Development of word recognition in preschoolers

Tristan Mahr

2018-07-24

[*draft*] Vocabulary size in preschool is a robust predictor of later language development, and early language skills predict early literacy skills at school entry. By studying the mechanisms that shape word learning, we can understand how individual differences in language ability arise and identify strategies for closing language gaps between children. Word recognition—the process of mapping incoming speech sounds to known or novel words—has been shown in toddlers to predict later language outcomes. We do not know how this ability develops over time, and we do not know when word recognition is most predictive of future outcomes. This dissertation reports the results for two word-recognition experiments administered during each year of a 3-year longitudinal study with 160 preschoolers. Ages were 2.5–3 years in year 1 and 4.5–5 years in year 3.

In the first experiment, four images of familiar nouns were presented onscreen followed by a prompt to view one of the images (e.g., *find the bell!*). Images included the target word (e.g., *bell*), a semantically related word (*drum*), a phonologically similar word (*bee*), and an unrelated word (*swing*). Early differences in word recognition were longitudinally stable so that children who were faster and more accurate at age 3 were relatively fast and accurate at age 5. Moreover, word recognition efficiency at age 3 was a stronger predictor of age-5 vocabulary size than concurrent (age-5) word recognition efficiency. These results confirm that word recognition behavior provides an important early predictor of vocabulary growth. Analysis of children’s looks to the competitors showed that children become more sensitive to the phonological and semantic foils, compared to the unrelated word, as they grew older. Children become better at recognizing familiar words by developing connections among words.

The second experiment used a mispronunciation study in which a child saw a familiar object and unfamiliar object and heard either a real word (*shoes*), a one-feature mispronunciation of the word (*suze*), or a nonword (*geeve*). Contrary to pre-analysis hypotheses, children recognized the real words and fast-selected novel-objects referents for the nonwords equally well, and in fact, performed better in the nonword condition. Children also became to associate the familiar object with the mispronunciations as they grew older. At age 5, children’s retention of the unfamiliar objects paired with the nonwords and mispronunciations was tested offline, and children showed better retention for the nonwords than mispronunciations.

# Dedication

For Penny

# Specific aims

Individual differences in language ability are apparent as soon as children start talking, but it is difficult to identify children at risk for language delay or disorder. Recent work suggests word recognition efficiency—that is, how well children map incoming speech to words—may help identify early differences in children’s language trajectories. Children learn spoken language by listening to caregivers, so children who are faster at recognizing words have an advantage for word learning. This view is borne out by some studies suggesting that children who are faster at processing words show greater vocabulary gains months later (e.g., Weisleder & Fernald, [2013](#ref-Weisleder2013)).

We do not know, however, how word recognition itself develops over time within a child. This is an important open question because word recognition may provide a key mechanism for understanding how individual differences emerge in word learning and persist into early language development. Without a developmental account of word recognition, we lack the context for understanding individual differences in lexical processing. Thus, even the big-picture questions are unclear: Do early differences persist over time so that faster processors remain relatively fast later in childhood? Or, is such a question ill-posed because the magnitude of the differences among children shrink with age? I plan to address this gap in knowledge by analyzing three years of word recognition data collected in a recently completed longitudinal study of 180 children.

In particular, I will examine the development of *familiar word recognition*, *lexical competition,* and *fast referent selection* (the ability to map novel words to novel objects in the moment). Through these analyses, I will develop a fine-grained description of how the dynamics of word recognition change year over year, and I will study how differences in word recognition performance relate to other child-level measures (such as vocabulary and speech perception).

## Specific Aim 1 (Familiar Word Recognition and Lexical Competition)

**To characterize the development of familiar word recognition and lexical competition, I will analyze data from a visual world paradigm experiment, conducted at age 3, age 4, and age 5.**

In these eyetracking experiments, children were presented with four images of familiar objects and heard a prompt to view one of the images. The four images included a target word (e.g., *bell*), a semantically related word (*drum*), a phonologically similar word (*bee*), and an unrelated word (*swing*). I will use a series of growth curve analyses to describe how children’s familiar word recognition develops year over year. Of interest is how individual differences at age 3 persist into age 5. I will also examine the children’s looks to the distractors to study the developmental course of lexical competition from similar sounding and similar meaning words. Changes in sensitivity to competing words can reveal how lexical competition emerges as a byproduct of learning new words.

## Specific Aim 2 (Referent Selection and Mispronunciations)

**To characterize how fast referent selection develops longitudinally, I will analyze data from a looking-while-listening mispronunciation experiment, conducted at age 3, age 4, and age 5.**

In these eyetracking experiments (Law & Edwards, [2015](#ref-MPPaper); based on White & Morgan, [2008](#ref-WhiteMorgan2008)), children saw an image of a familiar object and an unfamiliar object, and they heard either a correct production of the familiar object (e.g., *soup*), a one-feature mispronunciation of the familiar object (*shoop*), or a novel word unrelated to either image (*cheem*). The correct productions test familiar word recognition and the nonwords test fast referent selection. The mispronunciations test the child’s phonological categories by showing whether the child permits, rejects, or equivocates about mispronunciations.

I will use growth curve analyses to study how children’s responses to the three word types change over time. I will examine familiar word recognition and fast referent selection to determine which feature of lexical processing better predicts vocabulary growth. I plan to examine dissociations or asymmetries in these forms of processing within children as a way to empirically assess the claim that “novel word processing (referent selection) is not distinct from familiar word recognition” (McMurray, Horst, & Samuelson, [2012](#ref-McMurray2012)). Finally, I will examine how individual differences in vocabulary and speech perception predict responses to mispronunciations and novel words.

## Summary

This project investigates how word recognition develops during the preschool years. There has been no published research studying word recognition longitudinally after age two. Findings will show how individual differences in lexical processing change over time and can reveal how low-level mechanisms underlying word recognition mature longitudinally in children. These findings will have translational value by studying processing abilities that subserve word learning and by assessing the predictive relationships between early word recognition ability and later language outcomes.

# Research hypotheses

In this section, I outline the main hypotheses I plan to study for each specific aim. This section is intended to preregister the main analyses for this project.

## Specific Aim 1 (Familiar Word Recognition and Lexical Competition)

* Children’s accuracy and efficiency of recognizing words will improve each year.
* There are stable individual differences in lexical processing of familiar words such that children who are relatively fast at age 3 remain relatively fast at age 4 and age 5.
* However, the magnitude of these individual differences diminishes over time, as children converge on a mature level of performance for this paradigm.
* Consequently, individual differences in word recognition at age 3, for example, will be more discriminating and predictive of age 5 language outcomes than differences at age 4 or age 5.
* Children will become more sensitive to lexical competitors as they age, based on the hypothesis that children discover similarities among words as a consequence of learning more and more words.
* Children will differ in their sensitivity to lexical competitors, and these individual differences will correlate with other child-level measures.

## Specific Aim 2 (Referent Selection and Mispronunciations)

* Children’s accuracy and efficiency of recognizing real words and fast-associating nonwords will improve each year.
* Performance in real word recognition and fast association of nonwords will be highly correlated, based on the hypothesis that the same process (referent selection) operates in both situations.
* Under the alternative hypothesis, real word recognition and fast referent selection reflect different skills with different developmental trajectories. Thus, if there is any dissociation between recognition of real words and nonwords, it will be observed in younger children.
* Although these two measures will be correlated, I predict performance in the nonword condition will be a better predict of future vocabulary growth than performance in the real word condition. This hypothesis is based on the idea that fast referent selection is a more relevant skill for learning new words than recognition of known words.
* For the mispronunciations, I predict children with larger vocabularies (that is, older children) will be more likely to tolerate a mispronunciation as a production of familiar word compared to children with smaller vocabularies.
* Mispronunciations that feature later-mastered sounds (e.g., *rice*/*wice*) will be more likely to be associated to novel objects than earlier-mastered sounds (*duck*/*guck*).

# (PART\*) Aim 1: Familiar Word Recognition and Lexical Competition

# Familiar word recognition

## Lexical processing dynamics

Mature listeners recognize spoken words by continuously evaluating incoming speech for possible word matches. The first part of a word activates multiple candidate words in parallel. These candidates compete as more of the speech signal enters the system, and the best-fitting word is the favored interpretation. For example, the onset “bee” might activate phonologically compatible candidates like *bee*, *beam*, *beetle*, *beak*, *beaker*, *beginning*, and so on, but an additional “m” would narrow the candidates to just *beam*. Semantic relationships also influence lexical processing, and cascading phonological-semantic effects—for instance, where *castle* activates the phonologically similar *candy* which in turn activates the semantically related *sweet*—have been demonstrated (Marslen-Wilson & Zwitserlood, [1989](#ref-Marslen-Wilson1989)). Both low-level phonetic cues and high-level grammatical, semantic and pragmatic information can influence this process, but this *continuous processing of multiple competing candidates* is the essential dynamic underlying word recognition in adults (Magnuson, Mirman, & Myers, [2013](#ref-Magnuson2013)).

What about young children who know considerably fewer words? Eyetracking studies with toddlers have suggested a developmental continuity between toddlers and adult listeners. Children recognize words incrementally (Swingley, Pinto, & Fernald, [1999](#ref-Swingley1999)), match truncated words to their intended referents (Fernald, Swingley, & Pinto, [2001](#ref-Fernald2001)), and use information from neighboring words in a sentence to facilitate word recognition. This information can be high-level grammatical or semantic cues. Lew-Williams and Fernald ([2007](#ref-Lew-Williams2007)) found that Spanish-acquiring preschoolers can use grammatical gender on determiners (*el* or *la*) to anticipate the word named in a two-object word recognition task. Borovsky, Elman, and Fernald ([2012](#ref-Borovsky2012)) showed that children can use semantic information from an agent and a verb (e.g., *the dog chased*) to anticipate a plausible noun (*the cat*). The information can also be low-level phonetic variation: We found that toddlers look earlier to a named image when the coarticulatory formant cues on word *the* predicted the noun of the sentence, compared to tokens with neutral coarticulation (Mahr, McMillan, Saffran, Ellis Weismer, & Edwards, [2015](#ref-Mahr2015)).

There is some evidence for lexical competition where children are sensitive to phonological and semantic similarities among words. Ellis Weismer, Haebig, Edwards, Saffran, and Venker ([2016](#ref-EllisWeismer2016)) showed that toddlers (14–29 months old) look less reliably to a named image when the onscreen competitor was a semantically related word or perceptually similar image. Huang and Snedeker ([2011](#ref-Huang2011)) presented evidence of cascading semantic-phonological activation in five-year-olds such that for a target word like *log*, the children looked more to an indirect phonological competitor like *key* (competing through its activation of *lock*) than they looked to an unrelated image like *carrot*.

Priming studies also reveal that children are sensitive to phonological similarities among words. Mani and Plunkett ([2010](#ref-Mani2010)) demonstrated cross-modal phonological priming effects in 18-month-olds. In this study, a picture of prime word (e.g., cat or teeth) was presented in silence; then two images (cup and shoe) were presented, one of which was named (*cup*). Children on average looked more to the target word (*cup*) when it was primed by an image of a phonological neighbor (*cat*), and the children performed at chance when the prime was not related to the named word. Mani, Durrant, and Floccia ([2012](#ref-Mani2012)) found a similar result for cascading phonological-semantic priming with 24-month-olds: Children looked more to a target (e.g., *shoe*) compared to a distractor (*door*) when primed by an image (*clock*, assumed to activate *sock* which primed *shoe*).[[1]](#footnote-32)

Altvater-Mackensen and Mani ([2013](#ref-AltvaterMackensen2013)) demonstrated phonological-semantic priming even when the prime is a mispronunciations. German-learning two-year-olds heard a prime word, and 200 ms later two images appear on screen (a cow, Kuh and a fork, Gabel), one of which was labeled (“Kuh”).The prime word was a semantically related word (“Schaf”, sheep), an onset-mispronunciation of the related word (“Faf” or “Taf”), or an unrelated prime (“Buch”, book). Children looked to the target about equally well in the two prime conditions: approximately .62 proportion looks to target with a normal prime versus approximately .60 for a mispronounced prime. In contrast, they looked less to the target in the unrelated prime condition (approximately .55). Thus, the children in this study showed cascading activation where the mispronunciation activates the mispronounced word which in term activates a semantically related word.

Chow, Aimola Davies, and Plunkett ([2017](#ref-Chow2017)) performed a very similar study to the one I present for Aim 1. applied a very similar Visual World Paradigm English-learning 24- and 30-month-olds saw a 2 x 2 grid of images which included a phonological (cohort) competitor and a semantic competitor, and they heard a prompt to view one of the images (Look at the bee). On filler trials, the target word and an unrelated image onscreen alongside the competitors. On test trials, the target word (bee) was not pictured onscreen, so the display had two unrelated images (sandwich, dress), a phonological competitor (bus), and a semantic competitor (cat). They found an early temporary advantage (a growth curve bump) in looks to the phonological competitor, followed by a late, more stable advantage for the semantic competitor. They found that increased receptive vocabulary predicted more looks to the phonological competitor and fewer looks to the semantic competitor. (They interpret the semantic effect as a consequence of the early phonological effect causing children to fixate on the semantic word later.) Their results support a kind of cascading activation: “Toddlers, like adults, extract phonological information from the spoken word faster than they extract semantic information.”

The above studies involved young children of different ages tested under different procedures, sometimes in different dialects and languages. Averaging these results together, so to speak, the studies suggest that early word recognition demonstrates some hallmarks of adult behavior: Continuous processing of words, integration of information from different levels of representation, and the influence of similar, unspoken words on recognition of a word. Nevertheless, we only have a fragmented view of how familiar word recognition develops within children.

One open question is how lexical competition develops in young listeners. For example, how and when do phonological or semantically similar words exert their influence on word recognition? As a guiding hypothesis, we can think of word learning as a gradual process where familiarity with a word moves from shallow receptive knowledge to deeper expressive knowledge. In adult listeners, words compete and they inhibit one another, so that a word is truly “learned” (integrated into the lexicon) when it can influence the processing of other words (a line of reasoning reviewed by Kapnoula, Packard, Gupta, & McMurray, [2015](#ref-Kapnoula2015)). Increasing sensitivity to similar sounding or similar meaning words over time would reveal that children more thoroughly learn familiar words with age.

## Individual differences in word recognition

We have a rough understanding of the development of word recognition, and these gaps in knowledge matter because young children differ in their word recognition abilities. These differences are usually measured using *accuracy* (a probability of recognizing to a word) or *efficiency* (a reaction time or some measure of how quickly accuracy changes over time). These differences are consequential too, as word recognition differences correlate with other language measures concurrently and prospectively.

Many studies highlight the predictive power of word recognition ability. Marchman and Fernald ([2008](#ref-MarchmanFernald2008)) found that vocabulary size and lexical processing efficiency at age 2 jointly predicted working memory scores and expressive language scores at age 8. Fernald and Marchman ([2012](#ref-Fernald2012)) found that late talkers who looked more quickly to a named word at 18 months showed larger gains in vocabulary by 30 months compared to late-talkers who looked more slowly at 18 months. Weisleder and Fernald ([2013](#ref-Weisleder2013)) found that lexical processing and language input at 19 months predicted vocabulary size at 25 months and that lexical processing mediated the effect of language input—the basic idea being that rich language input builds up word recognition ability which in turn supports word learning. Lany ([2017](#ref-Lany2017)) found a direct link between lexical processing and word learning: 18-month-olds and 30-month-olds who were faster at recognizing familiar words were also more accurate at recognizing novel words in a word-learning task. Thus, children who are better at recognizing words learn more words over time and perform better at word-learning tasks.

Word recognition performance predicts future language outcomes, so we conclude that individual differences in word recognition are important. But we do not know how word recognition develops within children, so we have no context for evaluating these individual differences. Are these differences in lexical processing persistent over development? Is word recognition a skill where most children catch up and converge on a mature range of performance by a certain age?

## The current study

In the previous two sections, I outlined two gaps in knowledge. The first is that we do not have a clear understanding of how the mechanisms underlying word recognition change in early childhood. We know that children show plenty of adult-like features of word recognition, but each of these findings is an isolated fact. What we need is a coherent set of facts that show how specific features of word recognition change with age. The second gap is that although we know that individual differences in word recognition are predictive of later outcomes, we do not have a developmental picture of these individual differences.

In this study, I tackle these two lines of research: The development of lexical competition effects and individual differences in familiar word recognition. I report the results of a longitudinal study of word recognition in preschoolers at age 3, age 4, and age 5. The study is described in detail in [Chapter 4](#aim1-method). Briefly stated, this experiment tested word recognition by presenting prompts like “find the horse” and recording children’s looks to an array of four images. The array of images included the target, a phonological competitor, a semantic competitor, and an unrelated image. In [Chapter 5](#fam-rec), I analyze the development patterns of familiar word recognition (looks to the target) and how individual differences change over time. In [Chapter 6](#lex-competitors), I study how the phonological and semantic competitors influence word recognition and how the influence of the competitors changes over time. Finally, in [Chapter 7](#aim1-discussion), I link these two lines of research together and describe both sets of results in terms of lexical processing dynamics. [Chapter @ref(#aim1-h-check)](#aim1-h-check) reviews the results of my pre-analysis research hypotheses.

# Method

## Participants

The data were collected as part of a three-year longitudinal study.[[2]](#footnote-37) For convenience, I refer to the three years as Age 3, Age 4, and Age 5, although the participants on average were three months younger than those nominal ages. In particular, the participants were 28–39 months-old at Age 3, 39–52 at Age 4, and 51–65 at Age 5. Approximately, 180 children participated at Age 3, 170 at Age 4, and 160 at Age 5. Of these children, approximately 20 were identified by their parents as late talkers. Prospective families were interviewed over telephone before participating in the study. Children were not scheduled for testing if a parent reported language problems, vision problems, developmental delays, or an individualized education program for the child. Recruitment and data collection occurred at two Learning to Talk lab sites—one at the University of Wisconsin–Madison and the other at the University of Minnesota.

Table 1 summarizes the cohort of children in each year of testing. The numbers and summary statistics here are general, describing children who participated at each year, but whose data may have been excluded from the analyses. Some potential reasons for exclusion include: excessive missing data during eyetracking, experiment or technology error, developmental concerns not identified until later in study, or a failed hearing screening. Final sample sizes will depend on the measures needed for an analysis and the results from data screening checks. I disclose all data exclusions following guidelines by the Center for Open Science (Nosek et al., [2014](#ref-OSF_Statement)).

Table 1 Participant characteristics. Education levels: *Low*: less than high school, or high school; *Middle*: trade school, technical or associates degree, some college, or college degree; and *High*: graduate degree.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Year 1 (Age 3) | Year 2 (Age 4) | Year 3 (Age 5) |
| N | 184 | 175 | 160 |
| Boys, Girls | 94, 90 | 89, 86 | 82, 78 |
| Maternal education: Low, Middle, High | 15, 98, 71 | 12, 92, 71 | 6, 90, 64 |
| Dialect: MAE, AAE | 171, 13 | 163, 12 | 153, 7 |
| Parent-identified late talkers | 20 | 19 | 16 |
| Age (months): Mean (SD) [Range] | 33 (3) [28–39] | 45 (4) [39–52] | 57 (4) [51–66] |
| EVT-2 standard score: Mean (SD) | 115 (18) | 118 (16) | 118 (14) |
| PPVT-4 standard score: Mean (SD) | 113 (17) | 120 (16) | — |
| GFTA-2 standard score: Mean (SD) | 92 (13) | — | 91 (13) |

## Visual World Paradigm

This experiment used a version of the Visual World Paradigm for word recognition experiments (Law, Mahr, Schneeberg, & Edwards, [2016](#ref-RWLPaper)). In eyetracking studies with toddlers, two familiar images are usually presented: a target and a distractor. This experiment is a four-image eyetracking task that was designed to provide a more demanding word recognition task for preschoolers. In this procedure, four familiar images are presented onscreen followed by a prompt to view one of the images (e.g., *find the bell!*). The four images include the target word (e.g., *bell*), a semantically related word (*drum*), a phonologically similar word (*bee*), and an unrelated word (*swing*). Figure 1 shows an example of a trial’s items. This procedure measures a child’s real-time comprehension of words by capturing how the child’s gaze location changes over time in response to speech.

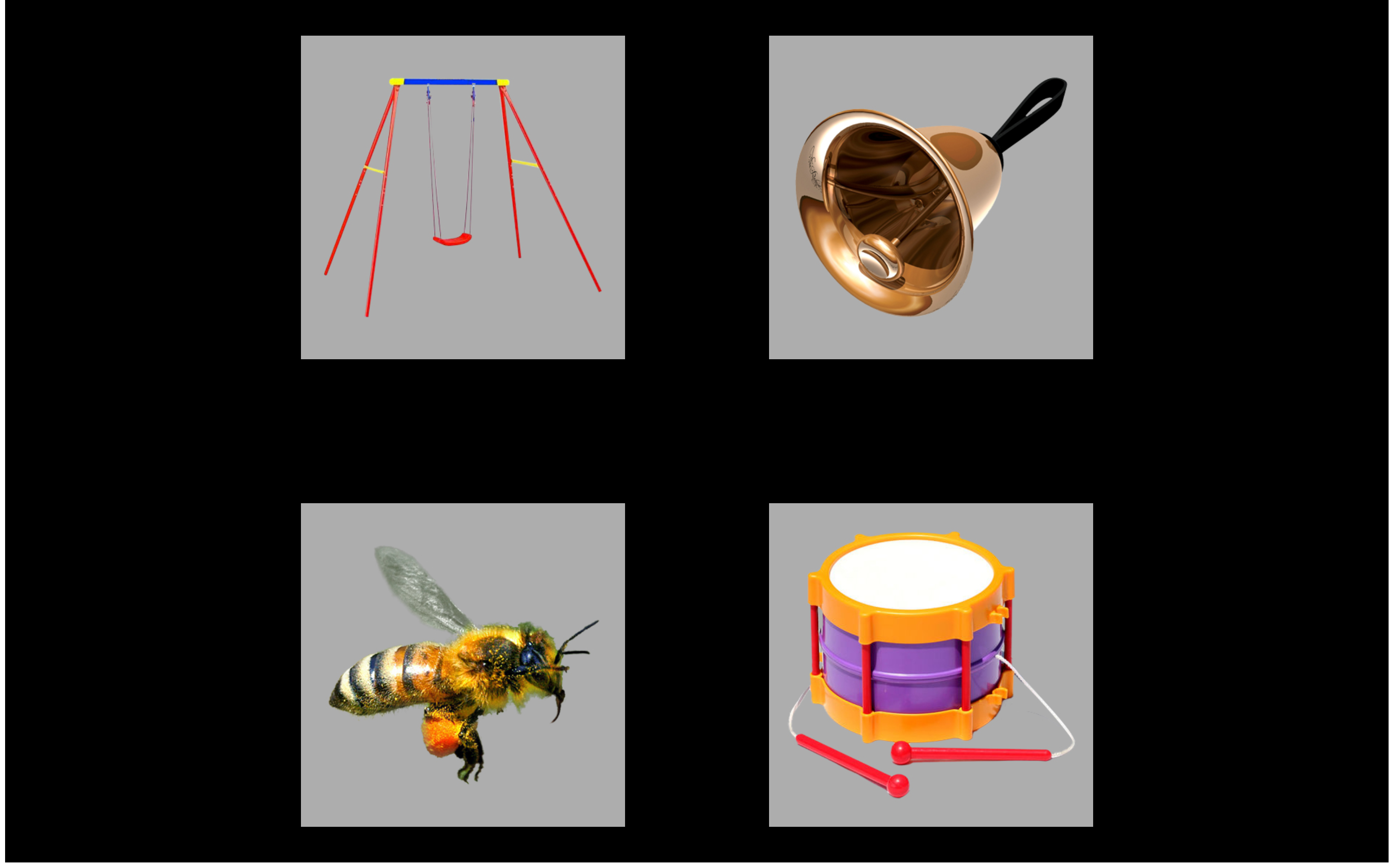


Figure 1 Example display for the target *bell* with the semantic foil *drum*, the phonological foil *bee*, and the unrelated *swing*.

## Experiment administration

Children participating in the study were tested over two lab visits (on different dates). The first portion of each visit involved “watching movies”—that is, performing two blocks of eyetracking experiments. A play break or hearing screening occurred between the two eyetracking blocks, depending on the visit.

Each eyetracking experiment was administered as a block of trials (24 for this experiment and 38 for a two-image task—see Chapter X). Children received two different blocks of each experiment. The blocks for an experiment differed in trial ordering and other features. Experiment order and block selection were counterbalanced over children and visits. For example, a child might have received Exp. 1 Block A and Exp. 2 Block B on Visit 1 and next received Exp. 2 Block A and Exp. 1 Block B on Visit 2. The purpose of this presentation was to control possible ordering effects where a particular experiment or block benefited from consistently occurring first or second.

Experiments were administered using E-Prime 2.0 and a Tobii T60XL eyetracker which recorded gaze location at a rate of 60 Hz. The experiments were conducted by two examiners, one “behind the scenes” who controlled the computer running the experiment and another “onstage” who guided the child through the experiment. At the beginning of each block, the child was positioned so the child’s eyes were approximately 60 cm from the screen. The examiners calibrated the eyetracker to the child’s eyes using a five-point calibration procedure (center of screen and centers of four screen quadrants). The examiners repeated this calibration procedure if one of the five calibration points for one of the eyes did not calibrate successfully. During the experiment, the behind-the-scenes examiner monitored the child’s distance from the screen and whether the eyetracker was capturing the child’s gaze. The onstage examiner coached the child to stay fixated on the screen and repositioned the child as needed to ensure the child’s eyes were being tracked. Every six or seven trials in a block of an experiment, the experiment briefly paused with a reinforcing animation or activity. During these breaks, the onstage examiner could reposition the child if necessary before resuming the experiment.

We used a gaze-contingent stimulus presentation. First, the images appeared in silence on screen for 2 s as a familiarization period. The experiment then checked whether the child’s gaze was being recorded. If the experiment could continuously track the child’s gaze for 300 ms, the child’s gaze was verified and the trial continued. If the experiment could not verify the gaze after 10 s, the trial continued. This procedure guaranteed that for most trials, the child was looking to the display before presenting the carrier phrase and that the experiment was ready to record the child’s response to the carrier. During year 1 (age 3) and year 2 (age 4), an attention-getter (e.g., *check it out*!) played 1 s following the end of the target noun. These reinforcers were dropped in year 3 (age 5) to streamline the experiment for older listeners.

## Stimuli

The four images on each trial consisted of a target noun, a phonological foil, a semantic foil, and an unrelated word. The phonological competitors shared a syllable onset (e.g., *flag*–*fly*, *bell*–*bee*), shared an initial consonant (*bread*–*bear*, *swing*–*spoon*), had a similar phonetically similar consonant onset (*kite*–*gift*), or shared a syllable rime (*van*–*pan*). The semantic competitors included words from the same category (e.g., *shirt*–*dress*, *horse*–*bear*), words that were perceptually similar (*sword*–*pen*, *flag*–*kite*), and words with less obvious relationships (*van*–*horse*, *swan*-*bee*). A complete list of the items used in the experiment in [Appendix 16](#vw-experiment-items).

The stimuli were recorded in both Mainstream American English (MAE) and African American English (AAE), so that the experiment could accommodate the child’s home dialect. Prior to the lab visit, we made a preliminary guess about the child’s home dialect, based on the recruitment channel, address, among other factors. If we expected the dialect to be AAE, then the lab visit was led by an examiner who natively spoke AAE and could fluently dialect-shift between AAE and MAE. At the beginning of the lab visit, the examiner listened to the interactions between the child and caregiver in order to confirm the child’s home dialect. Prompts to view the target image of a trial (e.g., *find the girl*) used the carrier phrases “find the” and “see the”. These carriers were recorded in the frame “find/see the egg” and cross-spliced with the target nouns to minimize coarticulatory cues on the determiner “the”. The stimuli were re-recorded after the first year of the study with the same speakers so that the average duration of the two dialect versions were more similar.

The images used in the experiment consisted of color photographs on gray backgrounds. These images were piloted with 30 children from two preschool classrooms to ensure that children consistently used the same label for familiar objects. The two preschool classrooms differed in their students’ SES demographics: One classroom (13 piloting students) was part of a university research center which predominantly serves higher-SES families, and the other classroom (17 piloting students) was part of Head Start center which predominantly serves lower-SES families. The images were tested by presenting four images (a target, a phonological foil, a semantic foil and an unrelated word) and having the student point to the named image. The pictures had to be recognized by at least 80% of students in each classroom.

## Data screening

To process the eyetracking data, I first mapped gaze *x*-*y* coordinates onto the onscreen images. I next performed *deblinking*. I interpolated short runs of missing gaze data (up to 150 ms) if the same image was fixated before and after the missing data run. Put differently, I classified a window of missing data as a blink if the window was brief and the gaze remained on the same image before and after the blink. I interpolated missing data from blinks using the fixated image.

After mapping the gaze coordinates onto the onscreen images, I performed data screening. I considered the time window from 0 to 2000 ms after target noun onset. I identified a trial as *unreliable* if at least 50% of the looks were missing during the time window. I excluded an entire block of trials if it had fewer than 12 reliable trials.

Table 2 shows the numbers of participants and trials at each year before and after data screening. There were more children in the second year than the first due to a timing error in the initial version of this experiment, leading to the exclusion of 27 participants from the first year.

Table 2 Eyetracking data before and after data screening. For convenience, the number of exclusions is included as Raw - Screened.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Study | N Children | N Blocks | N Trials |
| Raw | Age 3 | 178 | 332 | 7967 |
|  | Age 4 | 180 | 347 | 8327 |
|  | Age 5 | 163 | 322 | 7724 |
| Screened | Age 3 | 163 | 291 | 5951 |
|  | Age 4 | 165 | 305 | 6421 |
|  | Age 5 | 156 | 295 | 6483 |
| Raw − Screened | Age 3 | 15 | 41 | 2016 |
|  | Age 4 | 15 | 42 | 1906 |
|  | Age 5 | 7 | 27 | 1241 |

## Model preparation

To prepare the data for modeling, I downsampled the data into 50-ms (3-frame) bins, reducing the eyetracker’s effective sampling rate to 20 Hz. Eye movements have durations on the order of 100 or 200 ms, so capturing data every 16.67 ms oversamples eye movements and can introduce high-frequency noise into the signal. Binning together data from neighboring frames can smooth out this noise. I modeled the looks from 250 to 1500 ms. Lastly, I aggregated looks by child, study and time, and created orthogonal polynomials to use as time features for the model.

Figure 2 depicts the dataset following the data screening and model preparation steps. The lines start around .25 which is chance performance on four-alternative forced choice task. The lines rise as the word unfolds, and they peak and plateau around 1400 ms.

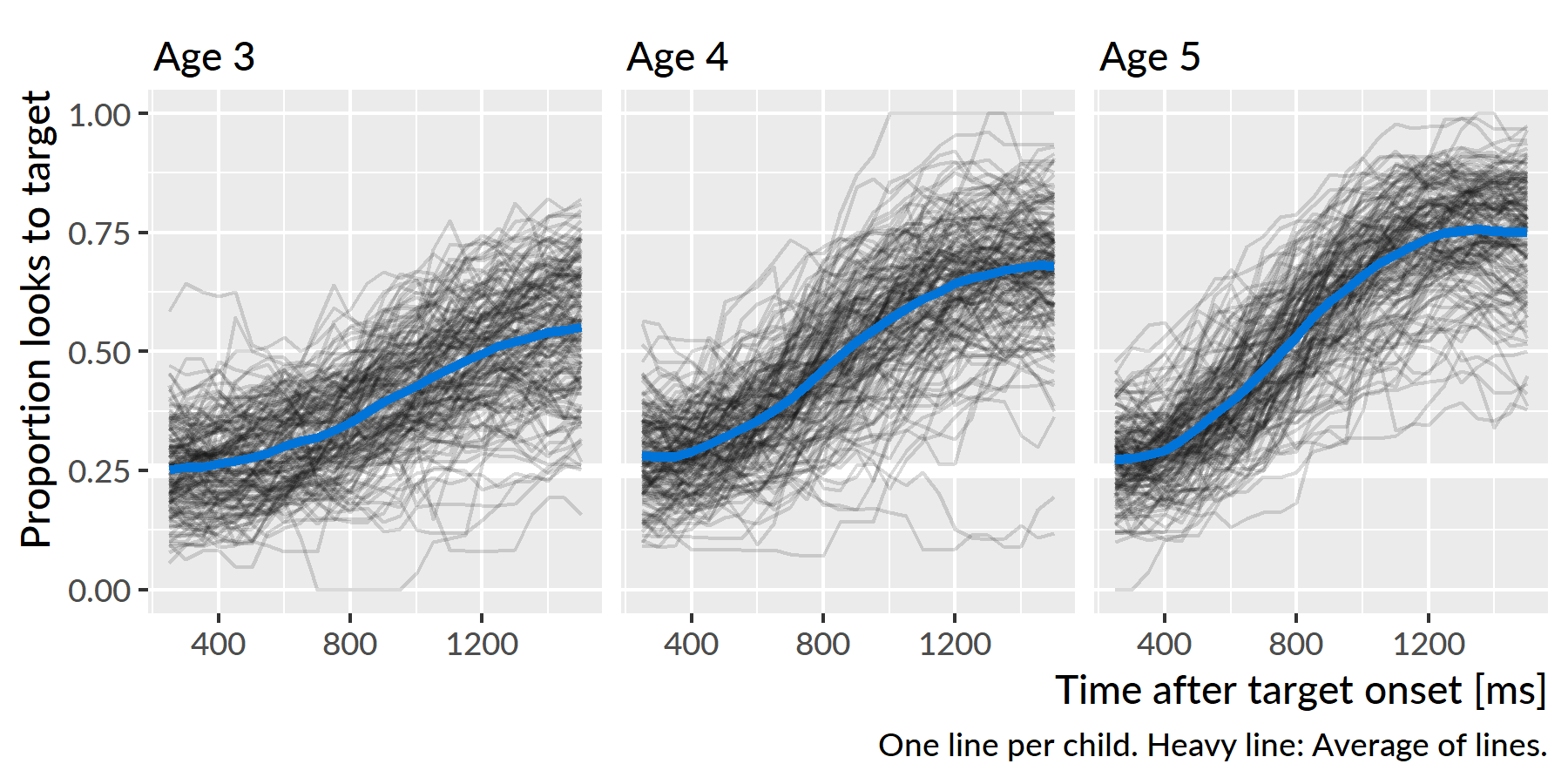


Figure 2 Empirical word recognition growth curves from each year of the study. Each line represents an individual child’s proportion of looks to the target image over time. The heavy lines are the averages of the lines for each year.

# Analysis of familiar word recognition

## Growth curve analysis

Looks to the familiar image were analyzed using Bayesian mixed effects logistic regression. I used *logistic* regression because the outcome measurement is a probability (the log-odds of looking to the target image versus a distractor). I used *mixed-effects* models to estimate a separate growth curve for each child (to measure individual differences in word recognition) but also treat each child’s individual growth curve as a draw from a distribution of related curves. I used *Bayesian* techniques to study a generative model of the data. Instead of reporting and describing a single, best-fitting model of some data, Bayesian methods consider an entire distribution of plausible models that are consistent with the data and any prior information we have about the models. By using this approach, one can explicitly quantify uncertainty about statistical effects and draw inferences using estimates of uncertainty (instead of using statistical significance—which is not a straightforward matter for mixed-effects models).[[3]](#footnote-47)

The eyetracking growth curves were fit using an orthogonal cubic polynomial function of time (a now-conventional approach; see Mirman, [2014](#ref-Mirman2014)). Put differently, I modeled the probability of looking to the target during an eyetracking task as:

That the time terms are *orthogonal* means that , and are transformed so that they are uncorrelated. Under this formulation, the parameters and have a direct interpretation in terms of lexical processing performance. The intercept, , measures the area under the growth curve—or the probability of fixating on the target word averaged over the whole window. We can think of as a measure of *word recognition reliability*. The linear time parameter, , estimates the steepness of the growth curve—or how the probability of fixating changes from frame to frame. We can think of as a measure of *processing efficiency*, because growth curves with stronger linear features exhibit steeper frame-by-frame increases in looking probability.[[4]](#footnote-48)

To study how word recognition changes over time, I modeled how the growth curves change over developmental time. This amounted to studying how the growth curve parameters changes year over year. I included dummy-coded indicators for Age 3, Age 4, and Age 5 and allowed these indicators to interact with the growth curve parameters. These year-by-growth-curve-feature terms captured how the shape of the growth curves changed each year. The model also included random effects to represent by-child and by-child-by-year effects to estimate a general growth curve for each child and to estimate how each child’s growth curve changed each year.

The models were fit in R (vers. 3.4.3) with the RStanARM package (vers. 2.16.3). [Appendix 17](#aim1-gca-models) contains the R code used to fit the model along with a description of the model specifications represented in the model syntax.

### Growth curve features as measures of word recognition performance

As mentioned above, two of the model’s growth curve features have straightforward interpretations in terms of lexical processing performance: The model’s intercept parameter corresponds to the average proportion or probability of looking to the named image over the trial window, and the linear time parameter corresponds to slope of the growth curve or lexical processing efficiency. I also was interested in *peak* proportion of looks to the target. I derived this value by computing the growth curves from the model and taking the median of the five highest points on the curve. Figure 3 shows three simulated growth curves and how each of these growth curve features relate to word recognition performance.

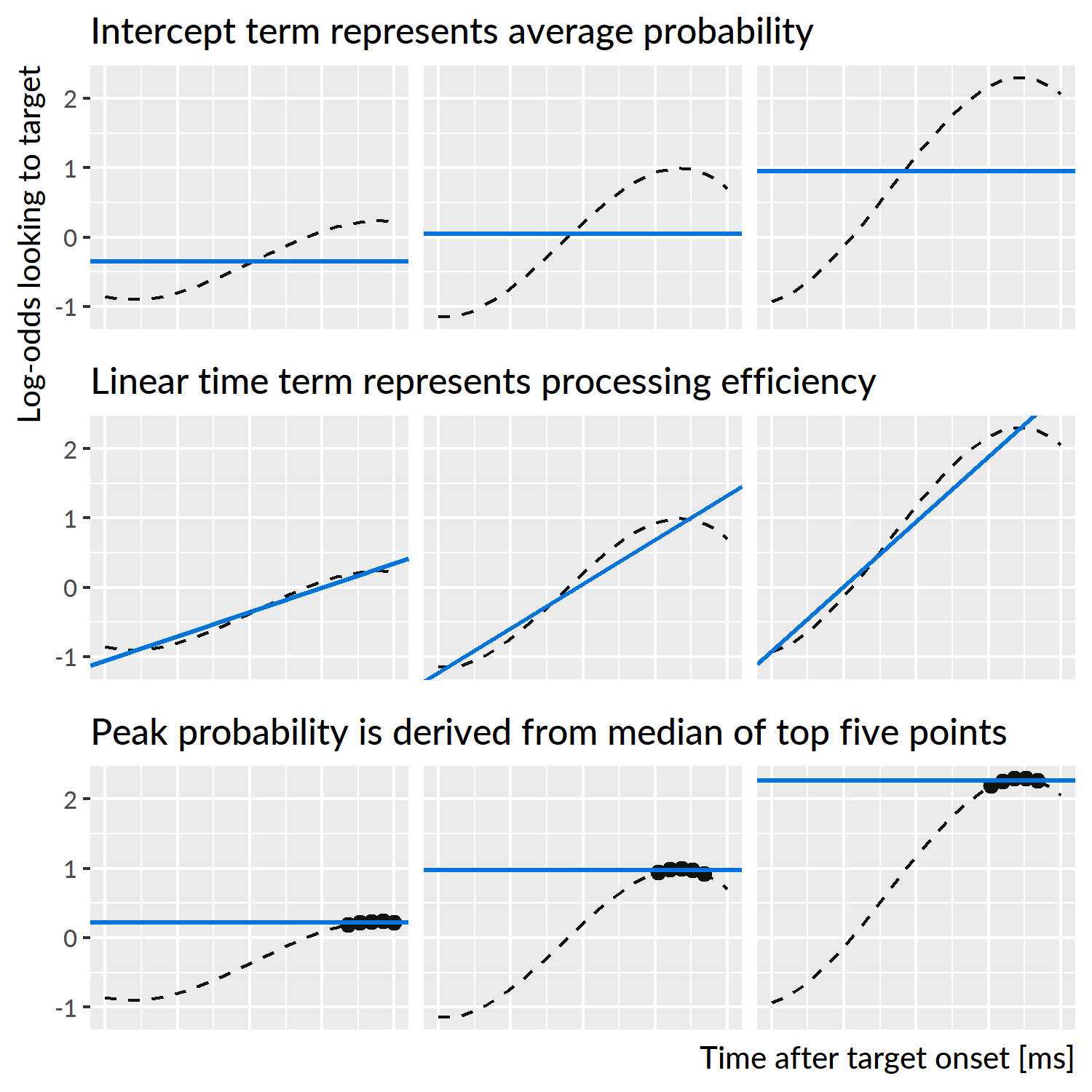


Figure 3 Illustration of the three growth curve features and how they describe lexical processing performance. The three curves used are simulations of new participants at Age 4.

## Year over year changes in word recognition performance

The mixed-effects model estimated a population-average growth curve (“fixed” effects) and how individual children deviated from average (“random” effects). Figure 4 shows 200 posterior samples of the average growth curves for each study. On average, the growth curves become steeper and achieve higher looking probabilities with each year of the study.

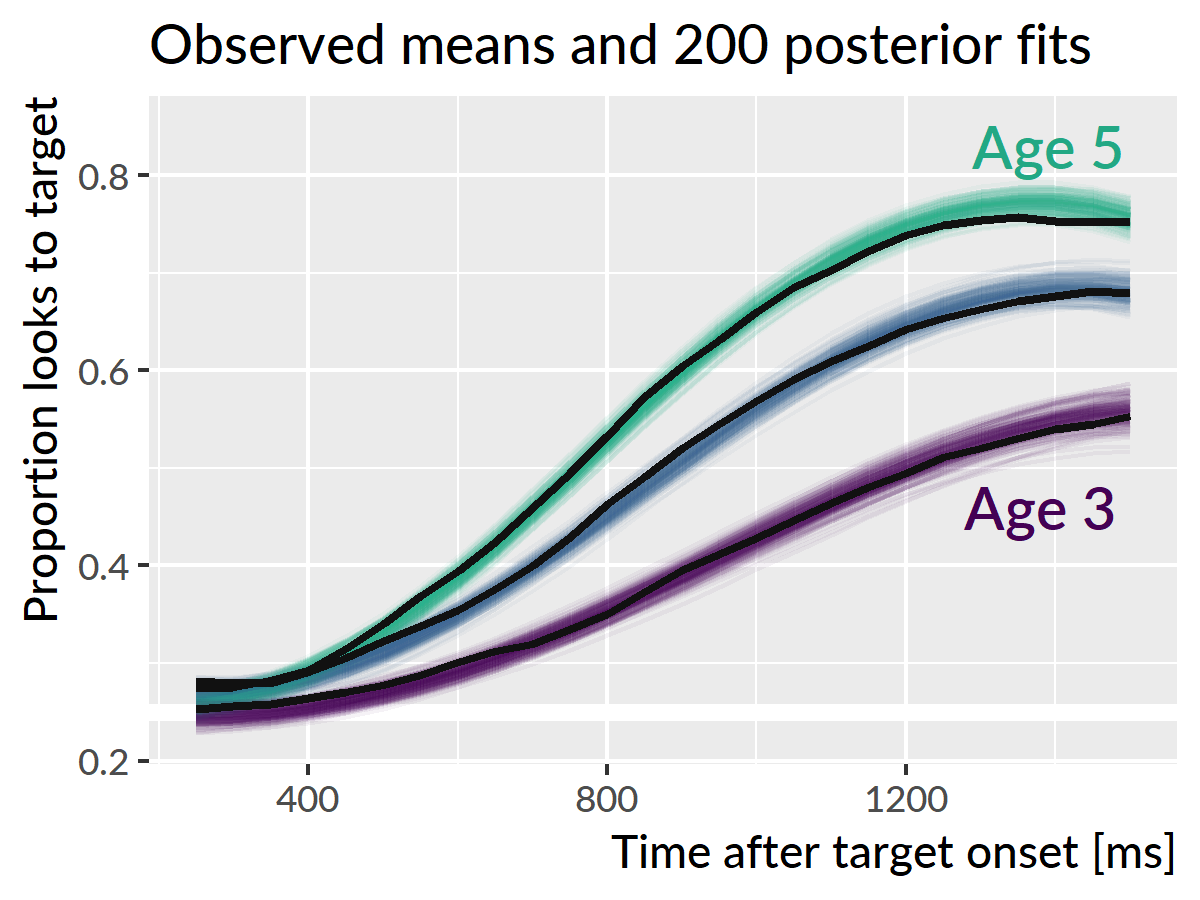


Figure 4 The model estimated an average word recognition growth for each study, and the colored lines represent 200 posterior samples of these growth curves. The thick dark lines represent the observed average growth curve in each study.

