Cognition

Experience Shapes Early Noun Comprehension from 8-18 Months: The Roles of Word Frequency and Referent Familiarity

--Manuscript Draft--

Manuscript Number:	COGNIT-D-25-01025
Article Type:	Full Length Article
Section/Category:	Fei Xu - Language development, infant cognition, comp. models of language development, word learning
Keywords:	word learning; language development; eyetracking; concept; frequency; infancy
Corresponding Author:	Lily Zihui Zhu Harvard University UNITED STATES OF AMERICA
First Author:	Lily Zihui Zhu
Order of Authors:	Lily Zihui Zhu
	Andrei Amatuni
	Shannon Egan-Dailey
	Hallie Garrison
	Evgenii Kalenkovich
	Sharath Koorathota
	Lillianna Righter
	Shaelise Tor
	Elika Bergelson
Abstract:	We investigated how experience with particular nouns and referents influences infant word comprehension by combining home recordings with in-lab eyetracking. In Study 1, infants (n=44) participated in a yearlong longitudinal study which included monthly home recordings (daylong audio and hourlong video, 6-17mo.) and hand-tailored eyetracking experiments (every two months, 8-18mo.). Word comprehension was operationalized as how much more infants looked at named images than distractors. Each session tested 16 words: 8 held constant across all infants at each timepoint ("generic" common nouns, e.g., "shoe"), and 8 hand-picked for each session from each child's recent home recordings. Of the hand-picked items, half appeared as images familiar to children from home (e.g., their own ball); the other half and all generic items appeared as prototypical child-friendly (but not individually-tailored) images. Results showed expected improvements with age, and an effect of word frequency (particularly at 12 and 14mo. timepoints) but not referent familiarity. In Study 2 (n=247), cross-sectional controls completed the same eyetracking experiments (i.e., those hand-tailored for Study 1 children) as a stronger test of the roles of frequency and referent familiarity. Here, we did find an effect of referent familiarity. Study 2 infants understood words less well than those in Study 1, particularly when tested with images of referents selected from Study 1 children's homes. Together, these findings highlight the role of experience in shaping early noun comprehension and suggest that once infants understand words, they readily generalize them to new instances beyond those they are familiar with.

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

Supplementary Materials

The sections below contain the Supplementary Materials for the main manuscript.

A. Computing and Comparing Frequency

B. Exclusion Criteria

C. R Packages

D. More Details on Comparing SEEDLingS and Control Participants

E. Adult Control Study

F. Timecourse Visualizations

Appendix A

Computing and Comparing Frequency

In this section, we present more details on how we computed the relative frequency measure and compared frequency across trial types within each participant group.

6.1 Annotating Noun Tokens in Recordings

We annotated the concrete noun tokens from infants' home recordings according to the following steps (Kalenkovich et al., 2024). First, each token was transcribed exactly as spoken (e.g., including affixes, compounds, diminutives, wordplay, etc.). To group related forms within a recording, each noun was then assigned a *basic level* form, defined as the most common version of that noun in that recording. These basic level forms were used in the eyetracking sessions. For example, if a family said 'bear', 'bears', and 'teddy bear', and 'teddy bear' appeared most frequently (with 'bear' in any form also ranking among the top nouns), then 'teddy bear' was selected as the test word. To ensure consistency across families, we then standardized all basic level nouns throughout the SEEDLingS corpus by assigning each one a *global basic label*, which functioned like a lemma (Moore & Bergelson, 2024). For instance, tokens labeled 'bear' or 'teddy bear' at different timepoints were consolidated under the global basic level 'bear'. To compare word frequencies across infants, each test word was converted to its global basic level. For example, if infant X was tested on 'teddy bear', the frequency of that word was calculated by counting all noun tokens with the global basic level 'bear', and dividing that by the total number of noun tokens in their recordings from the two months prior to the eyetracking session. Note that a highly similar procedure for computing frequency has been used in other work with the broader SEEDLingS dataset (e.g., Dong et al., 2025), so this description overlaps somewhat with others.

Table A1

Mean and standard error of untransformed test pair frequency across trial types (generic, audio-nouns, video-nouns) within each participant group at each month.

