DEVELOPMENT OF WORD RECOGNITION IN PRESCHOOLERS

by

Tristan Mahr

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Communication Sciences and Disorders)

at the

UNIVERSITY OF WISCONSIN–MADISON

2018

Date of final oral examination: August 7, 2018

The dissertation is approved by the following members of the Final Oral Committee:
Jan Edwards, Professor, Hearing and Speech Sciences (University of Maryland)
Susan Ellis Weismer, Professor, Communication Sciences and Disorders
Margarita Kaushanskaya, Professor, Communication Sciences and Disorders
Jenny Saffran, Professor, Psychology
David Kaplan, Professor, Educational Psychology
Bob McMurray, Professor, Psychology (University of Iowa)

ABSTRACT

Development of word recognition in preschoolers

by

Tristan Mahr

The University of Wisconsin–Madison, 2018

Under the supervision of Professor Jan Edwards and Professor Susan Ellis Weismer

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. 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. This dissertation reports the results for two word recognition experiments administered during each year of a 3-year longitudinal study with 160 preschoolers. Children were 2.5–3-years-old in year 1 and 4.5–5-years-old 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. Word recognition behavior thus provided 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 competitors, 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 an unfamiliar object and heard a real word (e.g., *shoes*), a one-feature mispronunciation (*suze*), or a nonword (*geeve*). Contrary to pre-analysis hypotheses, children recognized real words and fast-selected novel-object referents for nonwords equally well and even performed better in the nonword condition. Children became more likely to associate the familiar object with the mispronunciations as they grew older. At age 5, children showed better retention for novel objects labeled with nonwords than with mispronunciations.

For Penny

Contents

List of tables

List of figures

Acknowledgments

Thanks to my advisor and mentor Jan Edwards for recognizing my potential, plucking me from the Masters program, and giving me the opportunity to be a part of her lab. Jan changed my life by giving me a purpose and place to develop my skills.

Working in the Learning To Talk lab was an enriching experience. I had the freedom to explore and experiment, to fail and learn. I entered as a speech pathologist and left a data scientist. Thanks to Ben Munson, Mary Beckman, Franzo Law II, Matt Winn, Pat Reidy, Tatiana Thonesavanh, Alissa Schneeberg, Kayla Kristensen, Allie Johnson, Michelle Erskine, and Lizzy Hill. Special thanks to Nancy and Bob Wermuth for being role models.

A lot of smart and busy people gave me their time and thoughts. Thanks to the faculty who were always willing to advise me: Margarita Kaushanskaya, Audra Sterling, Jenny Saffran, Courtney Seidel, Katie Hustad. Thanks for Susan Ellis Weismer for taking me under her wing and helping me through some stressful periods. Thanks to Bob McMurray for giving an enthusiastic, eye-opening lecture at ASHA 2013 about how eyetracking data works as a measure of word recognition. Thanks to Tim Rogers for being on my prelim committee.

Thanks to my compatriots at Waisman and elsewhere whom I commiserated with often: Pierce Edmiston, Courtney Venker, Ron Pomper, Martin Zettersten, Liz Premo, Janine Mathee, Phoebe Natzke, whoever happened to be around in Rita's lab on a given day. Thanks to Lolo for telling me that graduate training for speech pathology is a career for a linguist.

This research would not have been possible without the families and children who participated in our research.

I am grateful for the financial support from the NIH grants that supported me as a trainee or as a research assistant: NICHD Grant 2-T32-HD049899 to Maryellen

MacDonald, NIDCD R01 DC012513 to Susan Ellis Weismer, Jan Edwards, and Jenny R. Saffran, and NIDCD T32 DC05359-10 to Susan Ellis Weismer. The data for this project were collected as part of the Learning to Talk project which was funded by NIDCD R01 DC002932 to Jan Edwards, Mary E. Beckman, and Benjamin Munson. Thanks to departmental administrators who kept the machine running: Maureen Garity, Amanda Talbert, Choutae Yang.

I spend all day working in R and RStudio. I want to thank Hadley Wickham for fixing the language and for working in open source so that I could learn from his work. Thanks to other RStudio people: Yihui Xie, Kevin Ushey, Jenny Bryan, and Mara Averick.

I've spent the last two years basically teaching myself Bayesian statistics. Thanks to the Jonah Gabry, Paul-Christian Bürkner and the Stan team for a great ecosystem; David Kaplan for talking me through some of my early doubts; and Richard McElreath for putting his *Statistical Rethinking* course on YouTube.

Recognition is due for the community of statisticians, data scientists, psychologists, and language scientists whom I spent idle moments with on Twitter. There are too many of you to name. Thanks to the fellow graduate students, post-docs and others early in their research careers for helping me feel a sense of community. I often used the medium as a way to think out loud as I worked through modeling, programming or plotting problems. Thanks to those who chimed in to answer questions, suggest references, or offer feedback.

I will finish by thanking the people closest to me. I grew up on dairy farm in rural Wisconsin and I was the seventh of ten children. That explains about 70% of who I am and whom I've become. My mom is nothing short of a superhero. My dad has been a paragon of steadiness and selflessness. Thanks to my siblings and their families for keeping me from taking myself too seriously. Thanks also to my-laws for being close to us.

Finally, thanks to my wife Amanda. None of this would have been possible without your love, support, humor, and nerdery. People sometimes thought I was a workaholic because I would get to my office at 8 am. But really, I just wanted to ride the bus with you in the morning. We started a family together! Shout out to Kiki and Nooper for the snuggles.

1 Overview and 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).

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? In this dissertation, I address this gap in knowledge by analyzing three years of word recognition data collected in a recently completed longitudinal study of 160 children.

In particular, I 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 develop a fine-grained description of how the dynamics of word recognition change year over year, and I document how differences in word recognition performance relate to other child-level measures (such as vocabulary and speech perception).

Study 1: Familiar word recognition and lexical competition

***Specific Aim:* To characterize the development of familiar word recognition and lexical competition, I 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*). In Chapter 5, I use a series of growth curve analyses to describe how children's familiar word recognition develops year over year. Children in this cohort cover a range of vocabulary scores at age 3, and this variability allows me to investigate individual differences in vocabulary and word recognition over time and assess the predictive value of these measures. Of interest was how individual differences at age 3 persisted into age 5 and how these differences related to vocabulary measures at later ages. In Chapter 6, I examine the children's looks to the distractors to study the developmental course of lexical competition from similar sounding and similar meaning words. Increases in sensitivity to competing words reveal how lexical competition effects emerge as a byproduct of learning new words and developing more efficient phonological and lexical representations. As I argue in Chapter 7, increased sensitivity to lexical competitors supports familiar word recognition because children become more efficient at activating a named word *and related words.* When children err, they become more likely to err on a lexically relevant alternative.

Study 2: Referent selection and mispronunciations

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

Not every word children hear are familiar to them. They may hear entirely new words, or they may hear variations and corruptions of familiar words. How children respond to both kinds of words is informative, as I review in Chapter 9. I describe this eyetracking experiment in detail—based on White and Morgan (2008) and Law and Edwards (2015)—in Chapter 10. 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 (*shoup*), or a novel word unrelated to either image (*cheem*). The correct productions tested familiar word recognition and the nonwords tested fast referent selection. The mispronunciations tested the child's phonological categories by showing whether the child permitted, rejected, or equivocated about mispronunciations.

I use growth curve analyses to study how children's responses to the three word types changed over time. In Chapter 11, I examine familiar word recognition and fast referent selection for novel words to determine which feature of lexical processing better predicts vocabulary growth. I compare these two conditions directly to look for 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). In Chapter 12, I examine how children interpreted mispronunciations of the familiar words at each age and study how individual differences in vocabulary and speech perception related to children's responses to the mispronunciations. I also report how children at age 5 are better able to retain nonwords than mispronunciations of familiar 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. Furthermore, this project also examines word recognition in two experimental tasks that tap into different aspects of word recognition. Specifically, a four-image experiment with semantic and phonological foils allows me to study how lexical competition develops, and a two-image experiment with nonwords and mispronunciations enables me to study how children's responses to unfamiliar words develop over time as well. Chapter 15 reviews the results from both studies in terms of lexical processing as well as the main contributions of this project. Findings show how individual differences in lexical processing change over time and reveal how low-level mechanisms underlying word recognition mature longitudinally in children. These findings 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.

2 Research hypotheses

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

Study 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.

Study 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 predictor 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*).

Study 1: Familiar word recognition and lexical competition

3 Familiar word recognition

3.1 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). 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).

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), match truncated words to their intended referents (Fernald, Swingley, & Pinto, 2001), 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) 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) 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).

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) 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) 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) 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) 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]

Altvater-Mackensen and Mani (2013) demonstrated phonological-semantic priming even when the prime is a mispronunciation. German-learning two-year-olds heard a prime word, and 200 ms later two images appeared onscreen (a cow, *Kuh*

[1] Arias-Trejo and Plunkett (2009) 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 outperformed 21-month-olds make me skeptical that semantic priming is the most plausible explanation of those results.