Figure 5 depicts uncertainty intervals with the model’s average effects of each timepoint on the growth curve features. The intercept and linear time effects increased each year, confirming that children become more reliable and faster at recognizing words as they grow older. The peak probability also increased each year. For each effect, the change from age 3 to age 4 is approximately the same as the change from age 4 to age 5, as illustrated in Figure 6.

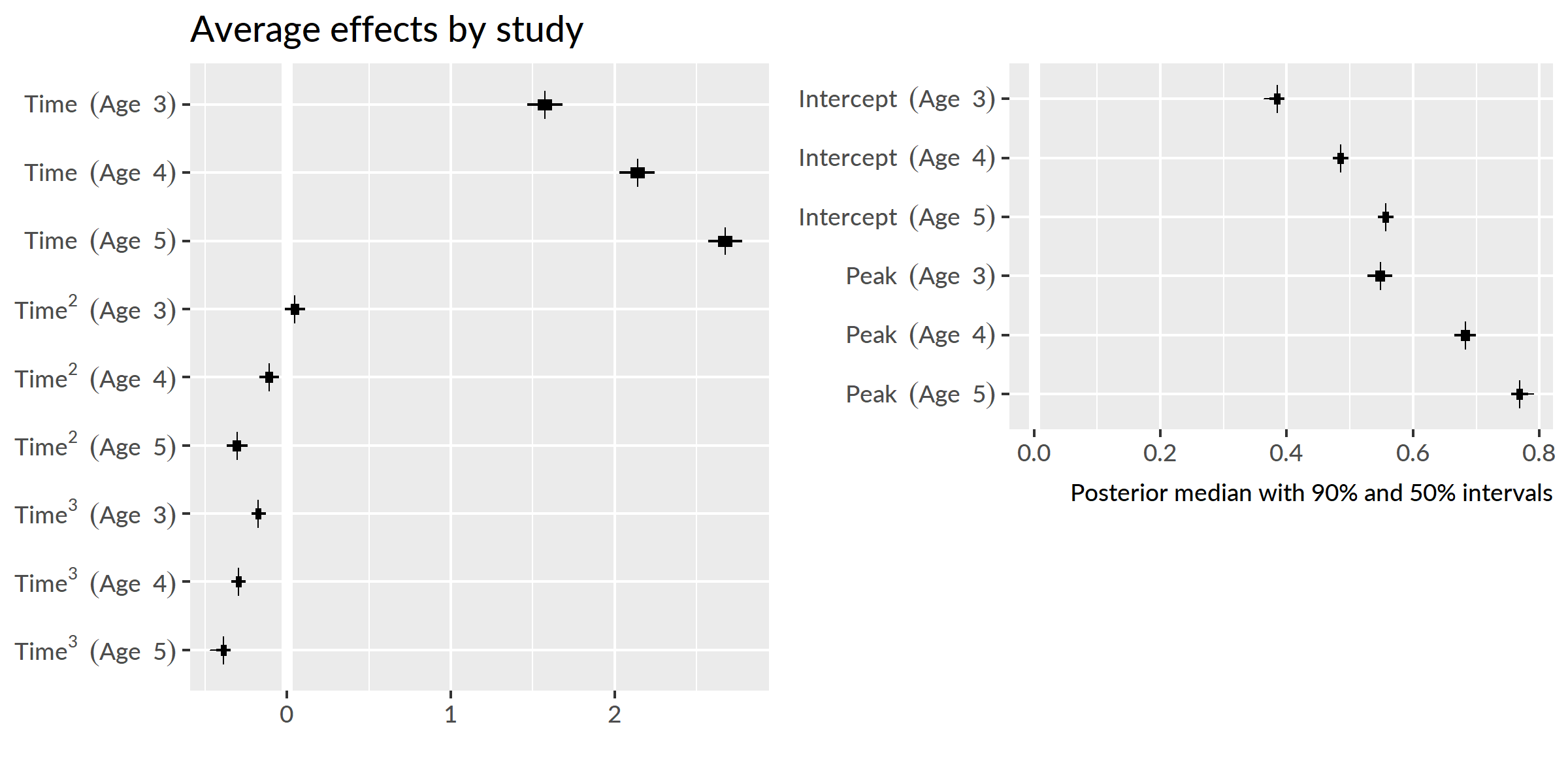


Figure 5 Uncertainty intervals for the effects of study years on growth curve features. The intercept and peak features were converted from log-odds to proportions to ease interpretation.

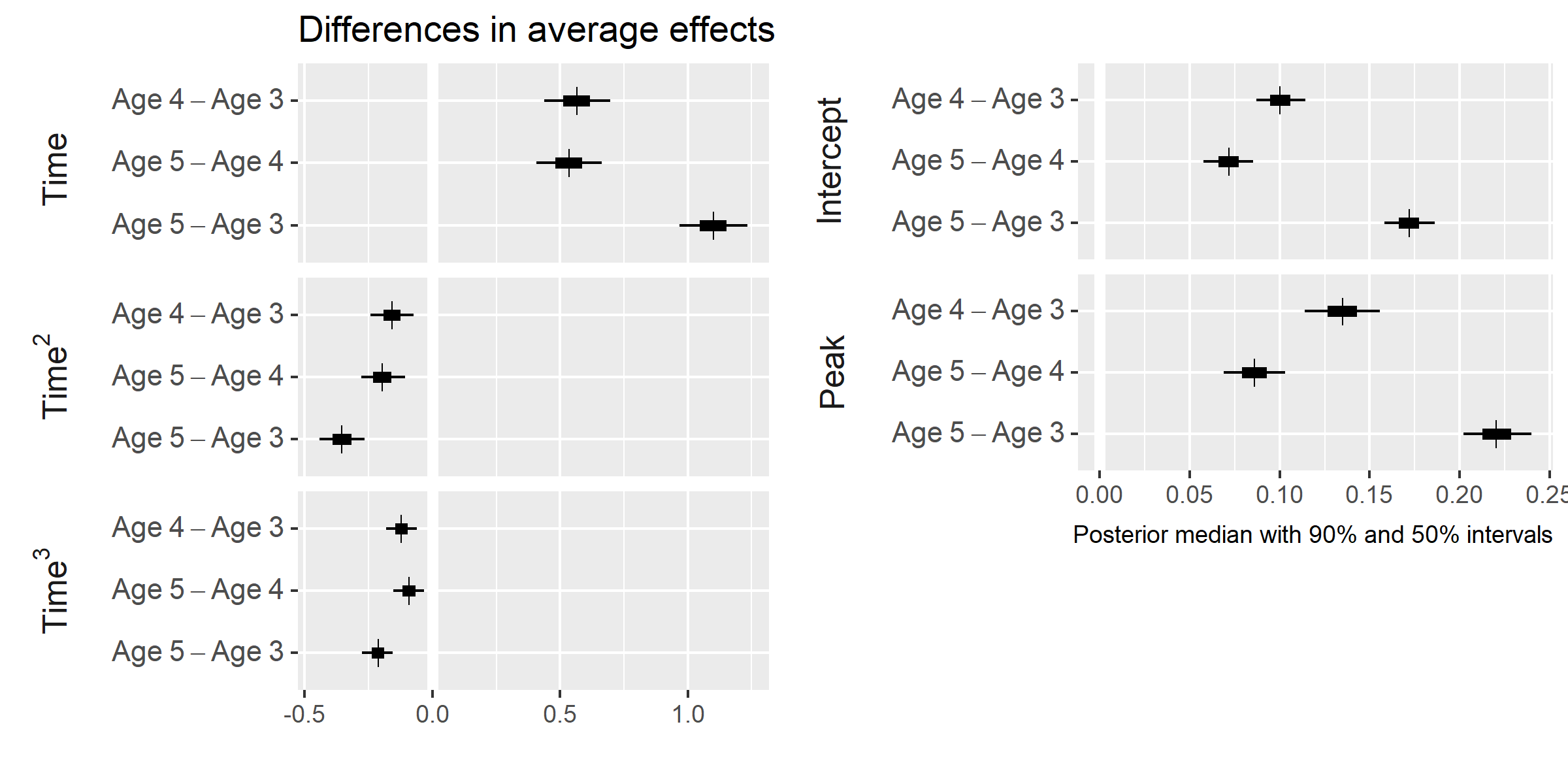


Figure 6 Uncertainty intervals for the differences between study timepoints. Again, the intercept and peak features were converted to proportions.

The average looking probability (intercept feature) was 0.38 [90% UI: 0.37, 0.40] at age 3, 0.49 [0.47, 0.50] at age 4, and 0.56 [0.54, 0.57] at age 5. The averages increased by 0.10 [0.09, 0.11] from age 3 to age 4 and by 0.07 [0.06, 0.09] from age 4 to age 5. The peak looking probability was 0.55 [0.53, 0.57] at age 3, 0.68 [0.67, 0.70] at age 4, and 0.77 [0.76, 0.78] at age 5. The peak values increased by 0.13 [0.11, 0.16] from age 3 to age 4 and by 0.09 [0.07, 0.10] from age 4 to age 5. These results numerically confirm the hypothesis that children would improve in their word recognition reliability, both in terms of average looking and in terms of peak looking, each year. The changes in peak probability were also rather large: children’s probability fixating on the target increased by approximately .1 each year. These growths indicate the task scaled with children’s development because they had room to improve each year.

**Summary**. The average growth curve features increased year over year, so that children looked to the target more quickly and more reliably as they grew older.

## Exploring plausible ranges of performance over time

Bayesian models are generative; they describe how the data could have been generated. This model assumed that each child’s growth curve was drawn from a population of related growth curves, and it tried to infer the parameters over that distribution. These two aspects—a generative model and learning about the population of growth curves—allow the model to simulate new samples from that distribution of growth curves. That is, we can predict a set of growth curves for a hypothetical, unobserved child drawn from the same distribution as the 195 children observed in this study. This procedure of studying model implications by having the model generate new data is called *posterior predictive inference*, and in this case, it allows one to explore the plausible degrees of variability in performance at each age.

Figure 7 shows the posterior predictions for 1,000 simulated participants, and it demonstrates how the model expects new participants to improve longitudinally but also exhibit stable individual features over time. Figure 8 shows uncertainty intervals for these simulations. The model learned to predict less accurate and more variable performance at age 3 with improving accuracy and narrowing variability at age 4 and age 5.

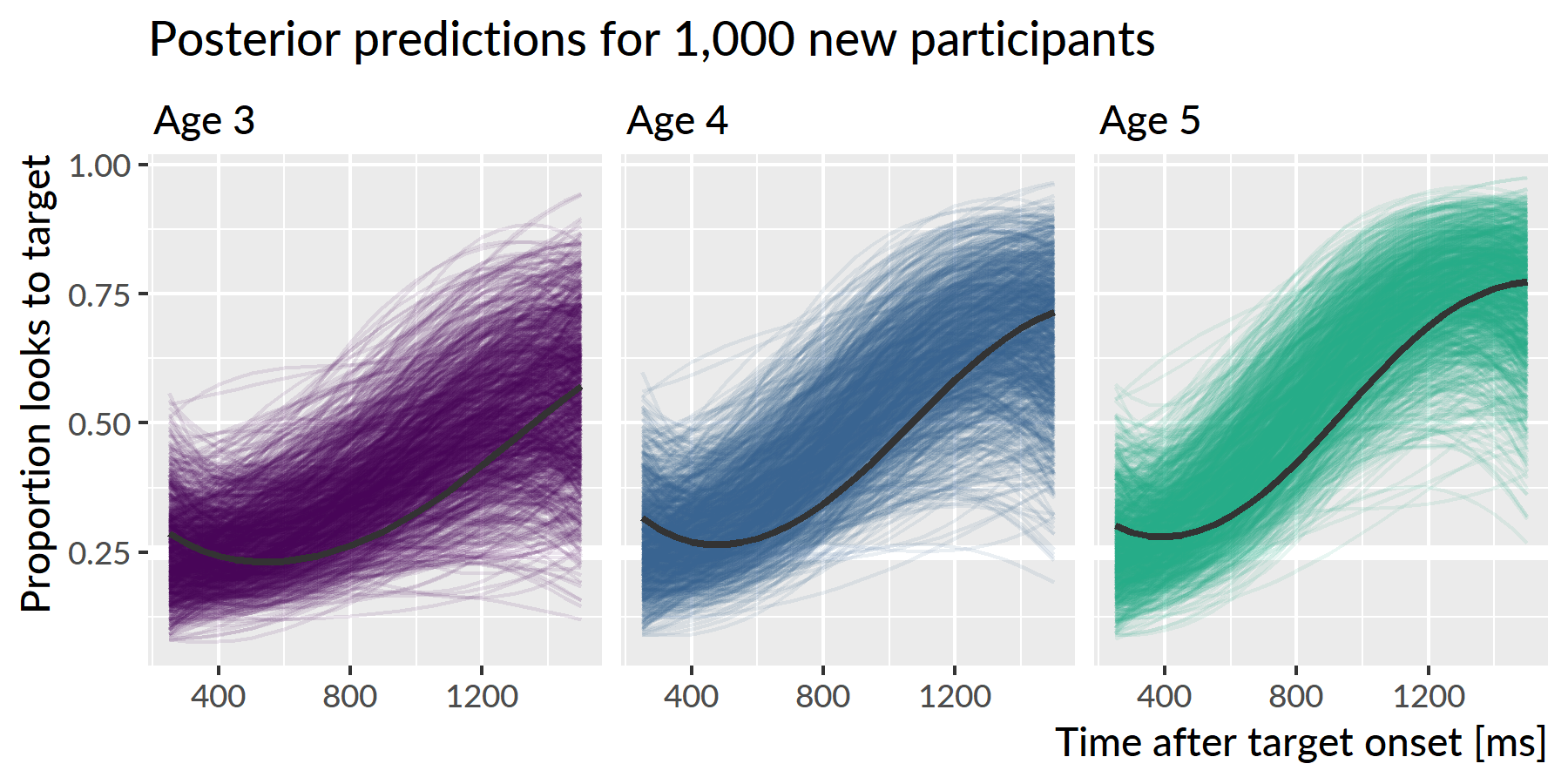


Figure 7 Posterior predictions for hypothetical *unobserved* participants. Each line represents the predicted performance for a new participant. The three dark lines highlight predictions from one single simulated participant. The simulated participant shows both longitudinal improvement in word recognition and similar relative performance compared to other simulations each year, indicating that the model would predict new children to improve year over year and show stable individual differences over time.

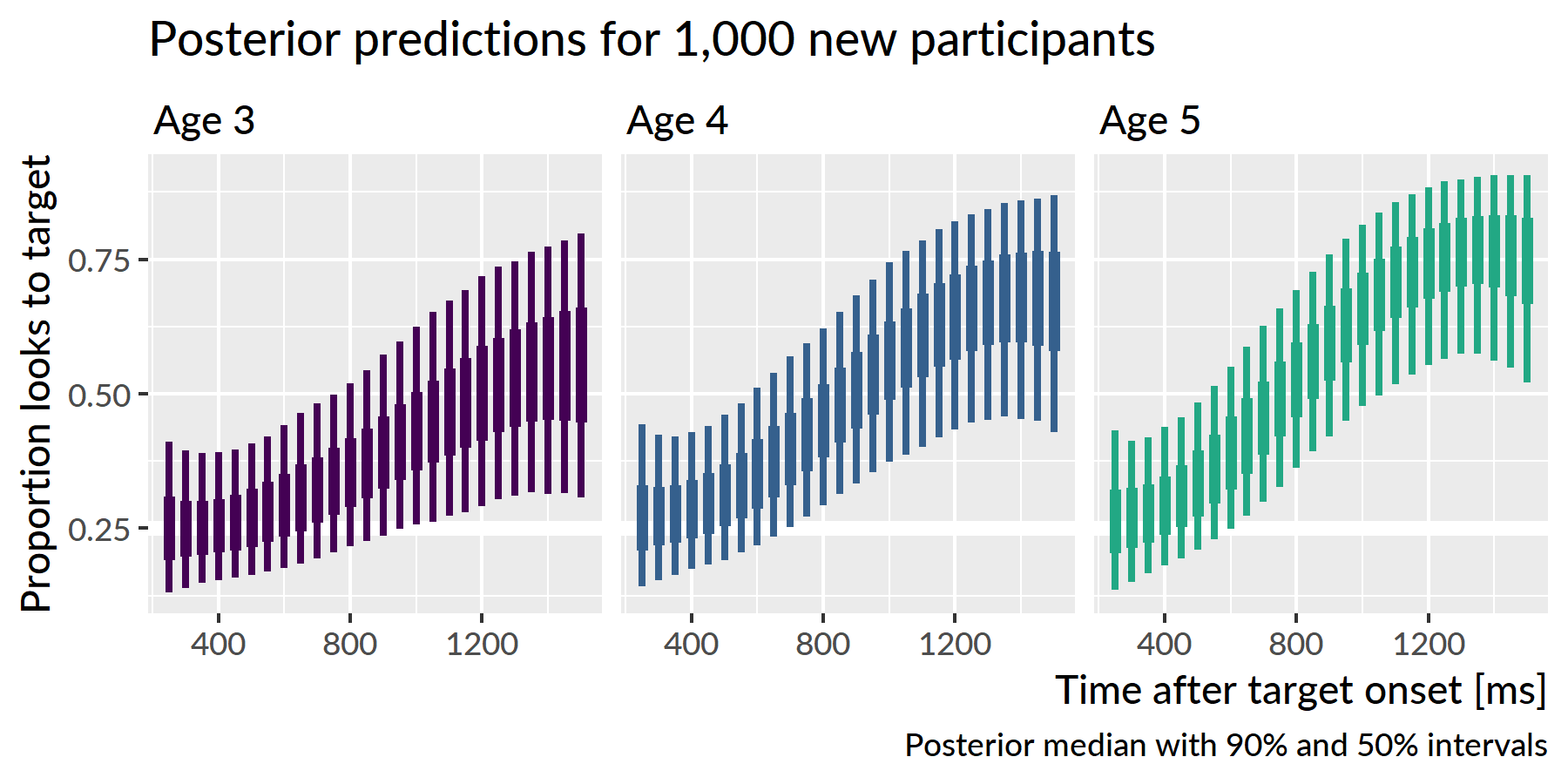


Figure 8 Uncertainty intervals for the simulated participants. Variability is widest at age 3 and narrowest at age 5, consistent with the prediction that children become less variable as they grow older.

I hypothesized that children would become less variable as they grew older and converged on a mature level of performance. To address this question, I inspected the ranges of predictions for the simulated participants. The claim that children become less variable would imply that the range of predictions should be narrower for age 5 than for age 4 and narrower for age 4 than for age 3. Figure 9 depicts the range of the predictions, both in terms of the 90 percentile range (i.e., the range of the middle 90% of the data) and in terms of the 50 percentile (interquartile) range. The ranges of performance decrease from age 3 to age 4 to age 5, consistent with the hypothesized reduction in variability.

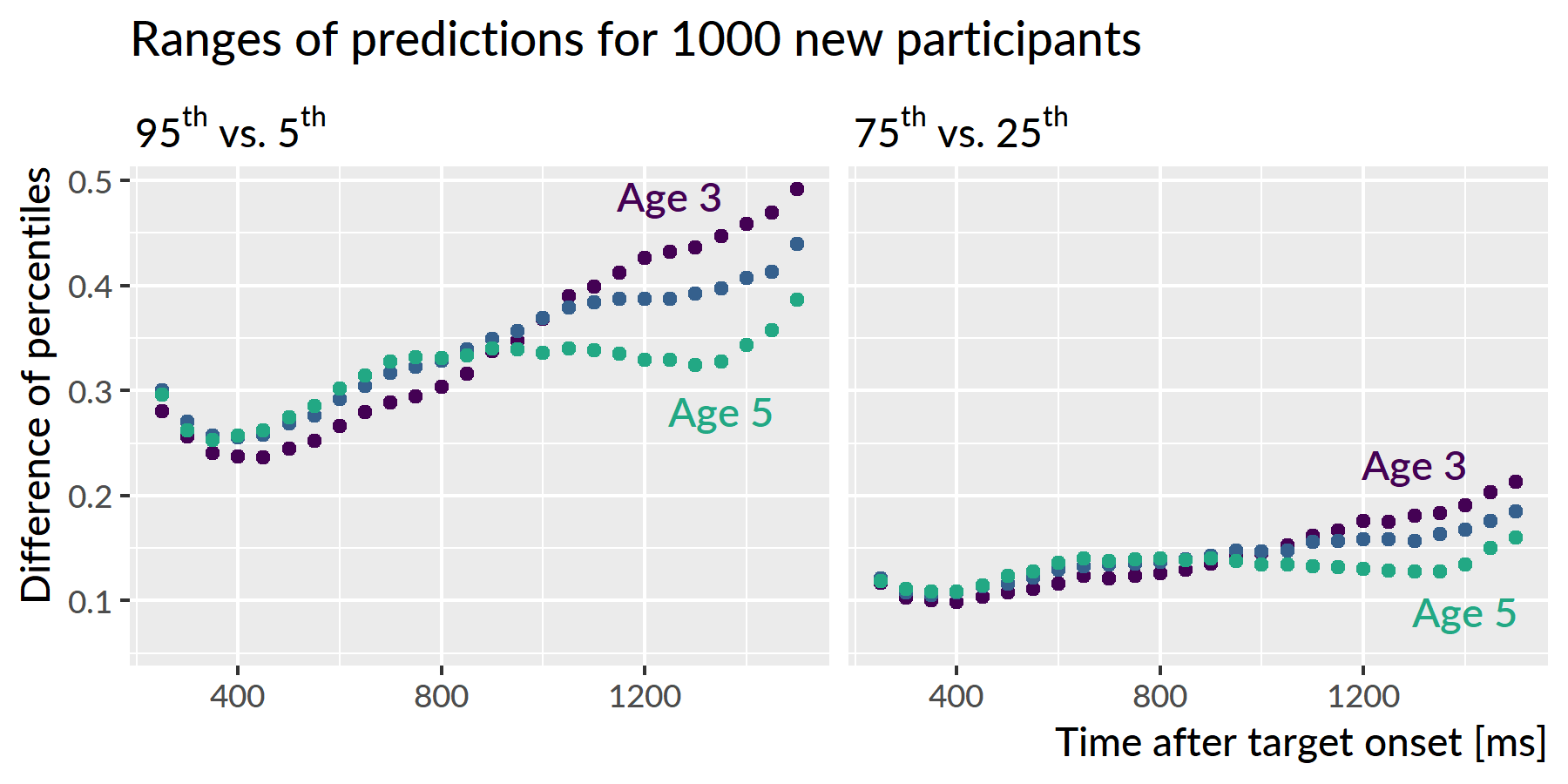


Figure 9 Ranges of predictions for simulated participants over the course of a trial. The ranges are most similar during the first half of the trial when participants are at chance performance, and the ranges are most different at the end of the trial as children reliably fixate on the target image. The ranges of performance decreases with each year of the study as children show less variability.

The developmental pattern of increasing reliability and decreasing variability was also observed for the growth curve peaks. For the synthetic participants, the model predicted that individual peak probabilities will increase each year, peak3 = 0.55 [90% UI: 0.35, 0.77], peak4 = 0.69 [0.48, 0.86], peak5 = 0.78 [0.59, 0.91]. Moreover, the range of plausible values for the individual peaks narrowed each for the simulated data. For instance, the difference between the 95th and 5th percentiles was 0.43 for age 3, 0.38 for age 4, and 0.32 for age 5.

**Summary**. I used the model’s random effects estimates to simulate growth curves from 1,000 hypothetical, unobserved participants. The simulated dataset showed increasing looking probability and decreasing variability with each year of the study. These simulations confirmed the hypothesis that variability would diminish as children began to demonstrate a mature degree of performance for this task.

## Are individual differences stable over time?

I predicted that children would show stable individual differences such that children who are faster and more reliable at recognizing words at age 3 remain relatively faster and more reliable at age 5. To evaluate this hypothesis, I used Kendall’s *W* (the coefficient of correspondence or concordance). This nonparametric statistic measures the degree of agreement among *J* judges who are rating *I* items. For these purposes, the items are the 123 children who provided reliable eyetracking for all three years of the study. (That is, I excluded children who only had reliable eyetracking data for one or two years.) The judges are the sets of growth curve parameters from each year of study. For example, the intercept term provides three sets of ratings: The participants’ intercept terms from year 1 are one set of ratings and the terms from years 2 and 3 provide two more sets of ratings. These three ratings are the “judges” used to compute the intercept’s *W*. Thus, I computed five groups of *W* coefficients, one for each set of growth curve features: intercept, Time1, Time2, Time3, and peak looking probability.

Because I used a Bayesian model, there is a distribution of ratings and thus a distribution of concordance statistics. Each sample of the posterior distribution fits a growth curve for each child in each study, so each posterior sample provides a set of ratings for concordance coefficients. This distribution of *W*’s lets us quantify our uncertainty because we can compute *W*’s for each of the 4000 samples from the posterior distribution.

One final matter is how to assess whether a concordance statistic is meaningful. To tackle this question, I also included a “null rater”, a fake parameter that assigned each child in each year a random number. I use the distribution of *W*’s generated by randomly rating children as a benchmark for assessing whether the other concordance statistics differ meaningfully from chance.

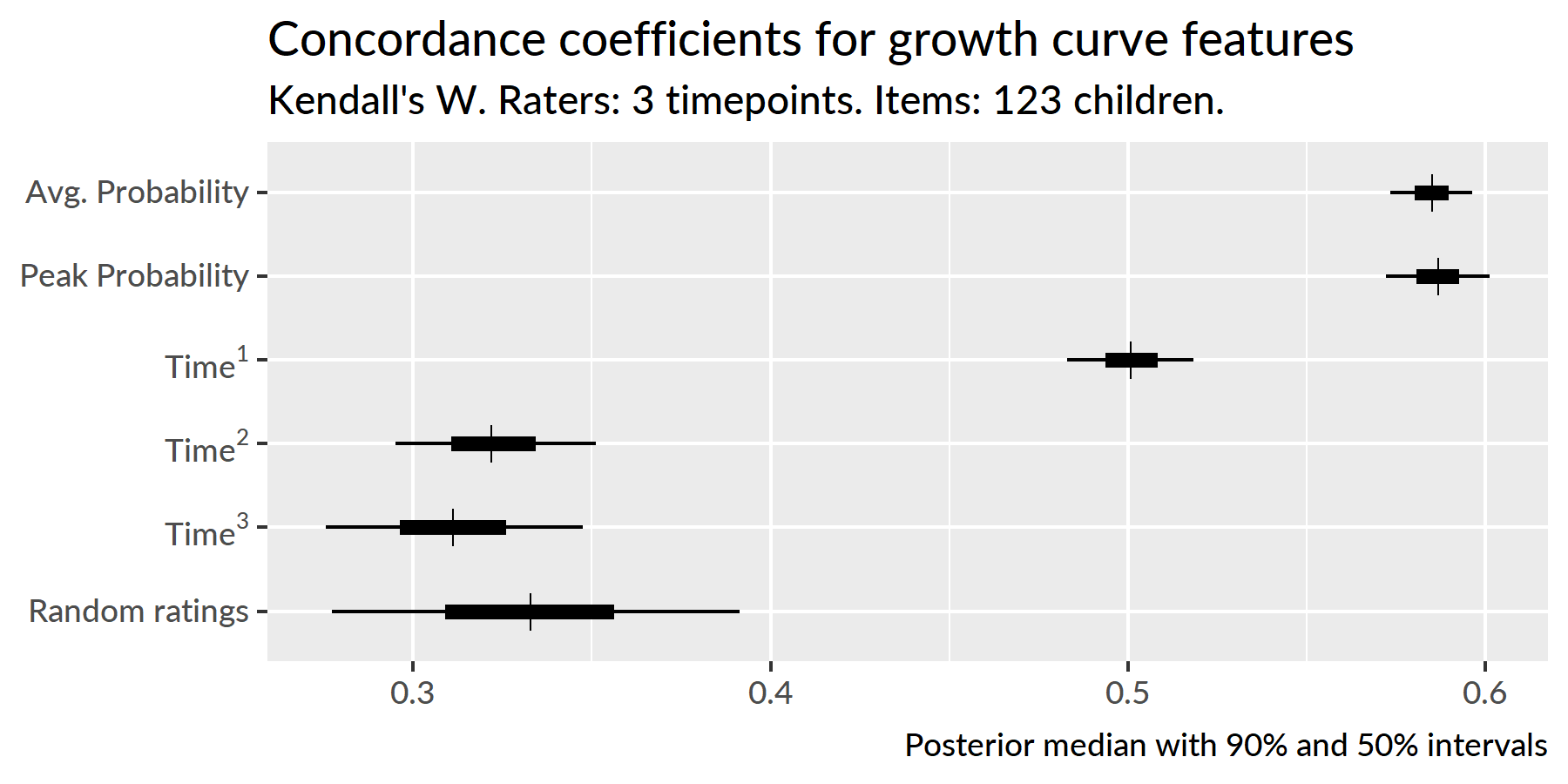


Figure 10 Uncertainty intervals for the Kendall’s coefficient of concordance. Random ratings provide a baseline of null *W* statistics. The peak, intercept and linear time features are decisively non-null, indicating a significant degree of correspondence in children’s relative word recognition reliability and efficiency over the three years of the study.

I used the kendall() function in the irr R package (vers. 0.84; Gamer, Lemon, & Singh, [2012](#ref-irr)) to compute concordance statistics. Figure 10 depicts uncertainty intervals for the Kendall *W*’s for these growth curve features. The 90% uncertainty interval of *W* statistics from random ratings, [.28, .39], subsumes the intervals for the Time2 effect [.30, .35] and the Time3 effect [.28, .35], indicating that these values do not differentiate children in a longitudinally stable way. Earlier, I claimed that only the intercept, linear time, and peak features have psychologically meaningful interpretations and that the higher-order time features mainly act to capture the curvature of the data. These null concordance statistics support that claim because the Time2 and Time3 features differentiate children across studies as well as random numbers.

Concordance is strongest for the peak feature, *W* = [] and the intercept term, *W* = [], followed by the linear time term, *W* = []. Because these values are far removed from the statistics for random ratings, I conclude that there is a credible degree of correspondence across studies when ranking children using their peak looking probability, average look probability (the intercept) or their growth curve slope (linear time).

**Summary**. Growth curve features measured individual differences in word recognition performance. By using Kendall’s *W* to measure the degree of concordance among growth curve features over time, I tested whether individual differences in lexical processing persisted over development. I found that the peak looking probability, average looking probability and linear time features were stable over time. Children who relatively fast (or reliable) at word recognition at one age were also relatively fast (or reliable) at other ages too.

## Predicting future vocabulary size

I hypothesized that individual differences in word recognition at age 3 will be more discriminating and predictive of future language outcomes than differences at age 4 or age 5. To test this hypothesis, I calculated the correlations of growth curve features with age 5 expressive vocabulary size and age 4 receptive vocabulary. (The receptive test was not administered during the last year of the study for logistical reasons.) As with the concordance analysis, I computed each of the correlations for each sample of the posterior distribution to obtain a distribution of correlations.

Figure 11 shows the correlations of the peak looking probability, average looking probability and linear time features with expressive vocabulary size at age 5, and Figure 12 shows analogous correlations for the receptive vocabulary at age 4. For all cases, the strongest correlations were found between the growth curve features at age 3.

Growth curve peaks from age 3 correlated with age 5 vocabulary with *r* = .52 [90% UI .50, .54], but the concurrent peaks from age 5 showed a correlation of just *r* = .31 [.29, .33], a difference between age-3 and age-5 correlations of *r*3−5 = .21 [.18, .24]. A similar pattern held for lexical processing efficiency values. Linear time features from age 3 correlated with age 5 vocabulary with *r* = .41 [.39, .44], whereas the concurrent lexical processing values from age 5 only showed a correlation of *r* = .28 [.26, .31], a difference of *r*3−5 = .13 [.10, .16]. For the average looking probabilities, the correlation for age 3, *r* = .39 [.39, .44], was probably only slightly greater than the correlation for age 4, *r*3−4 = .02 [−.01, .04] but considerably greater than the concurrent correlation at age 5, *r*3−5 = .08 [.05, .10].

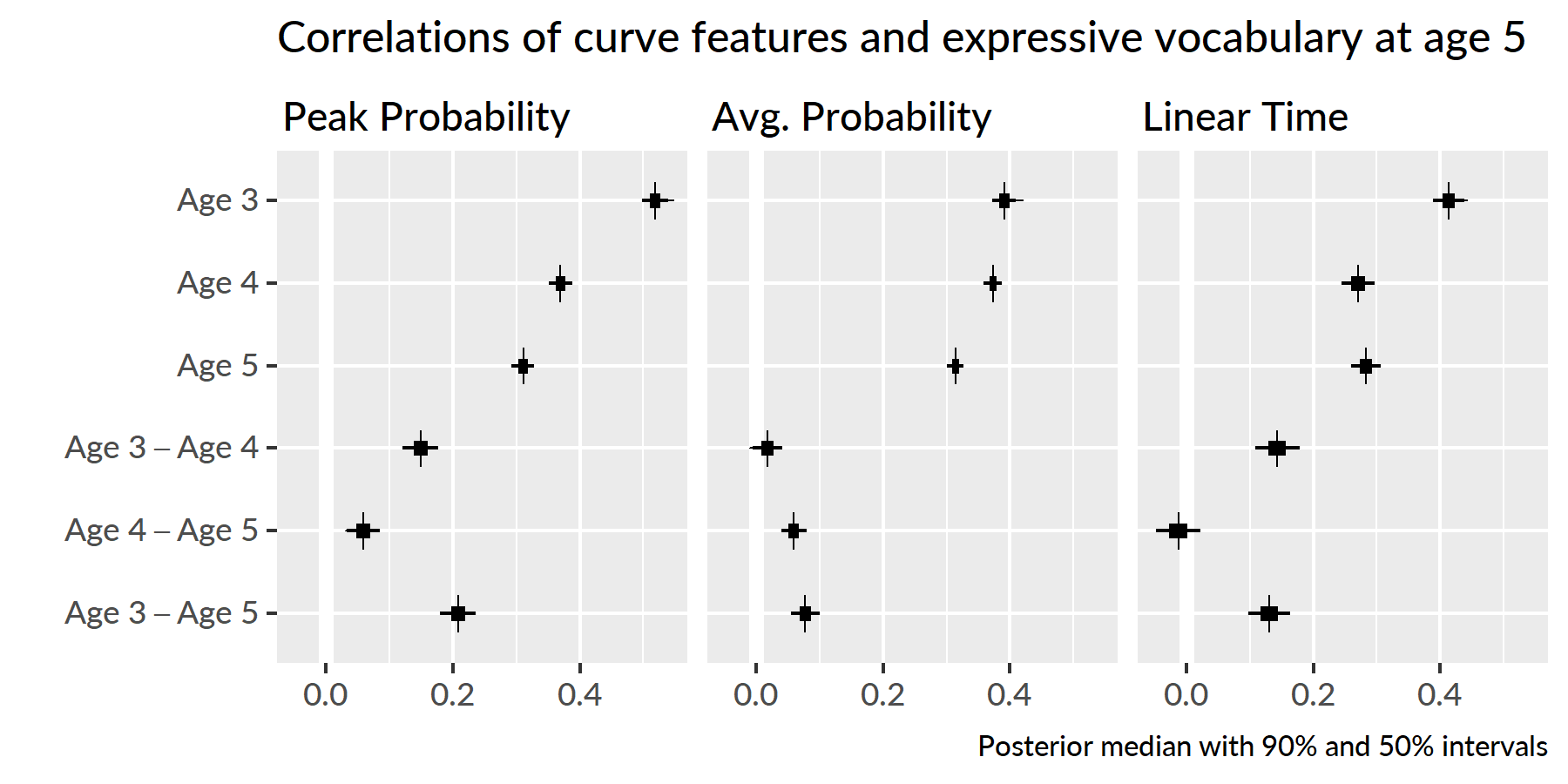


Figure 11 Uncertainty intervals for the correlations of growth curve features at each timepoint with age-5 expressive vocabulary (EVT-2 standard scores). The bottom rows provide intervals for the pairwise differences in correlations between timepoints. For example, the top row of the left panel is the correlation between age-3 peak probability and age-5 expressive vocabulary.

Peak looking probabilities from age 3 were strongly correlated with age 4 receptive vocabulary, *r* = .62 [.61, .64], and this correlation was much greater than the correlation observed for the age 4 growth curve peaks, *r*3−4 = .26 [.23, .29]. The correlation for age 3 average looking probabilities, *r* = .45 [.44, .47], was greater than the age 4 correlation, *r*3−4 = .08 [.06, .11], and the correlation for age 3 linear time features, *r* = .51 [.49, .54], was likewise greater, *r*3−4 = .22 [.19, .26].

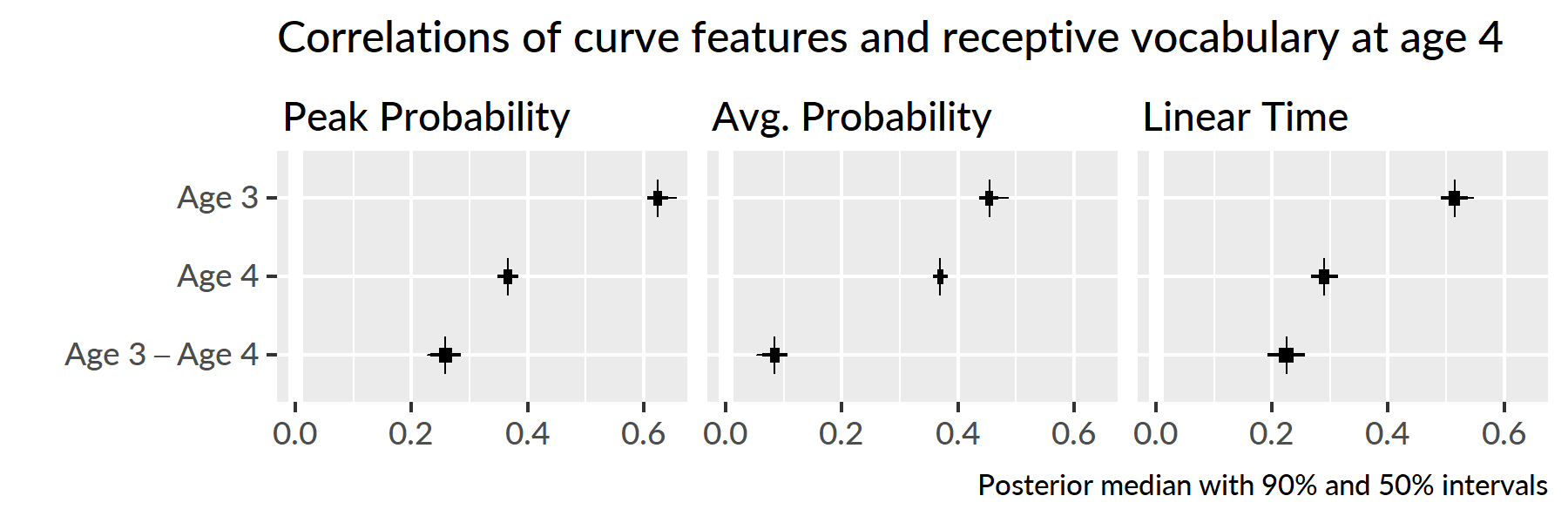


Figure 12 Uncertainty intervals for the correlations of growth curve features from age 43 and age 44 with age-4 receptive vocabulary (PPVT-4 standard scores). The bottom row shows pairwise differences between the age-3 and age-4 correlations.

**Summary**. Although individual differences in word recognition were stable over time, early differences were more significant than later ones. The strongest predictors of future vocabulary size were the growth curve features from age 3. Of these features, correlations were strongest for peak looking probabilities.

## Discussion

In the preceding analyses, I examined many aspects of children’s recognition of familiar words. First, I modeled how children’s looking patterns *on average* changed year over year. Children’s word recognition improved each year: The growth curves grew steeper, reached higher peaks, and increased in their overall average value each year. This result was unsurprising, but it was valuable because it confirmed that this word recognition task scaled with development. The task was simple enough that children could recognize words at age 3 but challenging enough for children’s performance to improve each year.

After establishing how the averages changed each year, I next asked how variability changed each year. To tackle this question, I used posterior predictive inference to have the model simulate samples of data, and in particular, to simulate new participants. The range of performance narrowed each year, so that children were most variable at age 3 and least variable at age 5. This result is consistent with a model of development children vary widely early on and converge on a more mature level of performance. From this perspective, word recognition is a skill where children “grow out” of immature and highly variable performance patterns. An alternative outcome would have been concerning: Word recognition differences that expanded with age with some children falling behind their peers.

Although the range of individual differences decreased with age, individual differences did not disappear over time. When children at each age were ranked using growth curve features, I found a high degree of correspondence among these ratings. Children who were faster or more accurate at age 3 remained relatively fast or accurate at age 5. Thus, differences in word recognition were longitudinally stable over the preschool years. Extrapolating forwards in time, these differences likely become smaller and smaller and become irrelevant for everyday listening situations. It is plausible, however, that under adverse listening conditions, individual differences might re-emerge and differentiate children’s word recognition performance.

Lastly, I analyzed how individual differences in word recognition features correlated with future vocabulary outcomes. The peak looking probabilities and growth curve slopes from age 3 showed the strongest correlations with future vocabulary scores. This finding was remarkable: Expressive vocabulary scores at age 5, for example, were more strongly correlated with word recognition data collected two years earlier than word recognition data collected during the same week.

We can understand the predictive value of age-3 word recognition performance from two perspectives. The first interpretation is statistical. Differences in children’s word recognition performance were greatest at age 3, so word recognition features at age 3 provide more variance and more information about the children and their future vocabulary size. The second interpretation is conceptual. Correlations were strongest for the growth curve peaks. We can think of this feature as measuring children’s maximum word recognition certainty. A child with a peak of .5, for example, looked the target image half of the time when they were most certain about the word. Although all of the words used were familiar to preschoolers, children with higher peaks knew those words *better*. These children had a stronger foundation for word-learning than children who show more uncertainty during word recognition, and as a result, these children had developed larger vocabularies two years later.

# Effects of phonological and semantic competitors

## Looks to the phonological competitor

The next question I asked was how children’s sensitivity to the phonological competitors changed over developmental time. Following our approach in Law et al. ([2016](#ref-RWLPaper)), I only examined trials for which the phonological foil and the noun shared the same syllable onset. For example, this criterion included trials with *dress*–*drum*, *fly*–*flag*, or *horse*–*heart*, but it excluded trials *kite*–*gift* (phonetic feature difference), *bear*–*bread* (onset difference), and *ring*–*swing* (rimes). I kept 13 of the 24 trials. [Appendix 16](#vw-experiment-items) provides a complete list of trials used.

The outcome measure for these analyses was the log-odds of fixating on the phonological competitor versus the unrelated word. Because children looked more to the target word with each year of the study, they necessarily looked less to the distractors each year. Figure 13 illustrates how the proportions of looks to the phonological foils declined each year. Therefore, I examined the effect of the phonological foil in comparison to the unrelated foil. For example, on the trials where the target is *fly*, we can study the effect of the phonological foil *flag* by looking at when and to what degree the children fixate on *flag* more than the unrelated image *pen*. If a window of time of shows a consistent advantage for the phonological foil over the unrelated image, we conclude that the children were sensitive to the phonological foil during that window. By studying the time course of fixations to the phonological competitor versus the unrelated word, we can identify when the phonological competitor affected word recognition most significantly.