Participant Group	Month	Generic	Audio-Nouns	Video-Nouns
SEEDLingS	8	0.57% (0.080%)	2.02% (0.178%)	2.14% (0.174%)
SEEDLingS	10	0.66% (0.054%)	1.92% (0.119%)	1.72% (0.122%)
SEEDLingS	12	0.60% (0.053%)	1.60% (0.107%)	1.72% (0.150%)
SEEDLingS	14	0.69% (0.051%)	1.67% (0.126%)	1.75% (0.134%)
SEEDLingS	16	0.47% (0.036%)	2.01% (0.145%)	1.60% (0.108%)
SEEDLingS	18	0.45% (0.028%)	1.46% (0.076%)	1.33% (0.080%)
Control (Proxy Estimates)	8	0.57% (0.025%)	0.93% (0.051%)	0.72% (0.060%)
Control (Proxy Estimates)	10	0.66% (0.019%)	0.82% (0.055%)	0.57% (0.047%)
Control (Proxy Estimates)	12	0.63% (0.019%)	0.62% (0.044%)	0.58% (0.056%)
Control (Proxy Estimates)	14	0.68% (0.027%)	0.66% (0.041%)	0.48% (0.042%)
Control (Proxy Estimates)	16	0.46% (0.019%)	0.70% (0.045%)	0.55% (0.048%)
Control (Proxy Estimates)	18	0.44% (0.013%)	0.43% (0.031%)	0.45% (0.038%)

6.2 Comparing Frequency Between Trial Types

Raw word frequencies are known to be highly skewed (e.g., following Zipfian distributions), and it is standard practice in psycholinguistics to apply a log transformation before entering frequency as a dependent variable in statistical models. However, as shown in Fig. 3, the distributions of log-transformed test pair frequencies remain visibly non-normal. This is particularly true for the proxy frequency estimates derived for control infants based on SEEDLingS infants' recordings, where the distributions are not even approximately bell-shaped. Therefore, we adopted nonparametric methods in this section to analyze untransformed test pair frequency. See Table A1 for summary statistics by participant group, month, and test pair type.

We first conducted Kruskal-Wallis tests (the nonparametric equivalent of one-way ANOVA) to examine whether frequencies differed across the three trial types (generic, audio-nouns, and video-nouns) within each participant group at each month (see Table A2). For the SEEDLingS group, we observed significant differences across trial types at all timepoints between 8-18 months (ps $< .001$). This pattern was expected given that the audio-nouns and video-nouns were selected based on their high frequency in SEEDLingS infants' recordings, whereas the generic words were not.

For the control group, the pattern was more nuanced. We again found significant differences found at all timepoints ($p < 0.01$) except at 18 months ($p = 0.22$). However, the test statistics for control infants were markedly smaller (χ^2s ranging from 9 to 29) compared to SEEDLingS infants (χ^2s > 100), indicating that even when differences existed among trial types, they were less pronounced in the control group.

To further explore which pairs of trial types differed in frequency, we conducted pairwise Wilcoxon Rank Sum Tests (nonparametric equivalent of two-sample t-tests) with Bonferroni corrections (Table A3). Results for the SEEDLingS group again aligned with our experimental design: For all months, both generic vs. audio-nouns and generic vs. video-nouns comparisons were highly significant (ps $< .001$), while audio- vs. video-nouns were statistically indistinguishable (ps $= 1$), with a single exception at 16 months ($p = 0.02$). However, even this exception had a higher p-value than the other two comparisons at 16 months, suggesting that the audio- and video-nouns were still more similar to each other than either was to the generic words.

By contrast, the control group exhibited no consistent pattern of differences in frequency. Some pairwise comparisons reached significance, but the effects were inconsistent across months and trial types. Importantly, when differences were detected, they were associated with higher p-values and smaller effect sizes than in the SEEDLingS group.

Together, these findings indicate that our frequency manipulation was effective and reliable for the SEEDLingS infants: Audio- and video-nouns were consistently more frequent than generic nouns, but not significantly different from each other. In contrast, the same high-frequency stimuli were not systematically more or less frequent for the control group.

Table A2

Results of Kruskal-Wallis tests (nonparametric equivalent of one-way ANOVA) that compared test pair frequency across trial types (generic, audio-nouns, video-nouns) within each participant group at each month.

Participant Group	Month	Chisq	df	p-value
SEEDLingS	8	124.71	2	$< .001$
SEEDLingS	10	100.48	2	$< .001$
SEEDLingS	12	103.72	2	$< .001$
SEEDLingS	14	107.13	2	$< .001$
SEEDLingS	16	177.30	2	$< .001$
SEEDLingS	18	181.26	2	$< .001$
Control (Proxy Estimates)	8	28.72	2	$< .001$
Control (Proxy Estimates)	10	14.14	2	$< .001$
Control (Proxy Estimates)	12	9.10	2	0.011
Control (Proxy Estimates)	14	20.00	2	$< .001$
Control (Proxy Estimates)	16	18.56	2	$< .001$
Control (Proxy Estimates)	18	3.00	2	0.223

Note. The Kruskal-Wallis tests were conducted using the 'stats::kruskal.test()' function in R.

Table A3

Results of pairwise Wilcoxon Rank Sum Tests (nonparametric equivalent of two-sample t-tests) with Bonferroni corrections that compared test pair frequency between each combination of trial types (generic vs. audio-nouns, generic vs. video-nouns, audio- vs. video-nouns) within each participant group at each month.