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) performed a very similar study to the one I present in the next chapters. They used the Visual World Paradigm with English-learning 24- and 30-month-olds. Children saw a 2 × 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 (e.g., *Look at the bee*). On *filler* trials, the target word and an unrelated image appear onscreen alongside the competitors. On *test* trials, the display had two unrelated images (*sandwich*, *dress*), a phonological competitor (*bus*), and a semantic competitor (*cat*) and children were prompted to look at an offscreen, unpictured target (*Look at the bee*). They found a temporary early advantage for the phonological competitor, so that the probability of looking to the phonological competitor was greater than the other competitors. This early advantage was followed by a late, more stable advantage for the semantic competitor. Moreover, they found that increased receptive vocabulary predicted more looks to the phonological competitor and fewer looks to the semantic competitor. (The looks to the semantic competitor were decreased because of the early advantage of the phonological competitor.) Their results support a kind of cascading activation in which phonological information comes online before 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 the 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 phonologically or semantically similar words exert their influence on word recognition? Chow et al. (2017) provide a promising first step, in which two-year-olds looked to the phonological and semantic relatives of a named word. (Yet I am skeptical of any word recognition study where a target word is absent and absent for many trials.) 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). Increasing sensitivity to similar sounding or similar meaning words over time would reveal that children more thoroughly learn familiar words with age.

3.2 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 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) 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) 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) 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) 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?

3.3 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 shows 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. 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, I analyze the development patterns of familiar word recognition (looks to the target) and how individual differences change over time. I hypothesized that children would show stable individual differences over time, but the range and magnitude of these differences would get smaller as children grew older. I also examine which word recognition measures correlate with future vocabulary size to test how word recognition behavior predicts later outcomes. In Chapter 6, I study how the phonological and semantic competitors influence word recognition, and I test the prediction that children will become more sensitive to competitors as they grow older. Finally, in Chapter 7, I link these two lines of research together and describe both sets of results in terms of lexical processing dynamics, and Chapter 8 reviews the results of my pre-analysis research hypotheses.

4 Method

4.1 Participants

The data were collected as part of a three-year longitudinal study.[1] 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 4.1 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 the study, or a failed hearing screening. Final sample sizes depend on the measures needed for an analysis and the results from data screening checks.

[1] Appendix F describes how this dissertation relates to other work from our lab.

Table 4.1: Participant characteristics. Education levels: *Low*: less than high school, or high school; *Mid*: trade school, technical or associates degree, some college, or college degree; and *High*: graduate degree. Dialects: *MAE*: Mainstream American English; *AAE*: African-American English.

	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 ed.: Low, Mid, 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)	33 (3)	45 (4)	57 (4)
Age (months): Range	28–39	39–52	51–66
EVT-2 standard: Mean (SD)	115 (18)	118 (16)	118 (14)
PPVT-4 standard: Mean (SD)	113 (17)	120 (16)	—
GFTA-2 standard: Mean (SD)	92 (13)	—	91 (13)

4.2 Visual World Paradigm

This experiment used a version of the Visual World Paradigm for word recognition experiments (Law, Mahr, Schneeberg, & Edwards, 2016). 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 4.1 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.

This experimental design—an eyetracking study of word recognition with four images—is referred to as the Visual World Paradigm throughout the literature. See Huettig, Rommers, and Meyer (2011) for a historical review and an overview of how it has been used to study syntactic, pragmatic, semantic, and phonological processing. The paradigm has been used extensively to study word recognition in adult listeners—and in preschool-age children (Borovsky et al., 2012; Chow et al.,

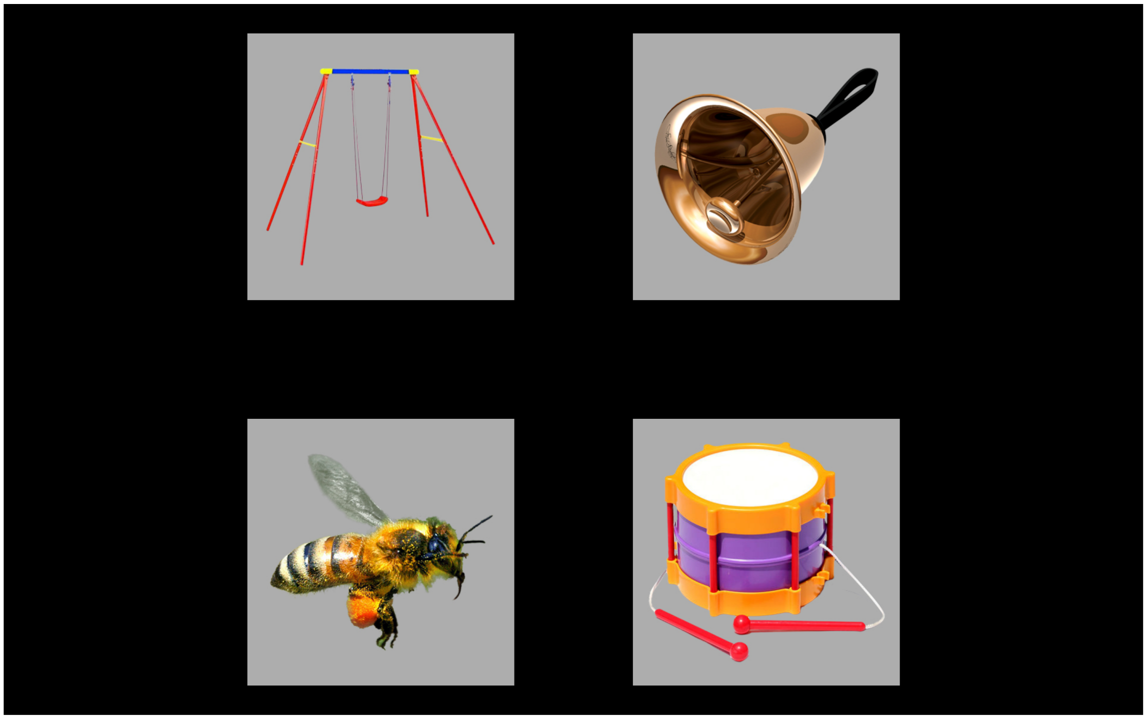


Figure 4.1: Example display for the target *bell* with the semantic foil *drum*, the phonological foil *bee*, and the unrelated *swing*.

2017; Huang & Snedeker, 2011; Law et al., 2016).

4.3 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 36 for a two-image task—see Chapter 10). 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 onscreen for 2 s as a familiarization period. The experiment’s software procedure then checked whether the child’s gaze was being recorded. If the procedure could continuously track the child’s gaze for 300 ms, the child’s gaze was verified and the trial continued. If the procedure could not verify the gaze after 10 s, the trial continued. This step 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.

4.4 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 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*). These different competitor types (phonological vs. semantic) and subtypes (e.g., shared syllable onset vs. rimes, shared category vs. perceptually similar) likely participate in word recognition to varying degrees and at different stages during lexical processing. For the analysis of familiar word recognition, I include all the competitors—they are aggregated together as *distractors*—but for the analysis of phonological and semantic competitors, I focus on subsets of competitors: the shared syllable onsets for phonological competitors and the category neighbors for semantic competitors. Appendix A provides a complete list of the items used in the experiment and in the analyses of competitor effects.

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 recruitment channel, address, and 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 durations 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 were recognized by at least 80% of students in each classroom.

4.5 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. The rationale for blockwise exclusion was that if the majority of trials were unreliable, then there was probably a problem during the session, such as a technical difficulty with the eyetracker or the child not complying with the task. As a result, all of the trials would be of questionable quality.

Table 4.2 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.

4.6 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. Fixations 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. I chose this window after visualizing the observed fixation probabilities and identifying when during a trial the probabilities started to rise and

Table 4.2: Eyetracking data before and after data screening. For convenience, the number of exclusions is included as Raw − Screened. *Percent Missing*: Percentage of looks offscreen during 0–2000 ms after target onset.

Dataset	Year	Children	Blocks	Trials	Percent Missing
Raw	Age 3	178	332	7967	24.4%
	Age 4	180	347	8327	22.9%
	Age 5	163	322	7724	17.8%
Screened	Age 3	163	291	5951	7.9%
	Age 4	165	305	6421	8.5%
	Age 5	156	295	6483	7.8%
Raw − Screened	Age 3	15	41	2016	16.5%
	Age 4	15	42	1906	14.3%
	Age 5	7	27	1241	10.1%

later plateaued. Lastly, I aggregated looks by child, year and time, and created orthogonal polynomials to use as time features for the model. Orthogonal polynomials are described in the next chapter.

Figure 4.2 depicts each child's proportion of looks to the target image following the data screening and model preparation steps. These are the observed or empirical growth curves; these are the probabilities that will be modeled with growth curve analysis. 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.

