	< .001
	Test Item	-1,783.7	726.7	[-3199.6, -360.7]	6.0	78.0	.016
	HPP	-261.3	586.8	[-1449.8, 991.7]	0.2	107.0	.667
	Test Item × *HPP*	969.5	481.8	[38.3, 2010.4]	4.0	78.0	.048

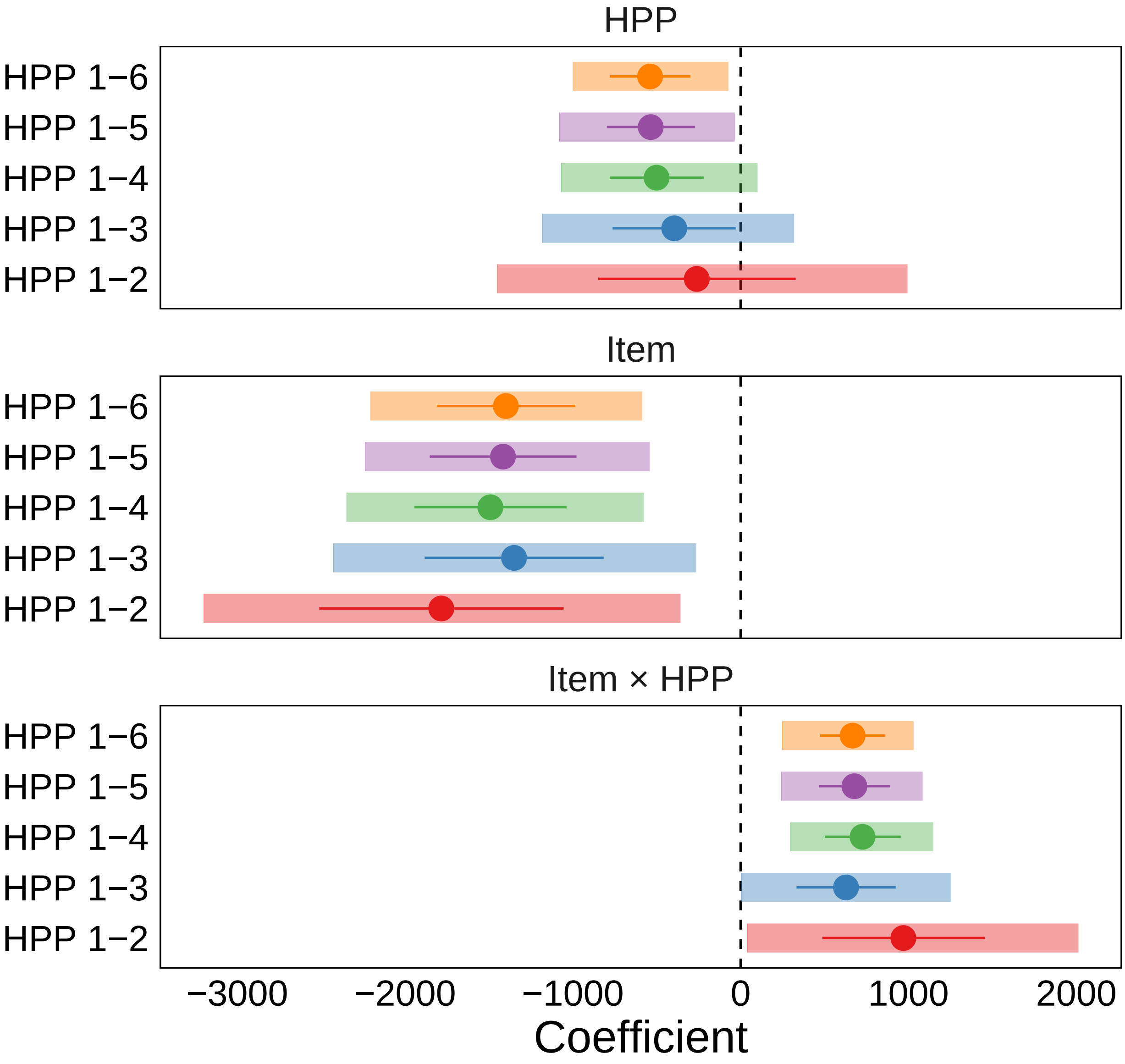

Figure A1. Estimated coefficients for the three predictors (Test Item, HPP, and their interaction) across the same linear mixed-effects model fitted on the overall sample (HPP 1-6, including all participants), and its subsets (including particpiants that completed less than 6, 5, 4, 3 HPP studies). Dots indicate point estimates, error bars indicate SEs, and shaded boxes indicate 95% CIs.

S5: Results of Saffran & Wilson (2003) and Saffran et al. (2008) only

We conducted this additional analysis to ensure that the results obtained on the entire dataset were not driven primarily by the two most recent datasets (Santolin & Saffran, 2019; Santolin et al., 2019), in which we first noticed the pattern of results (i.e., the flip in preference). Results closely mirrored those of the entire dataset, showing a statistically significant interaction between test item (novel vs. familiar) and number of HPP visits ($F(1{,}50.00) = 11.00$, $p = .002$). As shown in the figure below, there is a decline in familiarity preference as the number of HPP visits increases (Panel A), and an interaction between test item (novel vs. familiar) and number of HPP visits (Panel B).