Month	Generic vs. Audio-Nouns	Generic vs. Video-Nouns	Audio- vs. Video-Nouns
SEEDLingS			
8	$p < .001$	$p < .001$	$p = 1$
10	$p < .001$	$p < .001$	$p = 0.357$
12	$p < .001$	$p < .001$	$p = 1$
14	$p < .001$	$p < .001$	$p = 1$
16	$p < .001$	$p < .001$	$p = 0.017$
18	$p < .001$	$p < .001$	$p = 0.247$
Control			
8	$p < .001$	$p = 0.525$	$p = 0.023$
10	$p = 0.111$	$p = 0.022$	$p = 0.002$
12	$p = 0.696$	$p = 0.008$	$p = 0.378$
14	$p = 1$	$p < .001$	$p = 0.004$
16	$p < .001$	$p = 1$	$p = 0.02$
18	$p = 0.192$	$p = 1$	$p = 1$

Note. The pairwise Wilcoxon Rank Sum Tests were conducted using the 'stats::pairwise.wilcox.test()' function in R.

Appendix B

Exclusion Criteria

The text below provides a greater degree of detail regarding exclusions than the main manuscript for curious readers.

6.3 SEEDLingS

Our final SEEDLingS dataset included 6,165 trials from 254 infant visits (M = 24.27/32 trials, $SD = 6.94$, range: 8-32). During four visits, the participants had to restart the experiment due to technical errors; we retained only the most recent attempts. We originally planned for 8,448 trials in total (44 subjects $\times$ 6 visits $\times$ 32 planned trials per visit). However, 60 trials were skipped as the experiment ended early due to infant fussiness/crying.

Among the remaining 8,388 planned trials, 834 were attempted twice, and 15 were attempted three times, as the experimenter deemed the initial attempts likely unusable. This resulted in 9,252 trial attempts.

From these trial attempts, we applied several exclusion criteria. First, we removed 64 trials in which participants did not hear the planned stimuli due to infant fussiness/crying, experimenter error, parental failure to say the target word, or technical difficulties (as confirmed from real-time notes and videos of experiment sessions checked post-hoc). Another 2 trials were excluded because the experimenter failed to press a button when the parent said the target word, and we could not determine the target onset (due to noisy and/or missing experiment video). Additionally, 186 trials were removed due to a complete lack of fixation data, and another 2,640 trials were excluded because participants failed to fixate on either the target or distractor for more than 1/3 of the analysis window.

After these exclusions, we retained 6,360 trial attempts for 6,206 out of 8,448 originally planned trials. For trials with multiple attempts, we kept only the most recent

usable attempt. Finally, we excluded 10 participants who contributed too few data (< 25% of planned trials), along with an additional 41 remaining trials from these participants.

6.4 Control

Our final control dataset included 6,151 trials from 247 infant visits ($M = 24.90/32$ trials, $SD = 6.22$, range: 8-32). During 9 visits, the participants had to restart the experiment due to technical errors; we retained only the most recent attempts. Participants from another 25 visits were replaced due to fussiness ($n = 4$), technical issues ($n = 4$), experimenter error ($n = 4$), or ineligibility discovered after the session ($n = 13$). In two further visits, the participant was not replaced, and no further data were collected due to parental interference or inaudible parental speech. We originally planned for 8,448 trials in total (44 subjects × 6 months × 32 planned trials per visit). After excluding data from the two unusable but unreplaced visits (64 trials), another 178 trials were skipped as the experiment ended early due to infant fussiness or technical issues.

Among the remaining 8,206 planned trials, 107 were attempted twice, and 3 were attempted three times, as the experimenter deemed the initial attempts likely unusable. This resulted in 8,383 trial attempts.

From these trial attempts, we applied the same exclusion criteria as above. First, we removed 86 trials in which participants did not hear the planned stimuli due to infant fussiness/crying, experimenter error, parental failure to say the target word, or technical difficulties (as confirmed from real-time notes and videos of experiment sessions checked post-hoc). Another 78 trials were excluded because the experimenter failed to press a button when the parent said the target word, and we could not locate the target onset (due to noisy and/or missing experiment video). Additionally, 190 trials were removed due to a complete lack of fixation data, and another 1,717 trials were excluded because participants failed to fixate on either the target or distractor for more than 1/3 of the analysis window.

After these exclusions, we retained 6,312 trial attempts for 6,204 out of 8,448 originally planned trials. For trials with multiple attempts, we kept only the most recent usable attempt. Finally, we excluded 15 participants who contributed too few data ($< 25\%$ of planned trials), along with their additional 53 remaining trials.