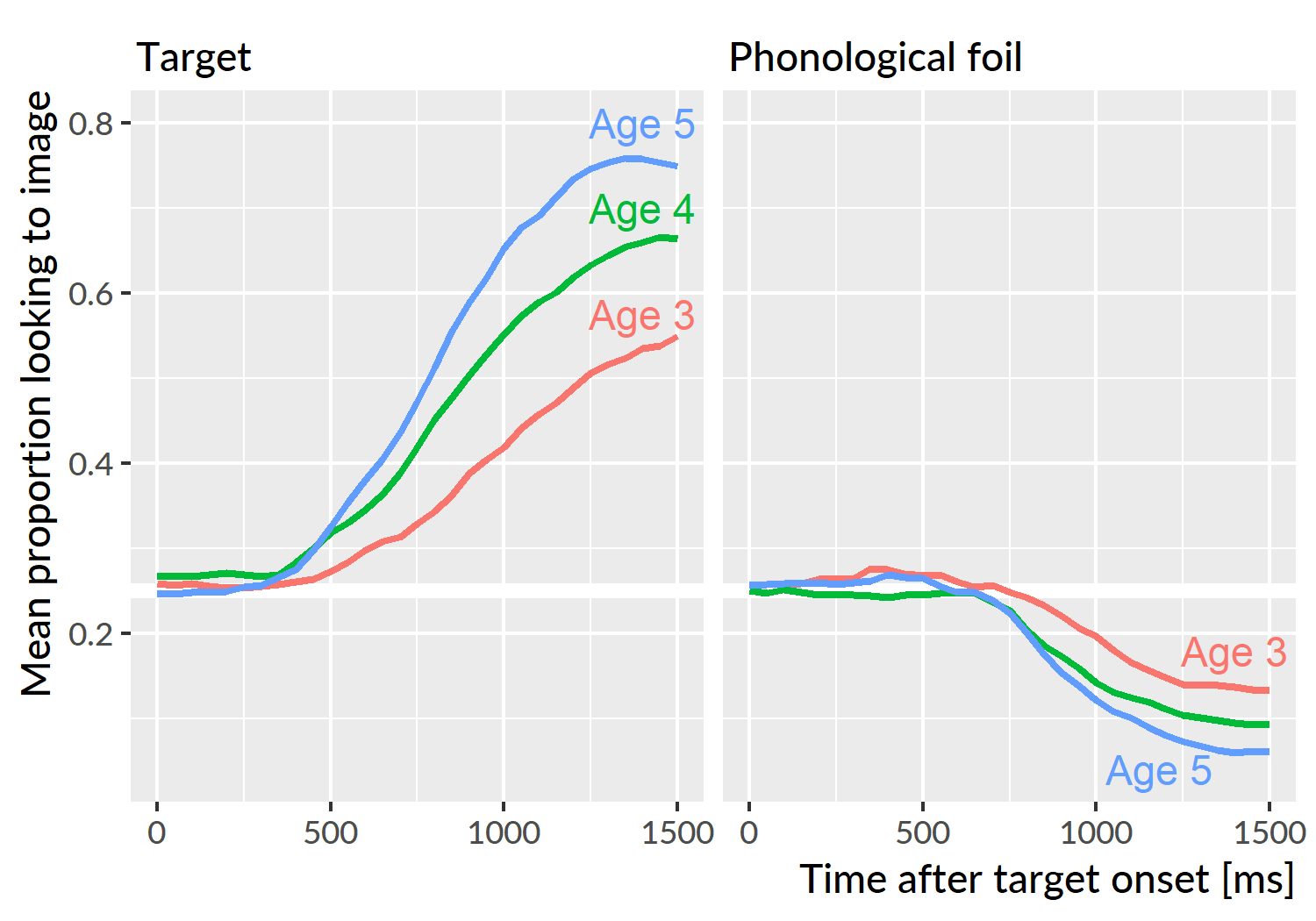


Figure 13 Because children looked more to the target as they grew older, they numerically looked less the foils too. This effect is why I evaluated the phonological and semantic foils by comparing them against the unrelated image.

As in the models from the previous chapter, I downsampled the data into 50-ms (3-frame) bins in order to smooth the data. For these trials, I modeled the looks from 0 to 1500 ms, and I aggregated looks by child, study and time bin. To account for the sparseness of the data, I used the empirical log-odds (or empirical logit) transformation (Barr, [2008](#ref-Barr2008)). This transformation adds .5 to the looking counts. For example, a time-frame with 4 looks to the phonological foil and 1 look to the unrelated image has a conventional log-odds of log(4/1) = 1.39 and empirical log-odds of log(4.5/1.5) = 1.10. This transformation fills in bins with 0 looks with .5/.5 (avoiding 0/0 problems), and it dampens the extremeness of some probabilities that arise in sparse count data.

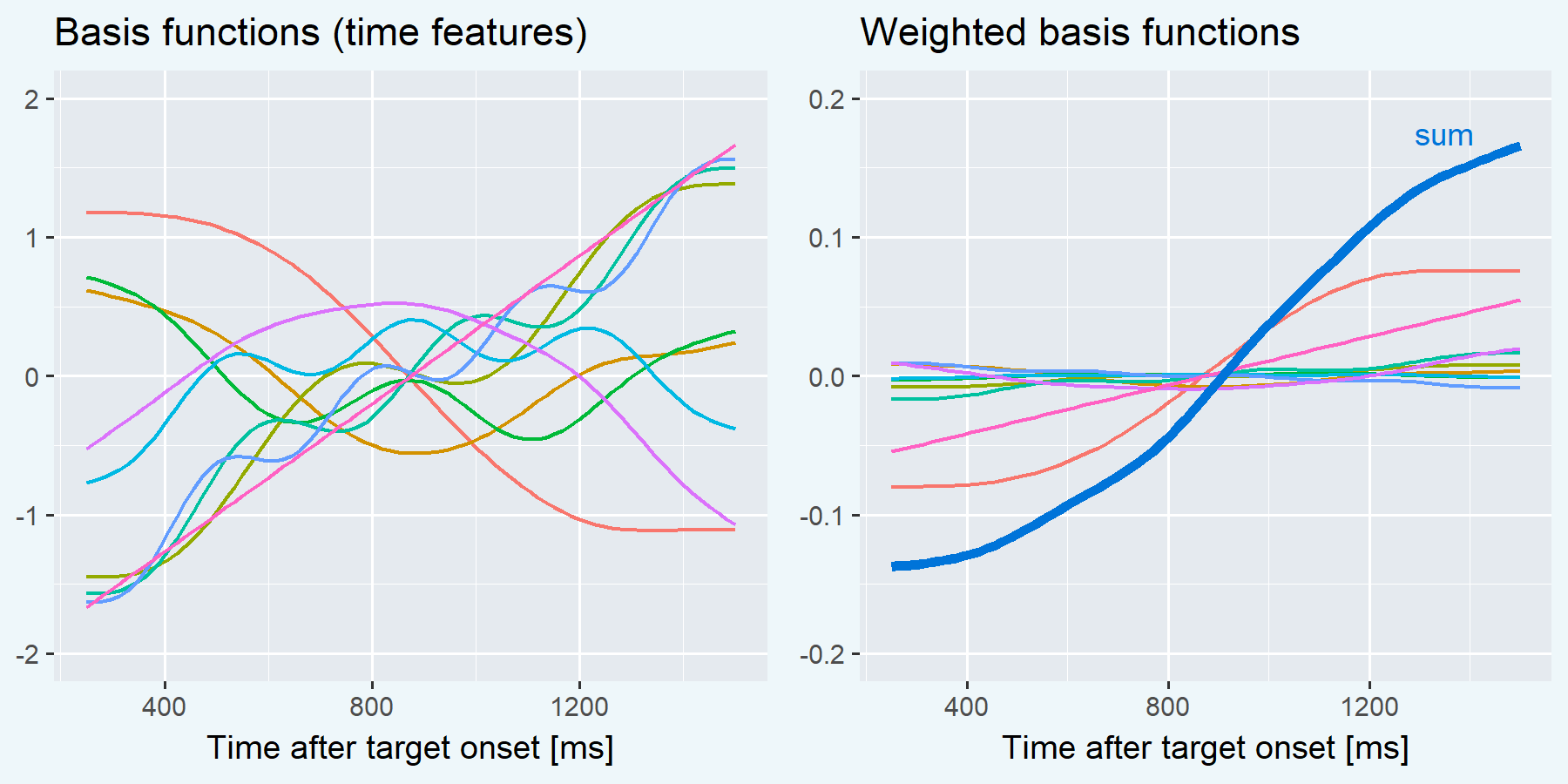
To model these data, I fit a generalized additive model with fast restricted maximum likelihood estimation (see Sóskuthy, [2017](#ref-Soskuthy2017) for a tutorial for linguists; Winter & Wieling, [2016](#ref-Winter2016); Wood, [2017](#ref-Wood2017)). Box 1 provides a brief overview of these models. I used the mgcv R package (vers. 1.8.24; Wood, [2017](#ref-Wood2017)) with support from the tools in the itsadug R package (vers. 2.3; van Rij, Wieling, Baayen, & van Rijn, [2017](#ref-itsadug)).[[5]](#footnote-68) [Appendix 17](#aim1-gca-models) contains the R code used to fit these models along with a description of the specifications represented by the model syntax.

**Box 1: Intuition behind generalized additive models**.

In these analyses, the outcome of interest is a value that changes over time in a nonlinear way. We model these time series by building a set of features to represent time values. In the growth curve analyses of familiar word recognition, I used a set of polynomial features which expressed time as the weighted sum of a linear trend, a quadratic trend and cubic trend. That is:

But another way to think about the polynomial terms is as *basis functions*: A set of features that combine to approximate some nonlinear function of time. Under this framework, the model can be expressed as:

This is the idea behind generalized additive models and their *smooth terms*. These smooths fit nonlinear functions of data by weighting and adding simple functions together. The figures below show 9 basis functions from a “thin-plate spline” and how they can be weighted and summed to fit a growth curve.



Each of these basis functions is weighted by a model coefficient, but the individual basis functions are not a priori meaningful. Rather, it is the whole set of functions that approximate the curvature of the data—i.e., *f*(Time)—so we statistically evaluate the whole batch of coefficients simultaneously. This joint testing is similar to how one might test a batch of effects in an ANOVA. If the batch of effects jointly improve model fit, we infer that there is a significant smooth or shape effect at play.

Smooth terms come with an estimated degrees of freedom (EDF). These values provide a sense of how many degrees of freedom the smooth consumed. An EDF of 1 is a perfectly straight line, indicating no smoothing. Higher EDF values indicate that the smooth term captured more curvature from the data.

The model included main effects of study year. These *parametric* terms work like conventional regression effects and determined the growth curve’s average values. The model used age 4 as the reference year, so the intercept represented the average looking probability at age 4. The year effects represented differences between age 4 vs. age 3 and age 4 vs. age 5.

The model also included *smooth* terms to represent the time course of the data. As with the parametric effects, age 4 served as the reference year. The model estimated a smooth for age 4 and it estimated *difference smooths* to capture how the curvature at age 3 and age 5 differed from the age-4 curvature. Each of these study-level smooths used 10 knots (9 basis functions). I also included child-level *random smooths* to represent child-level variation in growth curve shapes. Because there is much as less data at the child level than at the study level, these random smooths only included 5 knots (4 basis functions). We can think of these simpler splines as coarse adjustments in growth curve shape to capture child-level variation from limited data. Altogether, the model contained the following terms:

$$
\small
$$

The model’s fitted values are shown in Figure 14. These are the average empirical log-odds of fixating on the phonological foil versus the unrelated image for each year of the study. The model captured the trend for increased looks to the competitor image with each year of the study. At age 4 and age 5, the shape rises from a baseline to the peak around 800 ms. These curves slope downwards and eventually fall beneath the initial baseline. The shape at age 3 does not have a steady rise from baseline and shows a small peak around 800 ms. The peak proportions of looks to the phonological competitor versus the unrelated word were .57 at 800 ms for age 3, .61 at 750 ms for age 4, and .64 at 750 ms for age 5.

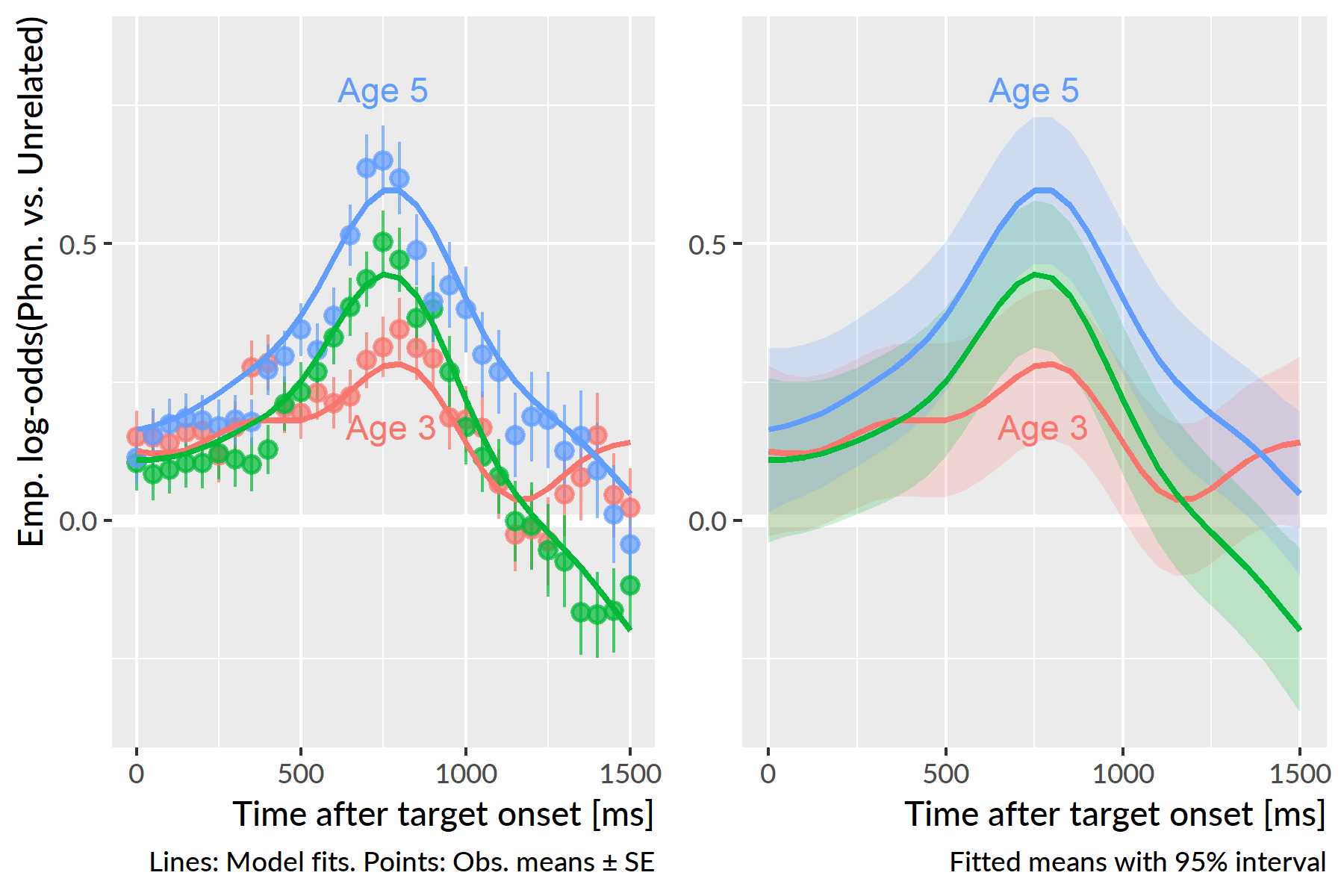


Figure 14 With each year of the study, children looked more to the phonological competitor (relative to the unrelated image) during and after the target noun. Both figures show means for each year estimated by the generalized additive model. The left panel compares model estimates to observed means and standard errors, and the right panel visualizes estimated means and their 95% confidence intervals.

The early peaks occur when one would expect if children are acting on partial phonological information. The similarity between the phonological competitor and the target noun occurs early on in the trial. Suppose a child acts on the first 400 ms of the phonological competitor. Assuming a 200–300 ms overhead to execute an eye movement in response to speech, the child would reach the phonological foil around 600–700 ms. This window is slightly before the observed peaks at 750–800 ms, but the age 4 and age 5 curves both are on the rise away from baseline during this window.

The average looks to the phonological foil over the unrelated image for age 4 was 0.16 emp. log-odds, .54 proportion units. The averages for age 3 and age 4 did not significantly differ, *p* = .85, but the average value was significantly greater at age 5, 0.31 emp. log-odds, .58 proportion units, *p* < .001. Visually, this effect shows up in the almost constant height difference between the age-4 and the age-5 curves.

There was a significant smooth term for time at age 4, estimated degrees of freedom (EDF) = 7.28, *p* < .001. Figure 15 visualizes how and when the smooths from other studies differed from the age-4 smooth.

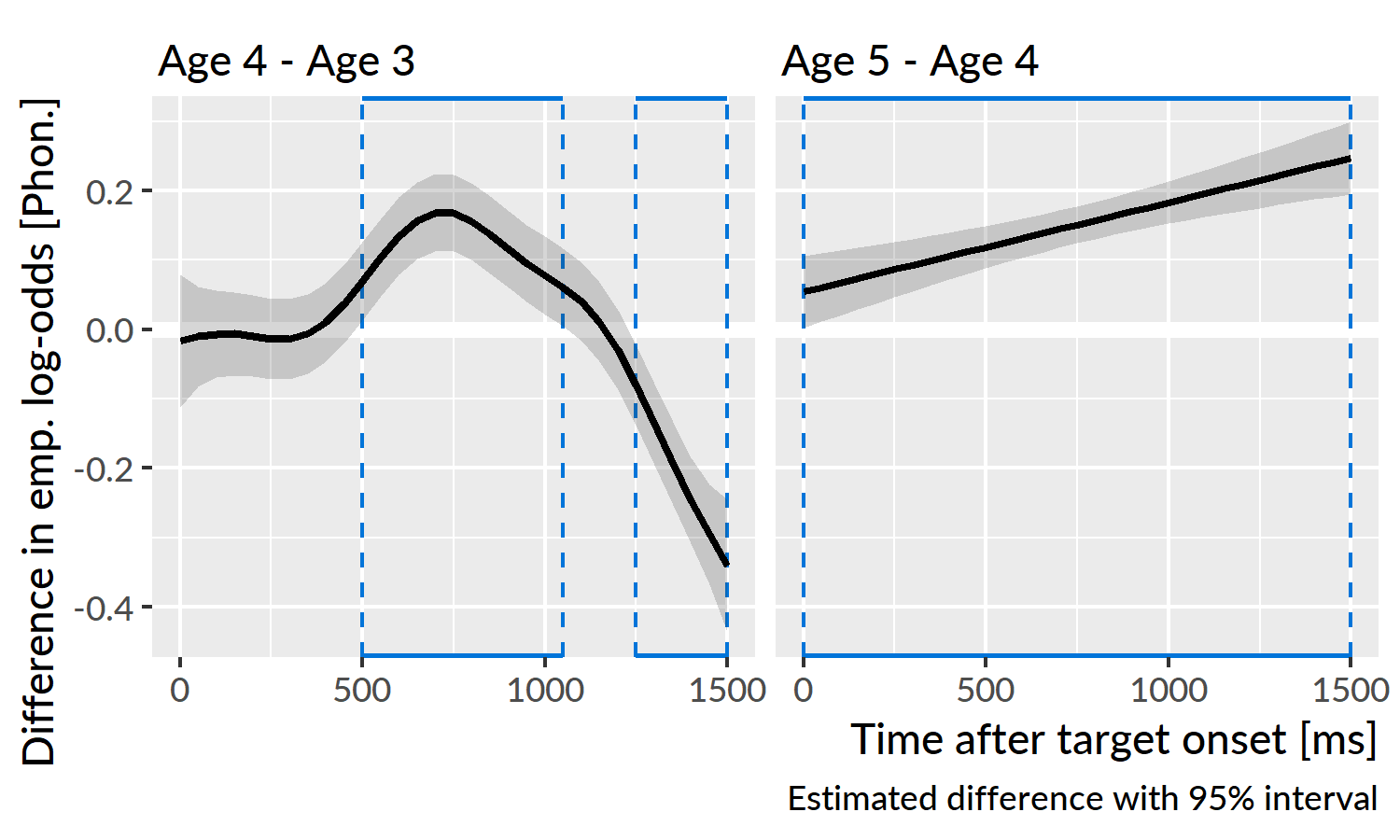


Figure 15 Differences in the average looks to the phonological competitor versus the unrelated image between age 4 and the other ages. Plotted line is estimated difference and the shaded region is the 95% confidence interval around that difference. Boxes highlight regions where the 95% interval excludes zero. From age 3 to age 4, children become more sensitive to the phonological foil during and after the target noun. The linear difference curve for age 4 versus age 5 indicates that the two years largely have the same curvature, but they steadily diverge over the course of the trial.

The age-3 and age-4 curves significantly differed, EDF = 5.48, *p* < .001. In particular, the curves are significantly different from 500 to 1050 ms. This result confirms that the looks to the phonological foil increased from age 3 and age 4 during the time window immediately following presentation of the noun and that children became more sensitive to the phonological similarities between the competitor and the target from age 3 to age 4.

The age-3 and age-4 curves also differed significantly after 1250 ms, so that at age 4 children looked less to the competitor compared to age 3. The effect reflects how the looks to phonological competitor decrease as a trial progresses. After an incorrect look to the foil, the children on average corrected their gaze and looked even less to the phonological foil. We do not observe this degree of correction during age 3, because children at age 3 looked less overall to the phonological foil early on.

The age-4 and age-5 smooths also significantly differed, EDF = 1.00, *p* < .001, although the low EDF values indicates that the shape of the difference was a flat line. Thus, the difference between the age-4 and age-5 smooths is driven primarily by the intercept difference and a linear diverging trend—that is, the distance between the two grows slowly over time. The same general curvature was observed for the two studies, suggesting the same general looking behavior at both time points: Children showed an early increase in looks to the phonological foil relative to the unrelated image but after receiving disqualifying information from the rest of the word, the looks to the phonological foil rapidly decrease. The primary difference between age-4 and age-5 is that the competitor effect becomes more pronounced at age 5.

**Summary**. Children looked more to the phonological competitor than the unrelated image early on in the trials. The advantage of the phonological competitor peaked on average around 800 ms after target onset, and the early timing indicates that children were shifting their gaze in response to the fleeting phonological similarity of the competitor to the target noun. The peak was small at age 3 but increased in height with each year of the study. Children became more sensitive to the phonological cohort competitors as they grew older.

## Looks to the semantic competitor

I asked how children’s sensitivity to the semantic competitor changed as they grew older. As in Law et al. ([2016](#ref-RWLPaper)), I only examined trials for which the semantic foil and the noun were part of the same category. For example, I included trials with *bee*–*fly*, *shirt*–*dress*, and *spoon*–*pan*, but I excluded trials where the similarity was perceptual (*sword*–*pen*) or too abstract (*swan*–*bee*). This criterion kept 13 of the 24 trials. [Appendix 16](#vw-experiment-items) provides a complete list of trials used.

For these trials, I used the same modeling technique as the one used for phonological competitors: Generalized additive models with study effects and a time smooth, time-by-study difference smooths, and time-by-child random smooths. I modeled the looks from 250 to 1800 ms. This window was 300 ms longer than the one used for the phonological competitors in order to capture late-occurring semantic effects.

The model’s fitted values are shown in Figure 16. The average empirical log-odds of fixating on the semantic competitor versus the unrelated word increased with each year of the study. All three years show the same general time course of effects: Looks begin to increase from a baseline around 750 ms and peak around 1300 ms. The peak proportions of looks to the semantic competitor versus the unrelated word increased as children grew older: The peaks were .65 at 1400 ms for age 3, .68 at 1400 ms for age 4, and .71 at 1350 ms for age 5. Moreover, the semantic competitor shows a decisive advantage over the unrelated image at age 3, in contrast to the limited advantage of the phonological competitor at age 3.



Figure 16 With each year of the study, children looked more to the semantic foil (relative to the unrelated image) with peak looking occurring after the target noun. Both figures show means for each year estimated by the generalized additive model. The left panel compares model estimates to observed means and standard errors, and the right panel visualizes estimated means and their 95% confidence intervals.

The average looks to the semantic foil over the unrelated image for age 4 was 0.44 emp. log-odds, .61 proportion units. Children looked significantly less to the semantic foil on average at age 3, 0.30 emp. log-odds, .57 proportion units, *p* < .001, and they looked significantly more to the semantic foil at age 5, 0.50 emp. log-odds, .62 proportion units, *p* < .001.

There was a significant smooth term for time at age 4, estimated degrees of freedom (EDF) = 7.04, *p* < .001. Figure 17 visualizes the time course of the differences between the smooths from each study.

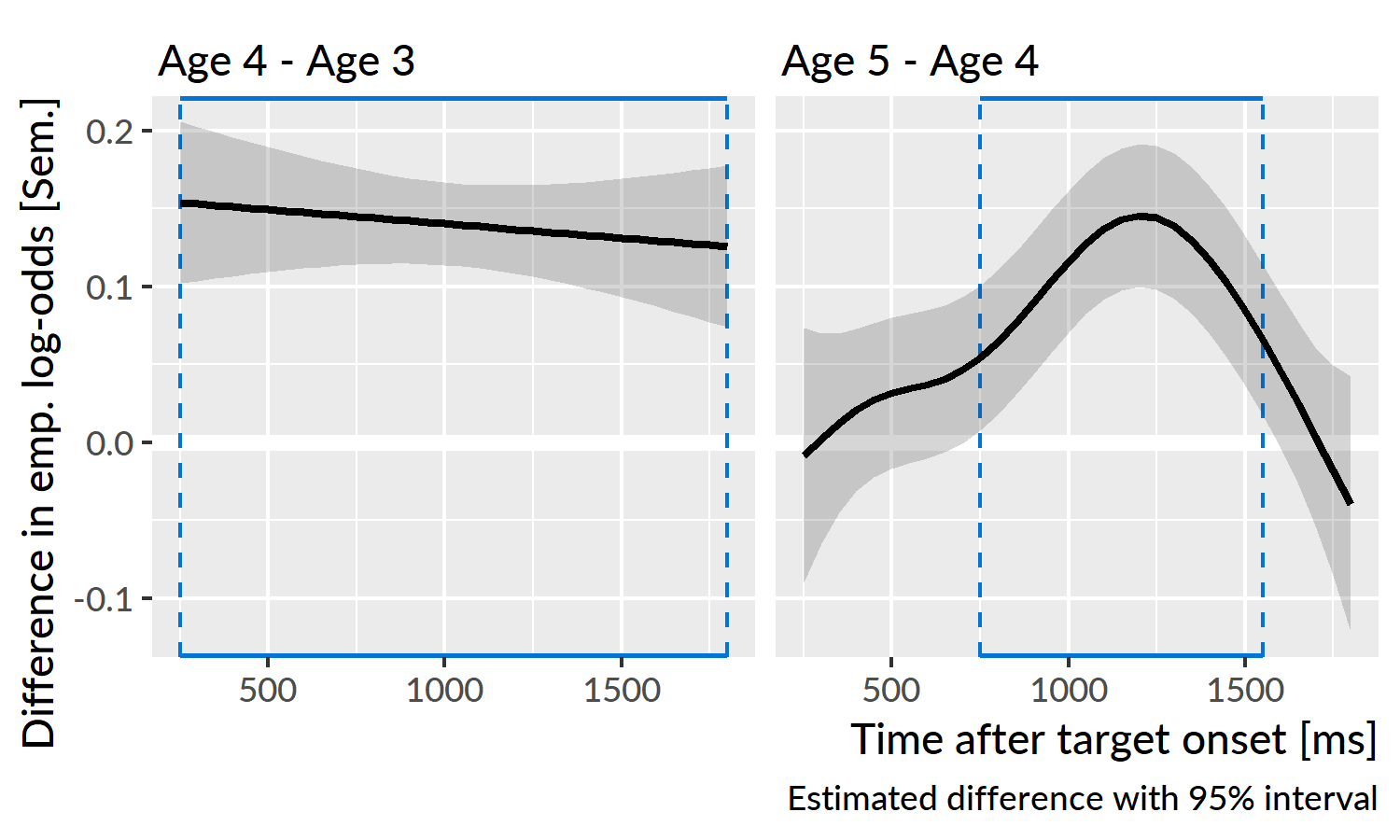


Figure 17 Differences in the average looks to the semantic competitor versus the unrelated word between age 4 and the other ages. Plotted line is estimated difference and the shaded region is the 95% confidence interval around that difference. Boxes highlight regions where the 95% interval excludes zero. The flat line on the left reflects how the shape of the growth curves remained the same from age 3 to age 4 and only differed in average height. From age 4 to age 5, the lines quickly diverge and the age-5 curve reaches a higher peak value.

The shapes of the age-3 and age-4 curves did not significantly differ, EDF = 1.00, *p* = .535. The age-3 curve begins to rise about 100 ms later, and it reaches a shallower peak value than the age-4 curve. These two features create a nearly constant height difference between the two curves, and thus the two curves show the same overall shape.

The age-4 and age-5 smooths significantly differed, EDF = 3.74, *p* < .001. The differences are greatest after the end of the target noun, in the window from 750 to 1500 ms. The two curves start from a similar baseline but quickly diverge as the age-5 curve reaches a higher peak value. After 1500 ms, the age-5 curve turns downwards to overlap with the age-4 curve. Children looked more to the semantic foil relative to the unrelated image, but they were also quicker to correct and look away from it.

**Summary.** Children became more sensitive to the semantic competitor, compared to the unrelated word, with each year of the study. The semantic foils clearly influenced looking patterns at age 3, in contrast to the muted effect observed for the phonological foils. The semantic effect also occurred when we would expect: After the end of the target noun, following the lexical activation of the target noun and its semantic neighbors.

## Child-level differences in competitor sensitivity at age 3

Next, I asked whether children differed reliably in their sensitivity to the phonological and semantic foils based on speech perception and vocabulary measures collected at age 3

As a measure of speech perception, I used scores from a minimal pair discrimination experiment administered during the first year of the study. [citations] The task is essentially an ABX discrimination task: A picture of a familiar object is shown and labeled (e.g., “car”), another object is shown and labeled (“jar”), and then both images are shown and one of the two is named. The child then indicated which word they heard by tapping on the image on a touch-screen.

I derived speech perception scores by fitting a hierarchical item-response model. This logistic regression model estimates the probability of child *i* correctly choosing word *j* on word-pair *k*. The equation below provides a term-by-term description of the model. The model’s intercept term represents the average participant’s probability of correctly answering for an average item. By-child random intercepts capture a child’s deviation from the overall average, so they estimate the child’s *ability*. By-word and by-word-in-pair random intercepts capture the relative *difficulty* of particular items on the experiment. The by-word-in-pair effects were necessary because four words appeared in more than one word pair (e.g., *juice*–*goose* and *juice*–*moose*). The model also controlled for the children’s ages and receptive vocabulary scores (PPVT-4 growth scale values). These predictors were transformed to have mean 0 and standard deviation 1, so the model’s intercept reflected a child of an average age and an average vocabulary level. Therefore, the by-child intercepts reflect a child’s ability after controlling for age and receptive vocabulary.

$$
\small
$$

I tested whether phonemic discrimination ability at age 3 predicted looks to the phonological competitor over the unrelated image by modifying the generalized additive model from earlier. In particular, I included a smooth term for the phonemic discrimination ability score and a “smooth interaction” between the smooth of time and phonemic ability. These smooth interaction terms are analogous to interaction terms in linear models. In this case, the interaction term allows the ability score to change the shape of the time trend. The additive model was therefore:

$$
\small
$$

The model included data from 144 participants; these were children with eyetracking data, receptive vocabulary and phonemic discrimination data at age 3. There was not a significant smooth effect for discrimination ability, EDF = 1.00, *p* = .551 or for an interaction smooth between time and ability, EDF = 8.37, *p* = .303.

To test the role of receptive vocabulary, I also fit analogous models using growth scale value scores from the PPVT-4, a receptive vocabulary test. I first adjusted these scores in a regression model to control for–that is, to partial out the effects of—age and predicted accuracy on the discrimination task. There was not a significant smooth effect for receptive vocabulary, EDF = 1.00, *p* = .868, or a significant interaction smooth between time and receptive vocabulary, EDF = 5.57, *p* = .610. Receptive vocabulary therefore was not related to looks to the phonological foil at age 3.

I tested the same two predictors on looks to the semantic foil at age 3. These child-level factors did not show any significant parametric effects, smooth effects or smooth interactions with time. Thus, children’s looks to the semantic foil were not reliably related to phonemic discrimination or receptive vocabulary.

**Summary**. These models tested whether two child-level factors—minimal-pair discrimination ability and receptive vocabulary—predicted looks to the phonological and semantic competitors at age 3. No significant effects were observed for all cases.

## Discussion

In the preceding analyses, I examined children’s fixation patterns to the phonological and semantic competitors and how these fixation patterns changed over developmental time. With each year of the study, children looked more to the target overall, so they consequently looked less to the competitor images each year. To account for this fact, these analyses examined the ratio of looks to the competitors versus the unrelated word. This ratio measured the relative advantage of a competitor over the unrelated word.

### Immediate activation of phonological neighbors

Developmentally, children became more sensitive to the phonological competitors with each year of the study. These words shared the same syllable onset as the target noun—for example, the pairs *dress*–*drum* or *fly*–*flag*. The competitors affected word recognition early on, with relative looks to the phonological foils peaking around 800 ms. The target nouns were approximately 800 ms in duration at age 3 and 550–800 ms at later ages. Assuming a 200–300 ms overhead for executing an eye movement in response to speech, this timing indicates that children shifted their gaze immediately, based on partial information. Moreover, the tendency to act on partial information became stronger with age, because the early advantage of the phonological competitor increased with each year of the study.

When children looked to the phonological competitor, they fixated on the wrong image and had to revise their interpretation of the noun. At ages 4 and 5, the early peaks of looks to the phonological competitor were followed by a steep, monotonic decrease in looks: Children rejected their initial interpretation of the word and considered other images. At age 3, the average pattern showed more wiggliness, suggesting that children were less decisive in rejecting the phonological competitor. The shapes of the looking patterns at age 4 and age 5 were essentially the same. In particular, the older children were not any faster in the rejecting the phonological competitor on average.

We can interpret these findings in terms of lexical processing dynamics. Under this view, incoming speech activates phonetic and phonemic and lexical representations. The word with the strongest activation is the favored interpretation and the object of the child’s fixations. The early looks to the phonological competitors reflect immediate activation of lexical units: Children activate words on the basis of partial acoustic information. This result is a hallmark of spoken word recognition. The activation of phonologically plausible words becomes stronger with age, as reflected in children’s increasing sensitivity to the phonological competitors. Some mechanisms that may explain this developmental pattern include changes in lexical organization so that neighborhoods of phonologically similar words coactivate and changes in lexical representations so that partial information can more eagerly activate compatible words.[[6]](#footnote-78)

Children at age 4 and age 5 did not show any changes in how quickly they rejected the phonological foil, and this result suggests that lexical inhibition may not change over the preschool years. The reasoning is as follows: If children developed stronger lexical inhibition with age, so that lexical competition resolves more quickly, then we would expect activation of the phonological competitors to decay more quickly and for children to reject the phonological competitor more quickly. But this pattern is not what we observed in the growth curve analyses.[[7]](#footnote-79) The developmental trajectory here is one of increased activation, of children learning words and learning similarities among them so that phonological similar words participate in word recognition.

### Late activation of semantic neighbors

The semantic competitors were from the same category as the target noun: for example, *bee*–*fly* or *shirt*–*dress*. Children showed year-over-year increases in their sensitivity to the semantic competitor, compared to the unrelated image. Looks the semantic foils started rising steadily 500–700 ms after target onset and peaked late in the trial, around 1300 ms. This time-course is more protracted than the immediate peaks observed for the phonological competitor.

In terms of lexical processing, this late timing is consistent with cascading activation: Spoken words immediately activate phonological neighborhoods with activation cascading onto semantically related words. As a particular word is favored, its semantic relatives receive more secondary activation. For example, children hear “find the shirt”, activate the target *shirt*, but also activate other pieces of clothing including *dress*. The late timing of looks to the semantic competitor therefore reflects late, secondary activation of the spoken word’s semantic relatives. In other words, the activation of a semantic neighbor (like *dress*) is greatest when the activation of the spoken word (*shirt*) is greatest which happens relatively late, once the competition among phonological alternatives resolves.

Under this account, children hear a word, activate it, and become increasingly likely to fixate on the semantic competitor, compared to the unrelated image. The late looks probably reflect a combination of behaviors: children considering the semantically related image to check their initial interpretation as well as children looking to the wrong image because of confusion, lack of knowledge, overriding activation from the semantic competitor, or lack of interest in the target.

Initially, I had subscribed to a confusion or lack-of-knowledge interpretation of the semantic competitor’s advantage. That is, children look to the semantic competitor because they do not know the difference between the target and the semantic competitor. After all, my thinking went, these were young children and decisions like *bee* vs. *fly* or *goat* vs. *sheep* can be difficult. But there are two objections to that line of reasoning. First, our lab piloted the set of words in preschool classrooms, so we confirmed that children could reliably and correctly point to *bee* even when *fly* is an alternative. Second, we would a priori expect that children’s confusion among words to be greatest when they are youngest and have much less experience with these semantic categories. (Indeed, children at age 3 looked less to the target overall, so in general, they were less successful at recognizing the target word.)

The late looks to the semantic competitor, relative to the unrelated image, however, were greatest at age 5. Children’s looks became more selective with age: They looked more to the semantic competitor because they had discovered the semantic connections among words. They had learned the similarity between *bee* and *fly* or *shirt* and *dress*. Put another way, to demonstrate confusion between two choices, children must learn some association that connects the two; they must use or activate some information that induces warranted uncertainty. Rather than confusion about the meaning of nouns, the late looks likely reflect a confirmatory behavior where children give some consideration to the semantic alternative. This is especially the case at age 5, where the advantage of semantic competitor quickly decreases after its peak, indicating rejection of the semantic competitor.

### Lexical competitors and child-level predictors

I asked whether offline child-level measures predicted sensitivity to the phonological and semantic competitors at age 3. I used children’s ability scores from a minimal-pair discrimination task as a measure of phonemic speech perception, and I also used scores from a receptive vocabulary test. For the phonological competitor, I expected that children with better phonemic discrimination would show increased looks to the phonological competitor because they had more detailed phonemic representations that would activate phonological neighborhoods more quickly. For the semantic competitor, I likewise expected children with larger receptive vocabularies to show increased looks to the semantic competitor because these children knew more words and likely developed more semantic connections among the words. I tested these effects by using the scores as parametric effects to see if they predicted average looks to the competitor, and alternatively, by using the scores for smooth effects to see if they influenced the time course of looks to the foils.

None of these expectations held: Neither of the child-level measures predicted average sensitivity to the phonological or semantic competitors at age 3. Part of the result may be artifactual: The data—looks to a subset of images on a subset of trials—may be too limited at the individual level for the models to pick up on child-level effects. Part of the result may be developmental too: Children were least sensitive to the competitors at age 3, so individual differences may be too small for the data or models to capture. Further work, with different experimental designs, may elaborate on whether offline measures can reliably detect differences in sensitivity to lexical competitors during word recognition.

# General discussion

This study examined the development of familiar word recognition over the preschool years. The word recognition data came from a visual-world eyetracking experiment which recorded children’s fixations to images in response to prompts like *see the bear*. The trials featured a target noun (e.g., *bear*) along with a phonological competitor (*bell*), a semantic competitor (*horse*), and an unrelated image (*ring*). To describe children’s word recognition ability, I analyzed how the probability of fixating on the target image changed over the time course of a trial. The presence of the competitor images also allowed additional analyses about children’s sensitivities to the phonological and semantic competitors. The experiment was conducted as part of a three-year longitudinal study; children were 28–39 months-old at the Age 3 visit, 39–52 at Age 4, and 51–65 at Age 5. The longitudinal design allowed me to describe developmental changes in word recognition.

## How to improve word recognition

Children showed year-over-year improvements in word recognition, as measured by average looking probabilities, peak looking probabilities, and the rate of change in looking probabilities. Children became more reliable, less uncertain, and faster at recognizing familiar words as they grew older. At the same time, children also became more sensitive to the phonological and semantic competitors, compared to the unrelated images. With each year, children looked more to the target image, but when they erred, they were more likely to err on a lexically relevant word.

We can interpret these developmental patterns in terms of lexical activation and processing dynamics. In this task, children hear a stream of speech and activate some phonetic, phonological, lexical, and semantic representations that match the speech input. As they hear more of a word, the activation builds until a particular word is favored, and children shift their gaze onto the named image. Let’s imagine that we have to engineer this system. To make word recognition more efficient, we have to find ways to increase the relative activation of the correct word. In particular, we can boost the strength of connections so that activation can propagate more quickly through the system, and we can also allow inhibition among competing words so that the correct word can win out over its competitors more quickly.

The results from these studies indicate that children become more efficient at activating the target word *and related words* over the preschool years. As they grew older, children were faster to look at a named image and more likely to fixate on the phonological competitor (compared to the unrelated image). These two findings reflect changes in how partial acoustic information can propagate to activate phonologically plausible words. The phonological competitors shared the same syllable onset as the target noun (e.g., *dress*–*drum*), so the early part of the word matched both words. That children became more sensitive to the phonological competitor means that they learned and somehow encoded the phonological similarities among words because part of a word could activate a neighborhood of phonologically plausible matches. This developmental change supports faster word recognition because the listener can channel activation to relevant words more quickly. A similar line of reasoning applies to the semantic competitors: Relative looks to the semantic competitors increased with age, suggesting that children had learned semantic connections among words and activated semantically related words during word recognition.

The other mechanism we might tune to improve word recognition is inhibition. Children’s looks to the phonological or semantic competitors were temporary: Looks increase to some peak level and then quickly decrease. Behaviorally, the drop in looking probability reflects the rejection of an interpretation: for example, a child hears “dr”, shifts looks to *dress*, but hears “um”, revises the interpretation and jumps to *drum*. We can read these corrections as evidence for an inhibitory process: Corrections indicate a change in relative activation where a different word overrides an initial interpretation. But the evidence for *developmental changes* in lexical inhibition from these data was scant. The rate of rejection of the phonological competitor—that is, how quickly looks fall from their peak value—did not change from age 4 to age 5, although the rate did increase for the semantic competitor from age 4 to age 5. Preschoolers did demonstrate inhibition by revising their interpretations of nouns, but there were no clear developmental changes in inhibition.

Previous simulation work can help identify more specific mechanisms at play. McMurray, Samelson, Lee, and Tomblin ([2010](#ref-McMurray2010)) used the TRACE model of word recognition (McClelland & Elman, [1986](#ref-TRACE)) to simulate looks to a target and phonological competitors (cohorts and rimes) in adolescents with specific language impairment. The authors tuned a number of model parameters and analyzed how those changes affected simulated looks to the target and competitors. In the current dataset, I observed a developmental trend where the relative looks to the phonological competitors peak higher each year. In those TRACE simulations, looks to the cohort competitor peak higher if 1) the rate of lexical activation increased, 2) the rate of lexical decay decreased, or 3) strength of lexical inhibition decreased. Of these options, the growth curve for the decrease in lexical inhibition best matches the shape of the current data. The similarity does not mean that children inhibited words any less as they grew older. That would be too simplistic: Developmental changes in preschoolers are the result of simultaneous changes in many mechanisms. But those simulation results suggest that an *increase* in lexical inhibition is *not* one of the key developmental changes in preschoolers’ word recognition.

## Learn words and learn connections between words

Preschoolers showed increased activation of the target noun and semantically and phonologically related words but little developmental change in lexical inhibition. Paired with the findings from older children, these results lead to a compelling developmental story. Rigler et al. ([2015](#ref-Rigler2015)) compared 9- and 16-year-olds on a visual-world word recognition experiment with phonological (cohort and rime) competitors. The younger children were slower to look to the target image and showed more looks to the competitors. The implications are that children’s word recognition is still developing in late childhood and that in particular, children’s inhibition of lexical competitors became stronger with age.

The current study with 3-, 4-, and 5-year-olds followed a different pattern: Relative looks to the competitor images increased with age. Taken together, these two studies suggest an interesting progression for the development of lexical processing. During the preschool years, children learn many, many words, and they establish phonological and semantic connections between these words. These connections support the immediate activation of neighborhoods of related words. Later childhood, based on the Rigler et al. ([2015](#ref-Rigler2015)) findings, then is a time for refinement of those connections so that sensitivity to the competitors decreases. This refinement could follow from more selective activation channels, increased lexical inhibition, changes in resting activation (to favor more frequent words), or likely a combination of these factors.

## Individual differences are most important at younger ages

Another dimension of this study concerned individual differences in word recognition. Some children were faster or more accurate during word recognition, and these children also were more likely to be faster or more accurate at later ages. The magnitude of these differences diminished over time, as children approached a more mature level of performance.

In terms of lexical processing dynamics, we might think of early differences as reflecting early differences in the burgeoning lexicon. Children may have different numbers of words, different degrees of experience with some words, less established connections among words, and at a lower level, different phonetic and speech perception abilities, given the links between speech perception in infancy and early vocabulary development (Cristia, Seidl, Junge, Soderstrom, & Hagoort, [2014](#ref-Cristia2014_Review)). Differences in word recognition are greatest early on in development because this is when the differences among children’s lexicons are greatest. The task of learning new words, and more importantly, of developing representations and associations to organize words normalizes the early differences among children’s lexicons. That pressure would make the overall variability among children decrease over time while still preserving a relative ordering among children.