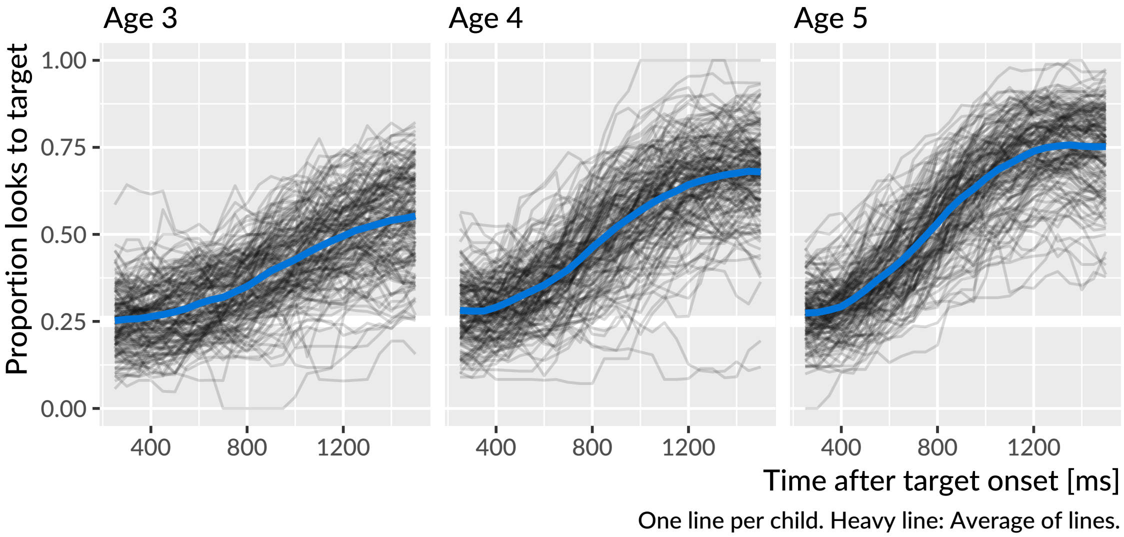


Figure 4.2: Empirical word recognition growth curves from each year of the study. Each line represents an individual child's proportion of looks to the target image over time. The heavy lines are the averages of the lines for each year.

5 Analysis of familiar word recognition

5.1 Growth curve analysis

The outcome measure of interest here is how the probability of fixating on the target image versus the distractors changes over time. There are many possible techniques one can employ for modeling time series data. In this chapter, I used growth curve analysis which uses polynomial functions of time (a linear trend, a quadratic trend, etc.) to estimate a time series. Barr (2008) and Mirman, Dixon, and Magnuson (2008) are important early tutorials for this technique of modeling looking probabilities. (Incidentally, the two articles were published together in a special issue of *Journal of Memory and Language* about "emerging" statistical techniques.) Mirman (2014) also provides a textbook treatment of growth curve analysis for eyetracking data. This approach is well suited for time series where the trajectory is relatively simple with one or two inflection points. Alternatively, one can forgo polynomial trends and use generalized additive (mixed) models to fit a more general nonlinear shape. I apply this now-emerging technique in Chapter 6 to model wigglier growth curve shapes. A third possibility is to use nonlinear, functional growth curves. For the polynomial and additive models, underlying time features are weighted and summed to fit a nonlinear shape. For the functional growth curve, the nonlinear shape is fixed in advance and the model estimates a set of curve parameters so the shape approximates the data. For example, Oleson, Cavanaugh, McMurray, and Brown (2017) and Seedorff, Oleson, and McMurray (2018) model eyetracking data by assuming an s-shaped curve (a logistic function) and then estimating the left and right asymptotes, the slope at the steepest point, and the point where the steepest rise occurs. These parameters are directly interpretable in terms of looking behaviors, but I have found that the technique is not flexible enough to handle the noisier shapes of children's eyetracking data.[1] For the following analyses, therefore, I used polynomial growth curves.

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 the distractors). 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 fit of some data, Bayesian methods consider an entire distribution of plausible fits that are consistent with the data and any prior information we have about the model parameters. 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).[2]

Word recognition growth curves—that is, looks to the target versus the distractors at 250 ms, 300 ms, etc.—were fit using an orthogonal cubic polynomial function of time. Put differently, I modeled the probability of looking to the target during an eyetracking task as:

$$\log \text{odds}(\text{looking}) = \beta_0 + \beta_1\text{Time}^1 + \beta_2\text{Time}^2 + \beta_3\text{Time}^3$$

That the time terms are *orthogonal* means that Time^1, Time^2 and Time^3 are transformed so that they are uncorrelated. See Box 1. Under this formulation, the parameters β_0 and β_1 have a direct interpretation in terms of lexical processing performance. The intercept, β_0, 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 β_0 as a measure of *word recognition reliability.* The linear time parameter, β_1, estimates the steepness of the growth curve—or how the probability of fixating

[1] More generally, I think of there being a flexibility–interpretability tradeoff with additive models being the most flexible but having the least interpretable parameters, functional curves being the least flexible but having the most interpretable parameters, and polynomials falling in between the two.

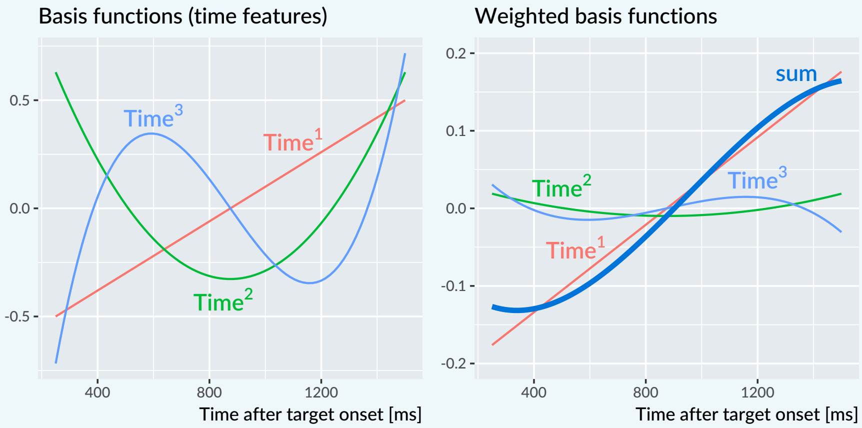
[2] My goals in using this method were simply to estimate model effects and quantify the uncertainty about those effects. This pragmatic, estimation-based approach of Bayesian statistics is illustrated in texts by Gelman and Hill (2007) and McElreath (2016).

changes from frame to frame. We can think of β_1 as a measure of *processing efficiency*, because growth curves with stronger linear features exhibit steeper frame-by-frame increases in looking probability.

Box 1: Orthogonal time.

I used orthogonal polynomial features of Time for these growth curve models. Unlike natural polynomials, these features are uncorrelated. This aspect makes these models more flexible: I do not have to worry about any collinearity between Time^1 and Time^2. Moreover, adding an orthogonal cubic Time^3 term to a quadratic model will not change any of the estimates for Time^1 or Time^2 because the added predictor is not correlated with the others. One disadvantage of this approach is that the features are not as straightforward to interpret.

The figure below shows the orthogonal polynomials used by the model and how they can be weighted and summed to fit a growth curve.



Note that the time features and weighted features are vertically centered around 0. The curves are adjusted up or down to their correct position by the model's intercept term. Conceptually, one can also think of the intercept as a Time^0 feature—that is, a horizontal line at $y = 1$ which is weighted to move the whole curve vertically. This is why in these models, the intercept is not the value at some time 0 but rather the average value of the fitted growth curve. To reiterate, for these word recognition models, the intercept is the average probability of the curve.

For all the polynomial growth curves models I used in this project, I scaled the features so that Time^1 ranges from $-.5$ to $.5$. In other words, a 1-unit change on Time^1 marks the whole traversal across the analysis window. After scaling, Time^2 ranges from $-.33$ to $.60$ and Time^3 ranges from $-.63$ to $.63$.

In my experience, only the intercept terms and linear time trends of an orthogonal polynomial model have a behaviorally straightforward interpretation. 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.

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:

$$\begin{aligned}
\log \mathrm{odds}(\mathrm{looking}) = {} & \beta_0 + \beta_1 \mathrm{Time}^1 + \beta_2 \mathrm{Time}^2 + \beta_3 \mathrm{Time}^3 + && \text{[age 3 growth curve]} \\
& (\gamma_0 + \gamma_1 \mathrm{Time}^1 + \gamma_2 \mathrm{Time}^2 + \gamma_3 \mathrm{Time}^3) * \mathrm{Age}\,4 + && \text{[age 4 adjustments]} \\
& (\delta_0 + \delta_1 \mathrm{Time}^1 + \delta_2 \mathrm{Time}^2 + \delta_3 \mathrm{Time}^3) * \mathrm{Age}\,5 && \text{[age 5 adjustments]}
\end{aligned}$$

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 B contains the R code used to fit the model along with a description of the model specifications represented in the model syntax.

