**Title**

The relationship between infant learning and measured effect size in preferential looking paradigms – Experiment 1

**Description**

In any given study, individual infants show a wide range of preferential responses. Our analytic techniques generally rely on group differences (looking times for familiar vs. novel stimuli, change scores vs. zero, etc.), and require sufficient power to wash out at least some of this individual variability. Importantly, the most common methodological and analytic approaches are all based on the premise that the size of individual infants’ looking times are meaningful. This interpretation is particularly important to meta-analyses (using the size of looking time differences to estimate effect sizes) and also correlational or longitudinal designs in which the size of looking times serves as the independent variable (e.g., Hoareau, Yeung, & Nazzi, 2019, who correlated performance on an infant SL task to speech input and language production). Some analytic approaches, however, are more conservative with respect to the meaningfulness of the size of looking time differences and focus only on the direction of preference as a meaningful index of perception/learning in longitudinal analyses (i.e., counting how many infants show a preference in the same direction, regardless of the size of the preference, e.g., Newman et al., 2006).

The aim of the proposed project is to systematically address the question of whether the magnitude of effect sizes is a meaningful index of the degree to which infants have learned. For this initial study, we propose to use a straightforward manipulation of amount of exposure as the independent variable. This is one factor that is known to influence degree of learning in related tasks. For example, Thiessen, Hill, & Saffran (2005) flipped infants’ direction of preference from familiarity to novelty in a statistical language learning task by doubling exposure to the target words.

Because our primary question is focused on size of preference, not direction of preference, we propose to use a very simple task in which we can reasonably expect infants to rapidly learn, in order to increase the probability that the vast majority of infants display a novelty preference. Infants will hear a list of nonce words, in citation form, during the exposure phase. They will then be tested on their preference for familiarized nonce words versus unfamiliarized nonce words. While all infants will hear the same number of words during familiarization, there will be a between-subjects manipulation of the frequency of the target words that each infant hears. Infants will then be tested on the two target words and two novel words (counterbalanced across infants).

The initial experiment described here will serve as a first test of the experimental design and the stimuli - a preliminary "fact-finding" mission of sorts ahead of an experiment testing our questions at a much larger scale. A second (future) project will aim to test our hypotheses with respect to infant effect sizes with a far larger sample size, while manipulating frequency of exposure in a more continuous fashion.

As a first test of the current manipulation of frequency of exposure to nonce words, we will investigate two boundary conditions: a condition where the nonce words are heard relatively infrequently during the training phase (4 occurrences in 80 total words, termed the Four Occurrences condition), and a condition where the nonce words are heard relatively frequently (16 occurrences in 80 total words, the Sixteen Occurrences condition). The full design (planned for subsequent data collection efforts) will include a continuous manipulation of nonce-word frequency across several additional conditions. By first measuring ends of the spectrum, we hope to gather data to inform our decision of where to set the upper and lower boundaries of frequency of exposure across conditions (e.g., whether 4 occurrences of a nonce word is a strong enough manipulation to elicit a preferential looking effect in the current design) in a large-scale test of the relationship between frequency of exposure and effect size.

**Hypotheses:**

Question 1: Do infants recognize familiar words in each frequency condition?

Our first goal is to verify that children demonstrate a preferential looking effect based on the nonce words heard during training (familiar nonce words vs. novel nonce words). We expect to observe a preferential looking effect (either familiarity or novelty preference) for both the Four Occurrences condition and the Sixteen Occurrences condition. However, since we are hoping to use this study to inform decisions about the strength of the frequency manipulation, we do not make a specific prediction regarding which direction of preference, or whether both conditions will evince the same or different directions of looking preference. It will also be informative if infants do not show an effect in one of the two conditions. A primary motivation for the current study is to establish the extremes of exposure that will yield the same direction of preference, in order to establish whether or not there is a continuous relationship between magnitude of looking preference and degree of learning/familiarity.

Question 2: Is the measured effect size larger in the Sixteen Occurrences condition than in the Four Occurrences condition?

Our main goal in future experiments is to assess the degree to which our frequency manipulation (which is intended to manipulate the strength of an infants' representation of the nonce words heard during training) leads to differences in effect size measured in a preferential looking paradigm. This experiment will also allow us to test this question by comparing the effect size measured in the Sixteen Occurrences condition to the Four Occurrences condition. We predict that, assuming the direction of looking preference is the same across both conditions (i.e., familiarity or novelty looking preference), the effect size will be larger in the Sixteen Occurrences condition than in the Four Occurrences condition.