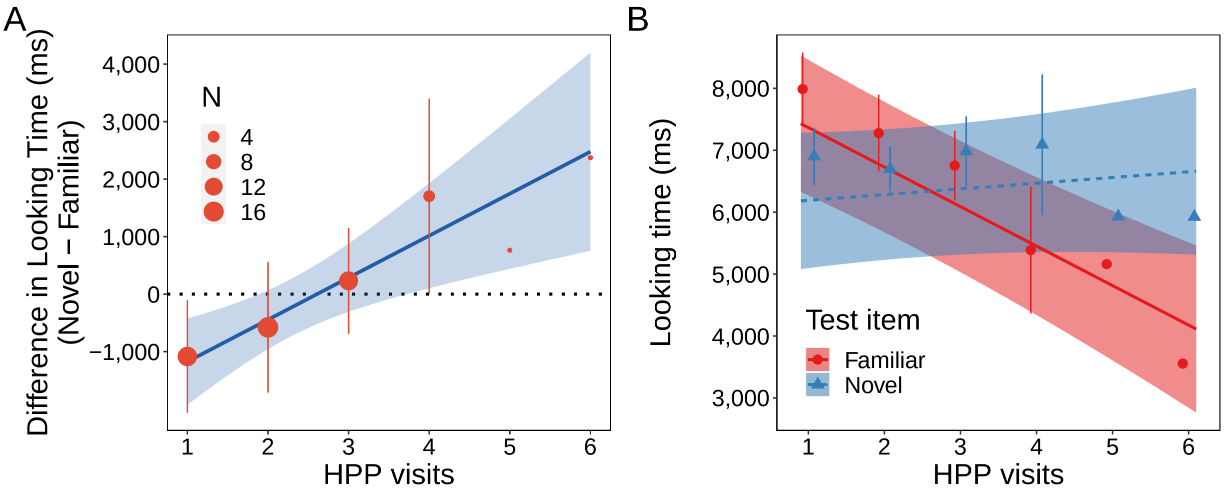

Figure A2. A: Difference in looking time between novel and familiar trials for data from Saffran & Wilson (2003) and Saffran et al. (2008) only, as a function of HPP visits. Shaded bands indicate 95% CIs. Points represent group means, with error bars representing 95% CIs. B: Predicted looking time (in ms) for familiar and novel test items plotted against number of HPP visits (older datasets only). Shaded bands represent +1/-1 SEs. Points represent group means with +1/-1 SEs as error bars.

S6. Session info

R version 3.6.3 (2020-02-29) Platform: x86_64-pc-linux-gnu (64-bit) Running under: Ubuntu 20.04.1 LTS

Matrix products: default BLAS: /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.9.0 LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0

locale: [1] LC_CTYPE=en_GB.UTF-8 LC_NUMERIC=C
[3] LC_TIME=es_ES.UTF-8 LC_COLLATE=en_GB.UTF-8
[5] LC_MONETARY=es_ES.UTF-8 LC_MESSAGES=en_GB.UTF-8
[7] LC_PAPER=es_ES.UTF-8 LC_NAME=C
[9] LC_ADDRESS=C LC_TELEPHONE=C
[11] LC_MEASUREMENT=es_ES.UTF-8 LC_IDENTIFICATION=C

attached base packages: [1] stats graphics grDevices utils datasets methods base

other attached packages: [1] purrr_0.3.4 kableExtra_1.2.1 ggplot2_3.3.2 here_0.1
[5] tibble_3.0.3 dplyr_1.0.2 magrittr_1.5 knitr_1.29
[9] papaja_0.1.0.9997

loaded via a namespace (and not attached): [1] pillar_1.4.6 compiler_3.6.3 highr_0.8 base64enc_0.1-3
[5] tools_3.6.3 digest_0.6.25 viridisLite_0.3.0 evaluate_0.14
[9] lifecycle_0.2.0 gtable_0.3.0 pkgconfig_2.0.3 rlang_0.4.7
[13] rstudioapi_0.11 yaml_2.2.1 xfun_0.16 xml2_1.3.2
[17] httr_1.4.2 withr_2.2.0 stringr_1.4.0 generics_0.0.2
[21] vctrs_0.3.2 webshot_0.5.2 rprojroot_1.3-2 grid_3.6.3
[25] tidyselect_1.1.0 glue_1.4.1 R6_2.4.1 rmarkdown_2.3
[29] bookdown_0.20 backports_1.1.9 scales_1.1.1 ellipsis_0.3.1

```
[33] htmltools_0.5.0 rvest_0.3.6 colorspace_1.4-1 stringi_1.4.6
[37] munsell_0.5.0 crayon_1.3.4
```

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., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models. *arXiv:1506.04967 [Stat]*. Retrieved from http://arxiv.org/abs/1506.04967

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

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.