I used Bayesian *uncertainty intervals* to draw statistical inferences. A Bayesian model is the posterior distribution: It is a distribution of plausible parameter values, given the data, a data-generating model and any prior information we have about those parameter values. In practice, these distributions are hard to calculate, so we use Markov Chain Monte Carlo to get a sample of thousands of values from the posterior distribution. Thus, rather than having a single best-fitting estimate of some effect β, we have a sample of, say, 4,000 plausible values for β. We can quantify our uncertainty about β by describing the distribution of those values. I use typically two statistics to describe that distribution. The median provides a *point estimate* for the distribution, and the 90% uncertainty interval provides bounds for the effect. These intervals have an intuitive interpretation. Suppose that for β we get a median of 8 and 90% uncertainty interval of [5, 21]. This interval means that we can be 90% certain that the "true" value of β is between 5 and 21, given the data, our model, and our prior information. Moreover, by inspecting the interval, we pinpoint areas of uncertainty. In this example, we can conclude that the effect is likely to be positive. The lower interval value of 5 tells us that 90% of the plausible values are greater than 5. A wide range of values are covered by the interval, however, so we would also conclude that we are not very certain about the size of the effect. It bears noting that one cannot interpret frequentist confidence intervals in this way. See Kruschke and Liddell (2017) for a recent review of frequentist versus Bayesian statistics.

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 5.1 shows three simulated growth curves and how each of these growth curve features relate to word recognition performance.

5.2 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 the average ("random" effects). Figure 5.2 shows 200 posterior samples of the population-average growth curves for each year. On average, the growth curves become steeper and achieve higher looking probabilities with each year of the study.

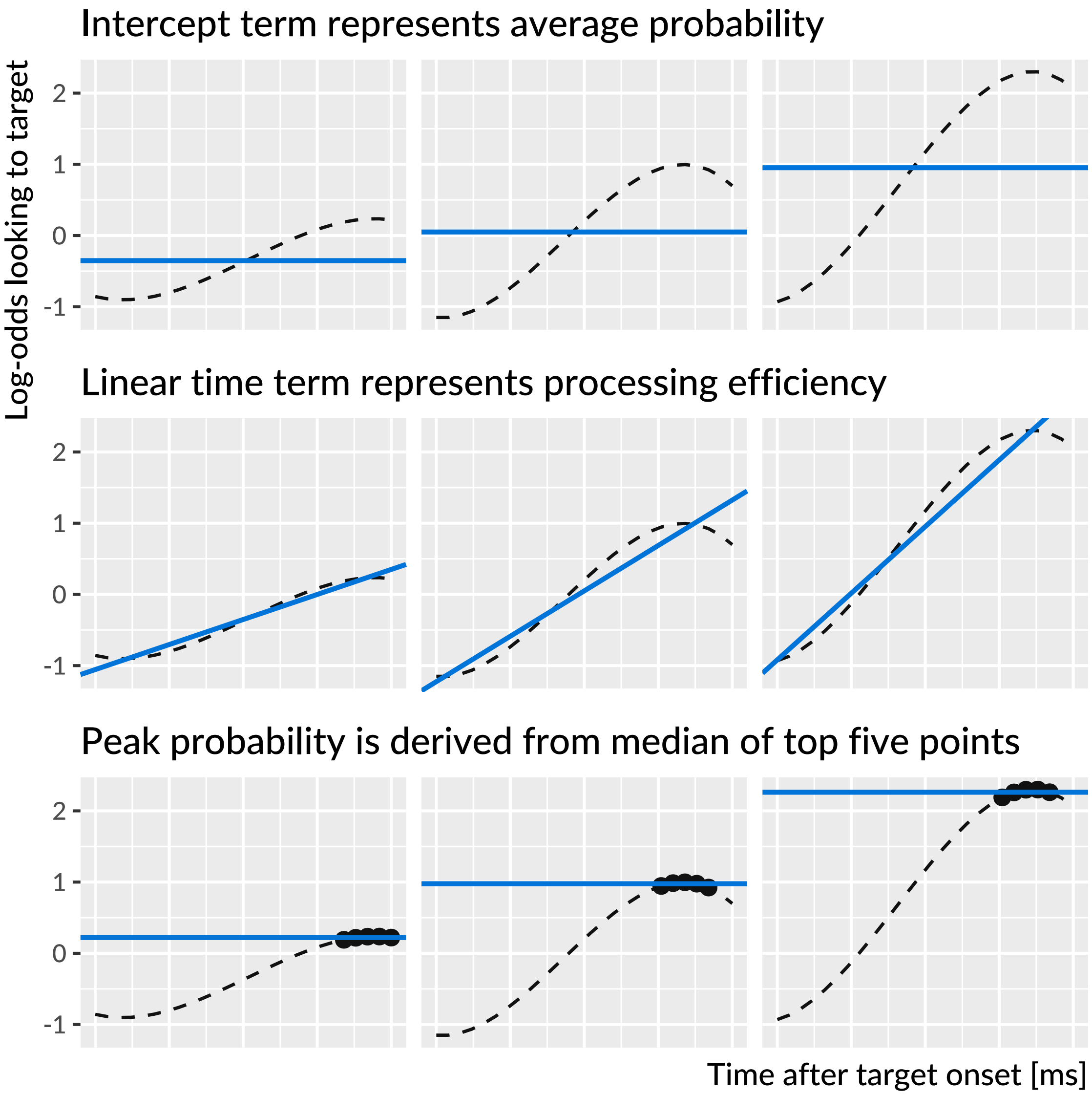


Figure 5.1: Illustration of the three growth curve features and how they describe lexical processing performance. The three curves used are simulations of new participants at age 4.

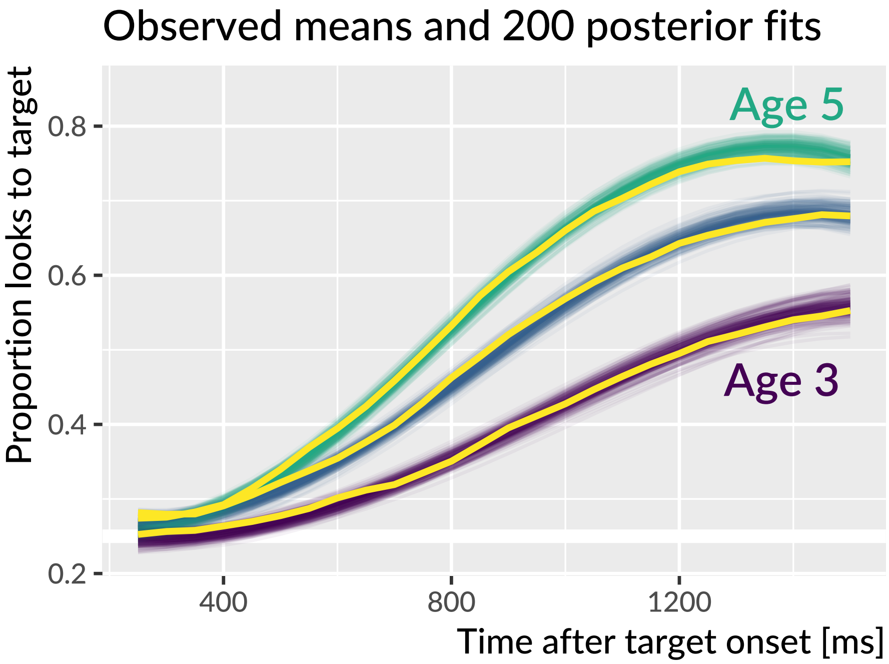


Figure 5.2: Population-average (“fixed effects”) word recognition growth curves at each age. Colored lines represent 200 posterior samples of these growth curves; these are included to visualize the uncertainty about the population averages. The thick light lines represent the observed average growth curve at each age.

Figure 5.3 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 5.4.

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

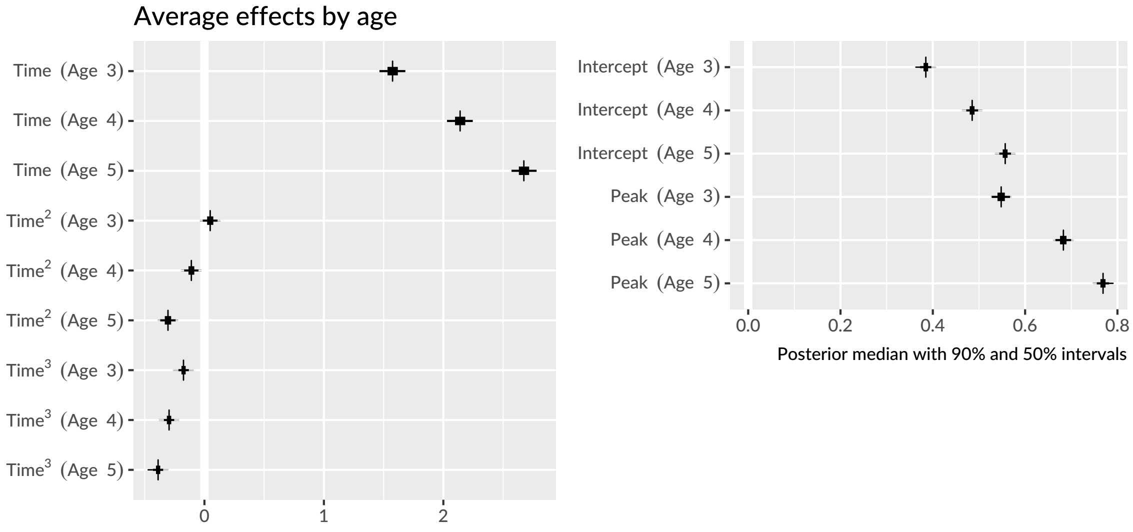


Figure 5.3: Uncertainty intervals for growth curve features at each age. The intercept and peak features were converted from log-odds to proportions to ease interpretation.