Appendix C

R Packages

We used R (Version 4.2.1; R Core Team, 2022) and the R-packages *assertr* (Version 3.0.1; Fischetti, 2023), *blabr* (Version 0.25.1; Lab, n.d.), *broom* (Version 1.0.5; Robinson et al., 2023), *dplyr* (Version 1.1.4; Wickham, François, et al., 2023), *forcats* (Version 1.0.0; Wickham, 2023), *ggh4x* (Version 0.3.1; van den Brand, 2025), *ggplot2* (Version 3.5.2; Wickham, 2016), *lme4* (Version 1.1.36; Bates et al., 2015), *lmerTest* (Version 3.1.3; Kuznetsova et al., 2017), *lubridate* (Version 1.9.2; Grolemund & Wickham, 2011), *MASS* (Version 7.3.60; Venables & Ripley, 2002), *Matrix* (Version 1.6.5; Bates et al., 2024), *papaja* (Version 0.1.2; Aust & Barth, 2023), *patchwork* (Version 1.3.1; Pedersen, 2025), *purrr* (Version 1.0.2; Wickham & Henry, 2023), *readr* (Version 2.1.4; Wickham, Hester, et al., 2023), *stringr* (Version 1.5.0; Wickham, 2022), *tibble* (Version 3.2.1; Müller & Wickham, 2023), *tidyr* (Version 1.3.0; Wickham, Vaughan, et al., 2023), *tidyverse* (Version 2.0.0; Wickham et al., 2019), and *tinylabels* (Version 0.2.4; Barth, 2023) for all our analyses.

Appendix D

More Details on Comparing SEEDLingS and Control Participants

This section provides further details on the statistical procedures and results comparing SEEDLingS and control participants in terms of the amount of retained data, vocabulary size, and word comprehension performance in the eyetracking experiments.

6.5 Amount of Retained Data

We first compared the two groups on the amount of retained data. If the longitudinal SEEDLingS participants became increasingly familiar with the test setup and benefited from repeated exposure, they should contribute a growing number of usable trials over time and more usable trials overall than the control group. However, Fig. 7 suggests that the two groups contributed a similar number of usable trials. This pattern was confirmed by a linear regression model predicting the number of usable trials per participant as a function of group and age. Results showed a significant main effect of age ($b = 1.14$, $SE = 0.16$, $t(498) = 6.94$, $p < .001$), indicating that both groups contributed more usable data as they got older, but the effect of group was not significant ($b_{\text{SEEDLingS}} =$ -0.65, $SE = 0.56$, $t(498) =$ -1.15, $p = 0.25$). Thus, we find no evidence that SEEDLingS participants contributed more usable trials than the control group, suggesting a limited role for repeated testing experience in data retention.

6.6 Vocabulary Size

We next compared the two groups on receptive and productive vocabulary sizes, as measured by the CDI. A linear model predicting CDI comprehension scores as a function of age, participant group, and their interaction revealed a significant main effect of age ($b =$ 37.85, $SE = 2.87$, $t(462) = 13.18$, $p < .001$), group ($b_{\text{SEEDLingS}} = 47.78$, $SE = 7.03$, $t(462)$ $= 6.80$, $p < .001$), and a significant age × group interaction ($b_{\text{SEEDLingS:age}} = 13.94$, $SE = 4$, $t(462) = 3.48$, $p < .001$). Specifically, according to the model's fixed effects, SEEDLingS participants were reported to comprehend 48 more words than controls at 12 months.

Receptive vocabulary grew with age for both groups: Controls gained about 38 words every two months, while SEEDLingS gained an additional 14 words, widening the gap over time.

In contrast, a similar linear model predicting CDI production scores revealed only a significant main effect of age ($b = 11.20$, $SE = 1.57$, $t(462) = 7.14$, $p < .001$), but not group ($b_{\text{SEEDLingS}} = 2.42$, $SE = 3.84$, $t(462) = 0.63$, $p = 0.53$), and no interaction ($b_{\text{SEEDLingS:age}} = 2.78$, $SE = 2.19$, $t(462) = 1.27$, $p = 0.20$). This suggests that for every two-month increase in age, both SEEDLingS and control participants were reported to produce 11 more words.

However, we interpret the group difference in CDI scores with caution for two main reasons. First, the difference may reflect variation in how parents completed the surveys. For parents of longitudinal SEEDLingS participants, all CDI surveys following the initial administration were pre-filled with responses from the previous form. In contrast, parents of cross-sectional control participants completed the survey only once, without any pre-filled responses. This pre-filling may have inflated CDI scores for SEEDLingS participants, as parents may be less likely to revise or remove previously endorsed words. Second, parents participating in the yearlong SEEDLingS study may have become more attuned to their children's vocabulary over time. Their extended involvement in a word learning study could make them more sensitive to their child's word knowledge and thus more accurate in reporting word comprehension. Supporting this possibility, we observed a stronger correlation between SEEDLingS participants' CDI comprehension scores and their eyetracking performance (0.48, 95% CI = [0.38, 0.57]) than between control participants' CDI comprehension scores and eyetracking performance (0.34, 95% CI = [0.22, 0.45]). However, the confidence intervals overlap, so this difference should be interpreted cautiously.