We can also interpret the predictive power of word recognition measures in terms of lexical processing and lexical organization. Correlations between word recognition performance and future vocabulary were strongest for the age-3 growth curve features, particularly for the peak probability of looking to the target word. The peak probability measures the overall certainty in word recognition and how strongly the target word is activated. Children with more efficient representations of familiar words at age 3 have a stronger foundation for encoding and integrating future words, and as a result, they showed larger vocabularies at age 4 and age 5.

Initially, I had expected processing *speed*—as approximated by growth curve slopes—to be the most predictive measure of vocabulary growth. Children who can more quickly recognize words, the reasoning goes, can take in information more quickly and devote extra processing resources towards learning.[[8]](#footnote-86) Processing speed was indeed correlated with future vocabulary size, yet peak probability was a stronger predictor of future vocabulary size. Granted, these two processing measures are highly related; to hit a higher peak by time *x*, a growth curve needs to start from higher baseline or have steeper slope. The idea of uncertainty suggests an alternative explanation of the predictive power of word recognition: Children who are more accurate (or less uncertain) during word recognition can extract and activate *more information* from the speech signal.

## Limitations and implications

The discussion of processing speed and word recognition certainty highlights one limitation of this research: The experiment’s four-image, eyetracking-based design meant that a clean measure of processing speed was not feasible. Other eyetracking studies with two images can use the latency of how long it takes the child to shift between images as a measure like reaction time. This approach does not translate to the four-image design, as children can visit multiple images on their way to the target. Visual world studies with older participants can obtain an explicit reaction time measure by means of a mouse click or tap on a touchscreen, but those additional task demands may not translate to young children like those in this study. Thus, this study could not address directly whether the predictive power of word recognition performance reflects a more developed lexicon, a general reaction-time-like speed advantage, or both.

The experimental design included semantic and phonological competitors on every trial, so isolating out the semantic and phonological competition effects required some subtlety. A more direct design would compare different types of a trials: for example, trials with a target vs. three unrelated images intermixed with trials with a target vs. a competitor vs. two unrelated images. The trials also used different kinds of phonological and semantic competitors. For example, two of the phonological competitors rhymed with the target, so they could not be included the analysis of phonological competitors (which focused on just competitors with the same onset as the target). The current design limited the number of trials that could be used in the analyses of the competitors and weakened the power of the analyses.

A final limitation includes the changes in the experiment procedure over the course of the longitudinal study. From age 3 to age 4, we re-recorded the stimuli (with the same original speakers) so that the noun durations between the two different dialect versions of the experiment were similar. From age 4 to age 5, we also shortened the duration of the trials by removing attention-getting prompts (e.g., *this is fun!*) from the ends of the trials. These small procedural changes mean that year-to-year differences do not reflect *pure* development differences. It is implausible, however, that the robust year-over-changes owe more to procedural changes than a year of learning and language development.

The findings from this study have implications for our understanding of word recognition and word learning. The first is the overall developmental narrative. Preschool children become better at recognizing words by learning similarities among words and using those similarities to activate neighborhoods of lexically relevant words. Rather than just measuring vocabulary size, word recognition reveals how well words have been integrated into the lexicon. The developmental trends here show that familiar words become more integrated and more connected over the preschool years. Even if a child knows a word at age 3 well enough to recognize or express it, their knowledge of the word will strengthen over time as the word develops connections to other similar words.

From this perspective, we can think of individual differences in word recognition as differences in lexical development. Variability in word recognition diminishes over time, so that differences are more predictive and discriminating at younger ages. Thus, if we wanted to intervene on word recognition, these results indicate that early intervention is better and that intervention should build connections among words and should target words that build onto existing semantic and phonological networks. The natural closing of gaps in word recognition performance with age, however, suggests that word recognition in and of itself may not be an important intervention target. Rather, word recognition measures should serve to supplement other vocabulary measures as an indicator of lexical processing and lexical integration.

# Hypothesis check (Specific Aim 1)

Here I revisit my pre-analysis hypotheses.

*Children’s accuracy and efficiency of recognizing words will improve each year.*

Yes. Their curves reached higher heights and showed steeper slopes each year.

*There are stable individual differences in lexical processing of familiar words such that children who are relatively fast at age 3 remain relatively fast at age 4 and age 5.*

Yes. The rankings of children by lexical processing measures (peak probability, average probability, linear slope) were concordant over the three years.

*However, the magnitude of these individual differences diminishes over time, as children converge on a mature level of performance for this paradigm.*

Yes. I simulated new longitudinal participants based on what the model learned about the observed children. The range of plausible looking proportions narrowed each year, so individual differences became less variable each year.

*Consequently, individual differences in word recognition at age 3, for example, will be more discriminating and predictive of age 5 language outcomes than differences at age 4 or age 5.*

Yes. Correlations between growth curve features with future vocabulary measures were strongest for the age 3 growth curve features.

*Children will become more sensitive to lexical competitors as they age, based on the hypothesis that children discover similarities among words as a consequence of learning more and more words.*

Yes. The advantage of the phonological competitor and semantic competitor over the unrelated word with development.

*Children will differ in their sensitivity to lexical competitors, and these individual differences will correlate with other child-level measures.*

No evidence to support or refute this hypothesis. I did not find a relationship between age-3 measures with the phonological or semantic competitors. In principle, however, one could derive a measure from the competitor looking curves that does correlate with child-level measures.

# (PART\*) Aim 2: Referent Selection and Mispronunciations

# Mispronunciations and referent selection

In the earlier chapters, I studied word recognition by examining how young listeners recognized familiar words. But children do not know all the words they encounter, and another avenue for studying word recognition is to examine how listeners respond to unfamiliar or novel stimuli. This study looks at how children responded to mispronunciations and novel words in a two-image word-recognition experiment.

## How phonetically detailed are children’s words?

There has been long, productive line of research examining how children respond to mispronunciations of familiar words. The motivation for this research was to determine how detailed children’s phonological representations are. One hypothesis held that infants and toddlers do not need to store words in much phonetic or phonological detail because they know so few words. In other words, their lexicon had *underspecified* or *holistic* representations.

A common argument for underspecified or holistic representations poses the building a lexicon as a design or engineering problem. It considers the question: What would be a smart way to build up a lexicon? One solution is that the word-learning system should be lazy, encoding words in just enough phonetic detail to differentiate among them and adding more details as the need arises. Early words should be underspecified and pick up details on demand. Why fully encode the phonetic form of “doggie” when there is no competition from similar words like “toggie” or “dokkie” or “tokkie”? An efficient solution would be to encode a minimal amount of phonetic detail. In fact, this strategy could be developmentally advantageous: “Perhaps children benefit from the sparseness of their lexicon by encoding only the detail necessary to distinguish words” (Swingley & Aslin, [2000](#ref-Swingley2000), p. 148).

Charles-Luce and Luce ([1990](#ref-charlesluceluce1990), [1995](#ref-charlesluceluce1995)) are commonly cited touchstones for this argument. Charles-Luce and Luce compared the expressive ([1990](#ref-charlesluceluce1990)) and receptive ([1995](#ref-charlesluceluce1995)) lexicons of 5- and 7-year-olds against those of adults. They observed that adults had much denser phonological neighborhoods than children, and they suggested that children may only have holistic representation of words given these sparser neighborhoods. Dollaghan ([1994](#ref-dollaghan1994)) rebutted the 1990 study, observing that kids have sparser neighborhoods because they have sparse lexicons. Dollaghan ([1994](#ref-dollaghan1994)) also showed that young children do indeed have dense neighborhoods in their lexicons. Coady and Aslin ([2003](#ref-coadyaslin2003)) elaborated this claim, observing that children’s lexicons would be comparatively denser early on in development if a child’s first words are made up of more common sounds and word shapes.

Structural studies of lexicons of this sort are rather limited. They describe the knowledge to be learned instead of the content of the representations throughout development. More direct evidence comes from studies where children have difficulty learning minimal pairs. For example, Barton ([1976](#ref-barton1976phonemic)) found that 27–35-month-olds who, say, knew the words *bear* and *pear* could differentiate them successfully (at approximately 90% accuracy), but when the children had to learn one of the words, then they were less successful at differentiating them (50–60%). The discrepancy invites a conclusion that newly learned words are underspecified.

The Switch Task, studied extensively by Werker and colleagues, also presented evidence where young children were unable to learn a minimal pair in the lab. In the classic switch paradigm, a child is habituated to the ostensive naming of two novel-object/novel-word associations. In other words, they see a novel object and hear a paired novel word (*liff*) and see different novel object with a different paired nonword (*neem*). Once the child is habituated, there is a critical switch trial where the *liff*-object is displayed but labeled with the other nonword *neem*. If the child looks longer on the switch, then we infer that they detected the change and pay more attention because their expectations were violated. This paradigm fits into the debate about representations because 14-month-olds could detect the *liff*-*neem* switch (Werker, Cohen, Lloyd, Casasola, & Stager, [1998](#ref-werker1998acquisition)) but *not* a minimal pair *bih*-*dih* switch (Stager & Werker, [1997](#ref-stager1997infants)).

One interpretation of these results is that children may have underspecified representations for recently learned words. That reading begs the question, however, of whether a child has actually “learned” the word or just has an inchoate, less-well learned representation from a few laboratory exposures to a nonword. In Fennell and Waxman ([2010](#ref-Fennell2010)), 14-month-olds were able to detect a *bin*/*din* switch when the nonwords were treated as words. That is, during familiarization the words were embedded in sentence prompts like *Look. It’s the din.* or *Do you see the din?* By changing the task in a way that makes word-learning easier, the children encoded more phonetic detail about the words. This suggests that the challenge of learning minimal-pair words seems to have more to do with the difficulty of word learning rather than with how the known words are stored.

Against that backdrop of lexicon design strategies and minimal-pair training studies, mispronunciation studies provide a rather direct way to study the phonetic representations of the words that children know. Swingley and Aslin ([2000](#ref-Swingley2000)) presented 18–23-month-olds two familiar images onscreen, like baby and dog, and heard a correct production (*where’s the baby*) or a mispronunciation (*where’s the vaby*). Children looked to the correct productions about 73% of the time and mispronunciations 61% of the time. They looked more than chance but less than the correct production, indicating they were sensitive to the mispronunciation. Children had encoded *baby* in enough phonetic detail that a small phonetic change made them less certain during word recognition. (Swingley and Aslin ([2002](#ref-Swingley2002)) found the same pattern of results for 14–15-month-olds with 60% looks to correct productions and around 53.5% looks for mispronunciations.)

A similar study by Bailey and Plunkett ([2002](#ref-Bailey2002)) tackled the representations of recently learned words. They created custom word-lists for children and included mispronunciations for words that the child purportedly learned long before testing and only recently before testing. They did not find a difference between the two types of words, suggesting that recently learned words were as well specified as earlier learned words.

One limitation with the Swingley and Aslin ([2000](#ref-Swingley2000)) design is that the child has no way to reject *vaby*. It could be that children might treat *vaby* a completely novel word, but they have to choose either *baby* or *dog* so they look at the image that rhymes with the *vaby*. White and Morgan ([2008](#ref-WhiteMorgan2008)) updated the paradigm to allow for these kind of rejections. They presented toddlers with images of a familiar object and a novel object, and children heard a correct production of the familiar object, mispronunciations of the familiar object of varying severity, or an unrelated nonword. Toddlers looked less to a familiar word when the first segment was mispronounced, so they did not treat the mispronunciations as nonwords. The children demonstrated graded sensitivity such that a 1-feature mispronunciation yielded more looks to a familiar image than a 2-feature mispronunciation, and a 2-feature mispronunciation yielded more looks than a 3-feature one. Finally, in the nonword condition, the children looked more to the novel object than the familiar one, demonstrating *fast referent selection* as they associated novel words to novel objects in the moment. In this case, mispronunciations can vary in severity and children’s responses to them will vary in turn.

Law and Edwards ([2015](#ref-MPPaper)) applied this study to preschool-age children, observing a similar pattern of effects: Preschoolers mapped real words to familiar objects, mapped nonwords to novel objects, and equivocated about mispronunciations of familiar words. They also found that the child’s vocabulary size was related to these looking behaviors such children with larger vocabularies looked more to the target in the real word and nonword trials and looked less to the familiar object in the mispronunciation trials.

## How to handle nonwords

Mispronunciations are not the only nonwords a young child might hear. In fact, if a child knows very few words, we expect them to bombarded by new and novel words. There has been a great deal of research on how children handle nonwords, especially when paired with a novel object as its referent (Novel Name–Nameless Category principle, B. & Jacquelyn, n.d.; mutual exclusivity principle, Markman & Wachtel, [1988](#ref-Markman1988)). The nonword trials in studies like Law and Edwards ([2015](#ref-MPPaper)) and White and Morgan ([2008](#ref-WhiteMorgan2008)) can shed light on other aspects of word recognition.

Children have a strong novelty bias when they hear nonwords. In Horst, Samuelson, Kucker, and McMurray ([2011](#ref-Horst2011)), two-year-olds were familiarized to novel objects. They were later tested with a prompt to select a novel object (*Can you get the fode?*) from three choices: two familiarized novel objects and one new unfamiliarized *super-novel* object. They demonstrated a clear preference for the super-novel object. Mather and Plunkett ([2012](#ref-Mather2012)) replicated the preference a super-novel during a word recognition eyetracking task. In this case, 22-month-old English-learning toddlers were pre-exposed to images of novel objects. Late, during a test trial, a familiar object, a familiarized novel object, and new unseen super-novel object appeared onscreen with a prompt to view a nonword (*Look at the gub! Look! Gub!*). Children looked to the novel object more than the other two objects. In a second experiment, they removed the familiar object leaving just the familiarized and super-novel objects. The advantage of the super-novel object was replicated but it only emerged after the third repetition of the trial.

A robust novelty bias raises the question of whether listener’s comprehension of familiar words and interpretation nonwords reflect different processes. McMurray et al. ([2012](#ref-McMurray2012)) propose that the same basic process is at play in both recognition of familiar words and fast association of nonwords. After all, in the lab, the observed behaviors are the same: Children hear a word (be it a real word or nonword) and direct their attention to an appropriate referent. “To the extent that a word links sound and meaning, any time that link is used to guide behavior, a word is being used. Thus, word use also includes processes such as comprehending known words, and even determining referents for new words” (McMurray et al., [2012](#ref-McMurray2012), p. 832).

Bion, Borovsky, and Fernald ([2013](#ref-Bion2013)) tested referent selection of nonwords and real words in 18-, 24-, and 30-month-olds. In the article’s second experiment, toddlers were trained on two novel words on disambiguation trials. They would see a familiar object and an unfamiliar object and heard a prompt with a nonword (*Where’s the dofa?*). Later retention trials presented the two unfamiliar objects were presented and prompted (*Where’s the dofa?*). Mixed in with these trials were familiar-object trials in which a familiar object was labeled (*Where’s the car?*). In that experiment, children looked more to the target on the familiar word trials than on the nonword disambiguation trials. For the 30-month-olds: 82% versus 68% percentage of looks to the target.

For Bion et al. ([2013](#ref-Bion2013)), toddlers performed better on the familiar-word trials than the novel-word trials. But if we think of the nonwords as just much less familiar words, then this result is wholly consistent with the idea that the same process operates in both familiar word recognition and nonword referent selection. The authors, interestingly, make a point to note that fast referent selection is not necessary for word-learning: “Those 18-month-olds whose accuracy scores on Disambiguation trials were lower than [*chance*] were reported to produce as many as 389 words …. those 24-month-olds who failed to show a disambiguation bias produced as many as 417 words”. In other words, a toddler may purportedly know hundreds of words but still not reliably look to a novel object given a novel label. This finding raises the possibility that nonword referent selection is not a guaranteed behavior in young children.

## The current study

As with lexical competition, it is unclear how children’s responses to mispronunciations and novel words change over time. For example, do children become more forgiving of mispronunciations as they mature and learn more words? Do familiar word recognition and nonword referent selection ever dissociate? Relatedly, is one of these behaviors more related to future word learning?

In this study, I report the results of a longitudinal study of word recognition in preschoolers at age 3, age 4, and age 5. The particular experiment here was a mispronunciation study following the paradigms of White and Morgan ([2008](#ref-WhiteMorgan2008)) and Law and Edwards ([2015](#ref-MPPaper)). Children saw a familiar object and unfamiliar object and heard either a real word (*shoes*), a one-feature mispronunciation of the word (*suze*), or a nonword (*geeve*). The study is described in detail in [Chapter 10](#aim2-method).

In [Chapter 11](#real-nonword-selection), I examine children’s development of referent selection in unambiguous contexts by comparing their performance in the real word and nonword conditions. Of interest is whether real word and nonword processing follow similar developmental trajectories. I expect the two to be highly related, but if they ever dissociate, it should happen with younger children. At face value, one might expect a child’s ability to associate new words with unfamiliar objects to be more direct measure of word-learning capacity than a child’s ability to process known words. Under this assumption, I predict that nonword referent selection will be a better measure of later vocabulary growth than familiar word recognition.

In [Chapter 12](#sensitivity-to-mispronunciations), I study how children’s responses to mispronunciations changed with age. From the literature review above, I expect preschoolers to treat the mispronunciations as passable but still flawed productions of known words. As for development, I expect children to become more tolerant of mispronunciations, based on the assumption that they become more experienced at listening to noisy, degraded, or misspoken speech. I also report data from age 5 where we tested children’s retention of the novel images paired with the nonwords and mispronunciations.

In [Appendix 20](#aim2-mp-items), I briefly present the results for specific mispronunciations, although item effects are not formally modeled. Finally, in [Chapter 13](#aim2-discussion), I describe the both sets of analyses together, and [Chapter 14](#aim2-h-check) reviews the results of my pre-analysis research hypotheses.

# Method

Data collection for this experiment occurred during the same longitudinal study as the familiar word recognition experiment of Aim 1. Thus, this experiment used the same participants, the same stimulus preparation and style, and the same general experimental procedure as those reported in detail in [Chapter 4](#aim1-method).

## Mispronunciation task

This experiment is an adaptation of the mispronunciation detection task by White and Morgan ([2008](#ref-WhiteMorgan2008)) and Law and Edwards ([2015](#ref-MPPaper)). In the experiment, two images are presented onscreen—a familiar object and an unfamiliar object—and the child hears a prompt to view one of the images. In the *correct pronunciation* (or *real word*) and *mispronunciation* conditions, the child hears either the familiar word (e.g., *soup*) or a one-feature mispronunciation of the first consonant of the target word (*shoup*). These conditions are designed to test whether children associate mispronunciations with novel objects. To encourage fast referent selection, there were also trials in a *nonword* condition where the label was an unambiguous novel word (e.g., *cheem* presented with images of a bed and a novel-looking pastry mixer). Each nonword was constructed to match the phonotactic probability of one of the mispronunciations. Figure 18 shows the screens used in two trials. Importantly, within a block of trials, the child never hears both the correct and mispronounced forms of the same word. A child hearing “duck” then a few trials later hearing “guck” would provide a basis of comparison so that the child can decide that “guck” is probably not “duck”—the design used here avoided this situation and is a change from the design of Law and Edwards ([2015](#ref-MPPaper)).

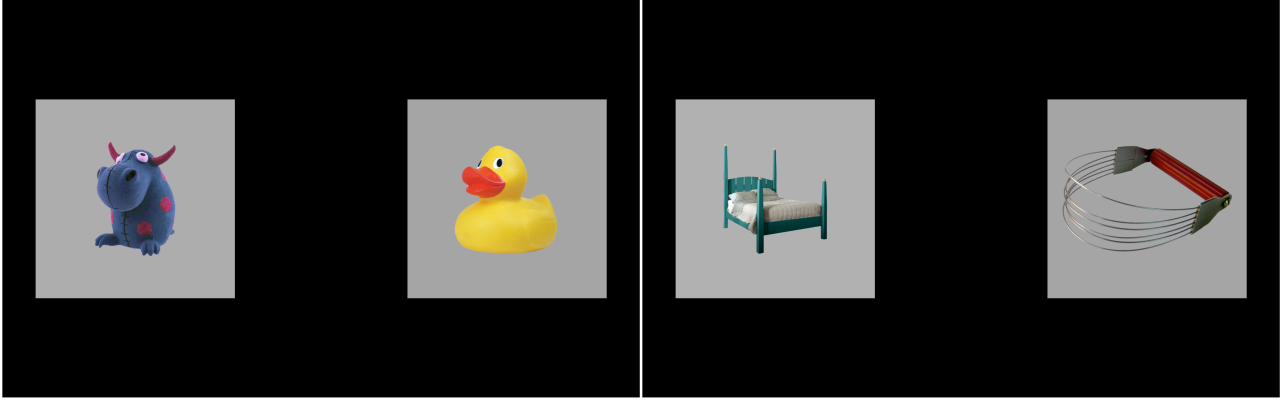


Figure 18 Example displays for a trial in which *duck* is mispronounced as “guck” (*left*) and a trial in which the nonword “cheem” is presented (*right*).

In a block of the experiment, there were 12 trials each from the correct production, mispronunciation, and nonword conditions, and children received two blocks of the task. A complete list of the items used in the experiment over the three years of the study is included in [Appendix 18](#mp-experiment-items).

## Visual stimuli

The images used in the experiment consisted of color photographs on gray backgrounds. As in the familiar word recognition experiment, these images were piloted in two preschool classrooms. Piloting confirmed that children consistently used the same label for familiar objects. For the novel objects, the children reported to not know a word for the object, or if they did name the object, they did not consistently use the same word for an object.

## Novel word retention tests

At age 5, following the second block of this task, we tested children’s retention of the labels for the novel objects. They were first tested using an open-set procedure: They were shown each of the images and asked to name it. I will not analyze or report those results, because children seldom named the novel objects using the labels from the task. For example, the rainbow-filled flasks used for *sooze* (mispronounced *shoes*) were called *science*, *potions*, *magic*, *bottles*, among other labels.

Following the open-set naming test, children had a closed-set recognition test. Two of the novel objects were paired. One of the objects was from a mispronunciation trial and the other was from a nonword trial. For example, the flasks (*sooze*) were paired with the trolley (*geeve*) from the nonword condition. The pairs were yoked, as each nonword was designed to have the same phonotactic probability as one of the mispronunciations. During the retention test, children saw two images of the novel objects (say, *trolley1.jpeg* and *flasks1.jpeg*), heard one of novel labels (*geeve*), and had to point to the named object. In a later trial during testing, the other label (*sooze*) was tested but using different images for the objects (*trolley2.jpeg* and *flasks2.jpeg*). In a block of testing, there were 12 trials, 6 for nonwords and 6 for mispronunciations.

## Data screening

Table 3 shows the numbers of participants and trials excluded during each of year of the study due to unreliable data. There were more children in the second year than the first year due to a timing error in the initial version of this experiment, leading to the exclusion of 30 participants from the first year.

Table 3 Eyetracking data counts before and after data screening.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Study | Children | Blocks | Trials |
| Raw | Age 3 | 177 | 341 | 12245 |
|  | Age 4 | 181 | 349 | 12600 |
|  | Age 5 | 164 | 325 | 11736 |
| Screened | Age 3 | 162 | 305 | 9062 |
|  | Age 4 | 170 | 320 | 10031 |
|  | Age 5 | 157 | 306 | 10113 |
| Raw − Screened | Age 3 | 15 | 36 | 3183 |
|  | Age 4 | 11 | 29 | 2569 |
|  | Age 5 | 7 | 19 | 1623 |

After mapping the gaze coordinates onto the onscreen images, I performed data screening following the same set of steps as in [Chapter 4](#aim1-method). To make data quality judgments, I only considered the window from 0 to 2000 ms after noun onset. Next, I identified a trial as *unreliable* if at least 50% of the looks were missing during the time window, and I excluded an entire block of trials if it had fewer than 18 reliable trials. As an additional criterion, I excluded participants who failed to provide at least 6 reliable trials per experimental condition.

### Classifying trials based on initial fixation location

During preliminary visualization of the study-level growth curves, I observed an increasing preference for the unfamiliar image for the nonword condition. The growth curves showed a typical pattern of a baseline at noun onset followed by a quick change in height as the word unfolded. For the nonword condition, this baseline level moved further from .5 (chance with two images) with each year of the study: Children became more likely to fixate on the novel object at the start of these trials.

Because this was a two-image task, I was able to account for the location of the child’s gaze at the onset of the target noun. For each trial, I counted the number of looks to the familiar object and the unfamiliar object during the first 250 ms after target noun onset (specifically, 0 ≤ *time* < 250 ms). If the majority of the looks landed on the familiar object, then the trial was a *familiar-initial* trial. An analogous rule labeled trials as *unfamiliar-initial* trials. Ties were broken by favoring the earlier fixated image. For example, a tie might be a trial with 7 frames of looking to the unfamiliar image, followed by 1 frame between the two images, followed by 7 frames to the familiar image. In this case, the unfamiliar image was viewed first, so the trial is classified as unfamiliar-initial. If there were no looks to either image during that window, the trial was not classified for either image and it was excluded.

Table 4 shows the counts and percentages of trial classification. About 5% of trials were excluded because the child looked to neither image during the first 250 ms of the noun onset. The table shows how the percentage of unfamiliar-initial trials increased with each year of the study. Accounting for this trend was the rationale for classifying trials based on the initial fixation location.

Table 4 Number of trials classified based on initial fixation location.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Study | Condition | Familiar initial | Unfamiliar initial | Neither/excluded |
| Age 3 | Real word | 1629 (53.7%) | 1250 (41.2%) | 154 (5.1%) |
|  | Nonword | 1284 (43.8%) | 1453 (49.6%) | 194 (6.6%) |
|  | Mispronunciation | 1608 (51.9%) | 1305 (42.1%) | 185 (6.0%) |
| Age 4 | Real word | 1561 (45.9%) | 1693 (49.8%) | 145 (4.3%) |
|  | Nonword | 1280 (39.2%) | 1799 (55.2%) | 183 (5.6%) |
|  | Mispronunciation | 1686 (50.0%) | 1552 (46.1%) | 132 (3.9%) |
| Age 5 | Real word | 1718 (50.5%) | 1558 (45.8%) | 125 (3.7%) |
|  | Nonword | 1172 (35.2%) | 1959 (58.9%) | 194 (5.8%) |
|  | Mispronunciation | 1752 (51.7%) | 1487 (43.9%) | 148 (4.4%) |

## Model preparation

To prepare the data for modeling, I downsampled the data into 50-ms (3-frame) bins. I modeled looks from 300 to 1,500 ms after noun onset. Lastly, I aggregated looks by child, study, condition, initial fixation location, and time, and I created orthogonal polynomials to use as time features for the model. Figure 19, Figure 20, and Figure 21 shows the empirical growth curves for each condition following the above-described data screening and preparation steps.

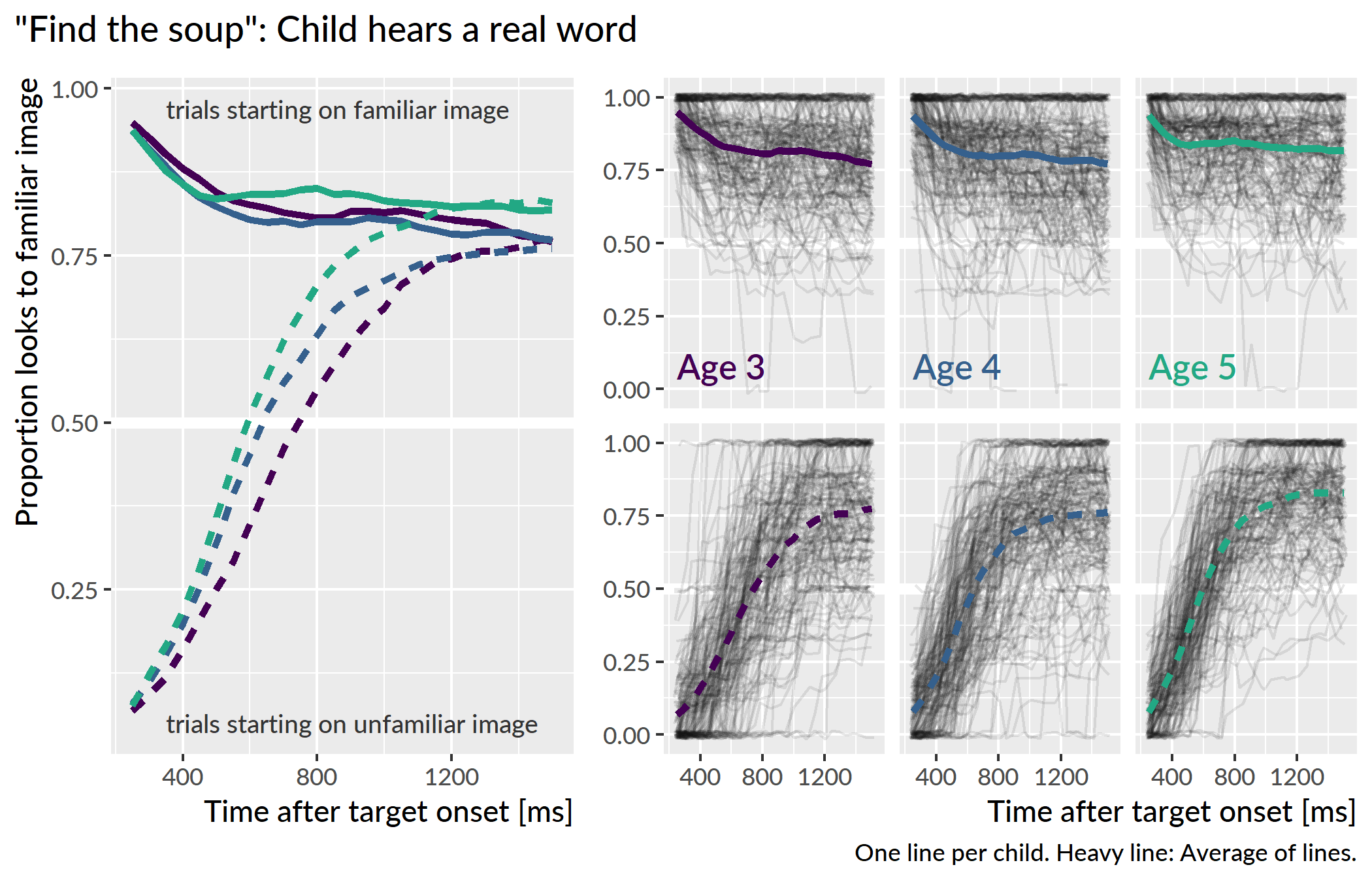


Figure 19 Empirical word recognition growth curves from each year of the study for the real word condition. Each line represents an individual child’s proportion of looks to the target image over time. The heavy lines are the averages of the lines for each year. Only the steep, upward growth curves from unfamiliar-initial trials are analyzed.

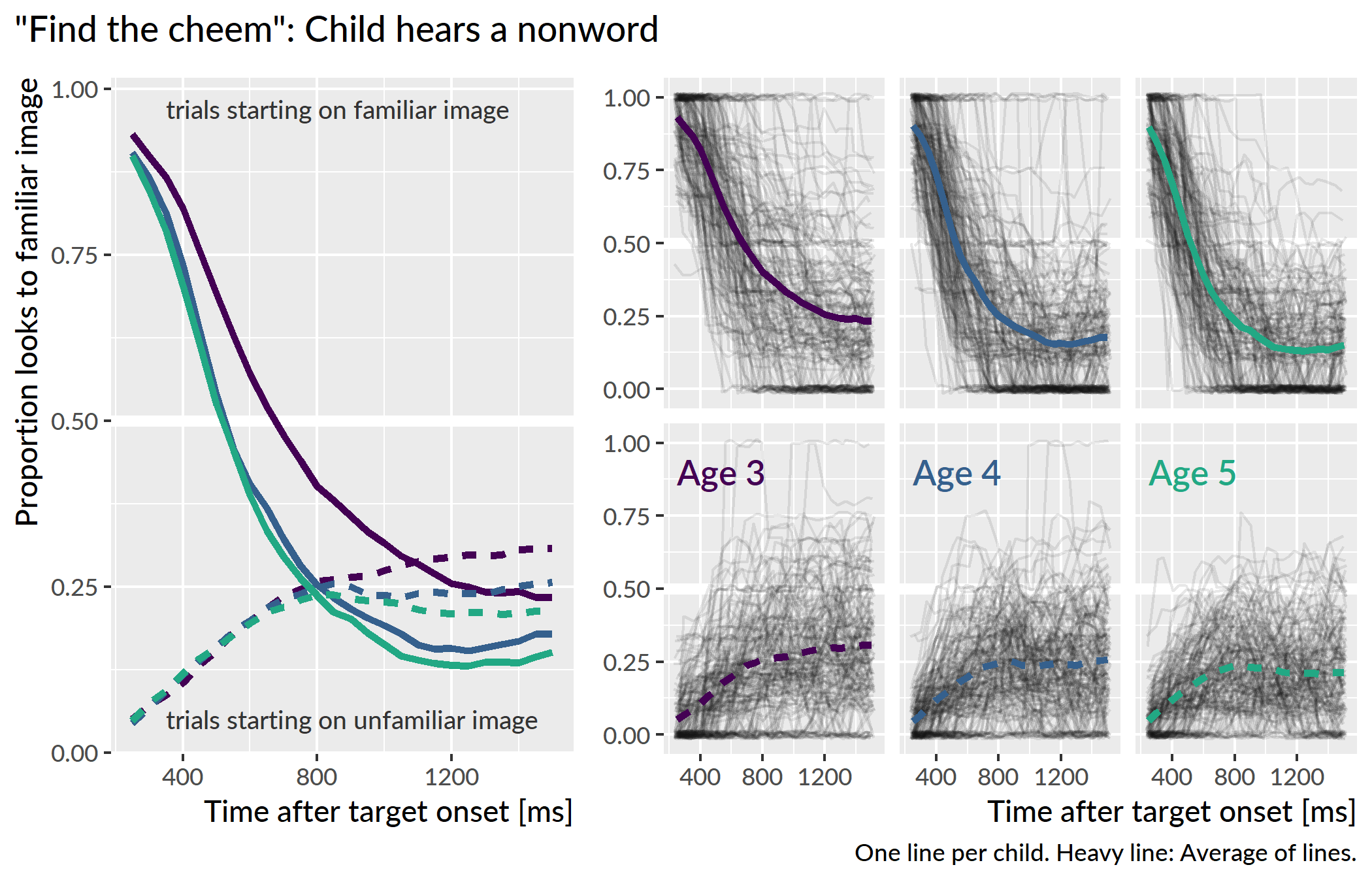


Figure 20 Empirical word recognition growth curves for nonword condition. Only the steep, downward growth curves from familiar-initial trials are analyzed.

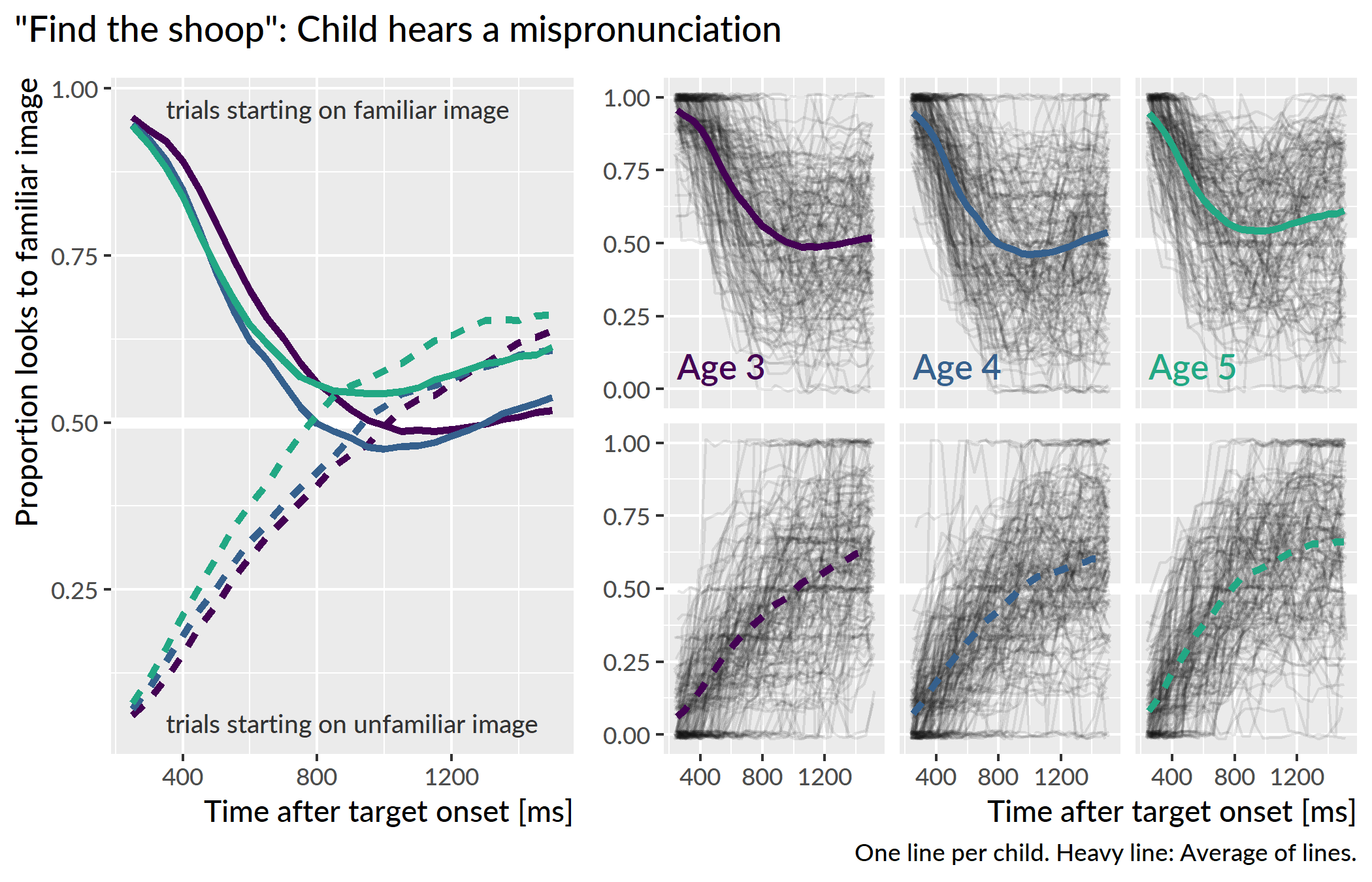


Figure 21 Empirical word recognition growth curves for mispronunciation condition. Both types of curves analyzed.

# Development of referent selection

## Nonwords versus familiar words

I asked whether the recognition of familiar words differed from the fast selection of referents for nonwords. I fit a Bayesian mixed-effects logistic regression growth-curve model, as in [Chapter 5](#fam-rec). For the real word and nonword conditions, there is a well-defined target image: the familiar image for real words and the novel/unfamiliar image for nonwords. The outcome measures were the probabilities of fixating to the target image in each condition:

* P(look to familiar image | hear a real word)
* P(look to unfamiliar image | hear a nonword)

Both the real word and nonword conditions measure referent selection as the probability of fixating on the appropriate referent when presented with a label. The important analytic question is whether and to what degree these two probabilities differ. The growth curve model is similar to the one in [Chapter 5](#fam-rec) with linear, quadratic and cubic time features but it adds a condition effect which interacts with these features. The linear model was:

$$
\small
$$

I fit a separate model for each year of the study. [Appendix 19](#aim2-gca-models) contains the R code used to fit these models along with a description of the specifications represented by the model syntax. The mixed model included by-child and by-child-by-condition random effects to allow some of a child’s growth curve features to be similar between conditions (by-child effects) and to differ between conditions (by-child-by-condition effects).

For these analyses, I limited focus to distractor-initial trials, and modeled the data from 300 to 1500 ms after target onset. I removed any Age × Child levels if a child had fewer than 4 fixations in a single time bin. In other words, children had to have at least 4 looks to one of the images in every 50 ms time bin. This screening removed 13 children at age 3, 15 at age 4, and 6 at age 5.

Figure 22 shows the group averages of the growth curves. For each condition and age, I computed the empirical growth curve for each participant, and I averaged the participants’ growth curves together to obtain group averages. I also applied this process to 100 model-estimated growth curves.

(ref:aim2-real-nonword-means) Averages of participants’ growth curves in each condition and age. The lines represent 100 posterior predictions of the group average.

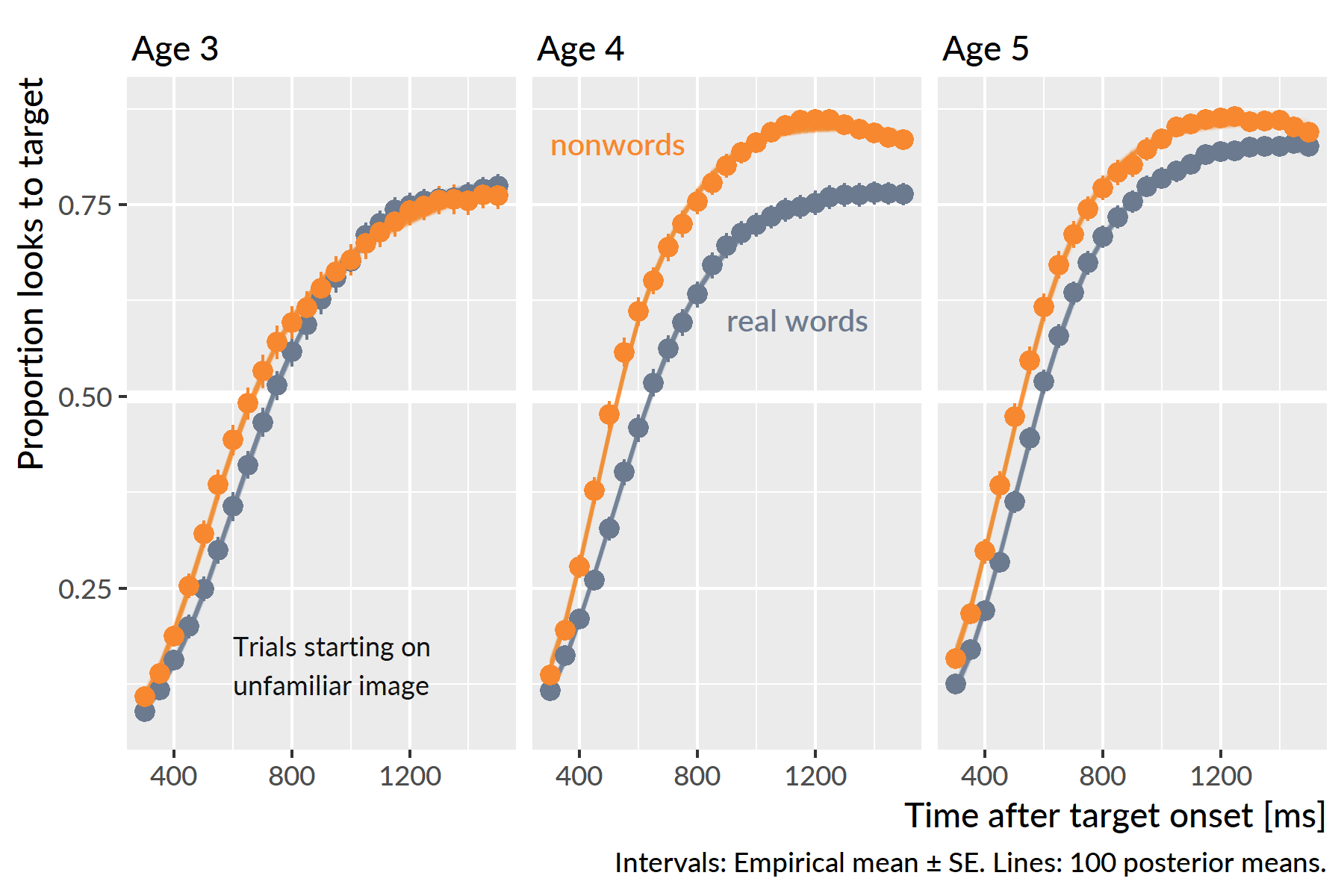


Figure 22 (ref:aim2-real-nonword-means)

In [Chapter 5](#fam-rec), I claim that for these growth curve models only the intercept and linear time terms are behaviorally meaningful model parameters. The intercept measures the average growth curve value so it reflects overall *looking reliability*, and the linear time term measures the overall steepness of the growth so it reflects *lexical processing efficiency*. I also derived a measure of peak looking probability by taking the median of top five points in a growth curve, and this peak provides a measure of *word recognition certainty*. Higher peaks indicate less uncertainty about a word.

I evaluated the general condition effects by looking at how the population-level (“fixed”) effects differed in each condition. Due to ceiling effects, where children’s growth curves saturated 100% looking probabilities, the population-level average growth curve outperformed the observed group averages in Figure 22. The condition differences described by these population-level effects, however, do qualitatively match the patterns in the group averages.