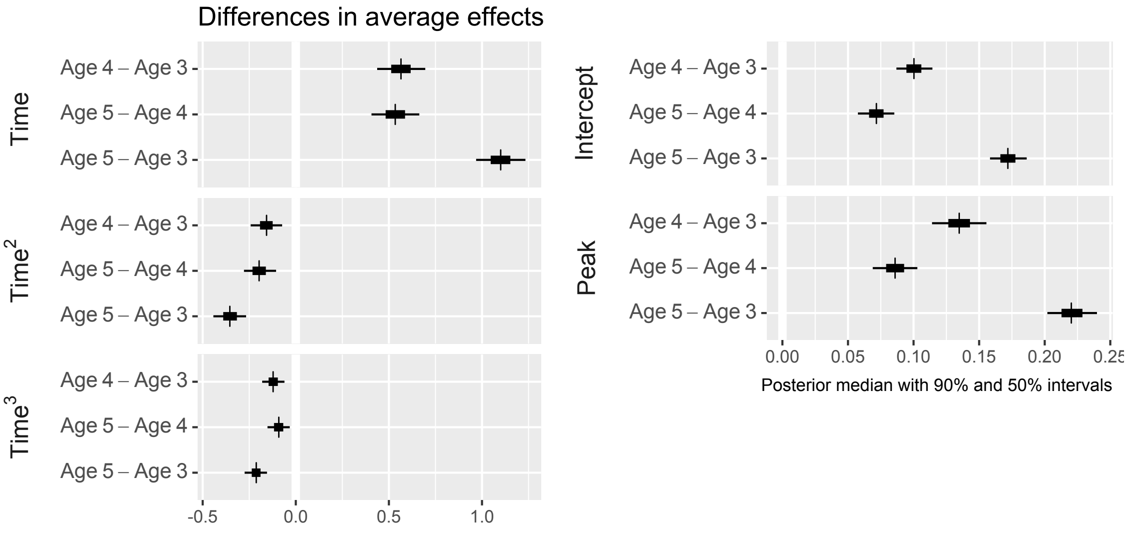


Figure 5.4: Uncertainty intervals for the differences in growth curve features between ages. Again, the intercept and peak features were converted to proportions.

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.

5.3 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 5.5 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 5.6 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.

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 5.7 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.

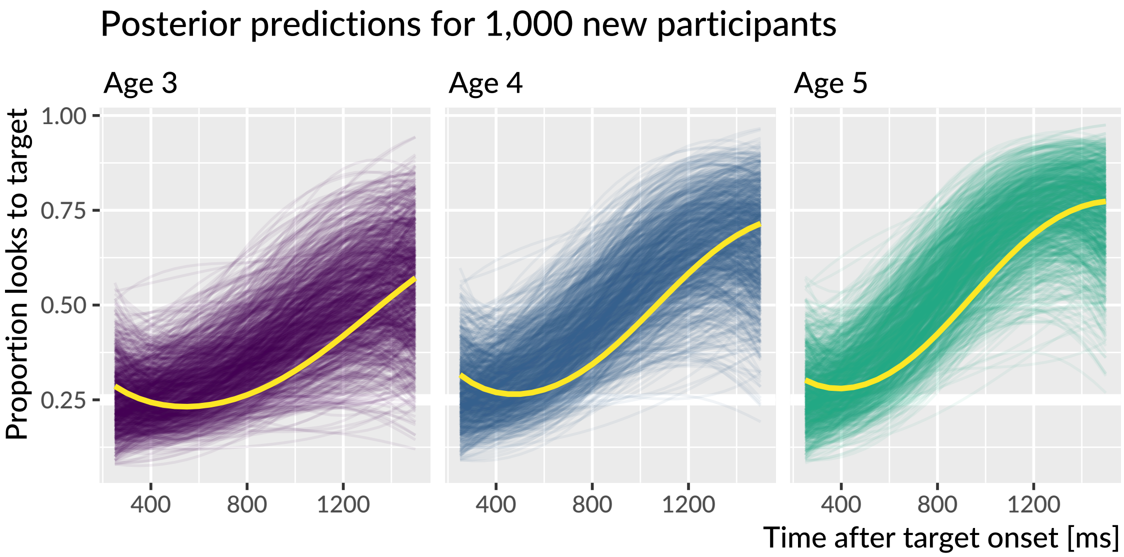


Figure 5.5: Posterior predictions for hypothetical *unobserved* participants. Each line represents the predicted performance for a new participant. The three light lines highlight predictions from one single simulated participant. The simulated participant shows both longitudinal improvement in word 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.

The ranges of performance decrease from age 3 to age 4 to age 5, consistent with the hypothesized reduction in 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, $\text{peak}_3 = 0.55$ [90% UI: 0.35, 0.77], $\text{peak}_4 = 0.69$ [0.48, 0.86], $\text{peak}_5 = 0.78$ [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 95^{th} and 5^{th} 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.

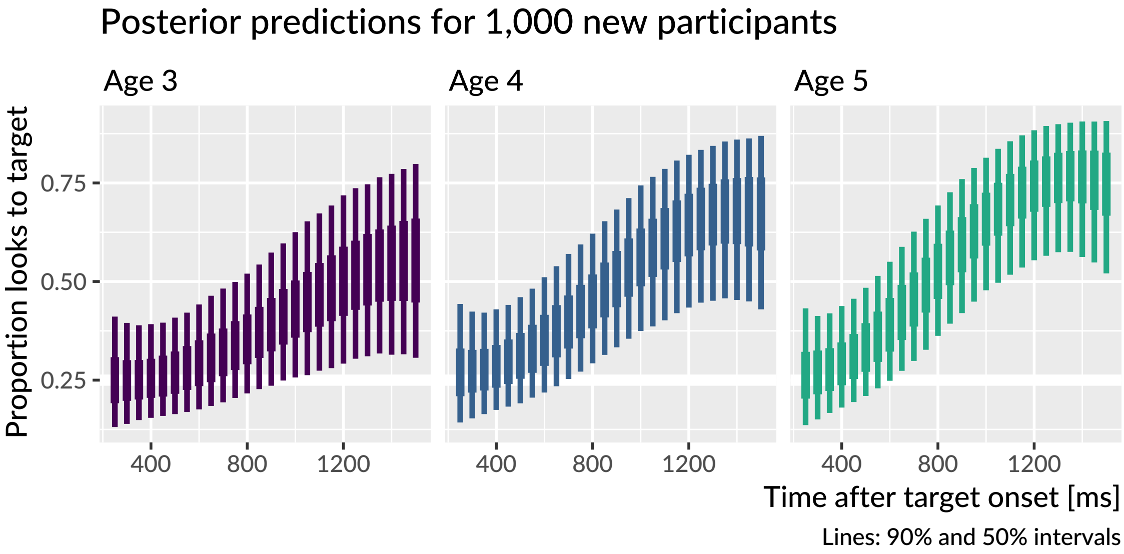


Figure 5.6: Uncertainty intervals for the simulated participants. Variability is widest at age 3 and narrowest at age 5, consistent with the prediction that children become less variable as they grow older.

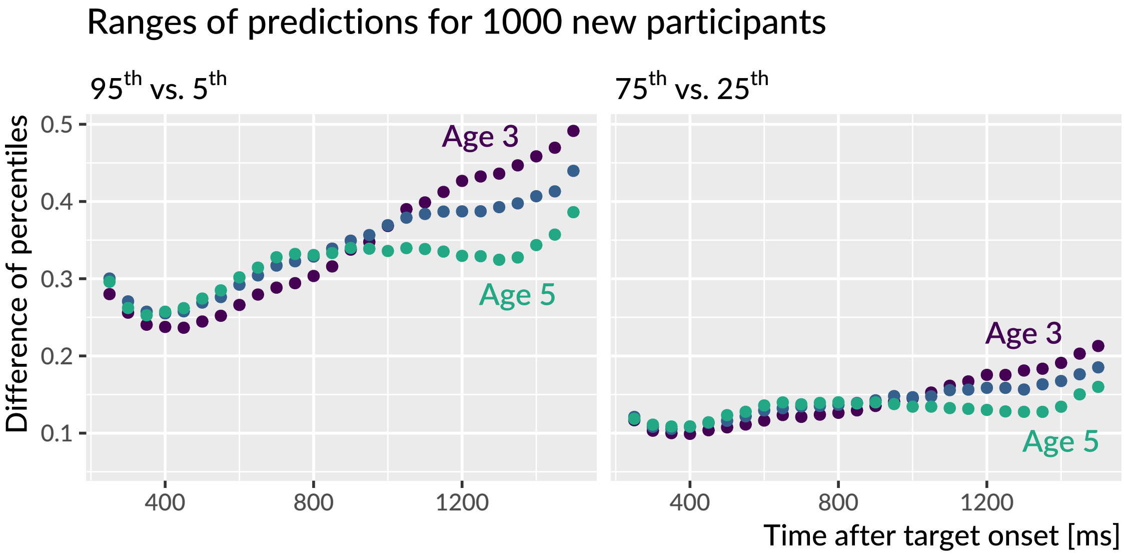


Figure 5.7: Ranges of predictions for simulated participants over the course of a trial. The ranges are most similar during the first half of the trial when participants are at chance performance, and the ranges are most different at the end of the trial as children reliably fixate on the target image. The ranges of performance decrease with each year of the study as children show less variability.