6.7 Word Comprehension

In the main text, we reported group differences in receptive vocabulary size (CDI comprehension) and word comprehension performance across the three trial types (generic,

audio-nouns, video-nouns) in eyetracking assessment. Here, we ask whether the observed group differences in word comprehension could be attributed to differences in receptive vocabulary. To test this, we fit linear mixed-effects models predicting word comprehension performance at 18 months, where group differences appeared largest, as a function of participant group, and then added CDI comprehension as a covariate to assess whether group effect remained. A caveat is that CDI data were only available for a subset of participants, so these analyses were conducted on a reduced sample (36 out of 44 SEEDLingS participants and 34 out of 41 control participants retained for eyetracking analyses at 18 months).

6.7.1 Generic Items

In the reduced sample with CDI data, we first fit a linear mixed-effects model predicting the increase in target looking for generic items at 18 months as a function of group. Model selection removed the fixed effect of group and returned an intercept-only model, suggesting that the originally observed group difference may have been weak or unstable in this subset.

We then fit a second model including both group and CDI comprehension as fixed effects. Model selection again dropped the group effect but retained a significant main effect of CDI comprehension (b = 5.87e-04, SE = 2.25e-04, $t(67.70) = 2.61$, $p = 0.01$), suggesting that for every 100-word increase in receptive vocabulary size, comprehension of generic words (operationalized as increase in target looking) would increase by 0.06 unit.

In sum, the group difference in comprehension of generic items was not robust. Among participants with CDI data, individual differences in receptive vocabulary better explained comprehension of generic items than participant group.

6.7.2 Audio-Nouns

Using the reduced sample, we first fit a linear mixed-effects model predicting the increase in target looking for audio-nouns at 18 months as a function of group. Model

selection removed the group effect and returned an intercept-only model, suggesting that the original group difference may not be robust within this smaller sample. We then fit a second model including both group and CDI comprehension as fixed effects. Model selection removed both the group effect and the effect of CDI comprehension, resulting in another intercept-only model. Thus, the group difference in audio-noun comprehension was not robust, and receptive vocabulary did not reliably predict performance on these items either.

6.7.3 Video-Nouns

For video-nouns, we first fit a linear mixed-effects model predicting the increase in target looking for video-nouns at 18 months as a function of group. Model selection retained the full model, and the group effect remained significant in this reduced sample ($b_{\text{SEEDLingS}} = 0.18$, $SE = 0.05$, $t(874.50) = 3.27$, $p < .01$). We then fit a second model including both group and CDI comprehension as fixed effects. This time, model selection removed the main effect of CDI comprehension but retained the group effect, yielding the same model structure as before. This indicates that the group difference in video-noun comprehension is robust and cannot be fully accounted for by differences in receptive vocabulary.

Appendix E

Adult Control Study

In order to assess how readily recognizable the hand-picked referent images (video-nouns) vs. prototypical images (audio-nouns and generic nouns) were to individuals with mature lexicons, we conducted an adult control study using the infant stimuli and a written prompt directing them to the target. A total of 113 native English-speaking adults were recruited via Prolific and compensated at an hourly rate of $15. An additional 2 participants were excluded for failing one or more attention checks, and another 6 were excluded for providing responses that indicated they were nonnative English speakers, based on an open-ended picture description task. Each session took an average of 9.01 minutes ($SD = 4.14$, range: 4.03-25.37).

Each participant completed half of the hand-picked trials (i.e., stimuli from 22 of the 44 SEEDLingS infants) and all of the generic trials from a specific infant test month (8, 10, 12, 14, 16, 18 months). This resulted in a total of 184 trials per participant: 8 generic, 88 audio-noun, and 88 video-noun trials. In each trial, participants saw yoked image pairs along with a prompt sentence (e.g., "Can you find the X?") and were instructed to click on the image corresponding to the labeled referent, X. We measured participants' reaction times (RTs) in making their selections. If the hand-picked referents in the video-noun trials were less recognizable than the prototypical referents in the audio-noun and generic trials, we would expect longer reaction times in the video-noun trials. We excluded 275 trials where RTs were more than 3 standard deviations from the mean for that trial type. Because RTs were heavily skewed, we applied a log transformation in subsequent statistical analyses. See Fig. E1 for a visualization of the RT distribution by trial type and Fig. E2 for RTs of the most extreme RT values on average in each condition.