The two conditions did not reliably differ at age 3. The population-level average proportion of looks to the target for nonwords was .60 [90% UI: .55, .65], compared to .56 [.51, .60] for real words—a difference (nonword advantage) of .05 [−.01, .11]. For the linear time feature, the nonword slope increases by 9% [−1%, 18%] in the real word condition. Both these 90% intervals include 0 as a plausible estimate for the condition difference, so there is uncertainty about the sign of the effect. I therefore conclude that the conditions did not credibly differ on average at age 3.

There was an advantage for the nonword condition at age 4 and age 5. The population-level average proportion of looks for the nonwords was .79 [90% UI: .76, .82], compared to .62 [.58, .66] for real words. On average, children looked less to target for the real words than the nonwords. There was a suggestive effect linear time effect where the nonword curve was 13% [1%, 25%] steeper than the real word one. The curve for real words was probably less steep at age 4 but small values near 0 remain plausible. At age 5, only the average probability difference was credible, .81 [90% UI: .78, .83] for nonwords compared to .72 [.68, .75] for real words. In general, children performed better in the nonword condition than the real word condition at age 4 and age 5. This difference shows up in the growth curve model through intercept effects, although it is plausible that children’s nonword growth curves were steeper than the real word curves at age 4.

I analyzed the children’s model-estimated growth curve peaks. Each posterior sample of the model represents a plausible set of growth curve parameters for the data, so for each of these samples, I calculated the growth curves for each child and the peaks of the growth curves. Figure 23 shows the the posterior averages of the growth curves peaks for each participant.

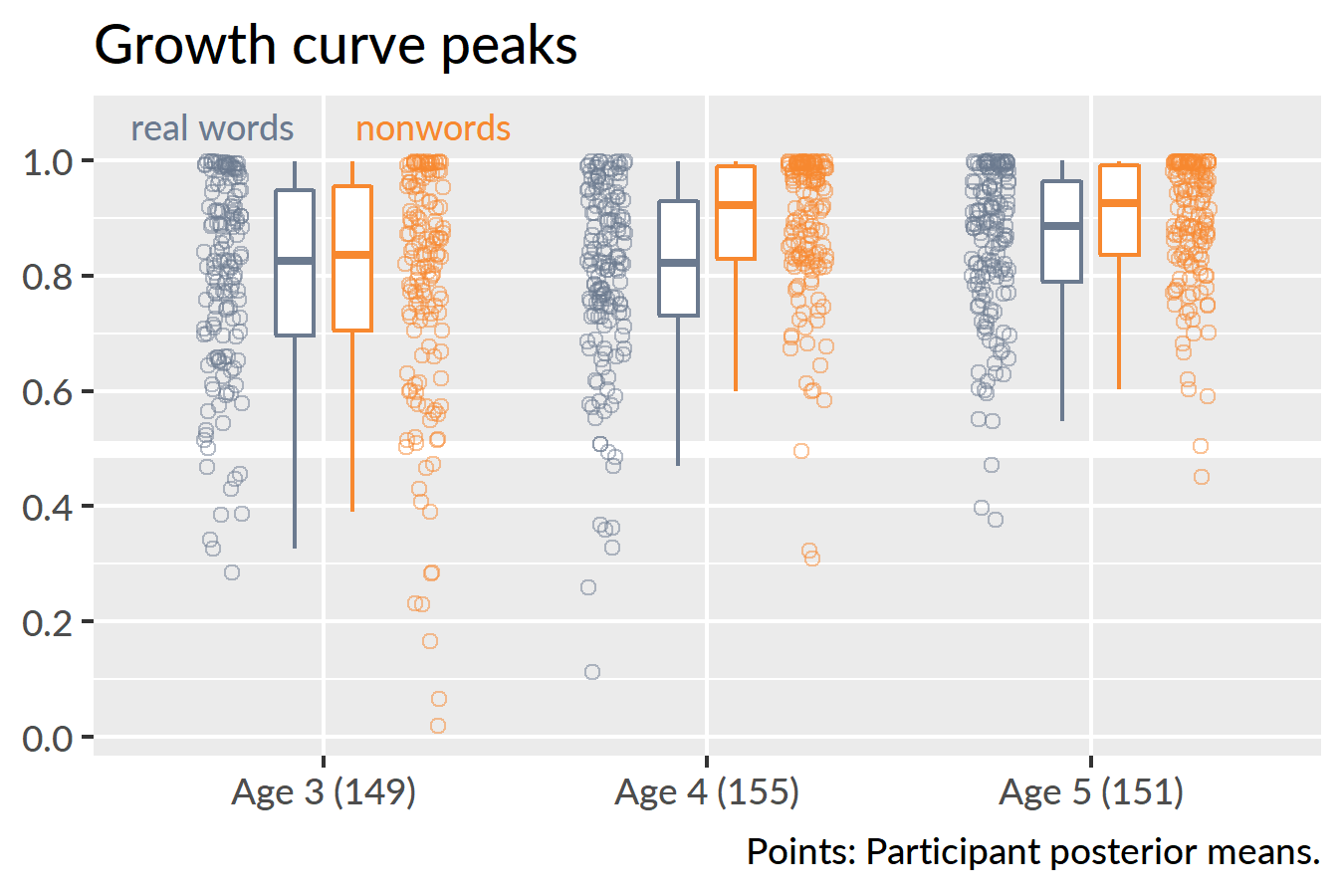


Figure 23 Growth curve peaks by child, condition and year of the study. The movement of the medians conveys how the nonword peaks effect increased from age 3 to age 4 and the real word peaks increased from age 4 to age 5. The piling of points near the 1.0 line depicts how children reached ceiling performance on this task.

Descriptive statistics reveal the developmental trends for this task. At age 3, the median peak values were similar for the two conditions: .84 for nonwords and .83 for real words. The peaks increased for the nonword condition in the following year with a median value of .92. It is worth emphasizing what this statistic tells us: At age 4, half of the children had a peak looking probability of .92 *or greater*. In other words, half the children performed near the ceiling on this task by age 4. At age 5, the median nonword peak was .93, essentially unchanged from age 4. For the real words, the median peak increased from .82 at age 4 to .89 at age 5.

To quantify the degree of ceiling performance, I calculated the number of children per condition with a growth curve peak greater than or equal to .99 over the posterior distribution. For the nonword condition, there were 23 [90% UI: 20, 26] children who performed at ceiling at age 3, 41 [36, 45] at age 4, 40 [36, 44] and at age 5. For the real word condition, the number of children attaining ceiling performance was more uneven: there were 20 [16, 24] ceiling-performers at age 3, 13 [10, 16] at age 4, and 26 [23, 29] at age 5.

To compare peaks looking probabilities between ages, I fit a linear mixed effects model with restricted maximum likelihood via the lme4 R package (vers. 1.1.17; Bates, Mächler, Bolker, & Walker, [2015](#ref-lme4)). I regressed the children’s average growth curve peaks onto experimental condition, age group, and the age × condition interaction. The model included randomly varying intercepts for child and child × age. This modeling software does not provide *p*-values for its effects estimates, so for these comparisons, I decided that an effect was significant when the *t* statistic for a population-level (“fixed”) effect had an absolute value of 2 or greater. In practical terms, this convention interprets an effect as “significant” when its estimate is at least 2 standard errors away from 0. (Gelman and Hill ([2007](#ref-GelmanHill)) use this approach when with mixed models.)

At age 3, the two conditions did not significantly differ, Breal-nonword = .01, *t* = 0.95. At age 4, nonword peaks were on average .09 proportion units greater than the real word peaks, *t* = 5.79, and at age 5, the nonword peaks were .04 proportion units greater than the real word peaks, *t* = 2.56. For the nonword condition there was a significant increase in the peaks from age 3 to age 4, *B*4-3 = .10, *t* = 5.99, whereas there was no improvement from age 4 to age 5, *t* = 0.37. In the real word condition, there was only a significant improvement from age 4 to age 5, *B*5-4 = .06, *t* = 3.25.

Finally, I asked whether expressive vocabulary size correlated with peak looking performance on the two conditions. Correlations among real word peaks, nonword peaks, expressive vocabulary and receptive vocabulary are given in Table 5. At all three years, children with larger vocabularies had higher nonword peak looking values. At age 3 and age 4, vocabulary positively correlated with real-word looking performance. Figure 24 illustrates the relationship between the peaks and expressive vocabulary. When there is more variability in the peaks, as at age 3, the vocabulary effect on the nonwords is strongest.

Table 5 Correlation between curve peaks and vocabulary measures

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Study |  | Real word peak | Nonword peak | PPVT-4 standard |
| Age 3 | Nonword peak | *r*(149) = .24\*\* |  |  |
|  | PPVT-4 standard | *r*(139) = .23\*\* | *r*(139) = .46\*\*\* |  |
|  | EVT-2 standard | *r*(137) = .15 | *r*(137) = .30\*\*\* | *r*(137) = .69\*\*\* |
| Age 4 | Nonword peak | *r*(155) = .29\*\*\* |  |  |
|  | PPVT-4 standard | *r*(152) = .15 | *r*(152) = .23\*\* |  |
|  | EVT-2 standard | *r*(153) = .23\*\* | *r*(153) = .20\* | *r*(152) = .78\*\*\* |
| Age 5 | Nonword peak | *r*(151) = .13 |  |  |
|  | EVT-2 standard | *r*(149) = −.06 | *r*(149) = .23\*\* |  |

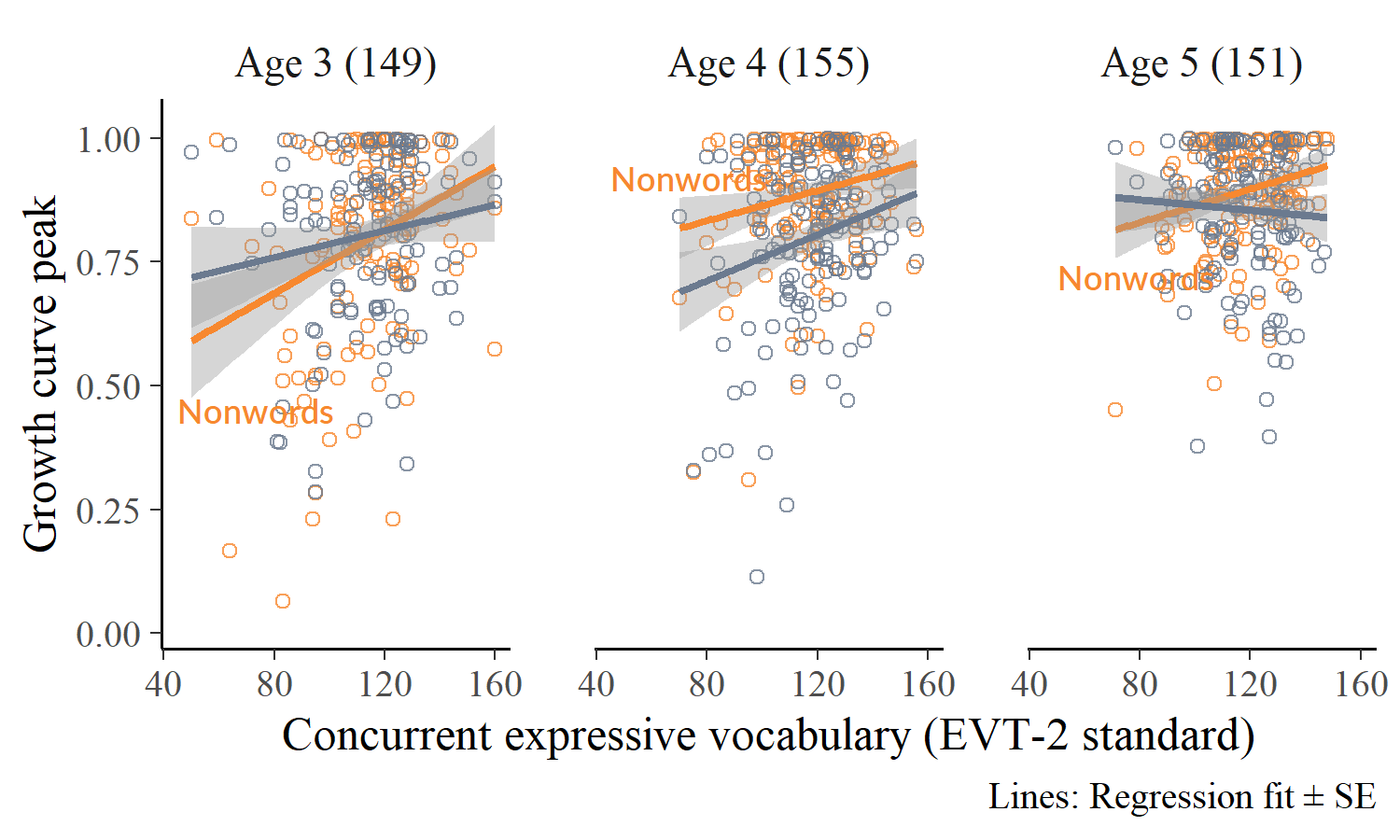


Figure 24 Text reference for caption

**Summary**. Children performed similarly for real words and nonwords at age 3. Children’s processing of nonwords improved at age 4. At this age, performance also began to saturate with the group average for peak looking probability greater than .9 for the nonword condition. Consequently, children did not improve in processing of nonwords from age 4 to age 5. For the real word condition, children’s performance did not change from age 3 to age 4 but it did improve from age 4 to age 5. At both age 4 and age 5, there was a decisive advantage for the nonword condition. Finally, children with larger expressive vocabularies looked more to the nonwords compared to children with smaller vocabularies. A comparable effect for real words was observed at age 3 and age 4 but only reliably observed at age 4.

## Does age 3 referent selection better predict age 5 vocabulary?

I hypothesized that performance on the nonword condition would be a better predictor of future vocabulary size than the real word condition. This hypothesis follows from the assumption that fast referent selection, as opposed to familiar word recognition, is a more relevant skill for word-learning. Put another way, a child’s ability to quickly map a novel word to a referent is more closely related to the demands of in the moment word-learning than familiar word recognition.

In [Chapter 5](#fam-rec), I found that peak looking probability at age 3 positively correlated with age 5 vocabulary. Pairing this finding with my hypothesis, I predicted that the growth curve peaks in the nonword condition at age 3 would be better predictors of vocabulary at age 5 than the real word peaks at age 3.

For these analyses, I regressed age-5 expressive vocabulary (EVT-2) standard scores onto age-3 expressive vocabulary score and onto age-3 real word peaks or age-3 nonword peaks. There were 116 children with data available for this analysis. There was an expectedly strong relationship between age 3 and age 5 vocabulary, *R*2 = .49. A 1-SD (18-point) increase in vocabulary at age 3 predicted an 0.7-SD (10-point) increase at age 5. There was no effect of age-3 real-word peak over and above age-3 vocabulary, *p* = .59. There was a significant effect of the nonword peak, *p* = .005, *ΔR*2 = .03, over and above age-3 vocabulary. A .1 increase in nonword peak probability predicted a 0.10-SD (1.4-point) increase in age-5 vocabulary. Figure 25 depicts the difference between the two conditions with a flat line for the real condition and small slope for the nonword condition.

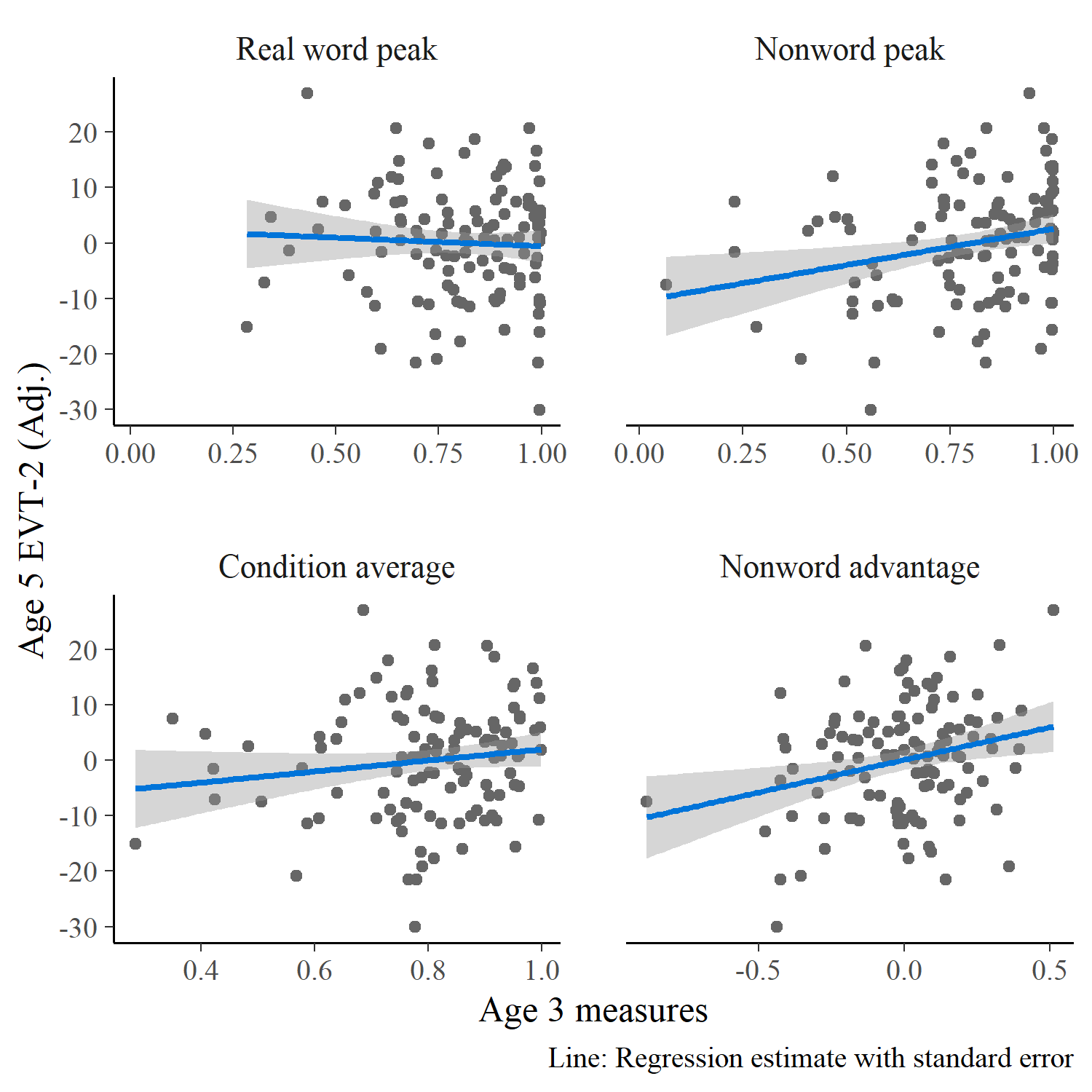


Figure 25 Marginal effects of age-3 referent selection measures on age-5 expressive vocabulary standard scores. The vocabulary scores were adjusted (residualized) to control for age-3 vocabulary, so these regression lines show the effects of the predictors over and above age-3 vocabulary.

Finally, I tested whether the difference between nonword and real word peaks within children predicted vocabulary growth. By themselves, differences do not convey much information about how well the child performed: A difference of 0 can happen if a child has peaks of .1 in both conditions or .9 in both conditions. To control for general referent selection performance, therefore, I also included the within-child averages of the two peaks. The model predicted age-5 vocabulary using the within-child average of the peaks, the nonword advantage, and age-3 vocabulary. In this case, condition-averaged performance did not significantly predict age-5 vocabulary, *p* = .22. The condition differences did predict age-5 vocabulary: A .1 increase in the nonword condition advantage predicted a 0.08-SD (1.1-point) increase in age-5 vocabulary, *p* = .009

**Summary**. A child’s performance in the nonword condition at age 3 positively predicted expressive vocabulary size at age 5. This effect held even when controlling age-3 vocabulary size, and the effect emerged when using the absolute growth curve peak or using the relative advantage of the nonword condition over the real word condition. Although the effects were significant, the effect sizes were small. The EVT-2 is normed to have an IQ-like scale with a mean of 100 and standard deviation of 15. An increase of .1 in age-3 growth curve peak predicted an increase in age-5 vocabulary of 1.4, approximately one tenth of the test norms’ standard deviation.

## Discussion

Children showed developmental improvements in referent selection for the real word and nonword trials. The changes were not consistent year-over-year improvements however. Nonword processing improved from age 3 to age 4 and real word recognition improved from age 4 and to age 5. One reason for these limited improvements is that the two-image word recognition task was too easy. At age 4, approximately 25% of children had nonword growth curve peaks of .99 or greater.

Despite the presense of ceiling effects, vocabulary size had a small-to-medium positive correlation with nonword growth curve peaks at all three ages. Children who knew more words had a higher probability of looking to the novel object when presented with a nonword. This replicates the vocabulary advantage in processing nonwords observed by Bion et al. ([2013](#ref-Bion2013)) and Law and Edwards ([2015](#ref-MPPaper)). This effect is probably bidirectional with larger vocabularies making fast referent selection easier, and fast referent selection being a crucial mechanism for word-learning. To further examine the direction of effect, I tested whether nonword performance at age 3 predicted expressive vocabulary size at age 5. There was a small effect predictive effective where children with high nonword peaks had a larger vocabulary size two years later. Although real word and nonword performance had a small-to-medium positive correlation, children’s processing of the real words had no predictive value. Real word peaks did not predict vocabulary, nor did the average of real word and nonword peaks did not have an effect over and above the difference of the peaks.

For these two conditions, I hypothesized were word recognition in the real word condition would be easier than in the nonword condition, or failing that, the two conditions would not reliably differ. I had discounted a third possibility of any overall advantage for nonwords over real words. The advantage of nonwords at age 4 and age 5 over real words was an therefore unexpected result.

Why might children perform better on the nonword trials than the real words? The results are consistent with a novelty bias in referent selection (Horst et al., [2011](#ref-Horst2011); Mather & Plunkett, [2012](#ref-Mather2012)). An additional factor may be the presense of the mispronunciation trials—reported in [Chapter 12](#sensitivity-to-mispronunciations). The mispronunciations may undermine familiar word recognition. For one-third of the trials, children hear an imperfect version of a familiar word, and they show more uncertain responses to them. This environment cause children to downweight the syllable-initial sounds. Such an adaptation would lead to lower overall activation of the real words. Alternatively, the novelty bias could interfere with processing of familiar words. For some amount of trials, children could have ignored the familiar word and focused attention on the novel object.

# Sensitivity to mispronunciations

For the mispronunciation trials, there is no correct “target”, as there is for the other conditions. The design of the task allows the child to associate a mispronunciation with an unfamiliar object or with the familiar object with a name that sounds like the mispronunciation. As a result, I analyzed the mispronunciation trials separately for both initial-fixation locations. One analysis handled trials where a child’s gaze started on the familiar object and another analysis handled trials starting on the unfamiliar object. For these models, I fit a Bayesian logistic regression growth curve model that included indicators for Age and Time × Age interactions, as in the model from [Chapter 5](#fam-rec). The linear model was therefore:

$$
\small
$$

The mixed effects model included by-child and by-child-by-age random effects so that it would capture how a child’s growth curve features may be similar over developmental time (by-child effects) and may differ at each age (by-child-by-age effects). [Appendix 19](#aim2-gca-models) contains the R code used to fit these models along with a description of the model’s specification/syntax.

For these analyses, I modeled the data from 300 to 1500 ms after target onset. As in the real word vs. nonword analyses, I removed any Age × Child levels if the child’s data had fewer than 4 fixations in a single time bin. As a result, children had to have at least 4 looks to one of the two images in every 50-ms time bin. For the unfamiliar-initial trials, this screening removed 6 children at age 3, 6 at age 4, and 2 at age 5, and for the familiar-initial trials, this screening removed 1, 4, and 0 children at ages 3, 4, and 5, respectively.

## Unfamiliar-initial trials: Move along now

When preschoolers started on the image of a novel object and heard a mispronunciation, they looked to the familiar image. Figure 26 shows the average of children’s growth curves along with the 100 model-estimated group averages. The growth curves all cross the .5 threshold, so the children on average looked more to the familiar than the unfamiliar image. Granted, the degree of referent selection is not as strong as that observed for the real words or nonwords. For those conditions, the average growth curve reached a peak of around .77 at age 3, but for the mispronunciations the age-3 peak is around .62. Children also were comparatively slower to process mispronunciations. For the real-word condition, the average age-3 growth curve crosses .5 looking probability around 775 ms after target onset, whereas in the mispronunciation condition, this threshold is crossed at 1000 ms. Children associate the mispronunciation with the familiar object, although they are slower and show greater uncertainty compared to real word trials.

(ref:unfam-initial-mp-trials) Averages of participants’ growth curves in each age. The lines represent 100 posterior predictions of the group average.

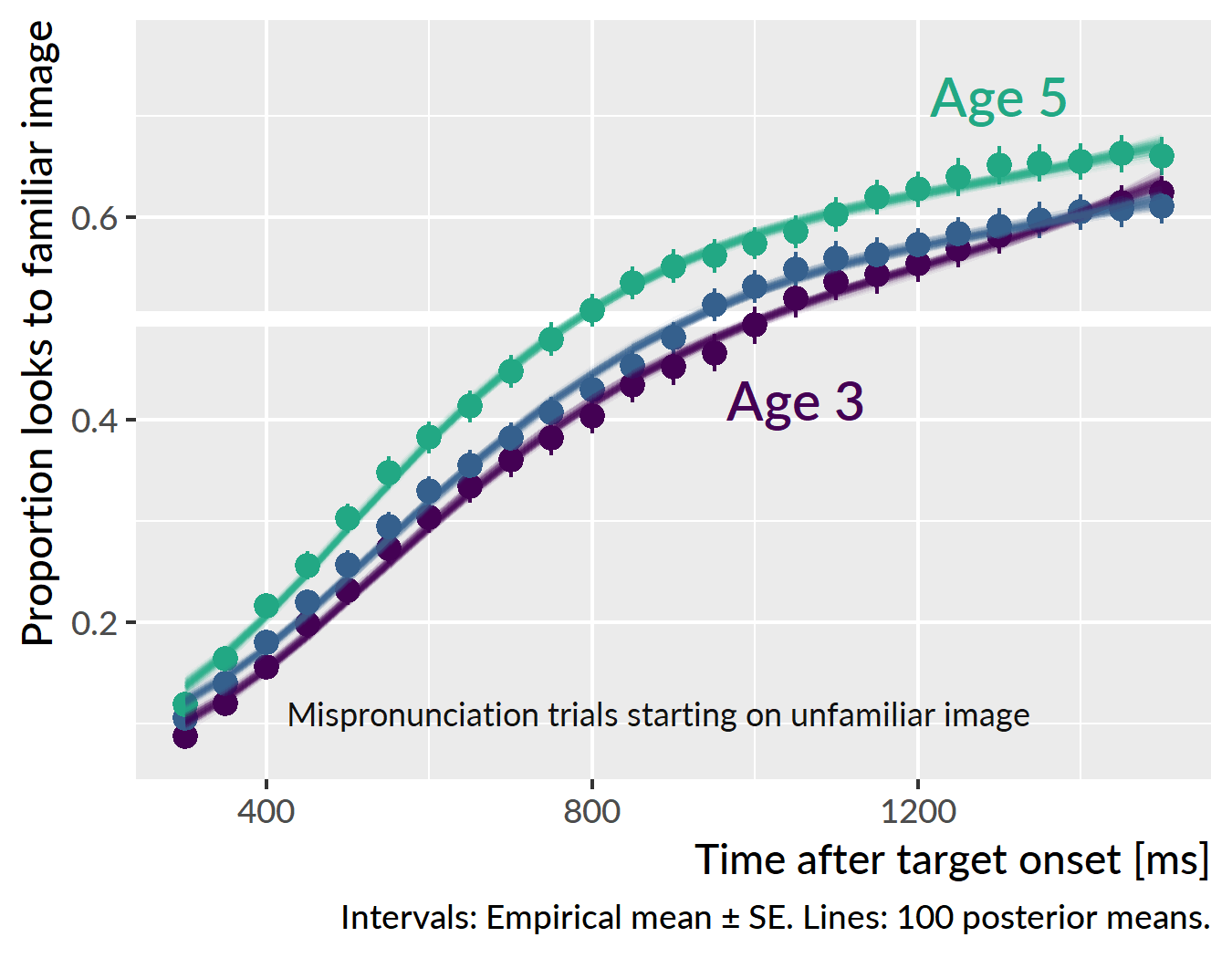
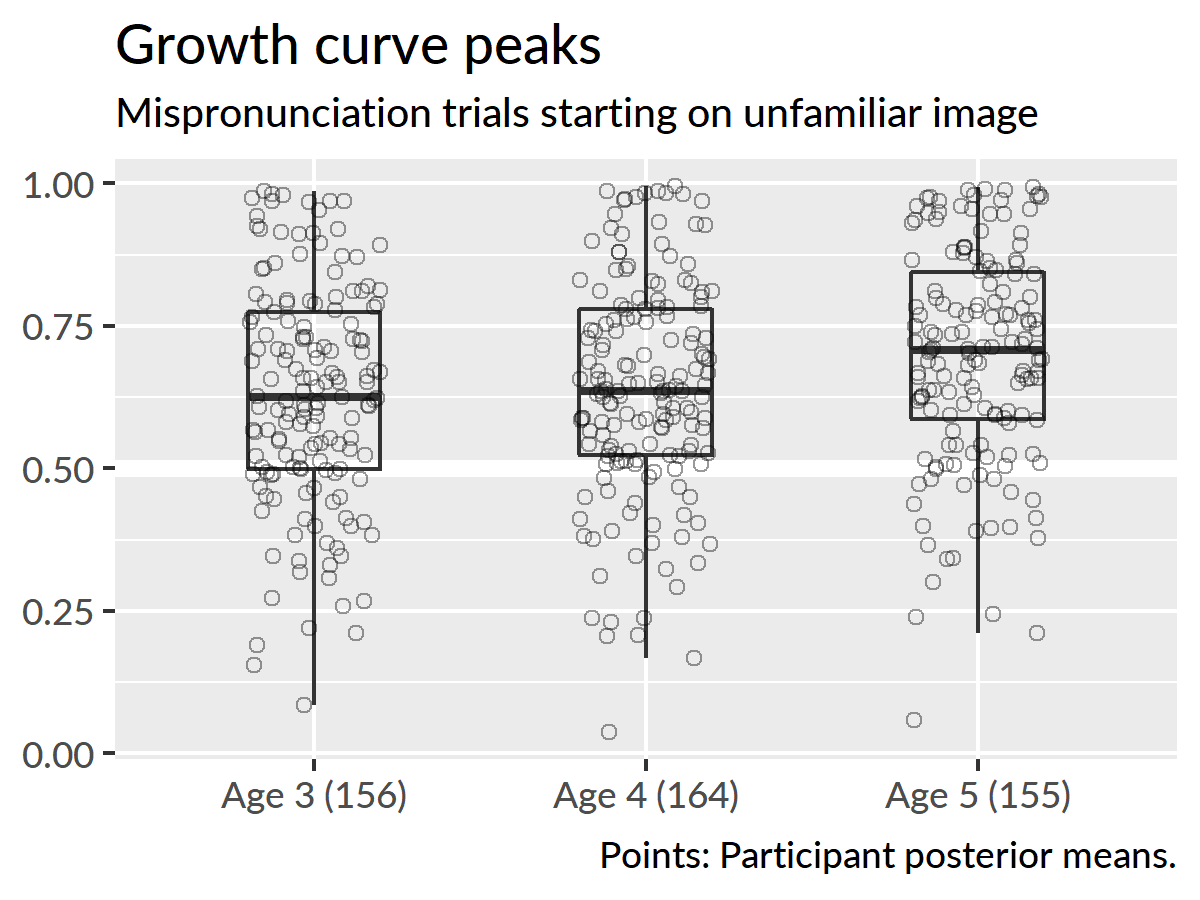


Figure 26 (ref:unfam-initial-mp-trials)

Of the growth curve features, developmental changes were only observed for the average probability (intercept) and peak probability features. At age 3, the average proportion of looks to the familiar image was .37 [90% UI: .34, .40]. At age 4, the looking proportion increased by .04 [−.01, .08] to .40 [.37, .44]. This year-over-year change was probably positive, but the the uncertainty interval still includes a change of 0 as a plausible result. Visually, this uncertainty appears in the growth curve plot by how close together the age-3 and age-4 growth curves appear. At age 4, the average proportion of looks increased by .07 [.03, .12] to .48 [.45, .51]. Here, there is more certainty that the year-over-year change was positive, and this result is consistent with the visual separation of the age-5 growth curve from the others. In short, performance was similar for age 3 and age 4 but there was a marked improvement at age 5.

Figure 27 shows participant’s growth curve peaks for each year of the study. The peaks were computed as in other chapters by taking the median of the five highest values on the curve. The average of the participants’ peak looking probabilities followed the same pattern as the average looking probabilities: similar levels at age 3 and age 4 (.63 versus .64) but a clear gain in looking peak probability at age 5 (.69).

 Figure 27 also indicates how most of the children at each age favored the familiar object over the unfamilar object. The bottom hinge of the boxplots mark the location of the 25th percentile. Therefore, approximately 75% of children at age 3 were on or above the .5 threshold. Unlike the other conditions, very few listeners achieve a peak of looking probability of .99: At age 5, only 3 [1, 5] children reached ceiling performance, compared to approximately 40 for nonwords and 13 for real words.

None of the other growth curve features showed developmental changes. That is, there were no credible year-over-year changes for the linear, quadratic or cubic time components of the growth curve. Although Figure 26 shows children’s probability of looking to the familiar image increasing more quickly at age 5, this effect cannot be clearly tied to any of the model’s polynomial time features. After 600 ms, the age-5 curve is almost parallel to other curves. This visual feature is consistent with the intercept effect: The curve is higher than the others on average, but it does not show any differences in shape.

### Child-level predictors

I tested whether child-level measures predicted looking behavior under these conditions. First, I asked if performance on a minimal pair discrimination task at age 3 predicted looking behavior at age 3. The rationale here is the hypothesis that children with better minimal pair discrimination may be especially sensitive to mispronunciations. Proportion of items correct on the task did not correlate with growth curve peaks, *r* = −.03 [90% UI: −.05, −.01], *n* = 138, nor with any other growth curve measures.

I also tested whether expressive vocabulary (EVT-2 standard score) predicted performance in this condition. In this case, there were significant effects at age 3 where a higher expressive vocabulary predicted higher peak probabilities and higher average probabilities. These effects, however, were very small. As shown in Figure 28, for example, a 15-point increase in expressive vocabulary predicted an increase of growth curve peak of .03, *R*2 = .03. Expressive vocabulary did not predict any of the growth curve features at age 4 or at age 5.

(ref:plot-evt-mp-unfam-peaks) Relationship between expressive vocabulary and growth curve peaks for mispronunciation trials starting on the unfamiliar image. I took 100 draws from the posterior distribution and computed participant’s growth curve peaks for each draw. Points represent the mean and standard error of 100 peaks. Lines represent regressions fit on each draw.

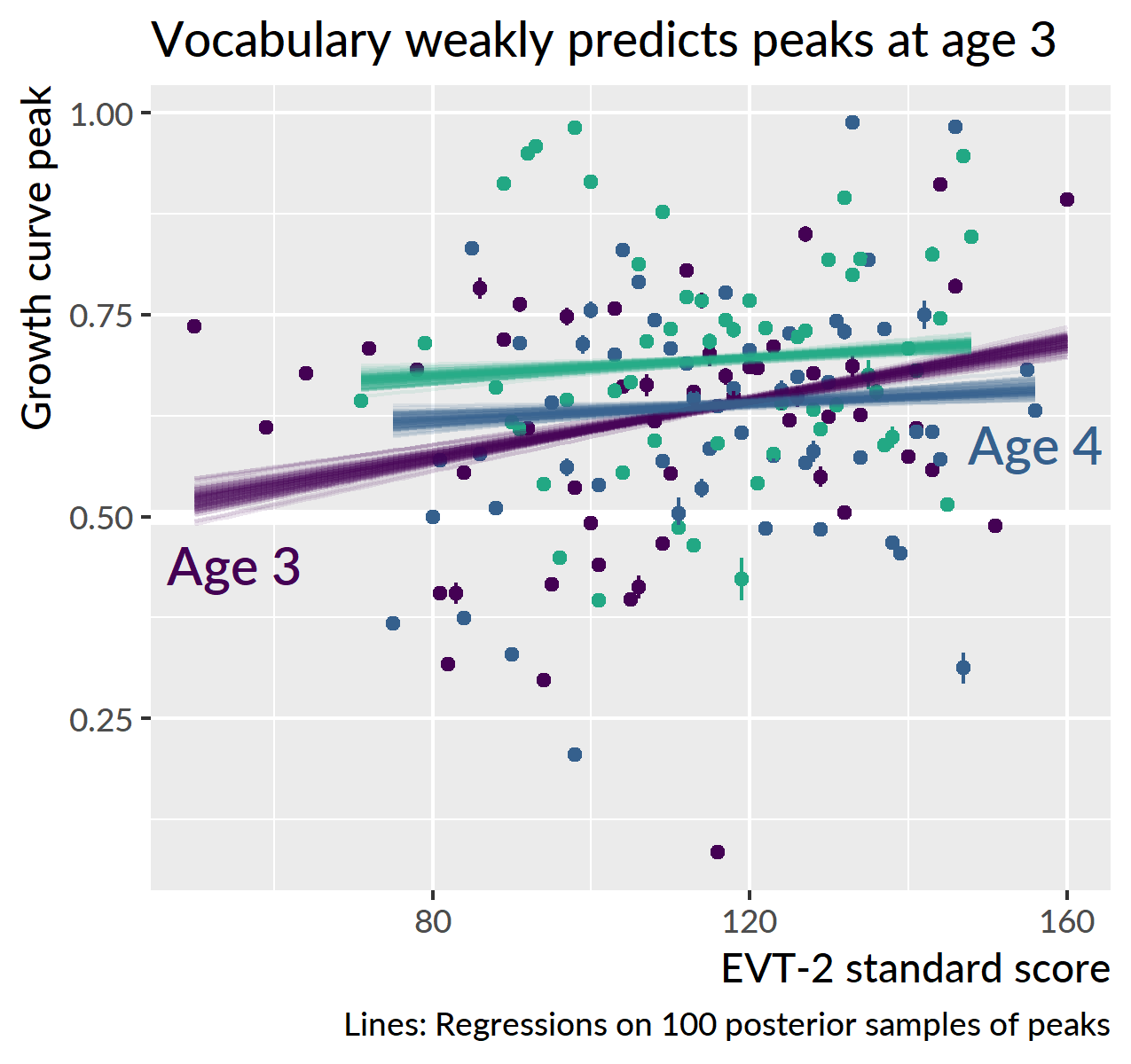


Figure 28 (ref:plot-evt-mp-unfam-peaks)

**Summary**. When children are looking at the unfamiliar object and hear a mispronunciation, they shift their looks on average to the familiar image that sounds like the mispronunciation. Children are much more uncertain in this condition, compared to the real-word and nonword conditions where the appropriate referent is more obvious. The only developmental changes observed were the increases in average looking probability and peak looking probability at age 5. Finally, there was a small effect of expressive vocabulary on looking probability at age 3, but no other effects of vocabulary were observed. Minimal pair discrimination at age 3 also did not predict looking behavior.

## Familiar-initial trials: Should I stay or should I go?

The preceding results showed that preschoolers associate one-feature onset-mispronunciations with the familiar word that matches the rime of the word. But that was only for trials where children start on the unfamilar object. I now consider the other situation, where children are fixating on a familiar object and hear a word that immediately mismatches with the name of that familiar object. On the basis of the first segment, children have information that supports switching to another image. But as the rest of the word unfolds, they hear a syllable rime that supports staying.

Figure 29 shows the growth curve averages for trials starting on the familiar image. The looking patterns show a sharp fall towards .5 which is chance-level performance. Behaviorally, children on average move quickly to look at both images equally. They rush into maximum uncertainty, especially at age 4. Patterns are somewhat more restrained at age 5. Here, the average of the growth curves never dips below .5, and in fact, it shows a late rise to .6 looking probability. At this age, children are overall more likely to stay on the familiar object. Finally, at age-3, the curve begins to fall later than the other curves, reflecting a slower change from the starting probability.

(ref:fam-initial-mp-trials) Averages of participants’ growth curves in each age. The lines represent 100 posterior predictions of the group average.

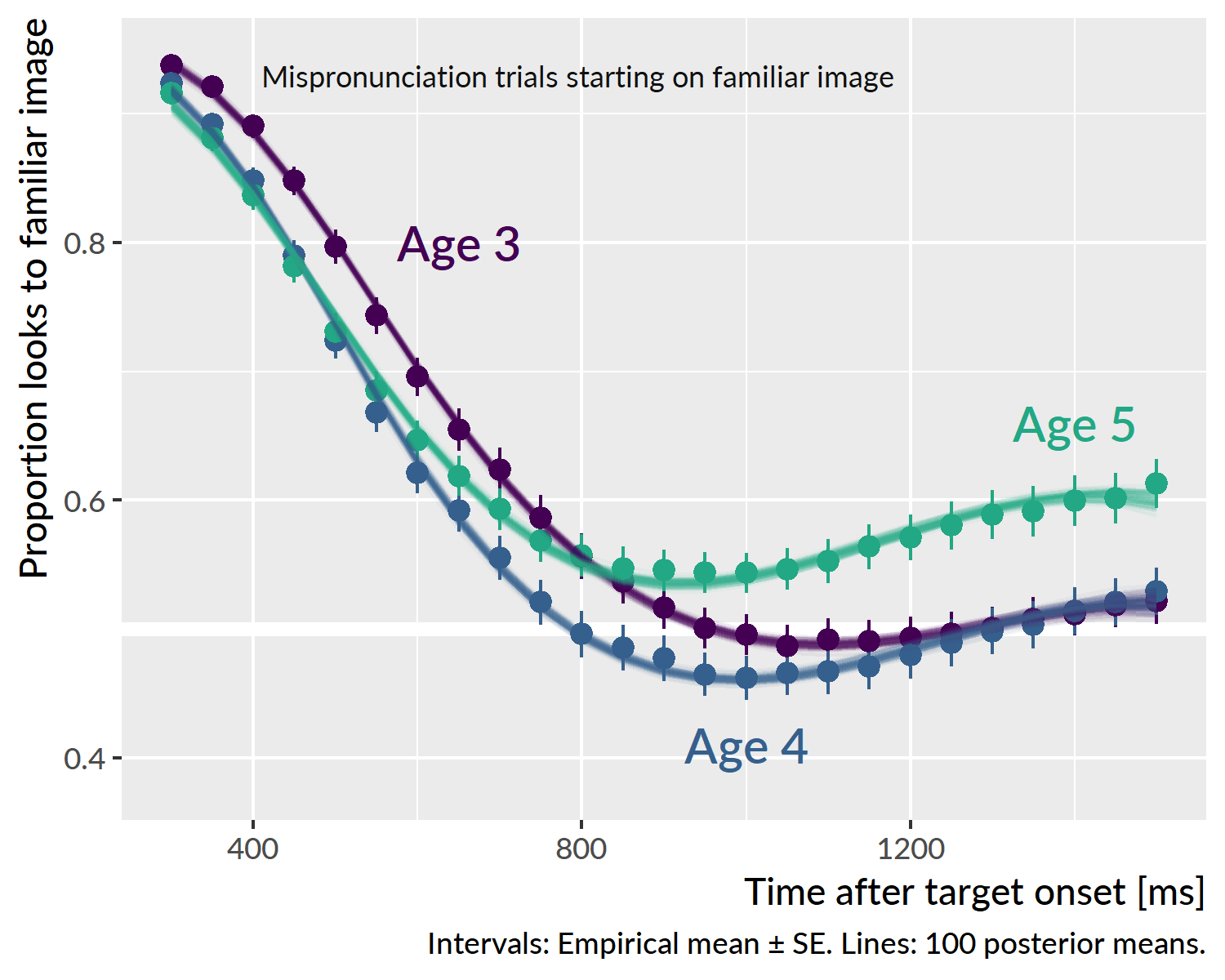


Figure 29 (ref:fam-initial-mp-trials)

In other analyses, growth curves rise and plateau, and age-related effects appear in how quickly the curves rise or the height at which they plateau. In those cases, it is straightforward to interpret how the intercept and linear time effects contribute to the curve’s shape over development. For this model, the curves *fall* and plateau, and there is not an obvious developmental, year-over-year change among the curves. Thus, more effort is required to interpret the model parameters and how they combine to form the growth curve shape.