5.4 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 age 3 are one set of ratings and the terms from ages 4 and 5 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: Time^1, Time^2, Time^3, average looking probability, 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 year, 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.

I used the `kendall()` function in the irr R package (vers. 0.84; Gamer, Lemon, & Singh, 2012) to compute concordance statistics. Figure 5.8 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 Time^2 effect [.30, .35] and the Time^3 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

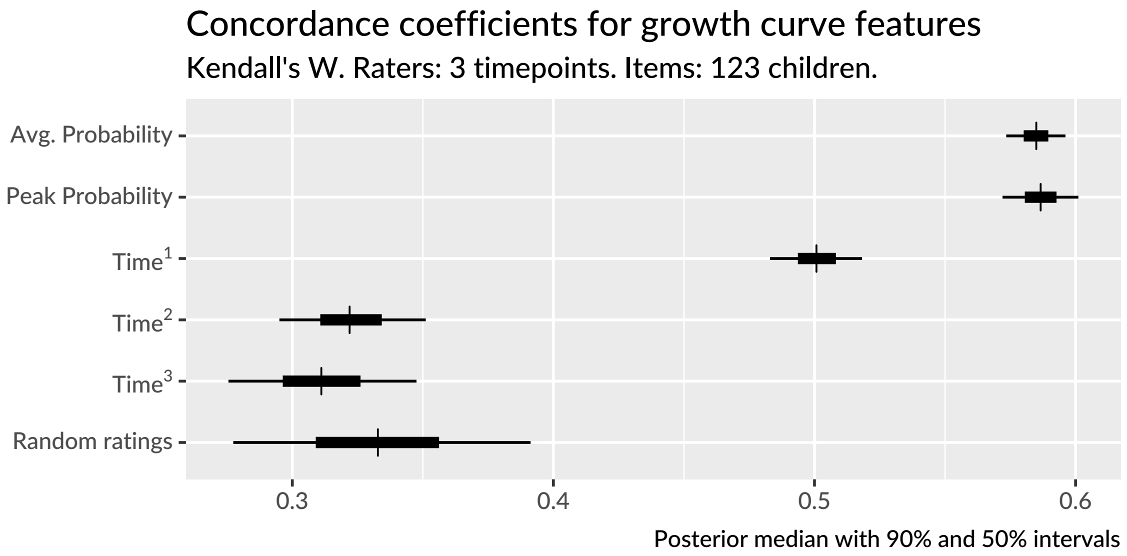


Figure 5.8: Uncertainty intervals for the Kendall's coefficient of concordance. Random ratings provide a baseline of null W statistics. The peak, intercept and linear time features are decisively non-null, indicating a significant degree of correspondence in children's relative word recognition reliability and efficiency over the three years of the study.

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 Time^2 and Time^3 features differentiate children across years as well as random numbers.

Concordance is strongest for the peak feature, $W = .59$ [.57, .60] and the intercept term, $W = .58$ [.57, .60], followed by the linear time term, $W = .50$ [.48, .52]. Because these values are far removed from the statistics for random ratings, I conclude that there is a credible degree of correspondence across years 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 were relatively fast (or reliable) at word recognition at one age were also relatively fast (or reliable) at other ages too.

5.5 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 5.9 shows the correlations of the peak looking probability, average looking probability and linear time features with expressive vocabulary size at age 5, and Figure 5.10 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].

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].

Summary. Although individual differences in word recognition were stable over time, early differences were more significant than later ones. The strongest predictors

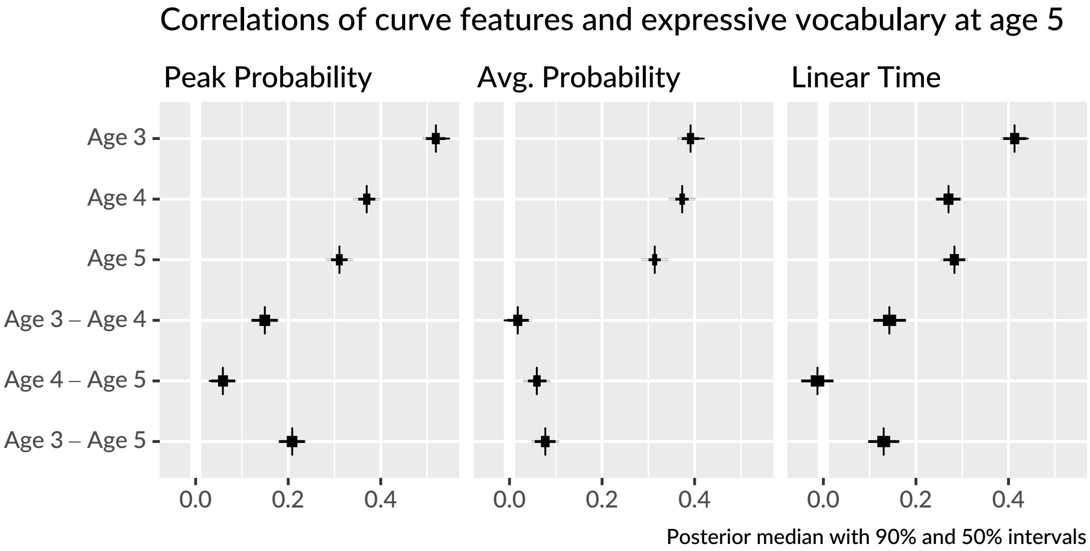


Figure 5.9: Uncertainty intervals for the correlations of growth curve features at each age with age-5 expressive vocabulary (EVT-2 standard scores). The bottom rows provide intervals for the pairwise differences in correlations between timepoints. For example, the top row of the left panel is the correlation between age-3 peak probability and age-5 expressive vocabulary.

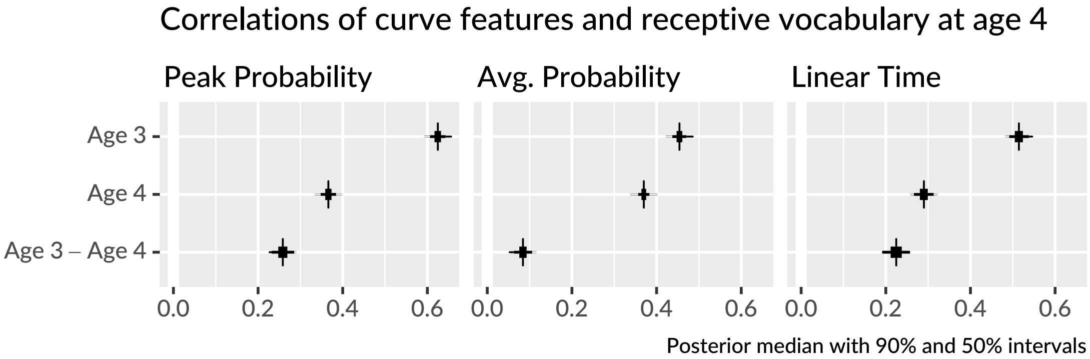


Figure 5.10: Uncertainty intervals for the correlations of growth curve features from age 3 and age 4 with age-4 receptive vocabulary (PPVT-4 standard scores). The bottom row shows pairwise differences between the age-3 and age-4 correlations.

of future vocabulary size were the growth curve features from age 3. Of these features, correlations were strongest for peak looking probabilities.

5.6 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 where 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.

6 Effects of phonological and semantic competitors

6.1 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), 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 A 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 three distractors each year. Figure 6.1 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.