To test whether RTs differed between trial types, and whether any observed differences were consistent across infant test months, we began by fitting the following full

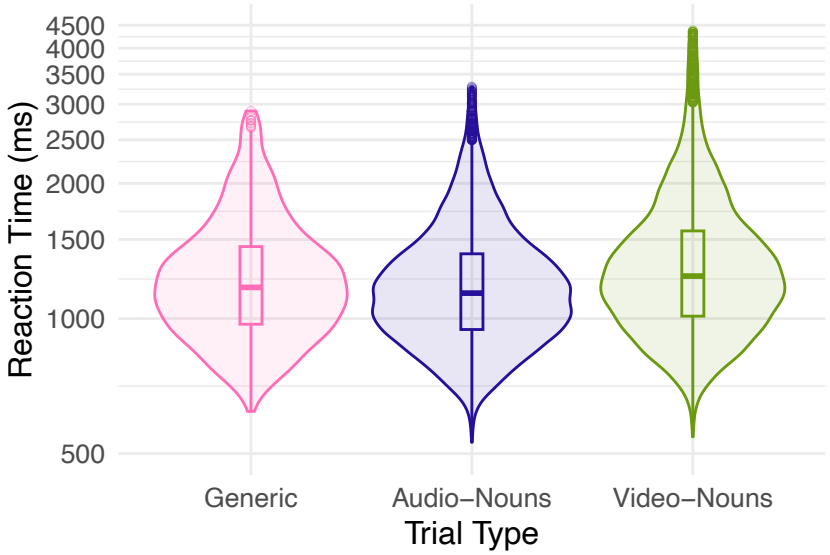

Figure E1

Adult reaction time distribution by trial type (generic, audio-nouns, video-nouns) across all infant test months. Boxes indicate the interquartile range and horizontal lines indicate medians. Empty dots indicate potential outliers with extreme reaction times. Adults were significantly slower to respond in video-noun trials than audio-noun and generic trials. Reaction times in audio-noun and generic trials were similar.

linear mixed-effects model: `logRT ~ month × trial type + (1|subject) + (1|test pair)`. Month was encoded as a categorical variable, as we were comparing between months rather than tracking developmental change over time. Trial type was dummy coded with video-nouns as the reference level, since our primary interest was in contrasting the hand-picked referents in video-nouns trials with the prototypical referents in audio-nouns and generic trials. Model selection dropped the interaction and the main effect of month but retained trial type, indicating that the adults processed different types of stimuli at different speeds, and that this effect was consistent across all infant test months (final model: `logRT ~ trial type + (1|subject) + (1|test pair)`). Results showed that video-noun trials took significantly longer for adults to process than both audio-nouns ($b_{\text{Audio-Nouns}}$ = -0.04, SE = 0.007, t(1,060.73) = -18.15, $p < .001$) and generic trials (b_{Generic} = -0.03, SE = 0.007, t(739.78) = -3.75, $p < .001$).

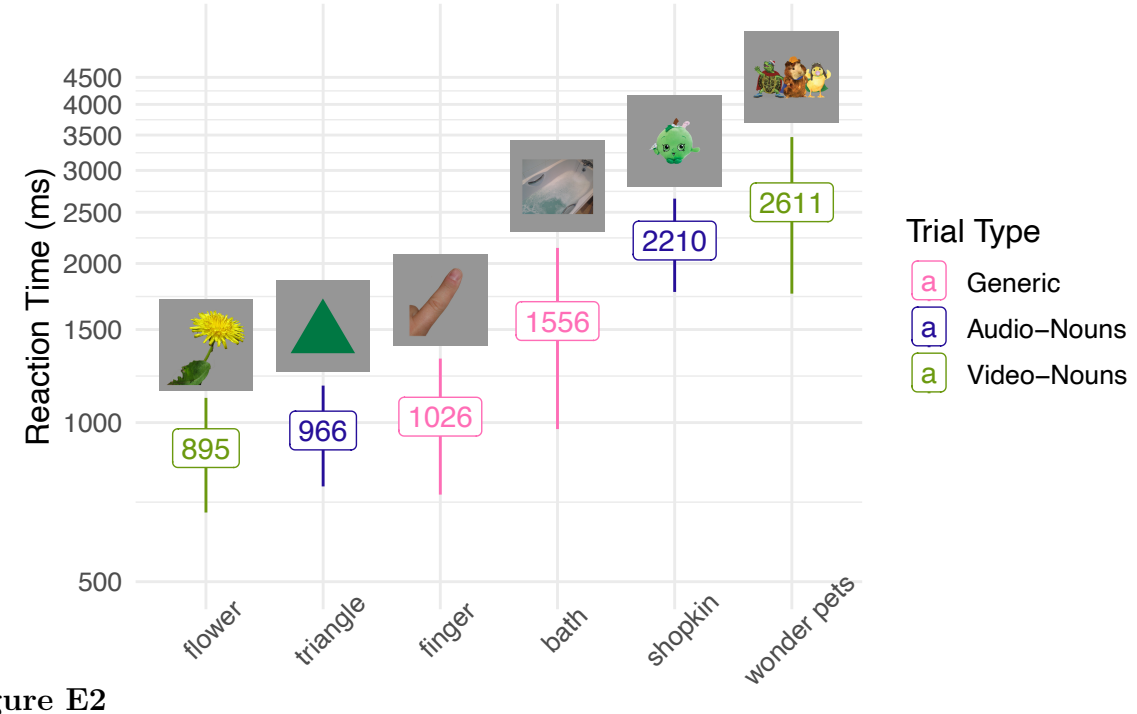

Figure E2

Examples of referents with the fastest and slowest adult reaction times by trial type (generic, audio-nouns, video-nouns). Numbers in each box indicate mean reaction times in milliseconds, and vertical lines represent standard deivations.