Figure 30 visualizes how the growth curve features are weighted at each year and how they contribute to the overall growth curve shape. At age 3, the intercept feature, or average proportion of looks to the familiar image, was .68 [90% UI: .65, .71]. The feature is less meaningful in this situation because the curves all start at a high probability which inflates the average value. That said, comparisons remain useful. At age 4, the average probability decreased by .05 [.02, .09] to .63 [.60, .66], and at age 5 the average probability returns to age-3 levels, .68 [.65, .71]. This intercept effect contributes to how the age-4 curve dips below the others and indeed briefly crosses the .5 probability threshold.

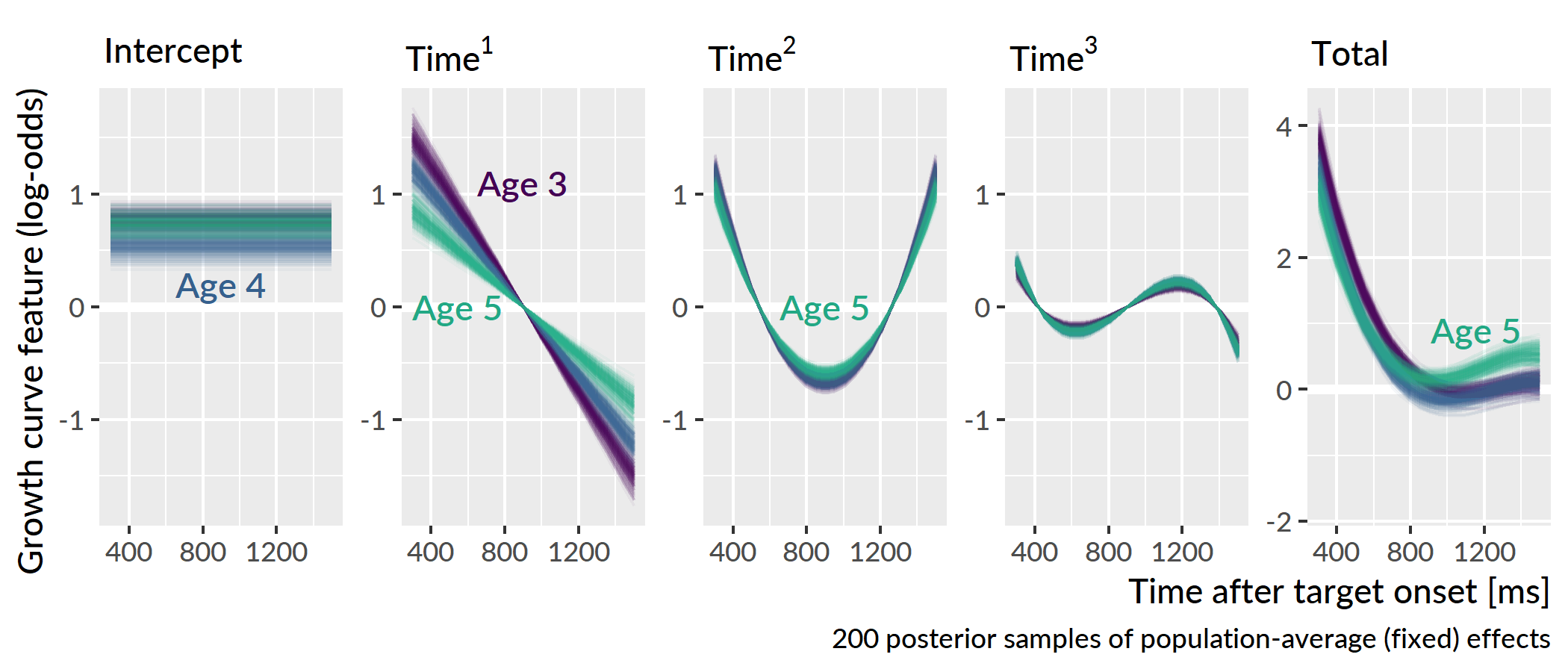


Figure 30 Weighted growth curve features. For the first four panels, the *y* axes are scaled to the same range. This scaling highlights how the cubic time component contributes less to the overall shape than the other features.

For the linear time feature, the slope becomes flatter year over year, decreasing by 19% [7%, 29%] from age 3 to age 4 and decreasing by 30% [17%, 42%] from age 4 to age 5. For these curves, however, the starting location is the highest value on the curve, so the linear time feature in this case mostly works to set the starting location of the curves. When the features are combined in Figure 30, the age-3 curve, which has the steepest linear time feature, starts at a higher value than the others.

There was a credible change in the quadratic time feature at age 5. One way to think of a positive quadratic trend is like a weight hanging on a string: It pulls and bends the whole curve downwards. At age 5, the quadratic feature is 12% [1%, 22%] smaller than at age 4, meaning that the age-5 curve has slightly less bend downwards. Finally, there were no credible differences in the cubic time feature. Compared to the other features, the cubic trend only a small amount to the overall shape of the curves.

The combination of these effects shows in the final panel of Figure 30. The age-4 curve dips down furthest beneath 0 log-odds (.5 probability)—this is driven by in the intercept feature. The age-5 curve stays above 0 log-odds and eventually starts to rise away from its minimum value, owning to the dampened linear and quadratic features.

**Summary**. The shape of the average growth curves changed with each year of the study. Given the interplay of the curve features, I will avoid assigning a developmental intepretation to individual features. There are two main noticeable developmental trends at play however. First, the age-3 curve starts to fall from its baseline probability a little later than the other curves. Second, the age-5 curve stays above .5 probability and starts to rise at the end of the trial. At age 5, children were more likely stay looking at the familiar object than look at both images equally.

### Child-level predictors and different listening behaviors

In other word recognition analyses, I derived a growth curve “peak” value as a measure of maximum looking probability or minimum word recognition uncertainty. For these trials, I asked whether analogous growth curve “valleys” provided a meaningful feature for looking behavior when children start a trial fixated on the familiar image. This value was defined as the median of the five smallest values of a growth curve. Intuitively, it reflects the maximum degree to which the novel image is considered as a referent for the mispronunciation.

Figure 31 shows the posterior means of participants’ growth curve valleys. Note that there is considerable variability at each age, with the 0–1 interval nearly covered at age 4. The median value is closer to .5 at age 5, and this difference is consistent with the growth curve trajectories where the average age 5 curve had did not dip as low as the other curves.

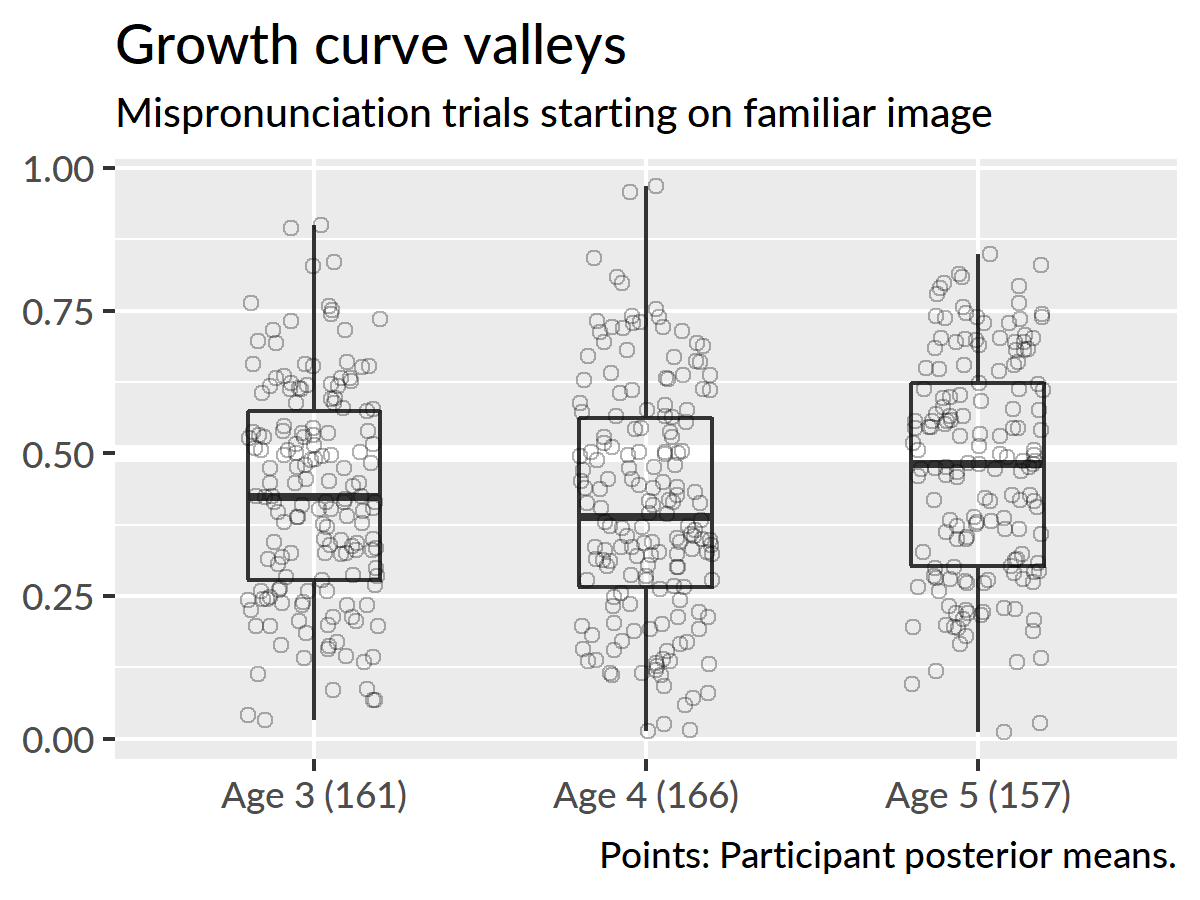


Figure 31 Growth curve peaks by age for mispronunciation trials starting on the familiar image.

The wide range of values for the growth curve valleys suggests that there are a few different listening behaviors that are being averaged over in the above analyses. The valleys above .6, for example, indicate that some children on average stay with the familiar image, and the valleys below .4 indicate children who favor the unfamiliar image.

To explore individual listening behaviors, I visualized children’s individual growth curves based on their growth curve valleys. Within each year, I grouped children into sextiles based on the posterior mean of the their valley and plotted their individual growth curves. Figure 32 shows the results from age 3. The final two bins show children who stayed with the familiar image throughout the mispronuciation trials. The first two bins mostly contain children who switched to the unfamilar image and stayed there. These are also children whose curves show a pronounced u-shaped trajectory. Specifically, the curves with the highest ending points in the first three bins highligh children with u-shaped trajectories. In these curves, the probability of fixating on the familiar image briefly decreases, as the child considers the other image.

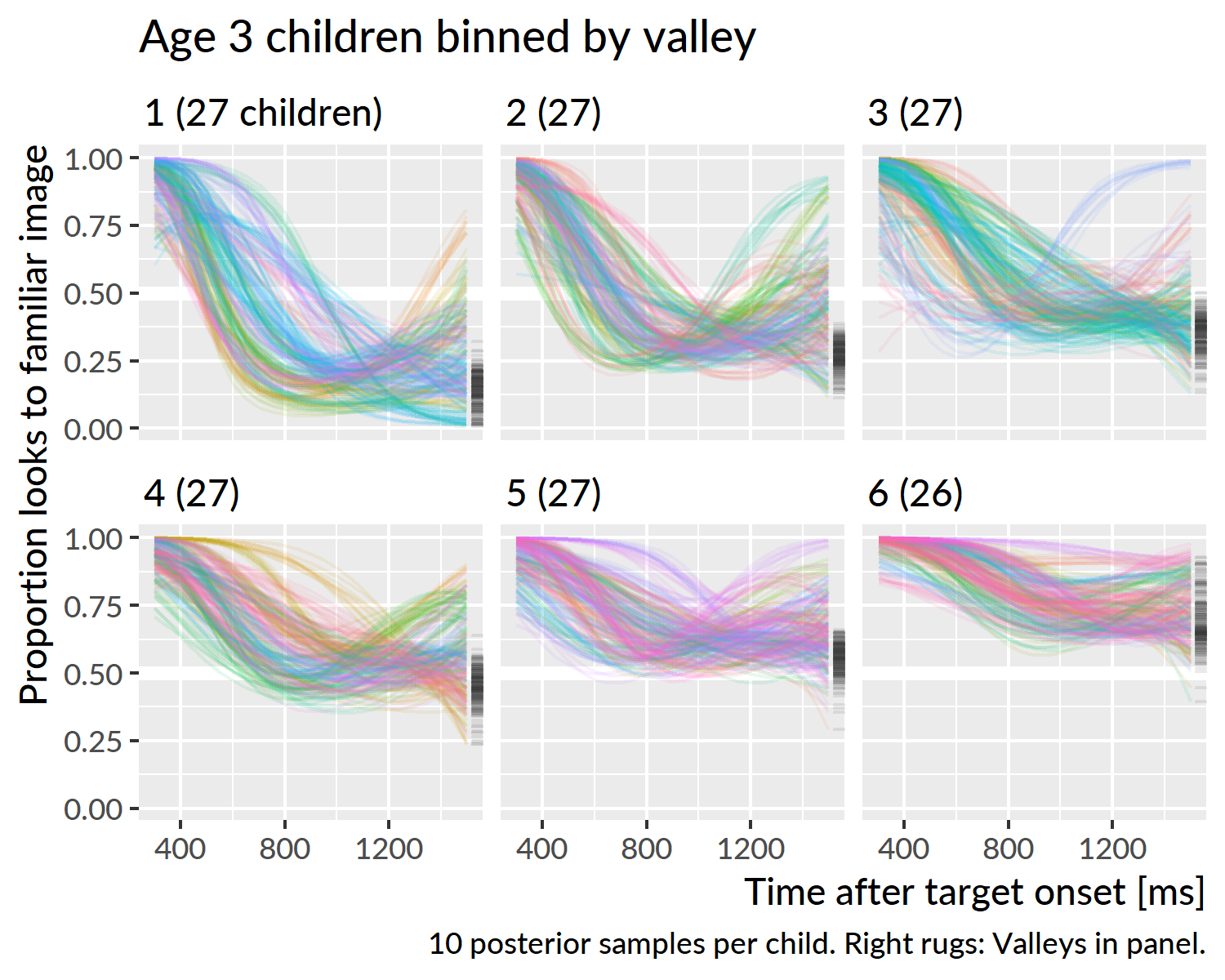


Figure 32 Growth curves for mispronunciations trials starting on the unfamilar object at age 3. Children were grouped into sextiles based on the posterior mean of the growth curve valley—that is, the lowest point on the growth curve. Ten lines are drawn per child to visualize uncertainty. Children were assigned colors arbitrarily. On the right side of each panel are “rugs” which mark the valleys in that panel.

I asked whether any child-level factors predicted children’s looking behaviors. I first regressed growth curve valleys on EVT-2 standard score. There was a small effect at age 3, *R*2 = .09, *n* = 145. A 15-point increase in expressive vocabulary predicted decrease in growth curve valley of .05. At the other ages, the effects are neglibly small, as shown in Figure 33.

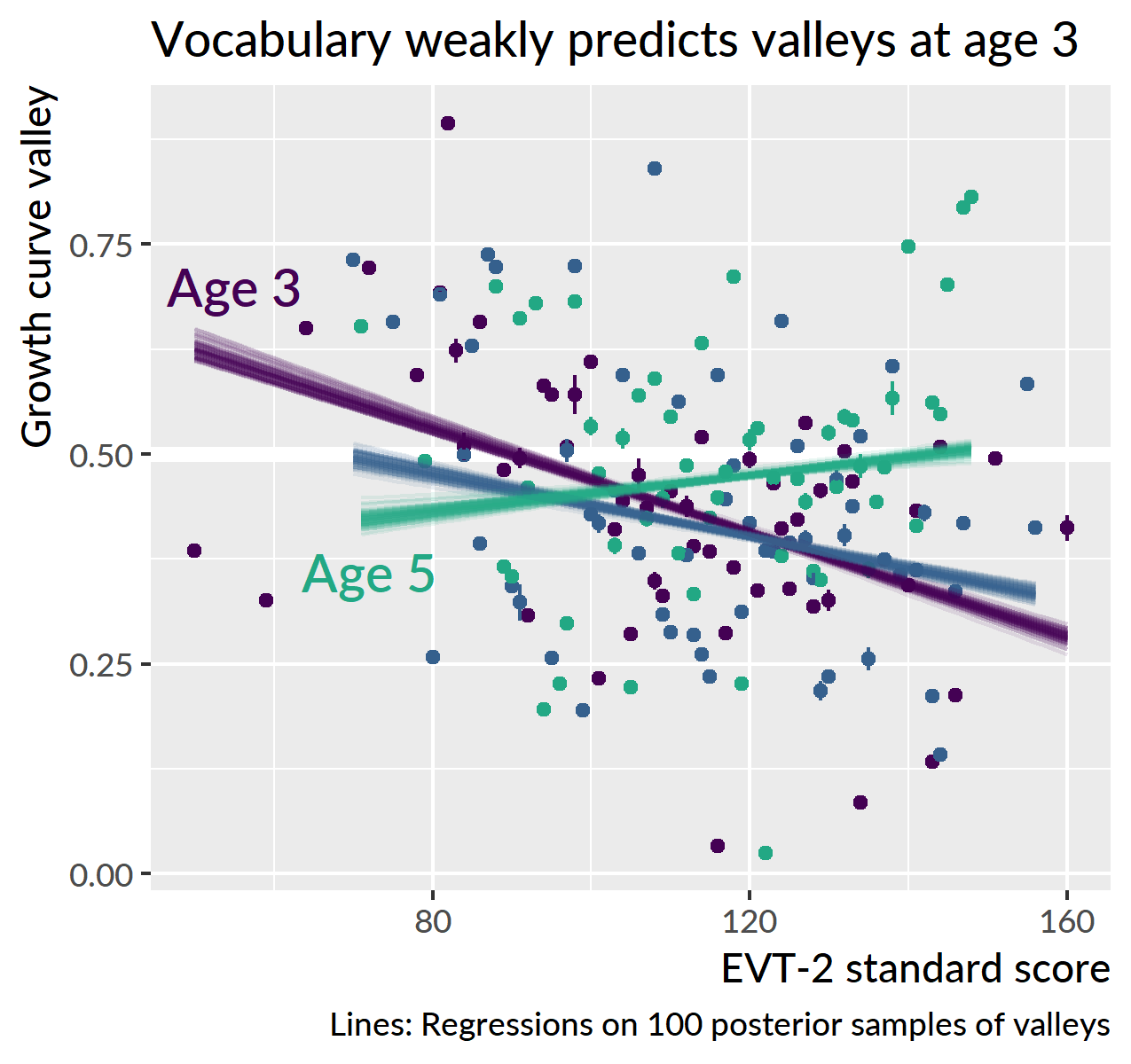


Figure 33 Relationship between expressive vocabulary and growth curve valleys for mispronunciation trials starting on the familiar image. Valleys were computed on 100 draws of the posterior distribution. Points represent the mean and standard error of the valleys. Lines represent the fit of the regression in each draw. Although the slopes of age-4 and age-5 lines are noticeably different from the age-3 one, they cover a tiny amount of the *y* axis so they represent a neglible effect.

I regressed age-3 valleys onto expressive vocabulary, minimal-pair discrimination accuracy, and their interaction. The two main effects and their interaction were all statistically significant, *R*2 = .17, *n* = 139. The effects of vocabulary and minimal-pair discrimination were both negative, so that higher scores on these measures predicted lower growth curve valleys—that is, a greater maximum probability of fixating on the unfamilar image. For an average participant, a 15-point increase in expressive vocabulary predicted decrease of .03, and an increase of minimal pair accuracy of .1 predicted a decrease in valley of .03. The interaction term, however, was positive, meaning that increasing one of the predictors simulataneously weakens the effect of the other. As one of the predictors increases, it can push the effect of the other closer to zero so that its simple effect is “no longer” statistically significant. In this case, the simple effect of expressive vocabulary was not significant when minimal pair accuracy was .71 or greater (that is, at the 60-percentile or greater). Conversely, the simple effect of minimal pair discrimination accuracy was not significant when expressive vocabulary standard score was 119 or greater (at the 60-percentile or greater). In summary, at age 3, both expressive vocabulary and minimal pair discrimination each predicted greater consideration of the unfamiliar image. But these effects was also redundant so that large changes in one predictor would weaken the effect of the other.

The growth curve valley feature measures the maximum extent to which the novel object is considered as the referent on these trials. But the u-shaped growth curves in Figure 32 suggest another listening response on this task: Confirmatory looks to the unfamiliar object. In these u-shaped curves, a child’s probability of fixating on the familiar object temporarily decreases as the novel object is considered, and the probability rises as that interpretation is rejected. To quantify this tendency, I computed each child’s growth curve on the probability scale and re-estimated the quadratic trends in these curves. Exploratory visualization showed that children with higher values on this quadratic trend were more likely to have a u-shape curve. This feature, however, also favored sigmoid or z-shaped curves that rapidly fell and plateaued. To avoid these kinds of curves, I weighted the quadratic trend using the median of the final five points of the curve. The weighted quadratic feature penalized curves that have a strong quadratic trend but end on a low probability. Figure 34 shows growth curves of age 5 children binned using this feature. The bottom row of panels illustrates how the u-shaped feature becomes stronger in each bin. The weighted quadratic feature was weakly correlated with the growth curve valleys, *r* = −.12, and the lack of correlation appears in the figure by how the curves in each panel reach different valleys.

(ref:mp-fam-prop-ot2-bins) Growth curves for mispronunciations trials starting on the unfamilar object at age 5. Children were grouped into sextiles based on the posterior mean of their curves quadratic trend weighted by the height of the curve in the final time bins. Ten lines are drawn per child to visualize uncertainty. Children were assigned colors arbitrarily.

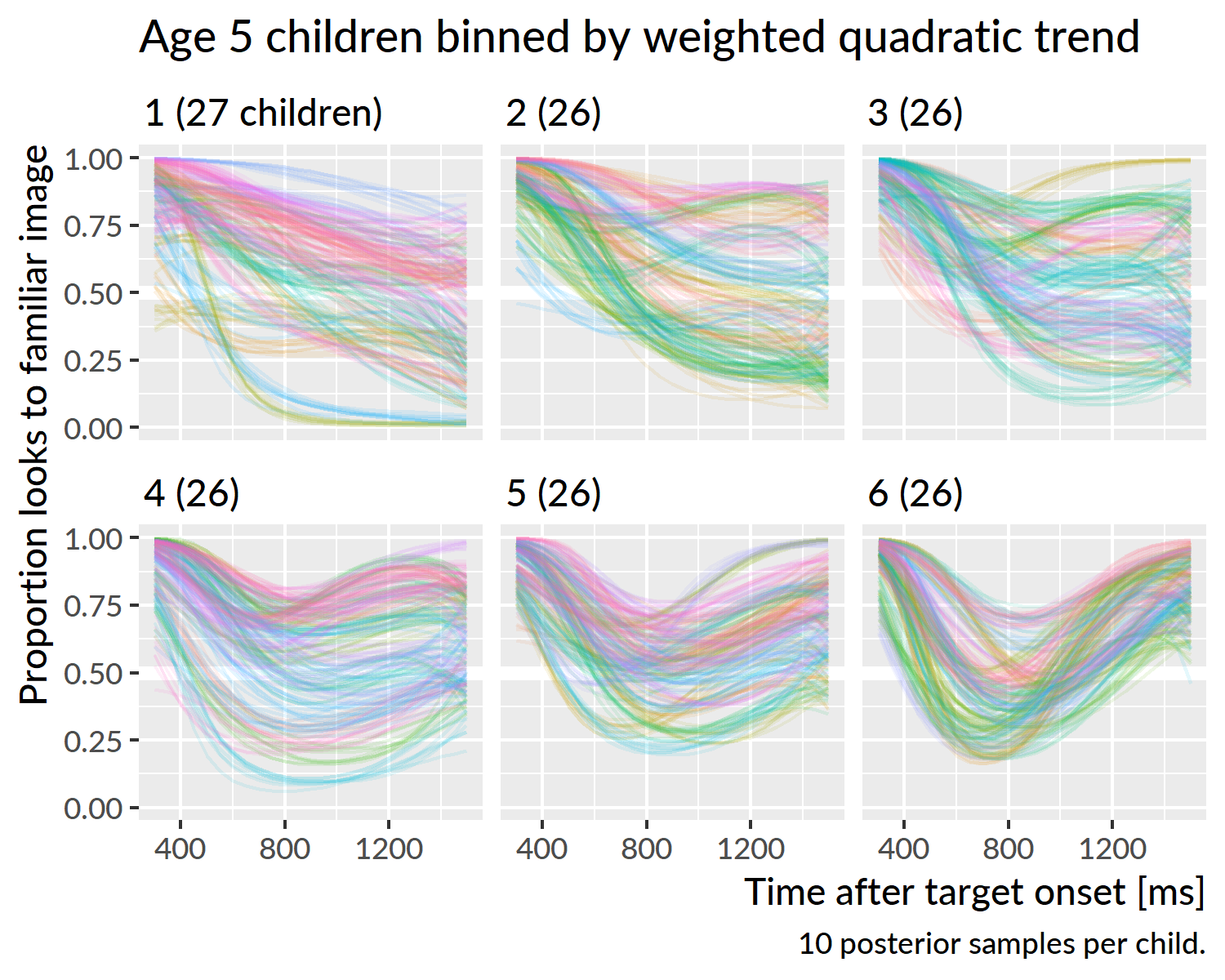


Figure 34 (ref:mp-fam-prop-ot2-bins)

I regressed the u-shaped curve feature onto expressive vocabulary standard score at each age and onto minimal pair discrimination for age 3. There was a tiny yet significant of vocabulary at age 5, *R*2 = .03, where children with lower vocabularies had a slightly stronger u-shaped trend in their growth curve. This effect, however, is so small as to be negligible. None of the effects were significant.

**Summary**. At age 3, children with larger expressive vocabularies or better minimal pair discrimination had lower growth curve valleys—that is, they looked more to the unfamiliar object when they heard a mispronunciation while fixated on an image of the mispronounced word. These two child-level measures significantly interacted so that increasing both measures simultaneously had diminishing returns. I devised a measure of the how u-shaped the growth curves were, but there were not any meaningful effects of vocabulary or expressive vocabulary on this measure.

## Looking behaviors and word learning

At age 5, we tested children’s retention of the novel objects paired with the mispronunciations and nonwords. Children saw the two novel objects and heard either the mispronunciation or the nonword, and they had to point to the objects that went with the image. All six nonwords and mispronunciations were tested.

Table 6 shows the results for each item. Overall, children performed better on the nonwords than the real words. Children performed decidedly better on the nonwords than the mispronunciations on four of pairs, and performed about equally well on the remaining two (*gake*-*pumm*, *wice*-*bape*).

Table 6 Results for the item recognition tests.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Word Group | Type | Item | Trials Correct | Percent Correct |
| cake | mispronunciation | gake | 61 / 107 | 57% ± 4.8 |
|  | nonword | pumm | 56 / 107 | 52% ± 4.8 |
| duck | mispronunciation | guck | 71 / 107 | 66% ± 4.6 |
|  | nonword | shan | 97 / 107 | 91% ± 2.8 |
| girl | mispronunciation | dirl | 71 / 107 | 66% ± 4.6 |
|  | nonword | naydge | 98 / 107 | 92% ± 2.7 |
| rice | mispronunciation | wice | 79 / 107 | 74% ± 4.2 |
|  | nonword | bape | 80 / 107 | 75% ± 4.2 |
| shoes | mispronunciation | sues | 60 / 107 | 56% ± 4.8 |
|  | nonword | geeve | 90 / 107 | 84% ± 3.5 |
| soup | mispronunciation | shoup | 63 / 107 | 59% ± 4.8 |
|  | nonword | cheem | 93 / 107 | 87% ± 3.3 |
| (all) | mispronunciation |  | 405 / 642 | 63% ± 1.9 |
|  | nonword |  | 514 / 642 | 80% ± 1.6 |

I performed an item-response analysis using a mixed-effects logistic regression model. [Appendix 19](#aim2-gca-models) reports the code used to specify the model. The model included varying intercepts for child, child × item type, item, and word-group. The first two effects capture information about a child’s general ability and their ability on each type of item. The second two effects capture information about an item’s difficulty and difficulty of object-pairs. I also asked whether growth curve peaks predicted novel word recognition accuracy, so I included growth curve peaks from each condition in the model. For the nonwords, I used the age-5 peak proportion of looks to the novel image for trials that started on the familiar object. For the mispronunciations, I used the peak looks to the familiar object on trials that started on the unfamiliar object. I chose those peaks based on the conclusion that children were reliably treating the mispronunciations as imperfect productions of the familiar word. The model included data from 101 children.

The model confirmed that children were much more successful on the nonword trials. For a child with an average mispronunciation peak (.72), the predicted proportion correct on the mispronunciation retention trials was .64 [.49, .76]. A 1-SD (.19) increase in the mispronunciation peak predicted a change in proportion correct of −.05 [−.10, −.01]. Children who looked more to the familiar object on these mispronunciation trials were less successful during the retention trials. For a child with an average peak on the nonword trials (.90), the predicted proportion correct on the nonword retention trials was .82 [90% UI: .70, .89]. A 1-SD increase (.10 to 1.00) in nonword peaks predicted a change in proportion correct of −.01 [−.05, .02]. The uncertainty interval here includes positive and negative values. It is uncertain whether the effect is positive or negative, so I conclude that there was not a reliable effect in the nonword case.

Figure 35 visualizes the model results. The difference in height between the two curves reflects the general advantage in the nonword condition. The negative slope for the mispronunciation line captures the effect of growth curve peaks. A change in misprounciation growth curve peak from .5 to 1 roughly predicts a change from 4/6 to 3/6 mispronunciation items correct. The nonword line hovers around 5/6 items correct: There is not enough information in the peaks or in the number of retention trials for a reliable effect to emerge.

Points were jittered by 1% to avoid overplotting. There were six trials per condition which is why the points fall into 6 bands.

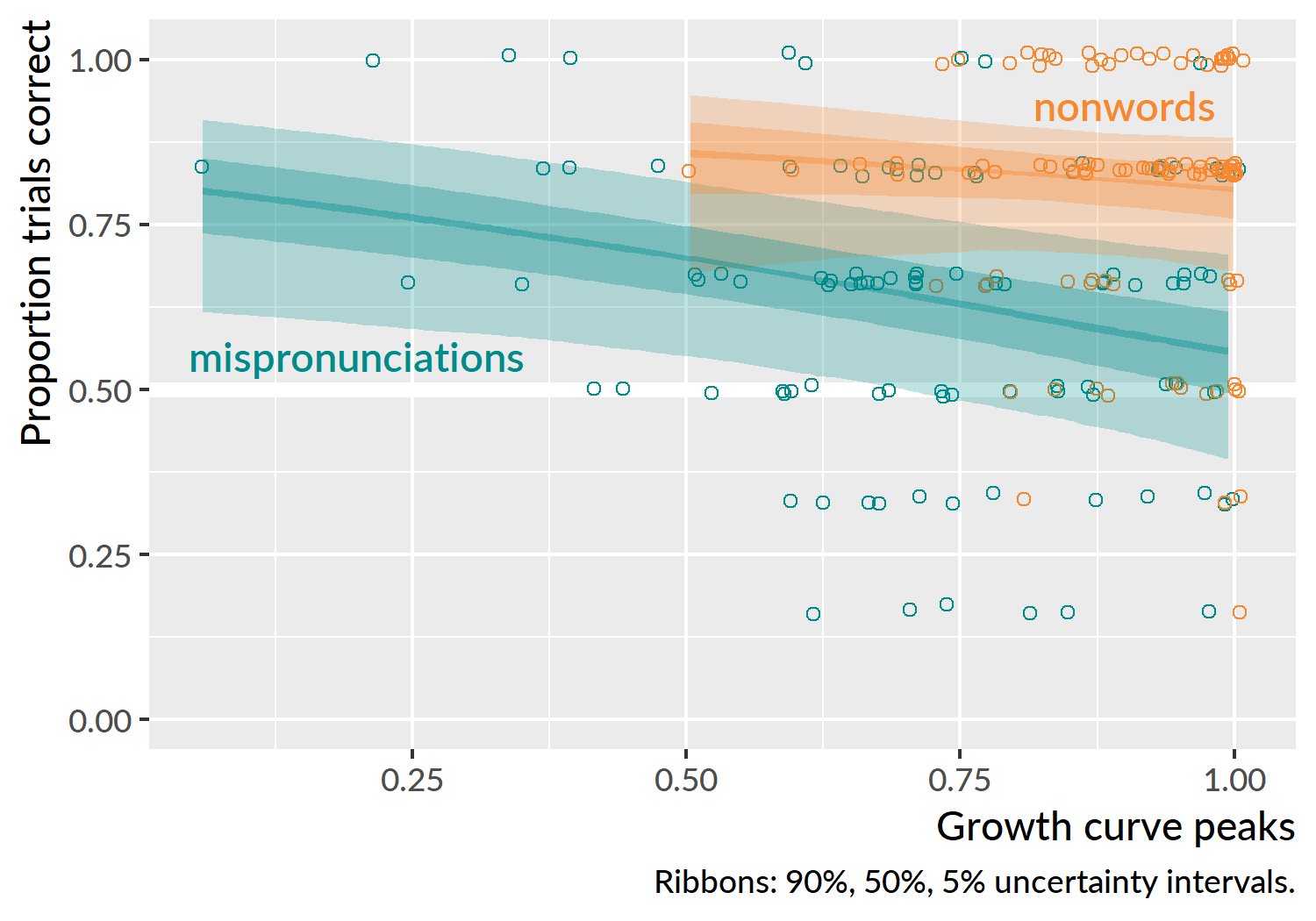


Figure 35 Effect of growth curve peaks on children’s accuracy on retention trials. For the mispronunciations, I used the peak looks to the familiar image on trials where the child started on the unfamilar image, so it represents, say, how much a child looked at *shoes* given “suze”. Thus, more permissive listeners looked performed more poorly on the retention trials. For the nonwords, I used the peak look to the unfamilar image on trials where the child started on the familiar image.

**Summary**. When 5-years-olds were tested on their retention of the unfamiliar images used on the mispronunciations and nonword trials, child were much more accurate for the nonwords than the mispronunciations. Children’s accuracy on the mispronunciations was related to their looking behaviors: Children looked more to the familiar image during mispronunciation trials had a lower accuracy on the mispronunciation retention trials.

## Discussion

In the lab, when preschoolers are looking at a novel object and hear the name of a different familiar object, albeit mispronounced, they look to the familiar object. Children do this reliably at age 3 and even more reliably at age 5. Thus, children recognized these mispronunciations as productions of familiar words, but this recognition was not without a penalty. They looked much less to the familiar object under these conditions, compared to trials where they hear a correct production of the familiar object—see Figure 36–or when they hear a nonword in a context that supports fast-referent selection. Therefore, preschoolers are unquestionably sensitive to mispronunciations of familiar words, as they show more uncertainty when hearing a mispronunciation.

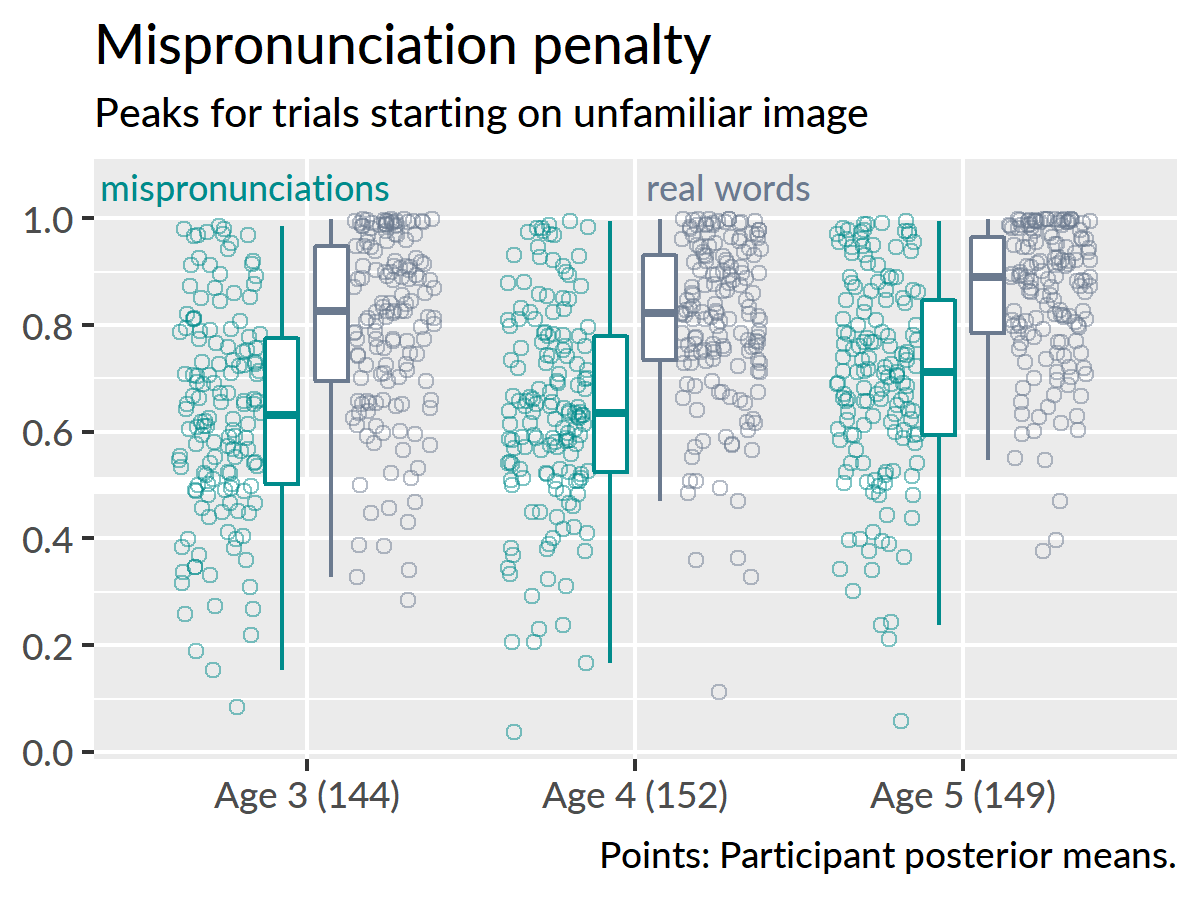


Figure 36 Differences in growth curve peaks for the real words and mispronunciations.

Child-level measures generally did not predict how well children tolerated mispronunciations. There was a very small effect of expressive vocabulary at age 3 such that children with larger vocabularies looked more the familiar image on these trials. Althought children differed in their tendency to look to a familiar image when given a mispronunciation, these differences could not be pinned to any child-level measures.

I also analyzed trials where children started on the familiar word and heard a mispronunciation. In this situation, there is no one clear strategy for referent selection, and children exhibit a few different patterns. Some listeners stay with the familiar image. Some reliably switch to the novel image. Some look at both equally. On average, the growth curve averages rush to .5—equal looking to both images and maximum uncertainty. At age 5, the curve does not reach quite as far down as the other curves, so they never fall into attain this degree of uncertainty. These sets of analyses mainly demonstrate that when children start on a familiar image and hear a mispronunciation, they have a few options for how to proceed.

Child-level predictors only were predictive at age 3 for curve valley. In this case, children with larger vocabularies or better minimal pair discrimination showed more consideration of the nonword object. I speculate that in this situation, the effect reflects that children with better abilities in these areas were more sensitive to the mispronunciation. These children were better at recognizing the mismatch from partial information and thus allocated more credibility to the alternative image.

One strategy to resolve the uncertainty in this situation would be to verify and reject the other image. Therefore, I also defined a weighted quadratic growth curve feature that measured how u-shaped the curves were. Such curves would reflect a child temporarily decreasing looks to the familiar image as a confirmatory looking behaviors. The u-shaped curve features were not reliably related to any child-level factors, outside of neglibly small effect of expressive vocabulary at age 5.

Children demonstrated different looking behaviors based on their initial fixation location. For unfamiliar-initial trials, the growth curves show a reliable shift to the familiar image and we infer that the children treat the mispronunciation as passable production of the familiar word. For the familiar-initial trials, the children show much more uncertainty and a reliable advantage for the familiar word only begins to appear by age 5.

What are children doing in this situation? I initially thought that some children might “be finished” with the trial when they hear the word. That is, the child fixates on the familiar word, hears the mispronunciation prompt, and notice that they have found the image, and then look other parts of the screen. The problem with this possibility is that it does not happen in other conditions. For the nonword and real word conditions, when children start on the named image, they stick with the target. The average empirical growth curves in Figure 19 and Figure 20 tend to stay around 70–80% looking to the target image. Rather than reflecting disengagement from the task, the looking patterns indicate increased uncertainty in these trials.

Another possibility is that children show increased uncertainty because the mispronunciation effect is greater when the child is fixating on the familiar word. That is, children fixated on the familiar image might internally name the object and have built up the word’s resting activation. The mispronunciation directly conflicts with the child’s pre-naming expectations for the word’s name, thus inducing more uncertainty after the word is named. For the unfamiliar-initial trials, on the other hand, a child’s attention is on the novel object and they have a less potent expectation about the words they might hear.

Visual attention did seem to influence how children retained information from the mispronunciation trials. At age 5, we tested children’s retention of the mispronunciations and nonwords. Children were much more likely to recall which unfamiliar object appeared on the nonword trials than the mispronunciation trials. This difference is not unexpected. In the nonword trials, children looked to an unfamiliar object when given an unambiguous novel word for a label. Thus, each trial worked to build an association between a new word and an unfamiliar object. But children generally treated the mispronunciations as productions of the familiar words. For the unfamiliar-initial trials, they looked more to the familiar object, a result that held at all three ages. Rather than developing a new object-label mapping, children are working on resolving an ambiguous and uncertain production on these trials. This idea is consistent with the effect of growth curve peaks on retention accuracy: Child who looked less to the unfamiliar object on mispronunciation trials were less likely to recall that unfamiliar object during retention testing.

# General discussion

This experiment tested how children responded to familiar words, one-feature mispronunciations of those familiar words, and unambiguous novel words. These three types of stimuli allow us to examine features of children’s representations of familiar words and their processing of unfamiliar words.

## A lexical processing account of the results

Children were sensitive to mispronunciations of familiar words. In terms of lexical processing, the mispronounced syllable onset leads a child down a phonological garden path. The “d” in *dirl* activates a neighborhood a d-inital words. The rest of the syllable, however, provides information needed to activate *girl*. Children therefore are slower to build up activation of the word because of the onset-mismatch, and they show less activation overall because the spoken word only matches the rime of the word. Put differently, they get a late start and have to work with a poor-fitting form of the word. The finding that children are better at processing mispronunciations at age 5 suggests that older children are better able to build up more activation to these candidate rime words.

Children also processed mispronunciations differently based on the image they were fixated on. One complication with a simple lexical processing interpretation is that this is a two-image task. Children have ample time to view each image before the noun onset and build up an expectations about the words they might hear named. This possibility leads me to speculate that children might be more uncertain in the trials where they hear “shoup” while fixated on *soup* because they build up the activation of *soup*. This prepotent activation would make the mismatch from the mispronunciation more severe leading to greater uncertainty. This explanation, however, implies some kind of inhibition where *soup* suppresses the activation of “sh”-initial words. The results from Specific Aim 1, which do not provide any evidence for changes in lexical inhibition, make me skeptical of this explanation. Thus, the effect of children’s fixation location on their immediate response to speech provides an avenue for further research in this area.

## A nonword is just a word you haven’t learned yet

I hypothesized that the nonword condition might be more difficult than the real word condition, particularly for younger children as seen in Bion et al. ([2013](#ref-Bion2013)). This prediction did not bear out at all, and indeed, there was an advantage in the nonword condition in later ages. Part of this advantage may be a novelty bias. Mayor and Plunkett ([2014](#ref-TRACE_Mispro)) used the TRACE model of word recognition (McClelland & Elman, [1986](#ref-TRACE)) to simulate these kinds of situations. In one set of simulations, the novel object recieves a novelty/salience boost to resting activation. During presentation of the nonword, none of the child’s known words build up enough activation to overtake the novel word. In another set of simulations, the novel word is added as to the lexicon as a low-frequency word, and the absense of competition of any familiar words causes the novel word to win out. In both of these accounts, children can quickly associate the novel object with a nonword because there are no familiar words to interfere with the process.

The data probably supports both processing accounts. During data reduction, I separated trials based on initial-fixation location because children become increasingly likely to start nonword trials on the novel object. This bias affected 50% of nonword trials at age 3, 55% at age 4, and 59% at age 5. There is a novelty preference that for these trials that gets stronger with development. Children are also developing inchoate lexical representations of these nonwords, consistent with low-frequency lexical item simulations. Children were on average were able to recall the unfamiliar image paired with 5/6 nonwords.