As in the models from the previous chapter, I downsampled the data into 50-

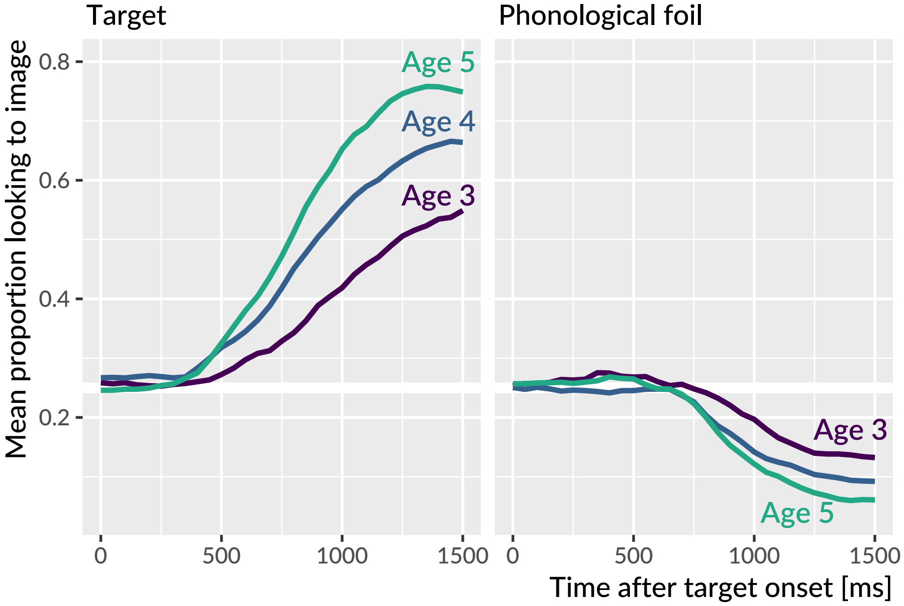


Figure 6.1: Because children looked more to the target as they grew older, they numerically looked less the foils too. This effect is why I evaluated the phonological and semantic foils by comparing them against the unrelated image.

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, year and time bin. To account for the sparseness of the data, I used the empirical log-odds (or empirical logit) transformation (Barr, 2008). 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 for a tutorial for linguists; Winter & Wieling, 2016; Wood, 2017). Box 2 provides a brief overview of these models. I used the mgcv R package (vers. 1.8.24; Wood, 2017) with support from the tools in the itsadug R package (vers. 2.3; van Rij, Wieling, Baayen, & van Rijn, 2017).[1]

[1] 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

Appendix B contains the R code used to fit these models along with a description of the specifications represented by the model syntax.

Box 2: Intuition behind generalized additive models.

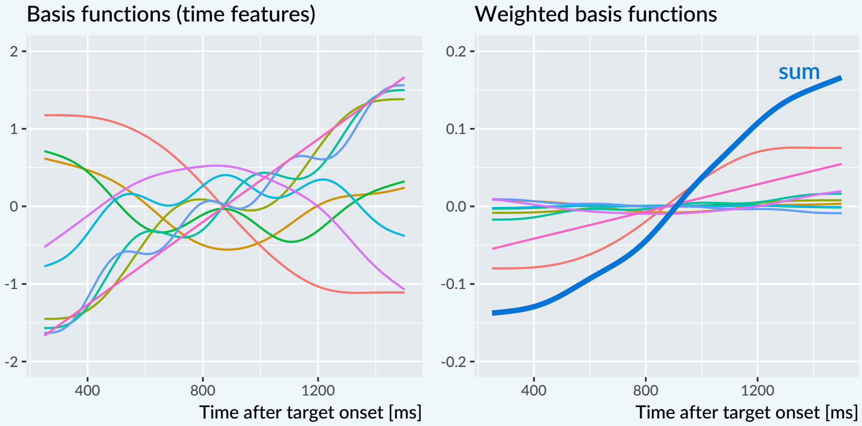
In these analyses, the outcome of interest is a value that changes over time in a nonlinear way. We model these time series by building a set of features to represent time values. In the growth curve analyses of familiar word recognition, I used a set of polynomial features which expressed time as the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

$$\log \text{odds}(\text{looking}) = \alpha + \beta_1 \text{Time}^1 + \beta_2 \text{Time}^2 + \beta_3 \text{Time}^3$$

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:

$$\log \text{odds}(\text{looking}) = \alpha + f(\text{Time})$$

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., $f(\text{Time})$—so we statistically evaluate the whole batch of coefficients simultaneously. This joint testing is similar to how one

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.

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 year-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 less data at the child level than at the year 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:

$$
\begin{aligned}
\text{emp. log odds(phon. vs. unrelated)} = {} & \alpha + \beta_1 \text{Age } 3 + \beta_2 \text{Age } 5 + && \text{[growth curve averages]} \\
& f_1(\text{Time}, \text{Age } 4)\ + && \text{[reference smooth]} \\
& f_2(\text{Time}, \text{Age } 4 - \text{Age } 3)\ + && \text{[difference smooths]} \\
& f_3(\text{Time}, \text{Age } 4 - \text{Age } 5)\ + && \\
& f_i(\text{Time}, \text{Child}_i) && \text{[by-child random smooths]}
\end{aligned}
$$

The model's fitted values are shown in Figure 6.2. 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

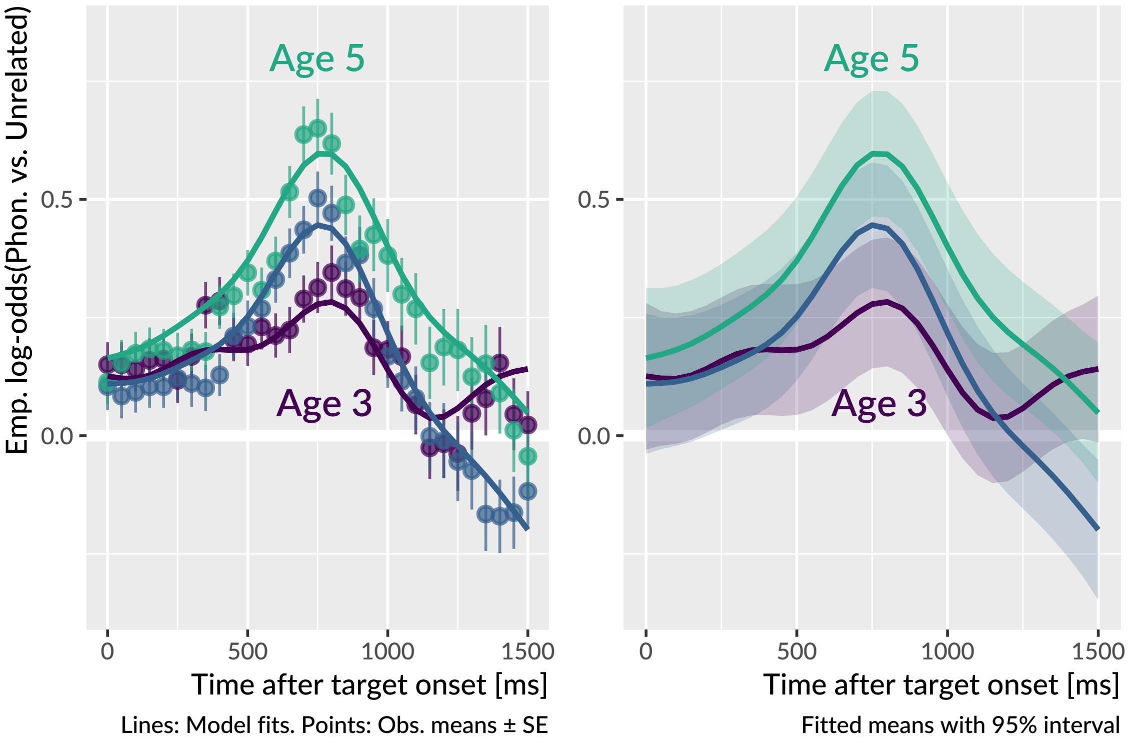


Figure 6.2: With each year of the study, children looked more to the phonological competitor (relative to the unrelated image) during and after the target noun. Both figures show means for each year estimated by the generalized additive model. The left panel compares model estimates to observed means and standard errors, and the right panel visualizes estimated means and their 95% confidence intervals.

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.

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.