**Blinding**

For studies that involve human subjects, they will not know the treatment group to which they have been assigned.

**Is there any additional blinding in this study?**

Infants participating in the study will sit on their parents' laps while listening to the auditory exposure and test stimuli. In order to prevent parents from influencing infants' behavior, we will ensure that parents wear sound-dampening headphones and listen to music during the experiment in order to mask the experimental stimuli.

In the headturn preference procedure, an experimenter sitting outside of the experimental booth codes infants looking to two side monitors based on a live videofeed of infants' behavior in the experimental booth. This experimenter is blind to the auditory stimuli that the infant hears during the test phase, i.e. in particular during periods in which they are coding infants' behavior (they will be unable to hear the sound that the infants are listening to in the booth and the experimental software does not display live information about the stimulus the infant is currently hearing).

**Study design**

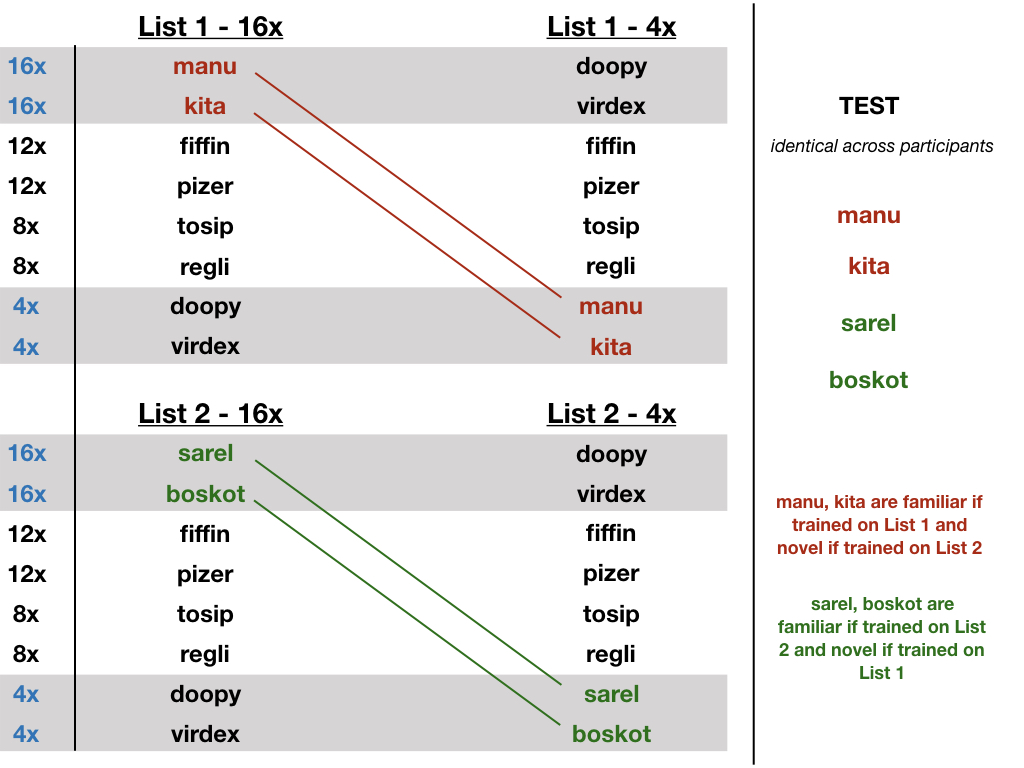
Infants' preferential looking to novel (not heard during training) vs. familiar (heard during training) nonce words is manipulated within participants. The experiment consists of two parts: a training phase and a test phase. During the training phase, infants listen to a word list consisting of 80 nonce words (see OSF materials for all stimuli). During the test phase, all participants are tested on the same four nonce words (kita, boskot, sarel, manu) in a headturn preference procedure (see below). Whether the word is familiar or novel for a given participant depends on their training "language", i.e., which words were heard during training: participants hear either kita and manu (L1) or sarel and boskot (L2). The training language is counterbalanced across participants (L1 vs. L2). For example, if participants are assigned to language L1 and hear the words kita and manu during training, then the words kita and manu are the familiar items and sarel and boskot are the novel items at test. If participants are assigned to L2 (and hear sarel and boskot), then the roles of the items at test is reversed (novel: kita, manu; familiar: sarel, boskot).