The findings of Swingley and Aslin ([2007](#ref-Swingley2007)) can help us understand the mispronunciation retention results. In that study, toddlers were better able to retain unneighbored nonwords (like *shang* or *meb*) compared to neighbored nonwords (mispronunciations like *tog* [*dog*] or *gall* [*ball*]). They concluded that part of the fast-referent selection involves a probability calculation in which children “evaluat[e] the likelihood that an utterance conveys a new word”.

Mayor and Plunkett ([2014](#ref-TRACE_Mispro)) simulated the nonword processing with TRACE in order to

-nonwords and real words

no dissociation.

-why the nonword advantage

novelty, plus penalized performance

Why do children prefer nonwords over familiar words? The most obvious answer would be a novelty effect. Another reason would be that these are unambiguous words. Because of the mispronunciation trials, a child hears familiar words or bad versions of familiar words. It could be that cost of processing or recovering from mispronunciations makes children more cautious or less certain for familiar words. Whereas for the nonwords, there is no ambiguity.

-mispronunciations

* referential looking predicted retention

## Limitations and implications

Limitations. The biggest limitation to this study is that it applied a procedure designed for toddlers on preschoolers. That is to say, it was too easy for there to be large year-over-year developmental changes in their performance. Children were successful at recognizing real words and fast-selecting referents for nonwords at age 3 and improved a litt

# Hypothesis check (Specific Aim 2)

*Children’s accuracy and efficiency of recognizing real words and fast-associating nonwords will improve each year.*

Yes and no. There were obvious gains for nonwords from 3 to 4 and from 4 to 5 for real words. There are gains, but not “each year”. One complication here is that many children started to hit ceiling performance by age 4, so there was not much of developmental gradient for these conditions.

*Performance in real word recognition and fast association of nonwords will be highly correlated, based on the hypothesis that the same process (referent selection) operates in both situations.*

Yes, there were correlated. But not “highly”.

*Under the alternative hypothesis, real word recognition and fast referent selection reflect different skills with different developmental trajectories. Thus, if there is any dissociation between recognition of real words and nonwords, it will be observed in younger children.*

No—not at all. I observed an dissociation between the two conditions, but it worked *in the opposite direction* and *at the older ages*. That is, children at age 4 and age 5 demonstrated an apparent nonword advantage.

*Although these two measures will be correlated, I predict performance in the nonword condition will be a better predict of future vocabulary growth than performance in the real word condition. This hypothesis is based on the idea that fast referent selection is a more relevant skill for learning new words than recognition of known words.*

Yes, peak looking probabilities for the nonword condition at age 3 predicted expressive vocabulary at age 5, and the real word looking probabilities did not. This size of this effect was rather small.

*For the mispronunciations, I predict children with larger vocabularies (that is, older children) will be more likely to tolerate a mispronunciation as a production of familiar word compared to children with smaller vocabularies.*

Yes, older children looked more to the familiar word on average on the mispronunciation trials.

*Mispronunciations that feature later-mastered sounds (e.g.,* rice*/*wice*) will be more likely to be associated to novel objects than earlier-mastered sounds (*duck*/*guck*).*

Not answered. I present the item-level results in in [Appendix @ref(#aim2-mp-items)](#aim2-mp-items). I chose not to formally analyze the items because there were only 6 mispronunciations each year, making it too difficult to generalize about different kinds of mispronunciations. Plus, children seemed to not know *rice* as well as the other familiar words which makes the problem of measuring a mispronunciation penalty more difficult.

# (PART\*) Afterword

# A needle in a self-organizing haystack

When I give my sales pitch for word recognition research, I sometimes employ a cheeky analogy of a needle in a haystack. Imagine, for the sake of discussion, that the information processing mechanism behind word recognition were just a naïve table search. Think of autocomplete on a phone or search engine text field. Over the course of development, children’s lexicons grow larger in size and yet they get *better* at recognizing words. Children with larger lexicons have to find a needle in a larger haystack—yet this apparent liability is an advantage. That is why the search analogy is naïve. Children become better at recognizing words as they learn more words because they extract regularities and discover similarities among words and develop more efficient lexical representations. The haystack develops regularity and becomes easier to search. Try to pull out a single needle and a tangled mess of related needles come with it.

But there’s a trap here. One promising route to efficient lexical representations is to have children lazily encode lexical information on demand. That’s the idea of holistic or underspecified representations, and it is safe to say that it has been discredited by mispronunciation studies and word recognition studies. So how we take for granted that lexical representations are well specified.

How to proceed?

Leach and Samuel (2007) make a distinction between *lexical configuration* and *lexical engagement*. If we adopt the analogy that the lexicon is a mental dictionary, a word’s configuration is the information found in its dictionary entry: sound, meaning, syntactic roles, and orthography. A word’s engagement is its connection to other words, like competitors or cohorts or semantic relatives. Children’s words are well-configured for the earliest stages of development.

# (APPENDIX) Appendix

# Items used in the visual world experiment

Table 7 lists the items used for the visual world experiment. Each row of the table represents a set of four images used in a trial. There were two blocks of trials with different images and trial orderings. For the two unrelated foils with more than one word listed, the two foils were used in different blocks.

Table 7 Sets of four images used for the Visual World Experiment.

|  |  |  |  |
| --- | --- | --- | --- |
| Target | Phonological | Semantic | Unrelated |
| bear | bell | horse | ring |
| bee | bear | fly | heart |
| bell | bee | drum | swing |
| bread | bear | cheese | vase |
| cheese | shirt | bread | van |
| dress | drum | shirt | swing |
| drum | dress | bell | sword |
| flag | fly | kite | pear |
| fly | flag | bee | pen |
| gift | kite | vase | bread |
| heart | horse | ring | bread/pan |
| horse | heart | bear | pan |
| kite | gift | flag | shirt |
| pan | pear | spoon | vase |
| pear | pen | cheese | ring/vase |
| pen | pear | sword | van |
| ring | swing | dress | flag |
| shirt | cheese | dress | fly |
| spoon | swan | pan | drum |
| swan | spoon | bee | bell |
| swing | spoon | kite | heart |
| sword | swan | pen | gift |
| van | pan | horse | sword |
| vase | van | gift | swan |

For the analysis of phonological competitors, I only used trials where the target and the phonological foil shared the same syllable onset (Table 8). For the analysis of semantic competitors, I only used trials where the target and the semantic foil belonged to the same category (Table 9).

Table 8 Items used for the analysis of phonological versus unrelated competitors.

|  |  |  |
| --- | --- | --- |
| Target | Phonological | Unrelated |
| bear | bell | ring |
| bee | bear | heart |
| bell | bee | swing |
| dress | drum | swing |
| drum | dress | sword |
| flag | fly | pear |
| fly | flag | pen |
| heart | horse | bread/pan |
| horse | heart | pan |
| pan | pear | vase |
| pear | pen | ring/vase |
| pen | pear | van |
| vase | van | swan |

Table 9 Items used for the analysis of semantic versus unrelated competitors.

|  |  |  |
| --- | --- | --- |
| Target | Semantic | Unrelated |
| bear | horse | ring |
| bee | fly | heart |
| bell | drum | swing |
| bread | cheese | vase |
| cheese | bread | van |
| dress | shirt | swing |
| drum | bell | sword |
| fly | bee | pen |
| horse | bear | pan |
| pan | spoon | vase |
| pear | cheese | ring/vase |
| shirt | dress | fly |
| spoon | pan | drum |

# Computational details for Specific Aim 1

## Growth curve analyses

These models were fit in R (vers. 3.4.3; R Core Team, [2018](#ref-R-base)) with the RStanARM package (vers. 2.16.3; Gabry & Goodrich, [2018](#ref-R-rstanarm)).

When I computed the orthogonal polynomial features for Time, they were scaled so that the linear feature ranged from −.5 to .5. Under this scaling a unit change in Time1 was equal to change from the start to the end of the analysis window. Table ?? shows the ranges of the time features.

Table 10 Ranges of the polynomial time features.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Min | Max | Range |
| Time1 | −0.50 | 0.50 | 1.00 |
| Time2 | −0.33 | 0.60 | 0.93 |
| Time3 | −0.63 | 0.63 | 1.26 |
| Trial window (ms) | 250.00 | 1500.00 | 1250.00 |

It took approximately 24 hours to run the model on four Monte Carlo sampling chains with 1000 warm-up iterations and 1000 sampling iterations. Warm-up iterations are discarded, so the model comprises 4000 samples from the posterior distribution.

The code used to fit the model with RStanARM is printed below. The variables ot1, ot2, and ot3 are the polynomial time features, ResearchID identifies children, and Study identifies the age/year of the study. Primary counts the number of looks to the target image at each time bin; Others counts looks to the other three images. cbind(Primary, Others) is used to package both counts together for a logistic regression.

library(rstanarm)  
  
# Run chains on different cores  
options(mc.cores = parallel::detectCores())  
  
m <- stan\_glmer(  
 cbind(Primary, Others) ~  
 (ot1 + ot2 + ot3) \* Study +  
 (ot1 + ot2 + ot3 | ResearchID/Study),  
 family = binomial,  
 prior = normal(0, 1),  
 prior\_intercept = normal(0, 5),  
 prior\_covariance = decov(2, 1, 1),  
 data = d\_m)  
  
# Save the output  
readr::write\_rds(m, "./data/stan\_aim1\_cubic\_model.rds.gz")

The code cbind(Primary, Others) ~ (ot1 + ot2 + ot3) \* Study fits a cubic growth curve for each year of the study. It uses R’s formula syntax to regress the looking counts onto an intercept term (implicitly included by default), ot1, ot2, ot3 along with the interactions of the Study variable with the intercept, ot1, ot2, and ot3.

The line (ot1 + ot2 + ot3 | ResearchID/Study) describes the random-effect structure of the model with the / indicating that data from each Study is nested within each ResearchID. Thus, for each child, we have a general intercept and general effects for Time1, Time2, and Time3. These child-level effects are further adjusted using Study:ResearchID effects. The effects in each level are allowed to correlate. For example, I would expect that participants with low average looking probabilities (low intercepts) to have flatter growth curves (low Time1 effects), and this relationship would be captured by one of the random-effect correlation terms.

Printing the model object reports the point estimates of the model fixed effects and point-estimate correlation matrices for the random effects.

print(m, digits = 2)  
#> stan\_glmer  
#> family: binomial [logit]  
#> formula: cbind(Primary, Others) ~ (ot1 + ot2 + ot3) \* Study + (ot1 + ot2 +   
#> ot3 | ResearchID/Study)  
#> observations: 12584  
#> ------  
#> Median MAD\_SD  
#> (Intercept) -0.47 0.03   
#> ot1 1.57 0.06   
#> ot2 0.05 0.04   
#> ot3 -0.18 0.03   
#> StudyTimePoint2 0.41 0.03   
#> StudyTimePoint3 0.70 0.04   
#> ot1:StudyTimePoint2 0.56 0.08   
#> ot1:StudyTimePoint3 1.10 0.08   
#> ot2:StudyTimePoint2 -0.16 0.05   
#> ot2:StudyTimePoint3 -0.35 0.05   
#> ot3:StudyTimePoint2 -0.12 0.04   
#> ot3:StudyTimePoint3 -0.21 0.04   
#>   
#> Error terms:  
#> Groups Name Std.Dev. Corr   
#> Study:ResearchID (Intercept) 0.3054   
#> ot1 0.6914 0.20   
#> ot2 0.4367 -0.11 0.02   
#> ot3 0.2938 -0.11 -0.44 -0.06  
#> ResearchID (Intercept) 0.2635   
#> ot1 0.4228 0.78   
#> ot2 0.1251 -0.75 -0.56   
#> ot3 0.0576 -0.23 -0.31 0.19  
#> Num. levels: Study:ResearchID 484, ResearchID 195   
#>   
#> Sample avg. posterior predictive distribution of y:  
#> Median MAD\_SD  
#> mean\_PPD 49.86 0.06   
#>   
#> ------  
#> For info on the priors used see help('prior\_summary.stanreg').

The model used the following priors:

prior\_summary(m)  
#> Priors for model 'm'   
#> ------  
#> Intercept (after predictors centered)  
#> ~ normal(location = 0, scale = 5)  
#>   
#> Coefficients  
#> ~ normal(location = [0,0,0,...], scale = [1,1,1,...])  
#> \*\*adjusted scale = [3.33,3.33,3.33,...]  
#>   
#> Covariance  
#> ~ decov(reg. = 2, conc. = 1, shape = 1, scale = 1)  
#> ------  
#> See help('prior\_summary.stanreg') for more details

The priors for the intercept and regression coefficients are wide, very weakly informative normal distributions. These distributions are centered at 0, so negative and positive effects are equally likely. The intercept distribution as a standard deviation of 5, and the coefficients have a standard deviation of around 3. On the log-odds scale, 95% looking to target would be 2.94, so effects of this magnitude are easily accommodated by distributions like Normal(0 [mean], 3 [SD]) and Normal(0, 5).

For the random-effect part of the model, I used RStanARM’s decov() prior which simultaneously sets a prior on the variances and correlations of the model’s random effect terms. I used the default prior for the variance terms and applied a weakly informative LKJ(2) prior on the random-effect correlations. Figure 37 shows samples from the prior distribution of two dummy models fit with the default LKJ(1) prior and the weakly informative LKJ(2) prior used here. Under LKJ(2), extreme correlations are less plausible; the prior shifts the probability mass away from the ±1 boundaries towards the center. The motivation for this kind of prior was *regularization*: I give the model a small amount of information to nudge it away from extreme, degenerate values.

(ref:lkj-prior) Samples of correlation effects drawn from *LKJ*(1) and *LKJ*(2) priors.

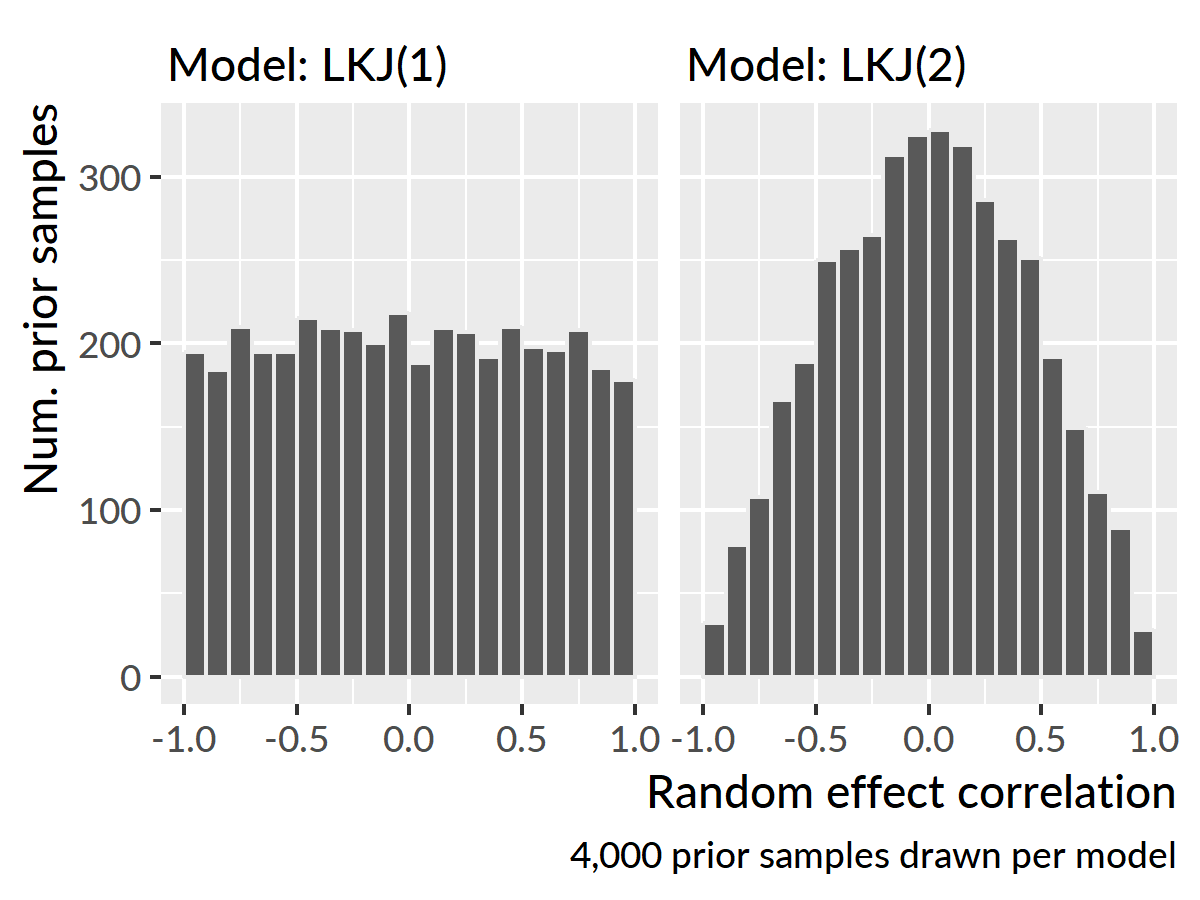


Figure 37 (ref:lkj-prior)

Summary of the familiar word recognition model with diagnostics and 90% uncertainty intervals:

# the last 20 column names are the random effects  
ranef\_names <- tail(colnames(as\_tibble(m)), 20)  
  
summary(  
 object = m,   
 pars = c("alpha", "beta", ranef\_names),  
 probs = c(.05, .95),  
 digits = 3)  
#>   
#> Model Info:  
#>   
#> function: stan\_glmer  
#> family: binomial [logit]  
#> formula: cbind(Primary, Others) ~ (ot1 + ot2 + ot3) \* Study + (ot1 + ot2 +   
#> ot3 | ResearchID/Study)  
#> algorithm: sampling  
#> priors: see help('prior\_summary')  
#> sample: 4000 (posterior sample size)  
#> observations: 12584  
#> groups: Study:ResearchID (484), ResearchID (195)  
#>   
#> Estimates:  
#> mean sd 5% 95%   
#> (Intercept) -0.469 0.032 -0.523 -0.419  
#> ot1 1.575 0.066 1.465 1.682  
#> ot2 0.048 0.038 -0.014 0.110  
#> ot3 -0.175 0.026 -0.218 -0.130  
#> StudyTimePoint2 0.410 0.035 0.355 0.468  
#> StudyTimePoint3 0.697 0.035 0.641 0.757  
#> ot1:StudyTimePoint2 0.565 0.079 0.437 0.695  
#> ot1:StudyTimePoint3 1.099 0.080 0.968 1.233  
#> ot2:StudyTimePoint2 -0.157 0.052 -0.242 -0.073  
#> ot2:StudyTimePoint3 -0.354 0.053 -0.443 -0.267  
#> ot3:StudyTimePoint2 -0.121 0.036 -0.181 -0.061  
#> ot3:StudyTimePoint3 -0.213 0.036 -0.275 -0.155  
#> Sigma[Study:ResearchID:(Intercept),(Intercept)] 0.093 0.008 0.081 0.107  
#> Sigma[Study:ResearchID:ot1,(Intercept)] 0.042 0.013 0.022 0.064  
#> Sigma[Study:ResearchID:ot2,(Intercept)] -0.015 0.008 -0.029 -0.001  
#> Sigma[Study:ResearchID:ot3,(Intercept)] -0.010 0.005 -0.019 -0.001  
#> Sigma[Study:ResearchID:ot1,ot1] 0.478 0.043 0.411 0.551  
#> Sigma[Study:ResearchID:ot2,ot1] 0.006 0.019 -0.026 0.036  
#> Sigma[Study:ResearchID:ot3,ot1] -0.089 0.013 -0.111 -0.069  
#> Sigma[Study:ResearchID:ot2,ot2] 0.191 0.015 0.166 0.217  
#> Sigma[Study:ResearchID:ot3,ot2] -0.007 0.008 -0.019 0.005  
#> Sigma[Study:ResearchID:ot3,ot3] 0.086 0.008 0.074 0.099  
#> Sigma[ResearchID:(Intercept),(Intercept)] 0.069 0.012 0.051 0.090  
#> Sigma[ResearchID:ot1,(Intercept)] 0.087 0.018 0.060 0.117  
#> Sigma[ResearchID:ot2,(Intercept)] -0.025 0.009 -0.040 -0.011  
#> Sigma[ResearchID:ot3,(Intercept)] -0.004 0.004 -0.011 0.003  
#> Sigma[ResearchID:ot1,ot1] 0.179 0.043 0.113 0.252  
#> Sigma[ResearchID:ot2,ot1] -0.030 0.015 -0.056 -0.006  
#> Sigma[ResearchID:ot3,ot1] -0.008 0.008 -0.022 0.004  
#> Sigma[ResearchID:ot2,ot2] 0.016 0.008 0.005 0.030  
#> Sigma[ResearchID:ot3,ot2] 0.001 0.002 -0.002 0.006  
#> Sigma[ResearchID:ot3,ot3] 0.003 0.002 0.001 0.008  
#>   
#> Diagnostics:  
#> mcse Rhat n\_eff  
#> (Intercept) 0.001 1.005 1086   
#> ot1 0.002 1.004 857   
#> ot2 0.001 1.006 842   
#> ot3 0.001 1.002 1156   
#> StudyTimePoint2 0.001 1.007 1034   
#> StudyTimePoint3 0.001 1.006 959   
#> ot1:StudyTimePoint2 0.003 1.014 674   
#> ot1:StudyTimePoint3 0.003 1.005 934   
#> ot2:StudyTimePoint2 0.002 1.003 836   
#> ot2:StudyTimePoint3 0.002 1.006 762   
#> ot3:StudyTimePoint2 0.001 1.003 1183   
#> ot3:StudyTimePoint3 0.001 1.001 1390   
#> Sigma[Study:ResearchID:(Intercept),(Intercept)] 0.000 1.002 1093   
#> Sigma[Study:ResearchID:ot1,(Intercept)] 0.001 1.009 475   
#> Sigma[Study:ResearchID:ot2,(Intercept)] 0.000 1.023 323   
#> Sigma[Study:ResearchID:ot3,(Intercept)] 0.000 1.003 792   
#> Sigma[Study:ResearchID:ot1,ot1] 0.002 1.003 547   
#> Sigma[Study:ResearchID:ot2,ot1] 0.001 1.013 277   
#> Sigma[Study:ResearchID:ot3,ot1] 0.000 1.005 806   
#> Sigma[Study:ResearchID:ot2,ot2] 0.001 1.010 665   
#> Sigma[Study:ResearchID:ot3,ot2] 0.000 1.001 1131   
#> Sigma[Study:ResearchID:ot3,ot3] 0.000 1.004 1220   
#> Sigma[ResearchID:(Intercept),(Intercept)] 0.000 1.004 913   
#> Sigma[ResearchID:ot1,(Intercept)] 0.001 1.008 636   
#> Sigma[ResearchID:ot2,(Intercept)] 0.001 1.026 307   
#> Sigma[ResearchID:ot3,(Intercept)] 0.000 1.006 711   
#> Sigma[ResearchID:ot1,ot1] 0.003 1.010 261   
#> Sigma[ResearchID:ot2,ot1] 0.001 1.021 242   
#> Sigma[ResearchID:ot3,ot1] 0.000 1.006 331   
#> Sigma[ResearchID:ot2,ot2] 0.000 1.037 257   
#> Sigma[ResearchID:ot3,ot2] 0.000 1.009 439   
#> Sigma[ResearchID:ot3,ot3] 0.000 1.007 340   
#>   
#> For each parameter, mcse is Monte Carlo standard error, n\_eff is a crude measure of effective sample size, and Rhat is the potential scale reduction factor on split chains (at convergence Rhat=1).

## Generalized additive models

To model the looks to the competitor images, I used generalized additive (mixed) models. The models were fit in R (vers. 3.4.3) using the mgcv R package (vers. 1.8.23; Wood, [2017](#ref-Wood2017)) with support from tools in the itsadug R package (vers. 2.3; van Rij et al., [2017](#ref-itsadug)).

I will briefly walk through the code used to fit one of these models in order to articulate the modeling decisions at play. I first convert the categorical variables into the right types, so that the model can fit difference smooths.

# Create a Study dummy variabe with Age 4 as the reference level  
phon\_d$S <- factor(phon\_d$Study, c("TimePoint2", "TimePoint1", "TimePoint3"))  
  
# Convert the ResearchID into a factor  
phon\_d$R <- as.factor(phon\_d$ResearchID)  
  
# Convert the Study factor (phon\_d$S) into an ordered factor.  
# This step is needed for the ti model estimate difference smooths.  
phon\_d$S2 <- as.ordered(phon\_d$S)  
contrasts(phon\_d$S2) <- "contr.treatment"  
contrasts(phon\_d$S2)

I fit the generalized additive model with the code below. The outcome elog is the empirical log-odds of looking to the phonological competitor relative to the unrelated word.

library(mgcv)  
  
phon\_gam <- bam(  
 elog ~ S2 +  
 s(Time) + s(Time, by = S2) +  
 s(Time, R, bs = "fs", m = 1, k = 5),  
 data = phon\_d)  
  
# Save the output  
readr::write\_rds(phon\_gam, "./data/aim1-phon-random-smooths.rds.gz")

There is just one parametric term: S2. The term computes the average effect of each study with Age 4 serving as the reference condition (and as the model intercept).

Next come the smooth terms. s(Time) fits the shape of Time for the reference condition (Age 4). s(Time, by = S2) fits the difference smooths for Age 3 versus Age 4 and Age 5 versus Age 4. s(Time, R, bs = "fs", m = 1, k = 5) fits a smooth for each participant (R). bs = "fs" means that the model should use a factor smooth (fs) basis (bs)—that is, a “random effect” smooth for each participant. m = 1 changes the smoothness penalty so that the random effects are pulled towards the group average; Winter and Wieling ([2016](#ref-Winter2016)) and Baayen, Rij, Cat, and Wood ([2016](#ref-Baayen2016)) suggest using this option. k = 5 means to use 5 knots (k) for the basis function. The other smooths use the default number of knots (10). I used fewer knots for the by-child smooths because of limited data. As a result, these smooths capture by-child variation by making coarse adjustments to study-level growth curves.

Summary of the phonological model:

m\_p <- readr::read\_rds("./data/aim1-phon-random-smooths.rds.gz")  
summary(m\_p)  
#>   
#> Family: gaussian   
#> Link function: identity   
#>   
#> Formula:  
#> elog ~ S2 + s(Time) + s(Time, by = S2) + s(Time, R, bs = "fs",   
#> m = 1, k = 5)  
#>   
#> Parametric coefficients:  
#> Estimate Std. Error t value Pr(>|t|)   
#> (Intercept) 0.159200 0.048807 3.262 0.00111 \*\*   
#> S2TimePoint1 -0.002641 0.013840 -0.191 0.84864   
#> S2TimePoint3 0.151073 0.013601 11.107 < 2e-16 \*\*\*  
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Approximate significance of smooth terms:  
#> edf Ref.df F p-value   
#> s(Time) 7.277 8.165 10.61 3.51e-15 \*\*\*  
#> s(Time):S2TimePoint1 5.478 6.590 17.10 < 2e-16 \*\*\*  
#> s(Time):S2TimePoint3 1.001 1.002 17.86 2.37e-05 \*\*\*  
#> s(Time,R) 852.928 974.000 12.97 < 2e-16 \*\*\*  
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> R-sq.(adj) = 0.311 Deviance explained = 33.1%  
#> fREML = 41726 Scale est. = 0.8629 n = 30008

Summary of the semantic model:

m\_s <- readr::read\_rds("./data/aim1-semy-random-smooths.rds.gz")  
summary(m\_s)  
#>   
#> Family: gaussian   
#> Link function: identity   
#>   
#> Formula:  
#> elog ~ S2 + s(Time) + s(Time, by = S2) + s(Time, R, bs = "fs",   
#> m = 1, k = 5)  
#>   
#> Parametric coefficients:  
#> Estimate Std. Error t value Pr(>|t|)   
#> (Intercept) 0.43878 0.04907 8.943 < 2e-16 \*\*\*  
#> S2TimePoint1 -0.13985 0.01352 -10.345 < 2e-16 \*\*\*  
#> S2TimePoint3 0.06486 0.01329 4.881 1.06e-06 \*\*\*  
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Approximate significance of smooth terms:  
#> edf Ref.df F p-value   
#> s(Time) 7.038 7.988 11.018 1.16e-15 \*\*\*  
#> s(Time):S2TimePoint1 1.001 1.001 0.387 0.534636   
#> s(Time):S2TimePoint3 3.739 4.623 4.909 0.000323 \*\*\*  
#> s(Time,R) 867.572 974.000 15.750 < 2e-16 \*\*\*  
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> R-sq.(adj) = 0.379 Deviance explained = 39.7%  
#> fREML = 42860 Scale est. = 0.85001 n = 30976

# Items used in the mispronunciation experiment

The stimuli changed between Year 1 and Year 2 so that dog/tog was replaced with rice/wice.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Time Points** | **Word Group** | **Condition** | **Audio (IPA)** | **Familiar Object** | **Unfamiliar Object** |
| 1 | dog | Correct Production | /dɔg/ | dog | wombat |
|  |  | Mispronunciation | /tɔg/ | dog | wombat |
|  |  | Nonword | /vef/ | ball | sextant |
| 1, 2, 3 | cake | Correct Production | /kek/ | cake | horned melon |
|  |  | Mispronunciation | /gek/ | cake | horned melon |
|  |  | Nonword | /pʌm/ | book | churn |
| 1, 2, 3 | duck | Correct Production | /dʌk/ | duck | toy creature |
|  |  | Mispronunciation | /gʌk/ | duck | toy creature |
|  |  | Nonword | /ʃæn/ | cup | reed |
| 1, 2, 3 | girl | Correct Production | /gɜ˞l/ | girl | marmoset |
|  |  | Mispronunciation | /dɜ˞l/ | girl | marmoset |
|  |  | Nonword | /nedʒ/ | car | work holder |
| 1, 2, 3 | shoes | Correct Production | /ʃuz/ | shoes | flasks |
|  |  | Mispronunciation | /suz/ | shoes | flasks |
|  |  | Nonword | /giv/ | sock | trolley |
| 1, 2, 3 | soup | Correct Production | /sup/ | soup | steamer |
|  |  | Mispronunciation | /ʃup/ | soup | steamer |
|  |  | Nonword | /tʃim/ | bed | pastry mixer |
| 2, 3 | rice | Correct Production | /ɹaɪs/ | rice | anise |
|  |  | Mispronunciation | /waɪs/ | rice | anise |
|  |  | Nonword | /bep/ | ball | sextant |

# Computational details for Specific Aim 2

## Real words versus nonwords growth curves

These models were fit in R (vers. 3.5.0; R Core Team, [2018](#ref-R-base)) with the brms package (vers. 2.3.1; Bürkner, [2017](#ref-brms)).

The orthogonal polynomial features for Time, they were scaled as in Specific Aim 1, so that the linear feature ranged from −.5 to .5. Under this scaling a unit change in Time1 was equal to change from the start to the end of the analysis window.

The model formula used to specify the model with brms is printed below. The variables ot1, ot2, and ot3 are the polynomial time features, ResearchID identifies children, and Condition identifies the experimental condition (either, the nonword or real word condition). Target counts the number of looks to the target image at each time bin; trials() is a flag that tells brms the number of trials for the binomial process. Here, it is the variable Trials, which is equal to the number of looks to target and distractor in each bin. The syntax (1 + ot1 + ot2 + ot3) \* Condition specifies a Time x Condition interaction; it says to estimate an intercept and the three time feature effects for each condition. The line (ot1 + ot2 + ot3 | ResearchID/Condition) describes the random-effect structure of the model with the / indicating that data from each Condition is nested within each ResearchID.

library(brms)  
  
# Fit a hierarchical logistic regression model  
formula <- bf(  
 Target | trials(Trials) ~   
 (1 + ot1 + ot2 + ot3) \* Condition +   
 (1 + ot1 + ot2 + ot3 | ResearchID/Condition),   
 family = binomial)

The priors for the model are described below. The regression effects (class = "b") have a moderately informative prior of Normal(0, 1). Because most of the action in the growth curves is a sharp rise, the linear time effect ot1 has a slightly wider prior of Normal(0, 2). A weakly informative LKJ(2) prior is put on the random-effect correlations. I review the role of the LKJ prior in  
[Appendix 17](#aim1-gca-models). A weakly informative prior is put on the random-effect standard deviations Student-*t*([df] 7, [mean] 0, [sd] 3). The Student-*t* distribution is like the normal distribution but it provides slightly thicker tails which allow extreme or outlying values.

priors <- c(  
 # Population-average intercept  
 set\_prior(class = "Intercept", "normal(0, 1)"),  
 # Population-average slopes  
 set\_prior(class = "b", "normal(0, 1)"),  
 # ... expect somewhat larger range of effects for linear time  
 set\_prior(class = "b", coef = "ot1", "normal(0, 2)"),  
 # Correlations for random effect terms  
 set\_prior(class = "cor", "lkj(2)"),  
 # Standard deviation of the distribution from  
 # which random-intercepts are drawn  
 set\_prior(class = "sd", "student\_t(7, 0, 3)"))

I originally tried a single model containing all three years with corresponding year effects, year × time interactions, and year × condition × time interactions, but this model took 30 hours to run and did not converge. Therefore, I fit separate models for each year of the study using syntax like the following.

m\_age3 <- brm(  
 formula = formula,  
 data = d\_age3,  
 prior = priors,  
 chains = 4,  
 iter = 2000,  
 cores = 4,  
 control = list(adapt\_delta = .99))  
  
# Save the output  
readr::write\_rds(m\_age3, "age3\_mp.rds.gz")

This code fits the model using four sampling chains in parallel over four processing cores. Early attempts at the model produced warnings, so I increased the adapt\_delta control option to make the sampling more robust and eliminate the warnings.

Model summary for real words versus nonwords at age 3:

m1 <- readr::read\_rds("./data/aim2-real-vs-nw-tp1.rds.gz")  
summary(m1, priors = TRUE, prob = .9)  
#> Family: binomial   
#> Links: mu = logit   
#> Formula: Target | trials(Trials) ~ (ot1 + ot2 + ot3) \* Condition + (ot1 + ot2 + ot3 | ResearchID/Condition)   
#> Data: test\_data\_1 (Number of observations: 7450)   
#> Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;  
#> total post-warmup samples = 4000  
#>   
#> Priors:   
#> b ~ normal(0, 1)  
#> b\_ot1 ~ normal(0, 2)  
#> Intercept ~ normal(0, 1)  
#> L ~ lkj\_corr\_cholesky(2)  
#> sd ~ student\_t(7, 0, 3)  
#>   
#> Group-Level Effects:   
#> ~ResearchID (Number of levels: 149)   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> sd(Intercept) 0.73 0.13 0.51 0.93 290 1.00  
#> sd(ot1) 0.64 0.40 0.05 1.31 68 1.03  
#> sd(ot2) 0.34 0.19 0.04 0.65 130 1.02  
#> sd(ot3) 0.14 0.10 0.01 0.33 233 1.01  
#> cor(Intercept,ot1) 0.25 0.33 -0.36 0.72 176 1.02  
#> cor(Intercept,ot2) -0.26 0.32 -0.73 0.32 838 1.00  
#> cor(ot1,ot2) -0.08 0.37 -0.64 0.56 380 1.01  
#> cor(Intercept,ot3) 0.12 0.34 -0.47 0.65 1338 1.00  
#> cor(ot1,ot3) 0.06 0.38 -0.57 0.66 740 1.00  
#> cor(ot2,ot3) -0.01 0.37 -0.60 0.61 1226 1.00  
#>   
#> ~ResearchID:Condition (Number of levels: 298)   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> sd(Intercept) 1.24 0.08 1.12 1.37 611 1.00  
#> sd(ot1) 2.66 0.16 2.40 2.93 320 1.01  
#> sd(ot2) 1.40 0.09 1.25 1.55 572 1.00  
#> sd(ot3) 0.83 0.05 0.74 0.92 923 1.00  
#> cor(Intercept,ot1) 0.29 0.08 0.15 0.41 407 1.01  
#> cor(Intercept,ot2) 0.06 0.10 -0.10 0.21 694 1.00  
#> cor(ot1,ot2) 0.01 0.09 -0.14 0.15 631 1.00  
#> cor(Intercept,ot3) -0.10 0.09 -0.25 0.06 945 1.00  
#> cor(ot1,ot3) -0.06 0.09 -0.21 0.10 1202 1.00  
#> cor(ot2,ot3) -0.06 0.08 -0.20 0.07 1128 1.00  
#>   
#> Population-Level Effects:   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> Intercept 0.41 0.12 0.21 0.61 1392 1.00  
#> ot1 4.59 0.24 4.20 4.99 744 1.01  
#> ot2 -1.36 0.13 -1.57 -1.15 1167 1.00  
#> ot3 0.39 0.08 0.25 0.52 1828 1.00  
#> Conditionreal -0.19 0.15 -0.43 0.05 1550 1.00  
#> ot1:Conditionreal 0.45 0.31 -0.05 0.94 673 1.01  
#> ot2:Conditionreal -0.02 0.17 -0.30 0.26 1077 1.00  
#> ot3:Conditionreal -0.07 0.11 -0.25 0.12 1674 1.00  
#>   
#> Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample   
#> is a crude measure of effective sample size, and Rhat is the potential   
#> scale reduction factor on split chains (at convergence, Rhat = 1).

Model summary for real words versus nonwords at age 4:

m2 <- readr::read\_rds("./data/aim2-real-vs-nw-tp2.rds.gz")  
summary(m2, priors = TRUE, prob = .9)  
#> Family: binomial   
#> Links: mu = logit   
#> Formula: Target | trials(Trials) ~ (ot1 + ot2 + ot3) \* Condition + (ot1 + ot2 + ot3 | ResearchID/Condition)   
#> Data: test\_data\_2 (Number of observations: 7750)   
#> Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;  
#> total post-warmup samples = 4000  
#>   
#> Priors:   
#> b ~ normal(0, 1)  
#> b\_ot1 ~ normal(0, 2)  
#> Intercept ~ normal(0, 1)  
#> L ~ lkj\_corr\_cholesky(2)  
#> sd ~ student\_t(7, 0, 3)  
#>   
#> Group-Level Effects:   
#> ~ResearchID (Number of levels: 155)   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> sd(Intercept) 0.64 0.09 0.49 0.79 461 1.01  
#> sd(ot1) 0.65 0.35 0.09 1.22 65 1.04  
#> sd(ot2) 0.31 0.19 0.03 0.63 155 1.03  
#> sd(ot3) 0.15 0.10 0.01 0.34 218 1.01  
#> cor(Intercept,ot1) 0.03 0.30 -0.51 0.50 666 1.01  
#> cor(Intercept,ot2) -0.25 0.32 -0.72 0.35 518 1.00  
#> cor(ot1,ot2) 0.04 0.36 -0.57 0.62 570 1.00  
#> cor(Intercept,ot3) -0.00 0.34 -0.55 0.57 1435 1.00  
#> cor(ot1,ot3) 0.01 0.37 -0.61 0.61 779 1.00  
#> cor(ot2,ot3) -0.02 0.37 -0.61 0.60 599 1.01  
#>   
#> ~ResearchID:Condition (Number of levels: 310)   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> sd(Intercept) 1.02 0.06 0.92 1.12 900 1.00  
#> sd(ot1) 2.54 0.16 2.28 2.81 296 1.01  
#> sd(ot2) 1.51 0.09 1.37 1.66 538 1.01  
#> sd(ot3) 0.87 0.05 0.78 0.96 1082 1.00  
#> cor(Intercept,ot1) 0.53 0.06 0.42 0.62 589 1.00  
#> cor(Intercept,ot2) -0.17 0.08 -0.30 -0.03 841 1.01  
#> cor(ot1,ot2) 0.10 0.08 -0.04 0.23 1010 1.00  
#> cor(Intercept,ot3) -0.22 0.09 -0.36 -0.08 922 1.00  
#> cor(ot1,ot3) -0.23 0.08 -0.36 -0.09 1372 1.00  
#> cor(ot2,ot3) -0.03 0.08 -0.16 0.10 1171 1.00  
#>   
#> Population-Level Effects:   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> Intercept 1.32 0.10 1.16 1.49 1613 1.00  
#> ot1 4.57 0.22 4.21 4.94 1397 1.00  
#> ot2 -1.71 0.14 -1.94 -1.49 1201 1.00  
#> ot3 0.41 0.09 0.27 0.55 1662 1.00  
#> Conditionreal -0.82 0.12 -1.01 -0.62 1401 1.00  
#> ot1:Conditionreal -0.51 0.29 -0.96 -0.04 1313 1.00  
#> ot2:Conditionreal 0.17 0.18 -0.13 0.47 1149 1.01  
#> ot3:Conditionreal -0.07 0.11 -0.25 0.11 1563 1.00  
#>   
#> Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample   
#> is a crude measure of effective sample size, and Rhat is the potential   
#> scale reduction factor on split chains (at convergence, Rhat = 1).

Model summary for real words versus nonwords at age 5:

m3 <- readr::read\_rds("./data/aim2-real-vs-nw-tp3.rds.gz")  
summary(m3, priors = TRUE, prob = .9)  
#> Family: binomial   
#> Links: mu = logit   
#> Formula: Target | trials(Trials) ~ (ot1 + ot2 + ot3) \* Condition + (ot1 + ot2 + ot3 | ResearchID/Condition)   
#> Data: test\_data\_3 (Number of observations: 7550)   
#> Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;  
#> total post-warmup samples = 4000  
#>   
#> Priors:   
#> b ~ normal(0, 1)  
#> b\_ot1 ~ normal(0, 2)  
#> Intercept ~ normal(0, 1)  
#> L ~ lkj\_corr\_cholesky(2)  
#> sd ~ student\_t(7, 0, 3)  
#>   
#> Group-Level Effects:   
#> ~ResearchID (Number of levels: 151)   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> sd(Intercept) 0.22 0.15 0.02 0.49 45 1.07  
#> sd(ot1) 0.48 0.31 0.05 1.02 52 1.06  
#> sd(ot2) 0.26 0.17 0.02 0.57 102 1.03  
#> sd(ot3) 0.20 0.11 0.03 0.39 219 1.03  
#> cor(Intercept,ot1) 0.21 0.38 -0.47 0.76 135 1.02  
#> cor(Intercept,ot2) -0.16 0.38 -0.72 0.52 241 1.01  
#> cor(ot1,ot2) -0.04 0.37 -0.65 0.59 443 1.01  
#> cor(Intercept,ot3) 0.14 0.36 -0.48 0.69 478 1.00  
#> cor(ot1,ot3) -0.03 0.37 -0.63 0.60 739 1.01  
#> cor(ot2,ot3) -0.05 0.37 -0.64 0.57 626 1.00  
#>   
#> ~ResearchID:Condition (Number of levels: 302)   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> sd(Intercept) 1.18 0.07 1.07 1.28 210 1.02  
#> sd(ot1) 2.78 0.16 2.51 3.05 429 1.01  
#> sd(ot2) 1.54 0.09 1.41 1.69 920 1.00  
#> sd(ot3) 0.88 0.06 0.79 0.98 771 1.01  
#> cor(Intercept,ot1) 0.59 0.05 0.50 0.67 845 1.00  
#> cor(Intercept,ot2) -0.13 0.08 -0.26 0.01 753 1.02  
#> cor(ot1,ot2) 0.09 0.08 -0.04 0.23 953 1.01  
#> cor(Intercept,ot3) -0.36 0.08 -0.48 -0.23 860 1.00  
#> cor(ot1,ot3) -0.39 0.08 -0.52 -0.27 1492 1.00  
#> cor(ot2,ot3) 0.15 0.08 0.01 0.28 1570 1.00  
#>   
#> Population-Level Effects:   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> Intercept 1.42 0.10 1.26 1.57 646 1.01  
#> ot1 4.67 0.23 4.30 5.07 602 1.01  
#> ot2 -1.70 0.14 -1.94 -1.47 1228 1.00  
#> ot3 0.28 0.09 0.13 0.42 1598 1.00  
#> Conditionreal -0.48 0.13 -0.70 -0.27 556 1.01  
#> ot1:Conditionreal -0.13 0.30 -0.64 0.37 702 1.00  
#> ot2:Conditionreal 0.19 0.19 -0.13 0.49 846 1.01  
#> ot3:Conditionreal 0.03 0.12 -0.16 0.22 1242 1.00  
#>   
#> Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample   
#> is a crude measure of effective sample size, and Rhat is the potential   
#> scale reduction factor on split chains (at convergence, Rhat = 1).