Overall, consistent with our prediction, it took adults significantly longer to select the correct referents in video-noun trials than in audio-noun or generic trials. This suggests that the difficulties observed in control infants on video-nouns may stem from properties of referents (e.g., low typicality) rather than overly narrow semantic representations.

Appendix F

Timecourse Visualizations

To complement our primary analyses, we provide two sets of timecourse visualizations that offer a finer-grained look at infants' gaze behavior during the eyetracking experiments. The first set (Fig. F1) displays the proportion of target looking over trial time by age and trial type. The second set (Fig. F2) presents onset-contingent switch plots, which captures infants' gaze shifts between the target and distractor images as a function of their starting position.

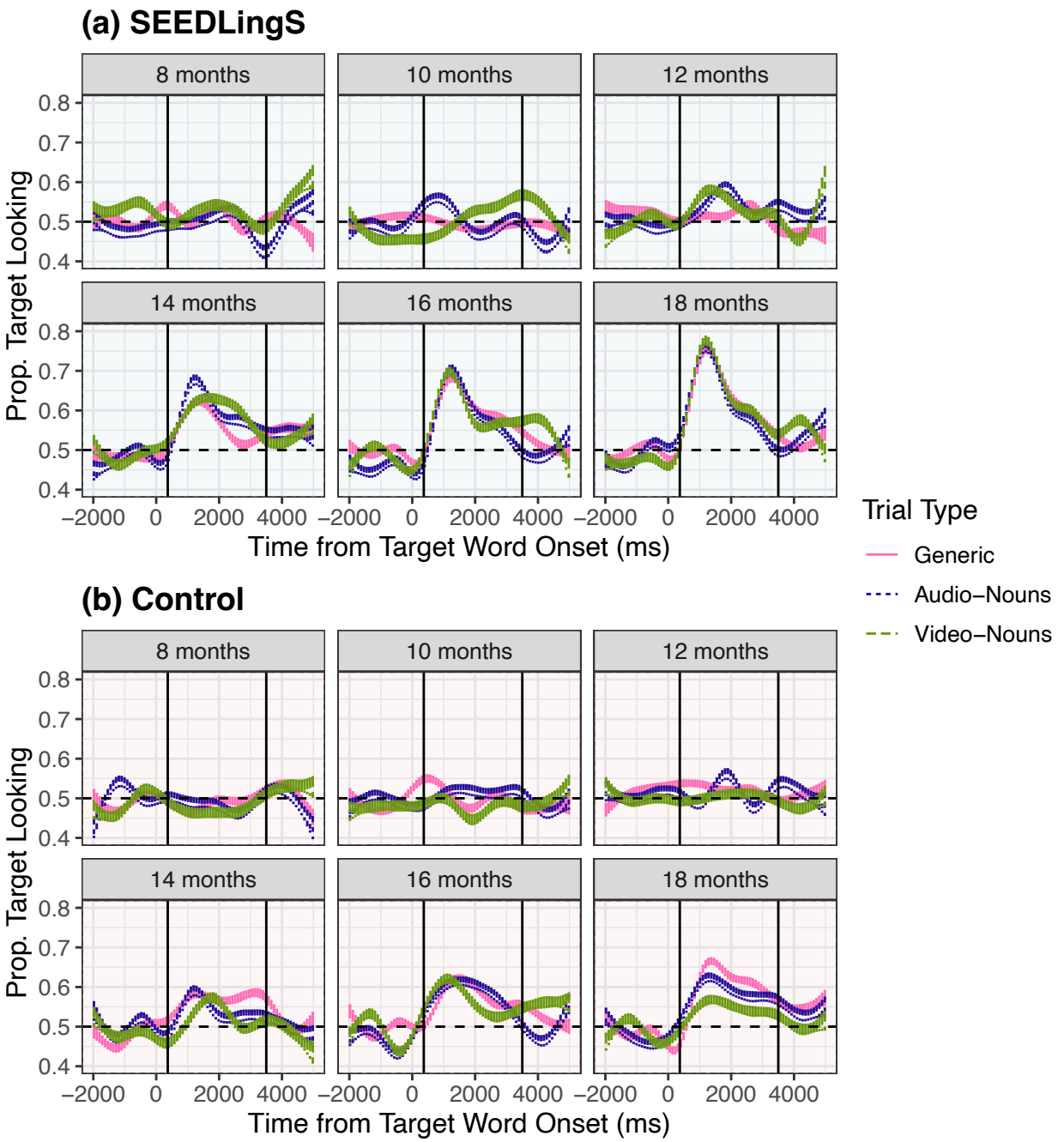

Figure F1

Time course of proportion of target looking among (a) SEEDLingS (top) and (b) control participants (bottom), split by age in months and trial type (generic, audio-nouns, video-nouns). Target word onset (e.g., "kitty" in "Do you see the kitty?") is marked as Time 0. Vertical lines indicate the analysis window (367–3,500 ms). The dashed horizontal line represents chance-level looking (0.5).