The main experimental condition is a between-subjects manipulation of the frequency with which participants hear the critical (familiar) nonce words during training. In the Four Occurrences condition, the familiar words (in L1: kita, manu; in L2: sarel, boskot) are heard four times across training. In the Sixteen Occurrences condition, the familiar words are heard 16 times across training. Participants are randomly assigned to the two frequency conditions.

Note that, in addition to the two critical items that infants will be tested on after training, the training language includes six other nonce words that occur at varying frequencies, such that the total training list includes 80 total nonce-word occurrences (see Figure). These items are included in the training list since subsequent experiments hope to test the effect of frequency in a more continuous fashion, including conditions testing infants’ recognition of items occurring at intermediate frequencies (e.g., an Eight Occurrences condition and a Twelve Occurrences condition). These conditions are currently planned for future experiments using the same paradigm.

The total length of the training language played during familiarization is 1m57s. During the first minute of the training phase, an attention-getting video will play on the central monitor (a video of clouds moving across the sky). This video is intended to provide infants with a visual stimulus to hold their visual attention on the central screen, without being so interesting as to distract infants from the auditory stimulus. After the first minute of the training phase has expired, infants will be familiarized with the visual contingencies of the headturn preference procedure while they listen to the remaining ~ 57s of the training language. On each trial, a visual attention-getting stimulus (a video of a rotating pinwheel) appears on the central monitor. Once the experimenter observes that the infant has directed their attention to the central monitor, they trigger a side event that appears randomly on the left- or the right-side monitor (balanced across trials). The same attention-getting pinwheel video appears on the left or the right monitor and the experimenter codes infants’ looking to the visual stimulus. The trial expires once the infant has looked away for more than 2 seconds (as in the headturn preference procedure used at test). Note that, unlike in the test phase, the audio presentation throughout the training phase is not infant-controlled (unlike the test phase): the training language continues to play independently of infants’ looking. The pinwheel video and contingent looking procedure is introduced during the training phase in order to familiarize infants with the visual contingency and smooth the transition to the headturn preference procedure of the test phase.

After the training phase, infants are tested using the headturn preference procedure. On each trial, a visual attention-getting stimulus (a video of a rotating pinwheel) appears on the central monitor. Once the experimenter observes that the infant has directed their attention to the central monitor, they trigger a side event that appears randomly on the left- or the right-side monitor (balanced across trials). The same attention-getting pinwheel video appears on the left or the right monitor. Once the experimenter begins registering that infants have turned their attention to the attention-getting video, infants hear the auditory recording of one of the four critical items played in a (repeated) loop (triggered by the experimenter). The auditory trial is infant-controlled: the test audio will play for as long as the infant looks towards the monitor presenting the auditory stimulus (as recorded by an experimenter observing from outside of the booth), until a maximum of 20s have elapsed or until the infant has looked away for two continuous seconds. Infants' looking preference will be tested on each critical word three times, for a total of 12 test trials split across three blocks. The order of the four critical items is randomized within each block. Trials to which infants did not look continuously for two seconds will be repeated at the end of the block.



**Existing data**

**Explanation of existing data**

**Data collection procedures**

Participants will be recruited from a volunteer database. Parents of monolingual English-learning infants between 6.0 - 9.0 months of age will be contacted via phone or email and asked whether they are interested in participating in the study. Participants will be compensated with the choice of a book, a t-shirt, or $10 in cash. Potential participants born more than 4 weeks early, who have had more than 4 ear infections in the past year, who have hearing or vision problems, or who hear a language other than English for more than 10 hours a week will be excluded, and not recruited for participation. Given our large age range and the difficulty of recruiting and scheduling infant participants, we will include infants who are up to one week younger than 6.0 months or one week older than 9.0 months. Participants will be excluded from the sample after participation has occurred if they did not contribute at least 8 test trials (out of 12), or for general fussiness/ inattentiveness during the experiment (as determined by the experimenter before analyzing the data). Recruitment and data collection will continue until a full sample (60 participants that meet the minimum criteria listed above) has been reached for each experiment.

The experiment will be presented and data will be collected using WISP, a MATLAB-based program designed for creating and running infant experimental paradigms such as the headturn preference procedure (<https://bitbucket.org/rholson1/wisp/src/default/>). Videos of each participants’ session will also be recorded, in case any questions with respect to a session need to be resolved (e.g., if the raw data show an unusual looking pattern or if there are questions about parental interference or infant fussiness).

**Sample size**

Our sample size will be 64 participants (6-9 months of age), 32 in the Four Occurrences condition and 32 in the Sixteen Occurrences condition.