## Mispronunication growth curves

Like the models above, these ones are Bayesian mixed-effects logistic regression growth curve models fit with brms. I used two separate models, one for unfamilar-initial trials and familiar-initial trials. Each model included data from all three years of the study. The code is essentially the same syntax with a Study variable replacing the Condition variable.

library(brms)  
  
# Fit a hierarchical logistic regression model  
formula <- bf(  
 Target | trials(Trials) ~   
 (1 + ot1 + ot2 + ot3) \* Study +   
 (1 + ot1 + ot2 + ot3 | ResearchID/Study),   
 family = binomial)  
  
priors <- c(  
 set\_prior(class = "Intercept", "normal(0, 1)"),  
 set\_prior(class = "b", "normal(0, 1)"),  
 set\_prior(class = "b", coef = "ot1", "normal(0, 2)"),  
 set\_prior(class = "cor", "lkj(2)"),  
 set\_prior(class = "sd", "student\_t(7, 0, 2)"))  
  
mp\_unfam <- brm(  
 formula = formula,  
 data = d\_u,  
 prior = priors,  
 chains = 4,  
 cores = 4,  
 # Run extra iterations to get a higher effective sample size  
 iter = 3000,  
 control = list(  
 adapt\_delta = .995,   
 max\_treedepth = 15))  
  
# Save the output  
readr::write\_rds(mp\_unfam, "./data/aim2-mp-unfam.rds.gz")  
  
mp\_fam <- brm(  
 formula = formula,  
 data = d\_f,  
 prior = priors,  
 chains = 4,  
 cores = 4,  
 control = list(  
 adapt\_delta = .99,   
 max\_treedepth = 15))  
  
# Save the output  
readr::write\_rds(mp\_fam, "./data/aim2-mp-fam.rds.gz")

The priors for this model are the same, except for a tighter prior on scale/standard deviations for the random effect distributions. The model had difficulty obtaining an effective number of samples for these parameters. Initially, I tried to tell the model to do more work on each sampling step (adapt\_delta = .995 and max\_treedepth = 15) and run the chains for 50% longer (iter = 3000). These changes did not solve the problem. By using a tighter prior, the model had a smaller search space meaning it could obtain samples more efficiently.

The revised prior was still weakly informative. Figure 38 llustrates the differences in the prior densities—that is, which values are plausible before seeing the data. It also shows posterior densities from the model and how those values are easily enclosed by the prior densities.

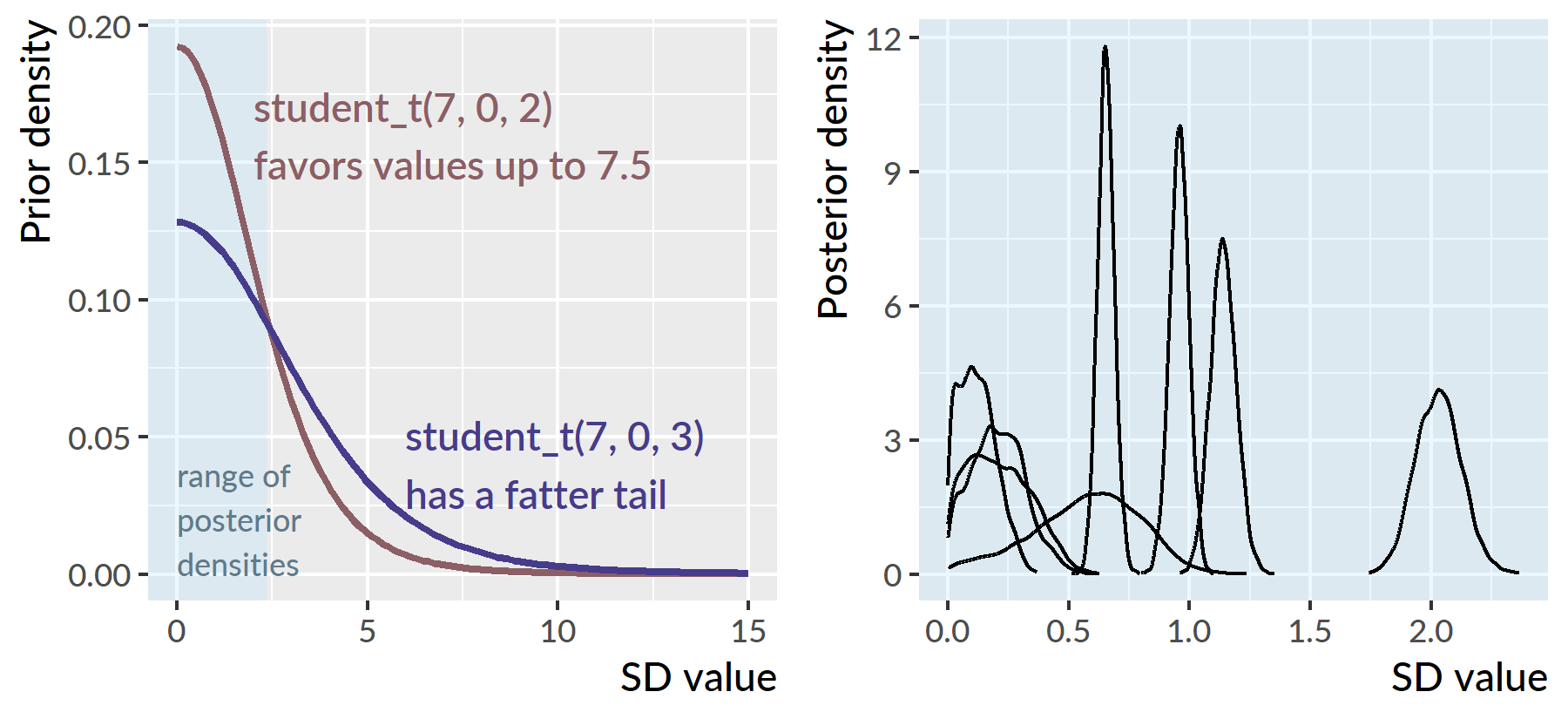


Figure 38 Prior densities (left) versus posterior densities (right) for the random-effect standard deviations. I changed the prior to be tighter, so that it favor values up to 7.5. This prior still turned out to be very conservative, given that the posterior samples for these values are all less than 3.

Model summary for unfamiliar-initial mispronunciation trials:

summary(mp\_unfam, priors = TRUE, prob = .9)  
#> Family: binomial   
#> Links: mu = logit   
#> Formula: Target | trials(Trials) ~ (1 + ot1 + ot2 + ot3) \* Study + (1 + ot1 + ot2 + ot3 | ResearchID/Study)   
#> Data: d\_u (Number of observations: 11875)   
#> Samples: 4 chains, each with iter = 3000; warmup = 1500; thin = 1;  
#> total post-warmup samples = 6000  
#>   
#> Priors:   
#> b ~ normal(0, 1)  
#> b\_ot1 ~ normal(0, 2)  
#> Intercept ~ normal(0, 1)  
#> L ~ lkj\_corr\_cholesky(2)  
#> sd ~ student\_t(7, 0, 2)  
#>   
#> Group-Level Effects:   
#> ~ResearchID (Number of levels: 193)   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> sd(Intercept) 0.21 0.11 0.03 0.40 66 1.04  
#> sd(ot1) 0.57 0.22 0.16 0.90 111 1.05  
#> sd(ot2) 0.21 0.13 0.02 0.44 77 1.05  
#> sd(ot3) 0.12 0.08 0.01 0.26 198 1.03  
#> cor(Intercept,ot1) -0.00 0.34 -0.55 0.55 284 1.00  
#> cor(Intercept,ot2) 0.04 0.37 -0.59 0.64 432 1.01  
#> cor(ot1,ot2) 0.11 0.35 -0.49 0.67 616 1.01  
#> cor(Intercept,ot3) -0.08 0.36 -0.66 0.52 792 1.00  
#> cor(ot1,ot3) 0.02 0.35 -0.55 0.60 1213 1.00  
#> cor(ot2,ot3) -0.10 0.36 -0.66 0.53 658 1.00  
#>   
#> ~ResearchID:Study (Number of levels: 475)   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> sd(Intercept) 0.96 0.04 0.89 1.02 185 1.01  
#> sd(ot1) 2.03 0.09 1.88 2.19 324 1.02  
#> sd(ot2) 1.14 0.05 1.05 1.23 388 1.01  
#> sd(ot3) 0.65 0.03 0.60 0.71 865 1.00  
#> cor(Intercept,ot1) 0.01 0.06 -0.09 0.10 763 1.00  
#> cor(Intercept,ot2) 0.10 0.06 -0.00 0.19 1309 1.00  
#> cor(ot1,ot2) -0.12 0.06 -0.22 -0.02 665 1.01  
#> cor(Intercept,ot3) 0.01 0.07 -0.10 0.12 2333 1.00  
#> cor(ot1,ot3) -0.16 0.06 -0.27 -0.06 1707 1.00  
#> cor(ot2,ot3) -0.29 0.06 -0.39 -0.19 2053 1.00  
#>   
#> Population-Level Effects:   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> Intercept -0.54 0.08 -0.67 -0.41 704 1.00  
#> ot1 3.36 0.17 3.07 3.63 1520 1.00  
#> ot2 -1.02 0.10 -1.20 -0.85 1360 1.00  
#> ot3 0.39 0.06 0.28 0.49 2244 1.00  
#> StudyTimePoint2 0.15 0.11 -0.02 0.33 492 1.00  
#> StudyTimePoint3 0.45 0.11 0.27 0.63 712 1.00  
#> ot1:StudyTimePoint2 -0.28 0.22 -0.65 0.09 1455 1.00  
#> ot1:StudyTimePoint3 -0.29 0.23 -0.67 0.09 1408 1.00  
#> ot2:StudyTimePoint2 0.03 0.14 -0.20 0.26 1202 1.00  
#> ot2:StudyTimePoint3 0.01 0.14 -0.21 0.24 1232 1.01  
#> ot3:StudyTimePoint2 -0.11 0.08 -0.25 0.02 2310 1.00  
#> ot3:StudyTimePoint3 0.01 0.09 -0.13 0.15 2622 1.00  
#>   
#> Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample   
#> is a crude measure of effective sample size, and Rhat is the potential   
#> scale reduction factor on split chains (at convergence, Rhat = 1).

Model summary for familiar-initial mispronunciation trials:

summary(mp\_fam, priors = TRUE, prob = .9)  
#> Family: binomial   
#> Links: mu = logit   
#> Formula: Target | trials(Trials) ~ (1 + ot1 + ot2 + ot3) \* Study + (1 + ot1 + ot2 + ot3 | ResearchID/Study)   
#> Data: d\_f (Number of observations: 12100)   
#> Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;  
#> total post-warmup samples = 4000  
#>   
#> Priors:   
#> b ~ normal(0, 1)  
#> b\_ot1 ~ normal(0, 2)  
#> Intercept ~ normal(0, 1)  
#> L ~ lkj\_corr\_cholesky(2)  
#> sd ~ student\_t(7, 0, 2)  
#>   
#> Group-Level Effects:   
#> ~ResearchID (Number of levels: 195)   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> sd(Intercept) 0.44 0.08 0.31 0.55 171 1.04  
#> sd(ot1) 0.73 0.24 0.25 1.06 70 1.05  
#> sd(ot2) 0.21 0.10 0.04 0.37 97 1.04  
#> sd(ot3) 0.08 0.05 0.01 0.18 240 1.02  
#> cor(Intercept,ot1) -0.01 0.25 -0.42 0.39 403 1.01  
#> cor(Intercept,ot2) 0.40 0.30 -0.15 0.81 376 1.01  
#> cor(ot1,ot2) -0.29 0.33 -0.74 0.36 252 1.02  
#> cor(Intercept,ot3) 0.13 0.35 -0.48 0.66 1555 1.00  
#> cor(ot1,ot3) -0.18 0.35 -0.71 0.45 1159 1.00  
#> cor(ot2,ot3) 0.08 0.36 -0.53 0.66 1358 1.00  
#>   
#> ~ResearchID:Study (Number of levels: 484)   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> sd(Intercept) 0.84 0.04 0.79 0.91 434 1.01  
#> sd(ot1) 1.95 0.10 1.79 2.11 195 1.01  
#> sd(ot2) 1.11 0.05 1.04 1.19 727 1.00  
#> sd(ot3) 0.62 0.03 0.57 0.66 1578 1.00  
#> cor(Intercept,ot1) -0.05 0.06 -0.16 0.05 678 1.00  
#> cor(Intercept,ot2) -0.02 0.06 -0.12 0.08 544 1.01  
#> cor(ot1,ot2) -0.21 0.06 -0.30 -0.11 705 1.00  
#> cor(Intercept,ot3) -0.06 0.07 -0.16 0.05 1189 1.00  
#> cor(ot1,ot3) -0.05 0.07 -0.16 0.06 1254 1.00  
#> cor(ot2,ot3) -0.49 0.05 -0.57 -0.40 1595 1.00  
#>   
#> Population-Level Effects:   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> Intercept 0.76 0.07 0.64 0.88 1660 1.00  
#> ot1 -3.01 0.17 -3.29 -2.74 1502 1.00  
#> ot2 2.00 0.09 1.85 2.15 1286 1.00  
#> ot3 -0.52 0.06 -0.61 -0.42 2008 1.00  
#> StudyTimePoint2 -0.23 0.09 -0.39 -0.08 1394 1.00  
#> StudyTimePoint3 -0.01 0.09 -0.17 0.14 1651 1.00  
#> ot1:StudyTimePoint2 0.56 0.21 0.21 0.92 1489 1.00  
#> ot1:StudyTimePoint3 1.30 0.21 0.94 1.64 1508 1.00  
#> ot2:StudyTimePoint2 -0.01 0.13 -0.23 0.20 1198 1.00  
#> ot2:StudyTimePoint3 -0.26 0.13 -0.48 -0.03 1179 1.00  
#> ot3:StudyTimePoint2 -0.11 0.08 -0.24 0.02 2008 1.00  
#> ot3:StudyTimePoint3 -0.10 0.08 -0.23 0.03 1939 1.00  
#>   
#> Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample   
#> is a crude measure of effective sample size, and Rhat is the potential   
#> scale reduction factor on split chains (at convergence, Rhat = 1).

## Item-response analysis for novel word retention

This is an item-response analysis carried out using a Bayesian mixed-effects logistic regression model. These models were fit in R (vers. 3.5.0; R Core Team, [2018](#ref-R-base)) with the brms package (vers. 2.3.1; Bürkner, [2017](#ref-brms)). I used weakly informative priors.

The linear model Correct ~ ItemType \* c\_peak\_10 says to estimate the log-odds of answering correctly using on mispronunciations (intercept) using the growth curve peaks (c\_peak\_10), adjust it for nonwords (ItemType) and for the nonword × peak interaction. The peaks were mean-centered within each type and multiplied by 10 so that the slope represents the effect of change of .1 from the mean peak value. The model includes four random intercepts: general child ability and child × condition ability adjustments (simultaneously requested using 1 | ResearchID/ItemType, i.e., ResearchID levels have ItemType levels nested under / them), plus specific item-level difficulties (1 | Item) and item-pair level difficulties (1 | WordGroup).

d <- readr::read\_csv("./data/mp-norming-data.csv.gz")  
  
priors <- c(  
 # Population-average intercept  
 set\_prior(class = "Intercept", "normal(0, 1)"),  
 # Population-average slopes  
 set\_prior(class = "b", "normal(0, 1)"),  
 # Standard deviation of the distribution from  
 # which random-intercepts are drawn  
 set\_prior(class = "sd", "student\_t(7, 0, 2)"))  
  
m\_norm <- brm(  
 Correct ~ ItemType \* c\_peak\_10 +  
 (1 | ResearchID/ItemType) +  
 (1 | WordGroup) +  
 (1 | Item),  
 prior = priors,  
 family = bernoulli,  
 chains = 4,  
 iter = 2000,  
 cores = 4,  
 control = list(adapt\_delta = .99),  
 data = d)  
  
readr::write\_rds(m\_norm, "./data/mp-norming-m2.rds.gz")

Model summary for retention trials at age 5:

summary(m\_norm, priors = TRUE, prob = .9)  
#> Family: bernoulli   
#> Links: mu = logit   
#> Formula: Correct ~ ItemType \* c\_peak\_10 + (1 | ResearchID/ItemType) + (1 | WordGroup) + (1 | Item)   
#> Data: d (Number of observations: 1200)   
#> Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;  
#> total post-warmup samples = 4000  
#>   
#> Priors:   
#> b ~ normal(0, 1)  
#> Intercept ~ normal(0, 1)  
#> sd ~ student\_t(7, 0, 2)  
#>   
#> Group-Level Effects:   
#> ~Item (Number of levels: 12)   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> sd(Intercept) 0.67 0.22 0.37 1.08 1468 1.00  
#>   
#> ~ResearchID (Number of levels: 101)   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> sd(Intercept) 0.34 0.17 0.05 0.62 454 1.01  
#>   
#> ~ResearchID:ItemType (Number of levels: 200)   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> sd(Intercept) 0.47 0.19 0.11 0.76 419 1.00  
#>   
#> ~WordGroup (Number of levels: 6)   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> sd(Intercept) 0.46 0.36 0.05 1.11 1152 1.00  
#>   
#> Population-Level Effects:   
#> Estimate Est.Error l-90% CI u-90% CI Eff.Sample Rhat  
#> Intercept 0.57 0.37 -0.03 1.16 1992 1.00  
#> ItemTypenonword 0.90 0.40 0.22 1.54 2180 1.00  
#> c\_peak\_10 -0.12 0.06 -0.22 -0.02 4000 1.00  
#> ItemTypenonword:c\_peak\_10 0.04 0.14 -0.20 0.28 4000 1.00  
#>   
#> Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample   
#> is a crude measure of effective sample size, and Rhat is the potential   
#> scale reduction factor on split chains (at convergence, Rhat = 1).

# Effects of specific mispronunciations

In this section, I briefly discuss item-level differences for the mispronunciation task.

I intended to formally analyze and model these effects as part of the main analysis. Prior to beginning this project, I expected that children would show different responses for different real-word–mispronunciation pairs. Not all mispronunciations are equally bad, and in fact, there could be some systematic tendencies in the badness of the mispronunciations. I hypothesized that children would show less of a penalty for later-acquired sounds. What I had in mind in particular was that *rice* and *wice* would be more similar than, say, *duck* and *guck* and other pairs.

The design of this experiment and the data collected, however, are not equipped to address this hypothesis. I came to this conclusion after visualizing the *rice* versus *wice* looking data and observing that children looked to *rice* less than the other real words. Figure 39 shows the growth curves for real word and mispronunciations.

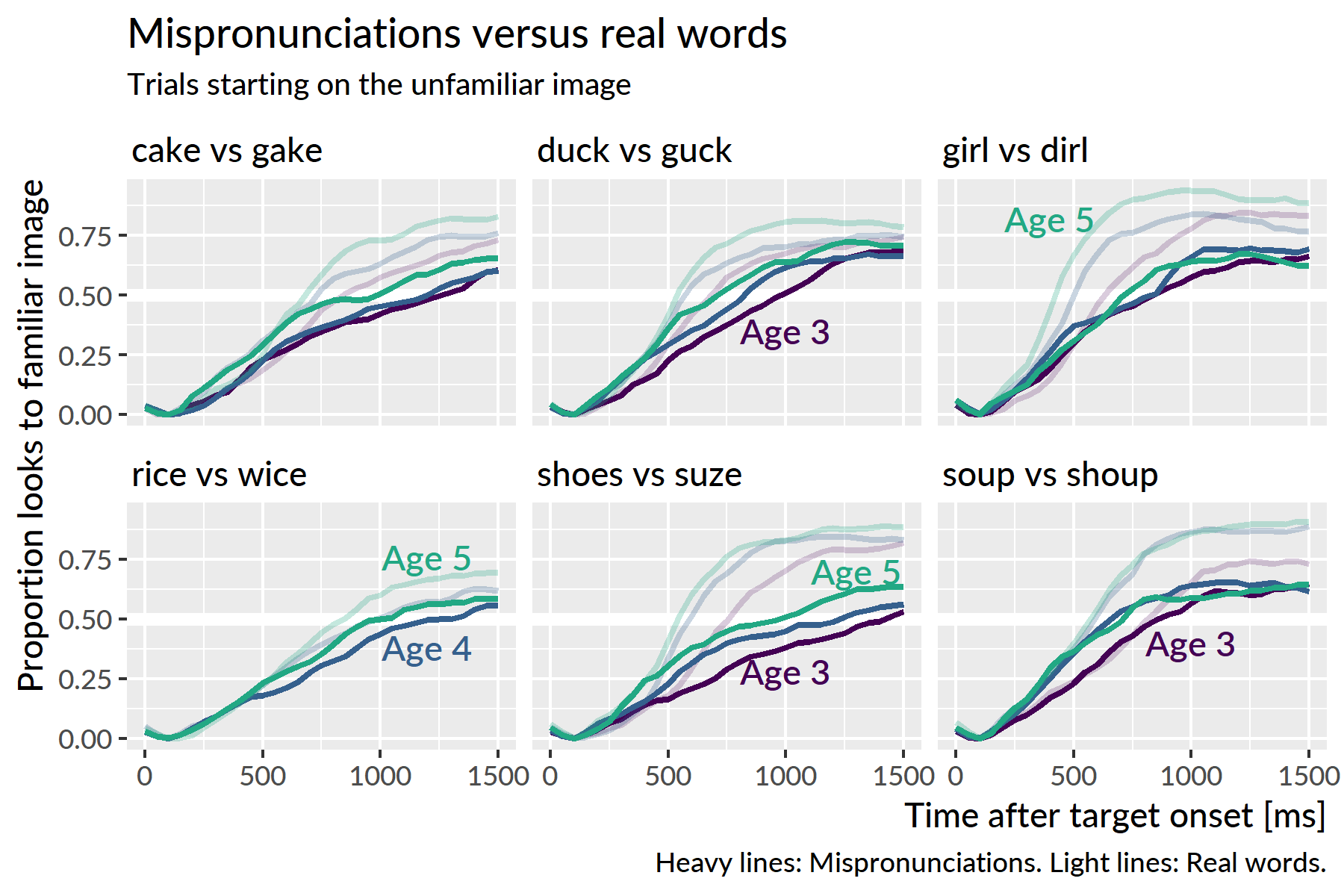


Figure 39 Average proportion of looks to the familiar object for real words and mispronunciations. A *dog*/*tog* pair was administered at age 3 but it was replaced by *rice*/*wice* at age 4.

*Rice* and *wice* are indeed very similar, but children do not know *rice* very well it seems. In the plot, for instance, age-5 *rice* has as many looks as age-4 *dirl*. Given that the data tested only 6 mispronunciations per year, compounded by the fact that some real words are harder than others, I decided that it would not feasible to draw conclusions about different kinds of mispronunciations. A more appropriate study would test many more mispronunciations and vary the familiarity of the paired real words to handle these limitations.

It is still informative to plot the differences between the real word and mispronunciation lines, as in Figure 40. This plot confirms that *rice* and *wice* are very similar. One interesting, and somewhat unexpected feature, is how the differences increase with age for some items. Namely, it appears that *dirl* becomes a worse and worse realization of *girl* as children grow older. A similar change happens with *shoup* for *soup*.

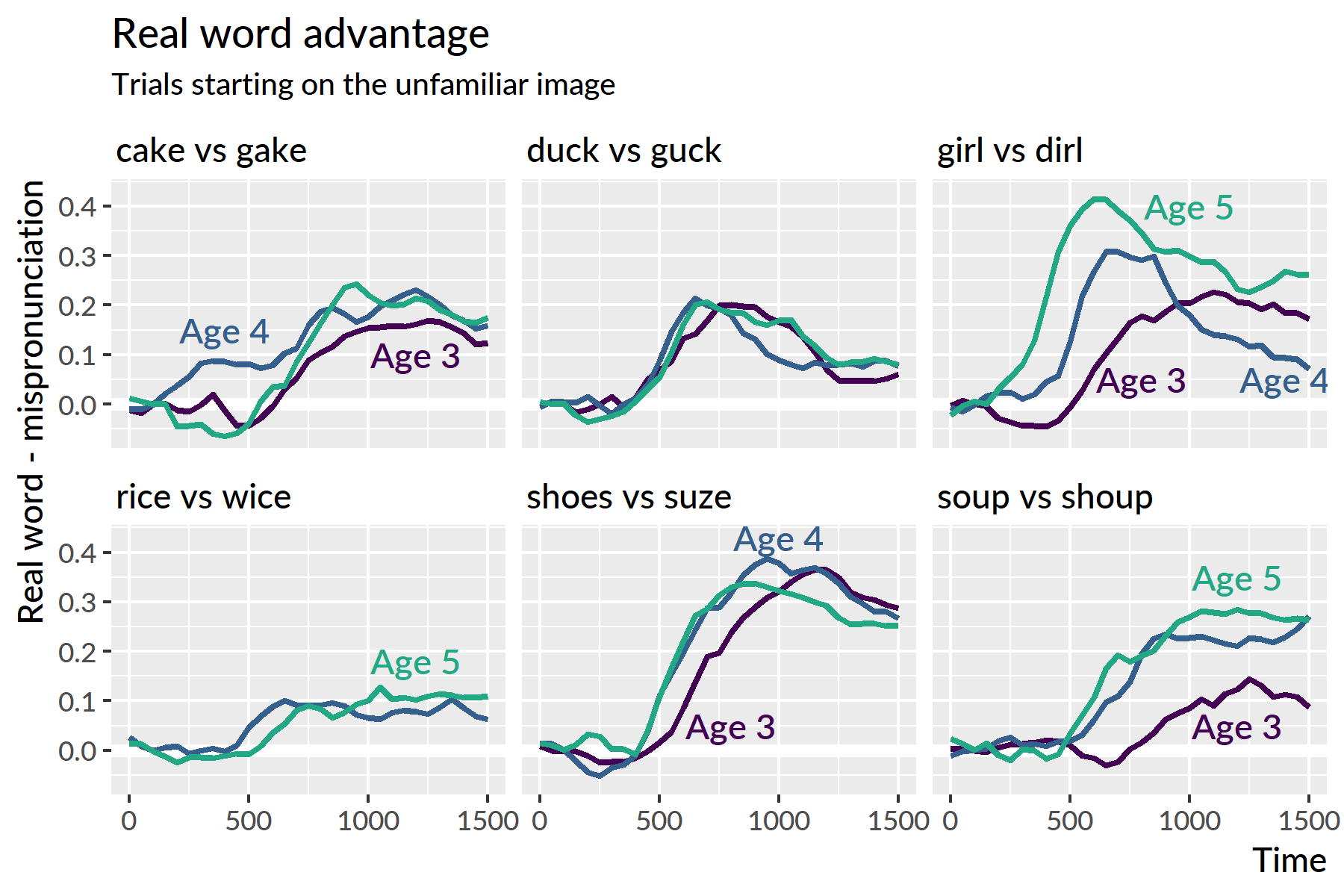


Figure 40 Difference between the average proportion of looks to real words and mispronunciations.

# Related work

In this section, I clarify relationships between this project and other word recognition research reported from our lab. In short, our lab has reported results about the two-image and four-image experiments from cross-sectional samples, describing child-level measures that predict performance in these tasks. In contrast, my dissertation 1) focuses on the longitudinal development of word recognition and 2) engages with the fine-grained details of lexical processing.

Law et al. ([2016](#ref-RWLPaper)) analyzed data from the four-image experiment in Specific Aim 1. This study featured a diverse cross-sectional sample of 60 children, half of whom received the experiment in African American English and half received it in Mainstream American English. The sample ranged in age from 28 to 60 months. The study included data from 23 participants from year 1 of the longitudinal study (i.e., what I refer to as age 3) in order to enrich parts of the sample demographics. For this manuscript, we analyzed how vocabulary and maternal education predicted looking patterns, including relative looks to the semantic and phonological foils, but with conventional polynomial growth curves. The use of generalized additive models is an innovation I developed for my dissertation.

Law and Edwards ([2015](#ref-MPPaper)) analyzed a different version of the mispronunciation experiment on a different sample of children (*n* = 34, 30–46 months old). This earlier version included both real word and the mispronunciation of the real word in the same block of trial. For example, a child would hear “dog” and “tog” during the same session of the experiment. This design might subtly temper the effect of mispronounced stimuli by allowing the listener to compare the mispronunciation to its correctly produced counterpart. The version of the experiment in Specific Aim 2 separates the real words and mispronunciations so that a child never hears a familiar word and its mispronunciation during the same block of trials. With this design, there is no explicit point of comparison for the mispronunciation, and the child has to rely on his or her own lexical representations when processing these words.

Mahr and Edwards (in press) was the manuscript I originally authored for my preliminary examinations. The paper analyzes the same kinds of relations as Weisleder and Fernald ([2013](#ref-Weisleder2013)) which showed that lexical processing efficiency mediated the effect of language input on future vocabulary size. Specifically, I asked whether word recognition performance on the four-image task of Specific Aim 1, vocabulary size, and home language input data from year 1 (age 3) predicted vocabulary size at year 2 (age 4). The paper only examined looks to the familiar image from one year of the study, so it did not analyze any lexical competition effects or the development of word recognition within children. In short, we found that receptive vocabulary was more sensitive to variability in lexical processing and home language input than expressive vocabulary.

# References

Altvater-Mackensen, N., & Mani, N. (2013). The impact of mispronunciations on toddler word recognition: Evidence for cascaded activation of semantically related words from mispronunciations of familiar words. *Infancy*, *18*(6), 1030–1052. doi:[10.1111/infa.12022](https://doi.org/10.1111/infa.12022)

Arias-Trejo, N., & Plunkett, K. (2009). Lexical-semantic priming effects during infancy. *Philosophical Transactions of the Royal Society of London*, *364*, 3633–3647. doi:[10.1098/rstb.2009.0146](https://doi.org/10.1098/rstb.2009.0146)

B., M. C., & Jacquelyn, B. (n.d.). Acquisition of the novel name–Nameless category (n3c) principle. *Child Development*, *65*(6), 1646–1662. doi:[10.1111/j.1467-8624.1994.tb00840.x](https://doi.org/10.1111/j.1467-8624.1994.tb00840.x)

Baayen, R. H., Rij, J. van, Cat, C. de, & Wood, S. N. (2016). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by generalized additive mixed models. Retrieved from <http://arxiv.org/abs/1601.02043>

Bailey, T. M., & Plunkett, K. (2002). Phonological specificity in early words. *Cognitive Development*, *17*(2), 1265–1282. doi:[10.1016/S0885-2014(02)00116-8](https://doi.org/10.1016/S0885-2014(02)00116-8)

Barr, D. J. (2008). Analyzing ‘visual world’ eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474. doi:[10.1016/j.jml.2007.09.002](https://doi.org/10.1016/j.jml.2007.09.002)

Barton, D. (1976). Phonemic discrimination and the knowledge of words in children under three years. *Papers and Reports on Child Language Development*, *11*, 61–68.

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

Bion, R. A. H., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word–object mappings in relation to vocabulary learning at 18, 24, and 30 months. *Cognition*, *126*(1), 39–53. doi:[10.1016/j.cognition.2012.08.008](https://doi.org/10.1016/j.cognition.2012.08.008)

Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one’s age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology*, *112*(4), 417–436. doi:[10.1016/j.jecp.2012.01.005](https://doi.org/10.1016/j.jecp.2012.01.005)

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi:[10.18637/jss.v080.i01](https://doi.org/10.18637/jss.v080.i01)

Charles-Luce, J., & Luce, P. A. (1990). Similarity neighbourhoods of words in young children’s lexicons. *Journal of Child Language*, *17*(1), 205–215. doi:[10.1017/S0305000900013180](https://doi.org/10.1017/S0305000900013180)

Charles-Luce, J., & Luce, P. A. (1995). An examination of similarity neighbourhoods in young children’s receptive vocabularies. *Journal of Child Language*, *22*(3), 727–735. doi:[10.1017/S0305000900010023](https://doi.org/10.1017/S0305000900010023)

Chow, J., Aimola Davies, A. M., & Plunkett, K. (2017). Spoken-word recognition in 2-year-olds: The tug of war between phonological and semantic activation. *Journal of Memory and Language*, *93*, 104–134. doi:[10.1016/j.jml.2016.08.004](https://doi.org/10.1016/j.jml.2016.08.004)

Coady, J. A., & Aslin, R. N. (2003). Phonological neighbourhoods in the developing lexicon. *Journal of Child Language*, *30*(2), 441–469. doi:[10.1017/S0305000903005579](https://doi.org/10.1017/S0305000903005579)

Cristia, A., Seidl, A., Junge, C., Soderstrom, M., & Hagoort, P. (2014). Predicting individual variation in language from infant speech perception measures. *Child Development*, *85*(4), 1330–1345. doi:[10.1111/cdev.12193](https://doi.org/10.1111/cdev.12193)

Dollaghan, C. A. (1994). Children’s phonological neighbourhoods: Half empty or half full? *Journal of Child Language*, *21*(2), 257–271. doi:[10.1017/S0305000900009260](https://doi.org/10.1017/S0305000900009260)

Ellis Weismer, S., Haebig, E., Edwards, J. R., Saffran, J. R., & Venker, C. E. (2016). Lexical processing in toddlers with ASD: Does weak central coherence play a role? *Journal of Autism and Developmental Disorders*, *46*(12), 3755–3769. doi:[10.1007/s10803-016-2926-y](https://doi.org/10.1007/s10803-016-2926-y)

Fennell, C. T., & Waxman, S. R. (2010). What paradox? Referential cues allow for infant use of phonetic detail in word learning. *Child Development*, *81*(5), 1376–83. doi:[10.1111/j.1467-8624.2010.01479.x](https://doi.org/10.1111/j.1467-8624.2010.01479.x)

Fernald, A., & Marchman, V. A. (2012). Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. *Child Development*, *83*(1), 203–222. doi:[10.1111/j.1467-8624.2011.01692.x](https://doi.org/10.1111/j.1467-8624.2011.01692.x)

Fernald, A., Swingley, D., & Pinto, J. P. (2001). When half a word is enough: Infants can recognize spoken words using partial phonetic information. *Child Development*, *72*(4), 1003–15. doi:[10.1111/1467-8624.00331](https://doi.org/10.1111/1467-8624.00331)

Gabry, J., & Goodrich, B. (2018). *Rstanarm: Bayesian applied regression modeling via stan*. Retrieved from <https://CRAN.R-project.org/package=rstanarm>

Gamer, M., Lemon, J., & Singh, I. F. P. (2012). irr: Various coefficients of interrater reliability and agreement. Retrieved from <https://CRAN.R-project.org/package=irr>

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Horst, J. S., Samuelson, L. K., Kucker, S. C., & McMurray, B. (2011). What’s new? Children prefer novelty in referent selection. *Cognition*, *118*(2), 234–244. doi:[10.1016/j.cognition.2010.10.015](https://doi.org/10.1016/j.cognition.2010.10.015)

Huang, Y. T., & Snedeker, J. (2011). Cascading activation across levels of representation in children’s lexical processing. *Journal of Child Language*, *38*(03), 644–661. doi:[10.1017/S0305000910000206](https://doi.org/10.1017/S0305000910000206)

Kapnoula, E. C., Packard, S., Gupta, P., & McMurray, B. (2015). Immediate lexical integration of novel word forms. *Cognition*, *134*, 85–99. doi:[10.1016/j.cognition.2014.09.007](https://doi.org/10.1016/j.cognition.2014.09.007)

Lany, J. (2017). Lexical-processing efficiency leverages novel word learning in infants and toddlers. *Developmental Science*. doi:[10.1111/desc.12569](https://doi.org/10.1111/desc.12569)

Law, F., II, & Edwards, J. R. (2015). Effects of vocabulary size on online lexical processing by preschoolers. *Language Learning and Development*, *11*(4), 331–355. doi:[10.1080/15475441.2014.961066](https://doi.org/10.1080/15475441.2014.961066)

Law, F., II, Mahr, T., Schneeberg, A., & Edwards, J. R. (2016). Vocabulary size and auditory word recognition in preschool children. *Applied Psycholinguistics*. doi:[10.1017/S0142716416000126](https://doi.org/10.1017/S0142716416000126)

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science*, *18*(3), 193–8. doi:[10.1111/j.1467-9280.2007.01871.x](https://doi.org/10.1111/j.1467-9280.2007.01871.x)

Magnuson, J. S., Mirman, D., & Myers, E. (2013). Spoken word recognition. In D. Reisberg (Ed.), *The oxford handbook of cognitive psychology*. Oxford University Press. doi:[10.1093/oxfordhb/9780195376746.013.0027](https://doi.org/10.1093/oxfordhb/9780195376746.013.0027)

Mahr, T., McMillan, B. T. M., Saffran, J. R., Ellis Weismer, S., & Edwards, J. (2015). Anticipatory coarticulation facilitates word recognition in toddlers. *Cognition*, *142*, 345–350. doi:[10.1016/j.cognition.2015.05.009](https://doi.org/10.1016/j.cognition.2015.05.009)

Mani, N., & Plunkett, K. (2010). In the infant’s mind’s ear: Evidence for implicit naming in 18-month-olds. *Psychological Science*, *21*(7), 908–913. doi:[10.1177/0956797610373371](https://doi.org/10.1177/0956797610373371)

Mani, N., Durrant, S., & Floccia, C. (2012). Activation of phonological and semantic codes in toddlers. *Journal of Memory and Language*, *66*(4), 612–622. doi:[10.1016/j.jml.2012.03.003](https://doi.org/10.1016/j.jml.2012.03.003)

Marchman, V. A., & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmental Science*, *11*(3), F9–16. doi:[10.1111/j.1467-7687.2008.00671.x](https://doi.org/10.1111/j.1467-7687.2008.00671.x)

Markman, E. M., & Wachtel, G. F. (1988). Children’s use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157. doi:[10.1016/0010-0285(88)90017-5](https://doi.org/10.1016/0010-0285(88)90017-5)

Marslen-Wilson, W. D., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 576–585. doi:[10.1037/0096-1523.15.3.576](https://doi.org/10.1037/0096-1523.15.3.576)

Mather, E., & Plunkett, K. (2012). The role of novelty in early word learning. *Cognitive Science*, *36*(7), 1157–1177. doi:[10.1111/j.1551-6709.2012.01239.x](https://doi.org/10.1111/j.1551-6709.2012.01239.x)

Mayor, J., & Plunkett, K. (2014). Infant word recognition: Insights from TRACE simulations. *Journal of Memory and Language*, *71*(1), 89–123. doi:[10.1016/j.jml.2013.09.009](https://doi.org/10.1016/j.jml.2013.09.009)

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. doi:[10.1016/0010-0285(86)90015-0](https://doi.org/10.1016/0010-0285(86)90015-0)

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: CRC Press/Taylor & Francis Group.

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*(4), 831–877. doi:[10.1037/a0029872](https://doi.org/10.1037/a0029872)

McMurray, B., Samelson, V. M., Lee, S. H., & Tomblin, J. B. (2010). Individual differences in online spoken word recognition: Implications for SLI. *Cognitive Psychology*, *60*(1), 1–39. doi:[10.1016/j.cogpsych.2009.06.003](https://doi.org/10.1016/j.cogpsych.2009.06.003)

Mirman, D. (2014). *Growth curve analysis and visualization using R*. Boca Raton, FL: Chapman & Hall/CRC.

Nosek, B. A., Simonsohn, U., Moore, D. A., Nelson, L. D., Simmons, J. P., Sallans, A., & LeBel, E. P. (2014, February). Standard reviewer statement for disclosure of sample, conditions, measures, and exclusions. Open Science Framework. Retrieved from <https://osf.io/hadz3>

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

Rigler, H., Farris-Trimble, A., Greiner, L., Walker, J., Tomblin, J. B., & McMurray, B. (2015). The slow developmental time course of real-time spoken word recognition. *Developmental Psychology*, *51*(12), 1690–1703. doi:[10.1037/dev0000044](https://doi.org/10.1037/dev0000044)

Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics: a practical introduction. Retrieved from <http://arxiv.org/abs/1703.05339>

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*(6640), 381. doi:[10.1038/41102](https://doi.org/10.1038/41102)

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*(2), 147–66. doi:[10.1016/S0010-0277(00)00081-0](https://doi.org/10.1016/S0010-0277(00)00081-0)

Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science*, *13*(5), 480–484. doi:[10.1111/1467-9280.00485](https://doi.org/10.1111/1467-9280.00485)

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children’s word learning. *Cognitive Psychology*, *54*(2), 99–132. doi:[10.1016/j.cogpsych.2006.05.001](https://doi.org/10.1016/j.cogpsych.2006.05.001)

Swingley, D., Pinto, J. P., & Fernald, A. (1999). Continuous processing in word recognition at 24 months. *Cognition*, *71*(2), 73–108. doi:[10.1016/S0010-0277(99)00021-9](https://doi.org/10.1016/S0010-0277(99)00021-9)

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2017). itsadug: Interpreting time series and autocorrelated data using GAMMs.

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143–52. doi:[10.1177/0956797613488145](https://doi.org/10.1177/0956797613488145)

Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., & Stager, C. L. (1998). Acquisition of word–object associations by 14-month-old infants. *Developmental Psychology*, *34*(6), 1289. doi:[10.1037/0012-1649.34.6.1289](https://doi.org/10.1037/0012-1649.34.6.1289)

White, K. S., & Morgan, J. L. (2008). Sub-segmental detail in early lexical representations. *Journal of Memory and Language*, *59*(1), 114–132. doi:[10.1016/j.jml.2008.03.001](https://doi.org/10.1016/j.jml.2008.03.001)

Winter, B., & Wieling, M. (2016). How to analyze linguistic change using mixed models, growth curve analysis and generalized additive modeling. *Journal of Language Evolution*, *1*(1), 7–18. doi:[10.1093/jole/lzv003](https://doi.org/10.1093/jole/lzv003)

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). CRC Press.

1. Arias-Trejo and Plunkett ([2009](#ref-Arias-Trejo2009)) is commonly cited as evidence of semantic priming effects. Toddlers heard sentences like “I saw a cat… dog”. During the word *dog*, two images (dog and door) are presented. The idea is that *cat* should prime looks to its semantic neighbor *dog*. The unnatural stimulus order (a sentence followed by an isolated single word) and a condition effect where 18 month-olds children outperformed 21 month-olds make me skeptical that semantic priming is the most plausible explanation of those results. [↑](#footnote-ref-32)
2. [Appendix 21](#related-work) describes how this dissertation relates to other work from our lab. [↑](#footnote-ref-37)
3. My goals in using this method were simply to estimate model effects and quantify the uncertainty about those effects. This pragmatic, estimation-based approach of Bayesian statistics is illustrated in texts by Gelman and Hill ([2007](#ref-GelmanHill)) and McElreath ([2016](#ref-RethinkingBook)). [↑](#footnote-ref-47)
4. The polynomial other terms are less important—or rather, they do not map as neatly onto behavioral descriptions as the accuracy and efficiency parameters. The primary purpose of quadratic and cubic terms is to ensure that the estimated growth curve adequately fits the data. In this kind of data, there is a steady baseline at chance probability before the child hears the word, followed a window of increasing probability of fixating on the target as the child recognizes the word, followed by a period of plateauing and then diminishing looks to target. The cubic polynomial allows the growth curve to be fit with two inflection points: the point when the looks to target start to increase from baseline and the point when the looks to target stops increasing. [↑](#footnote-ref-48)
5. Initially, I tried to use Bayesian polynomial growth curve models, as in the earlier analysis of the looks to the target image. These models however did not converge, even when strong priors were placed on the parameters. In principle, I could have used Bayesian generalized additive models, but the software ecosystem and available tools for model criticism and inference are currently rather limited. [↑](#footnote-ref-68)
6. I am not too committed to any particular mechanisms of *representation* or *organization*. Under a connectionist framework with distributed representations, for instance, a word is represented as a pattern of activation distributed over many shared units. (I think of numbers on a digital clock where seven lines turn on or off to make ten digits but exponentially more complicated.) In that case, representation and organization are inseparable, and it would make more sense to talk about the strength and number of connections instead. My point here is that the lexical mechanisms involved should be ones that enable stronger immediate activations as a result of learning more words. [↑](#footnote-ref-78)
7. Granted, there might be some subtle nonlinear effect at play where higher peak activations require a greater degree of inhibition to overcome, so changes in inhibition could be a plausible part of the developmental story. But there is no compelling reason from the data to make that assertion. [↑](#footnote-ref-79)
8. “The infant who identifies familiar words more quickly has more resources available for attending to subsequent words, with advantages for learning new words later in the sentence, as well as for tracking distributional information about relations among words… Being slow to identify the referent of a familiar word could interfere with lexical activation and impede success in tracking distributional regularities and managing attentional resources in real time (Evans, Saffran, & Robe-Torres, 2009)” (Fernald & Marchman, [2012](#ref-Fernald2012), p. 217). [↑](#footnote-ref-86)