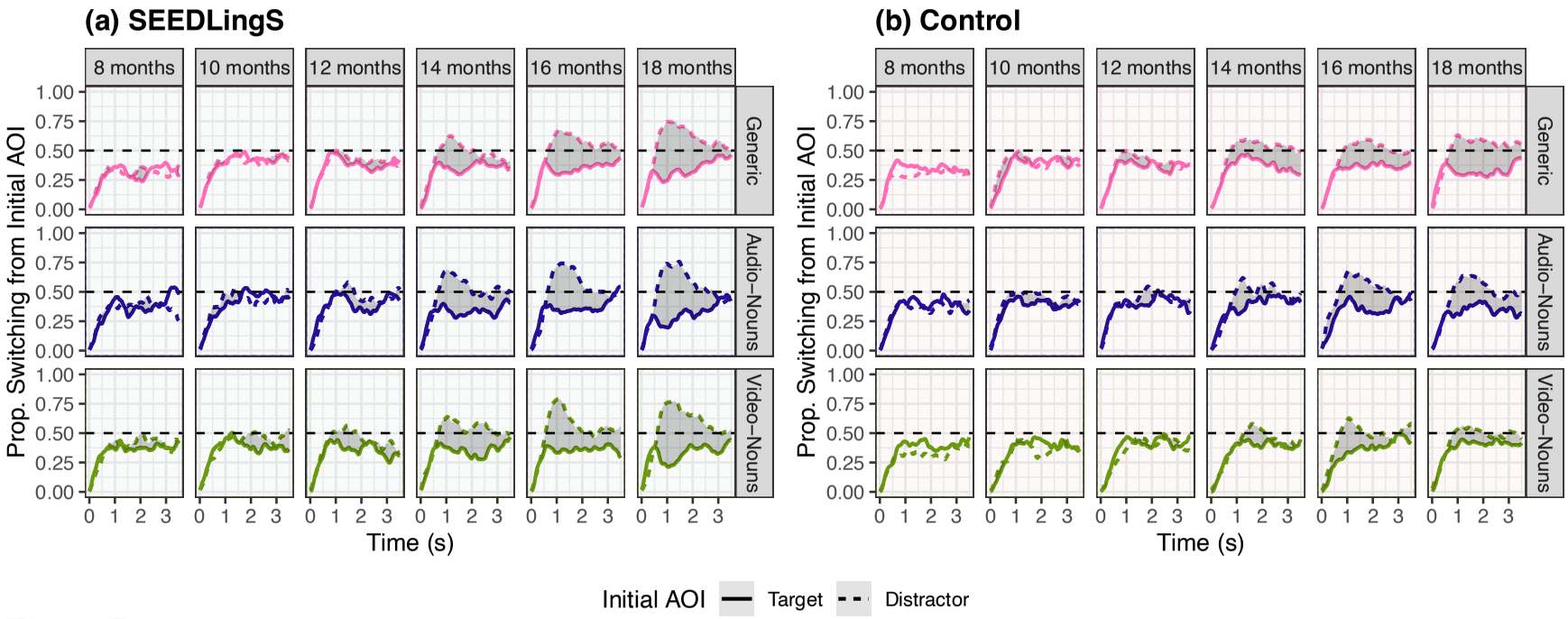

Figure F2

Onset-contingent plots comparing different types of test items (generic, audio-nouns, video-nouns) among (a) SEEDLingS (top) and (b) control participants (bottom). Each panel displays the proportion of shifts from the initial AOIs to the other over time, first averaged within participants across trials, then across participants. The solid line represents children who, by chance, began the trial looking at the correct target, and the dotted line represents children who began on the distractor. The gray shaded region highlights periods where more switches occurred from distractor to target than from target to distractor, reflecting how well children discriminated between the two test items based on the provided label. A larger gray region indicates better performance. For SEEDLingS participants, the size of the gray region appears consistent across test item types over months, suggesting that familiarity plays a small role in word recognition. In contrast, for control participants, the size of the gray region appears comparable between generic and audio-nouns, but the gap between these two and video-nouns widened over months, with video-nouns showing the poorest performance. This suggests that visual unfamiliarity may impede word recognition later in development.