**Sample size rationale**

We chose a sample size of n = 32 per condition in order to have roughly 80% power to detect a medium effect size (roughly a Cohen's d of 0.5 in a paired samples t-test) in each of our two conditions. This sample size decision also balances the fact that recruiting and collecting data from infant participants is a resource- and time-intensive process.

**Manipulated variables**

Experimental (Training) Condition - Four Occurrences condition vs. Sixteen Occurrences condition:

Participants' exposure frequency to the nonce words will be manipulated in two between-subjects conditions. In the Four Occurrences condition, the critical target words will be heard four times each across the training list. In the Sixteen Occurrences condition, the critical target words will be heard sixteen times each across the training list.

Training "Language" (L1 vs. L2) :

Participants will hear one of two sets of familiar words during the training phase (L1: kita, manu; L2: sarel, boskot). The training language will be counterbalanced within each experimental condition, i.e. training language will be crossed with experimental condition in a 2x2 between-subjects design. Participants hear the same additional six nonce words in each training language .

**Measured Variables**

Outcome measures:

We will measure looking time to familiar nonce words (heard during training) versus novel nonce words (not heard during training) during the testing phase.

Covariates:

We will also record infants' age and gender.

**Indices:**

Traditional headturn preference analysis:

In the main (traditional) analysis, we will compute the mean looking time to familiar nonce words and the mean looking time for novel nonce words for each participant, removing trials in which there was a trial-level error or in which participants did not look for the minimum required time. These values will be entered into a paired samples t-test.

**Statistical models**

Traditional headturn analysis:

1) Testing for an effect within each condition:

We will conduct two-tailed paired sample t-tests comparing participants' average looking times for familiar nonce words to their average looking times for novel nonce words within each condition. We predict that infants will show a preference for one set of words (familiar vs. novel) over the other. We do not have a strong prediction about the direction of preference (familiarity or novelty preference).

2) Testing the difference between conditions:

To test for a difference in preferential looking between the two experimental conditions, we will fit a linear model predicting children's difference in looking time (novel looking time - familiar looking time) from condition (mean-centered; Four Occurrences condition: 0.5; Sixteen Occurrences condition: 0.5). We predict that there will be an effect of condition, such that there is a larger difference in looking (assuming the direction of preference is consistent between the two conditions) in the Sixteen Occurrences condition compared to the Four Occurrences condition.

Trial-by-trial linear mixed-effects analysis:

As a complement to the traditional headturn preference procedure analyses described above, we will also conduct a parallel linear mixed-effects analysis on the trial-by-trial looking data testing the same two questions. We expect to find similar results in both analytic approaches. This analysis will help us to understand the robustness of the effect and infants' trial-by-trial variability in looking time across the test.

We will test both questions (individual preferential looking effects within each condition and the difference between conditions) within the same linear mixed-effects model. Following Csibra, Hernik, Mascaro, Tatone, and Lengyel (2016), looking times will be log-transformed. We will fit a model predicting infants' trial-by-trial log looking time from trial type (familiar vs. novel; centered), condition (Four Occurrences vs. 16 Occurrences; centered), and their interaction. We plan to include a maximal random effects structure (Barr et al., 2013), including all random effects that are appropriate given the experimental design (i.e., including random slopes for all predictors that are nested within participants and items). Specifically, we will include by-participant and by-item random intercepts, a by-participant random slope for trial type and a by-item random slopes for trial type, condition, and their interaction. We will specify the planned model using the lme4 package in R with the following command:

lmer(log\_looking\_time ~ trial\_type \* condition + (1 + trial\_type | participant) + (1 + trial\_type\*condition | item), data=data)

In cases of non-convergence, we will prune the random-effects structure in an iterative fashion, removing random slopes for items before random slopes for participant. We will obtain p values using the lmerTest package.

1) Testing for an effect within each condition:

In this analysis, effects of trial type (familiar vs. novel) indicate preferential looking to the novel (or the familiar) stimulus. To test infants' preferential looking within each condition, we will fit the model while re-coding the condition variable. To test for an effect of trial type in the Four Occurrences condition, we will re-code condition with the Four Occurrences condition coded as zero (Four Occurrences condition = 0; Sixteen Occurrences condition = 1). To test for an effect of trial type in the Sixteen Occurrences condition, we will re-code condition with the Sixteen Occurrences condition coded as zero (Four Occurrences condition = -1; Sixteen Occurrences condition = 0).

2) Testing the difference between conditions:

In the linear mixed-effects model specified above, a significant interaction between trial type (centered) and condition (centered) will indicate a significant difference in preferential looking between the two conditions.

Test number of children showing a specific direction of preference:

In addition to the tests above that treat looking time as a continuous measure of preference, we will also conduct chi-squared tests to investigate the number of children showing the dominant direction of preference. For each child, we will compute the difference in average looking time to the novel vs. the familiar words. Based on this difference score, children will be coded as showing either an overall novelty preference or an overall familiarity preference. Next, we will conduct chi-squared tests based on the number of children showing a novelty preference vs. a familiarity preference.

1) Testing for an effect within each condition:

For each condition, we will conduct a chi-squared test investigating whether the proportion of children showing the direction of preference in the dominant direction differs from chance.

2) Testing the difference between conditions:

We will conduct a chi-squared test investigating whether the number of children showing a novelty preference vs. a familiarity preference differs between the two conditions.

**Transformations**

Trial-by-trial linear mixed-effects analysis:

In the trial-by-trial linear mixed-effects analysis described above, looking times will be log-transformed, following Csibra, Hernik, Mascaro, Tatone, and Lengyel (2016).

**Inference criteria**

All hypotheses will be tested at a significance level of .05. Two-tailed tests will be used for all analyses.

**Data exclusion**

Trial-level exclusions:

During the experiment, test trials that last less than 2 seconds are repeated at the end of the phase, but are only repeated once. Any trial with looking times less than 2 seconds is excluded from data analyses. Trials may also be excluded for one of the following reasons:

1) experimenter error – the experimenter made an error while coding the infants’ looking

2) parent interference – the parent becomes unblinded or overtly directs infants’ attention to one of the two side monitors

3) fussiness – if the infant is very fussy/ crying during the trial

Participant-level exclusions:

We will remove participants who do not contribute at least 8 trials of looking time data. We also assess all comparisons by including the final sample, and then by excluding outliers. Outliers are identified as those whose looking times are above or below 2 standard deviations from the sample mean difference score. Participants will be excluded after data collection if they are judged to have been generally fussy and/ or inattentive throughout the entirety of the experiment or fussed out early during the experiment. Participants will also be excluded if an experimenter error occurs that affects the entire experiment (e.g., not being correctly blinded to infants’ auditory exposure) or if a technical error occurs (e.g., if the audio output locations do not function properly).

Participants will be excluded during recruitment if they have had more than 4 ear infections in the past year, were born more than 4 weeks early, or hear a language other than English for more than 10 hours a week.

**Missing data**

**Exploratory analysis:**

Since our analyses are intended to inform a larger-scale project, we will conduct a number of exploratory analyses that will help provide information for subsequent experiments:

a) Within the traditional headturn analyses, we will also fit similar models using a looking time difference DV that takes into account participants' relative preference for novel stimuli relative to familiar stimuli, rather than just the raw difference in overall looking time: (novel looking time - familiar looking time)/ (novel looking time + familiar looking time). We expect models using the difference between familiar and novel looking time as the DV and models using the difference between familiar and novel looking time relative to total looking time to yield comparable results.

b) Given our wide age range, we will explore the relationship between age and preferential looking, by testing whether age predicts looking time differences and whether it interacts with condition in the models described above.

c) We will also consider additional analyses comparing the effect sizes between the two conditions. One approach we will explore is computing the effect size within each condition (Cohen’s d\_z), along with a confidence interval, and then comparing the two effect sizes. Another approach we will consider is computing an effect size for each infant across the 12 test trials (mean difference in looking time / variance of the difference) and computing a t-test comparing the infant-wise effect sizes between the two conditions. In general, we expect these exploratory analyses to yield similar results to the main analytic approach described in the main analysis section. We will explore these alternative analytic approaches in order to gather information and tune our intuitions about the best approach to use in future experiments.

d) We will investigate whether the size of the effect interacts with block number (Blocks 1, 2, and 3) in either the Four Occurrences condition or in the Sixteen Occurrences condition.

e) We will also explore the effect of choosing more liberal or more stringent data exclusion criteria, specifically testing the same questions while retaining infants who contribute fewer trials (e.g., who contribute only 4 or 6 test trials) or only retaining infants who contribute more trials (e.g., who contribute at least 10 test trials).

f) We will also explore the use of equivalence tests to understand the degree to which a null effect in one of our two conditions or a null effect between our two conditions constrains the range of possible underlying effect sizes (i.e. how likely the effect size lies within a particular upper and lower bound).