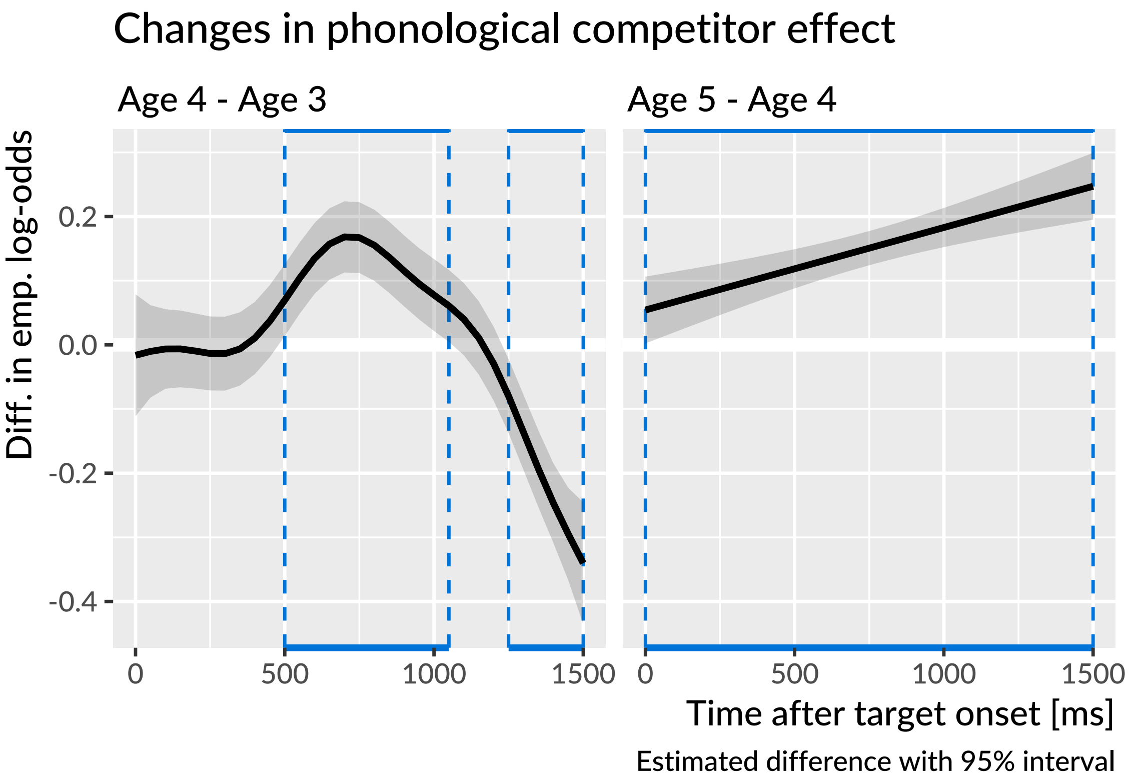


Figure 6.3: Differences in the average looks to the phonological competitor versus the unrelated image between age 4 and the other ages. Plotted line is estimated difference and the shaded region is the 95% confidence interval around that difference. Boxes highlight regions where the 95% interval excludes 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 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 6.3 visualizes how and when the smooths from other ages differed from the age-4 smooth.

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 age smooths, 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.

6.2 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), 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 A 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 year effects and a time smooth, time-by-year 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 6.4. 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.

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 6.5 visualizes the time course of the differences between the smooths from each year.

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

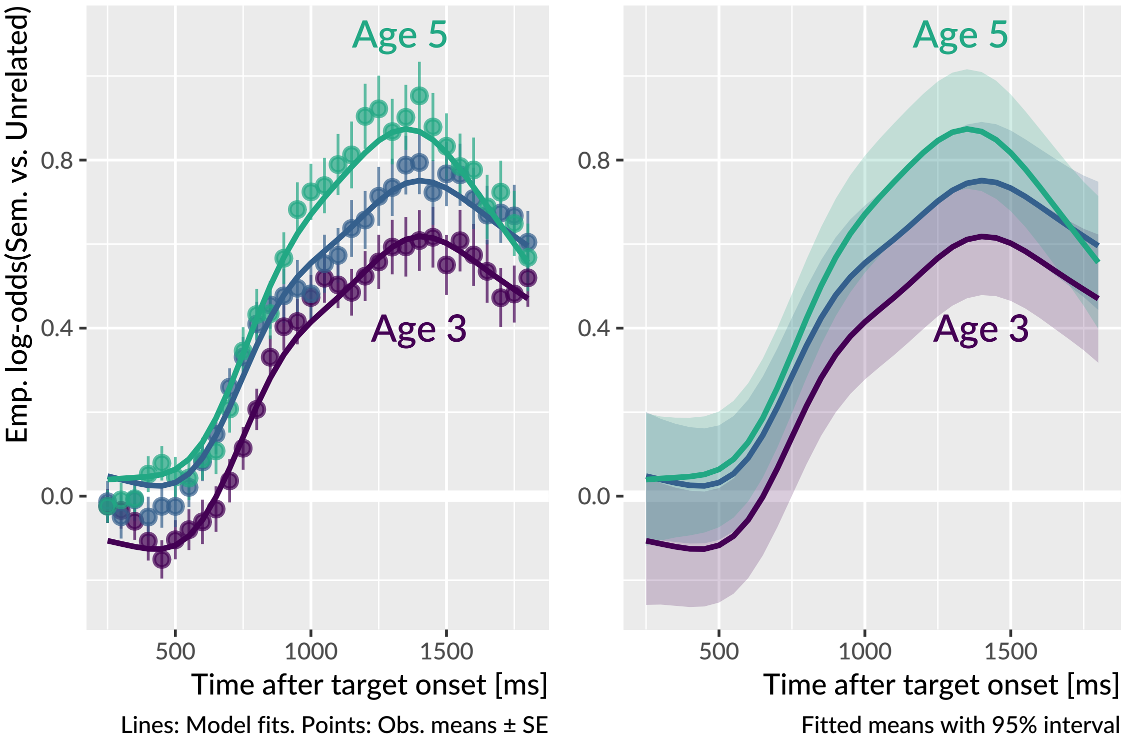


Figure 6.4: With each year of the study, children looked more to the semantic foil (relative to the unrelated image) with peak looking occurring after the target noun. Both figures show means for each year estimated by the generalized additive model. The left panel compares model estimates to observed means and standard errors, and the right panel visualizes estimated means and their 95% confidence intervals.

the 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.

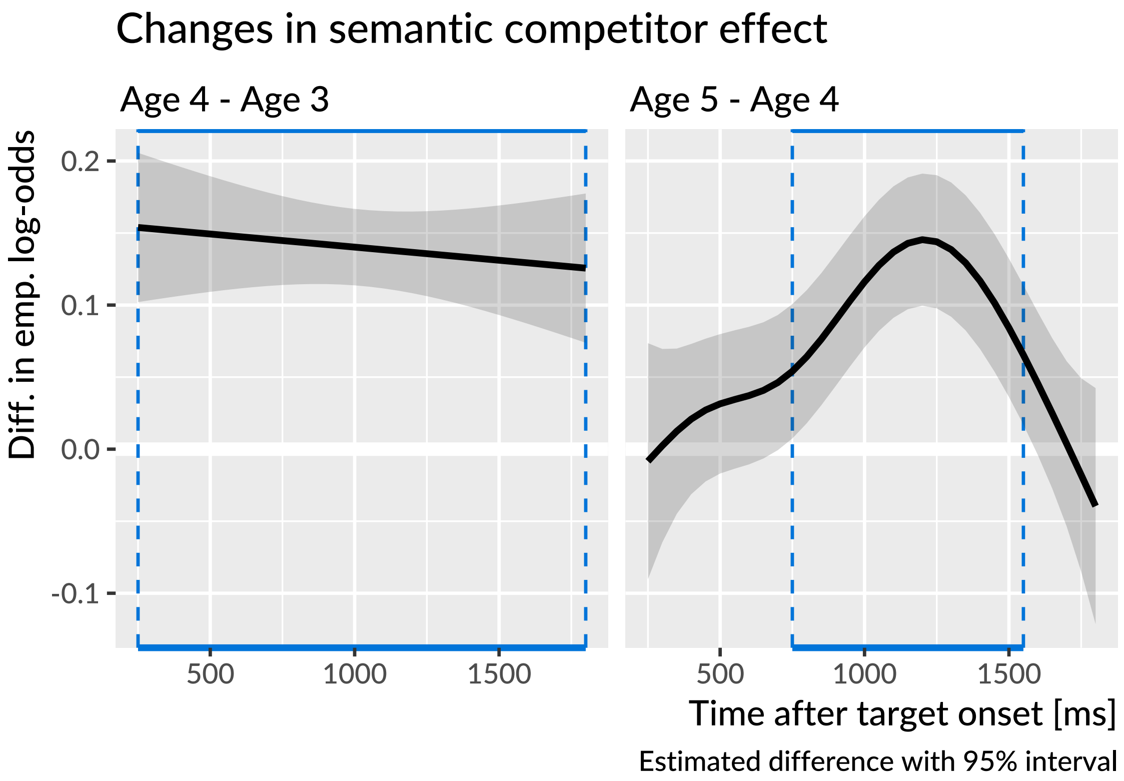


Figure 6.5: Differences in the average looks to the semantic competitor versus the unrelated word between age 4 and the other ages. Plotted line is estimated difference and the shaded region is the 95% confidence interval around that difference. Boxes highlight regions where the 95% interval excludes 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.

6.3 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. The task (based on Baylis, Munson, & Moller, 2008) 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.

$$
\begin{aligned}
\log \text{odds}(\text{choose target}) = {} & \alpha\ + && \text{[average child ability]} \\
& \alpha_i\ + && \text{[difference of child } i\text{'s ability from average]} \\
& \alpha_j\ + && \text{[word } j\text{'s difficulty]} \\
& \alpha_{j,k}\ + && \text{[word } j\text{'s difficulty in word-pair } k\text{]} \\
& \beta_1 \text{Age}\ + && \text{[child-level predictors]} \\
& \beta_2 \text{Vocabulary} &&
\end{aligned}
$$

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:

$$
\begin{aligned}
\text{emp.}\log\text{odds}(\text{phon. vs. unrelated}) = {} & \alpha\ + && \text{[growth curve average]} \\
& f_1(\text{Time})\ + && \text{[time smooth]} \\
& f_2(\text{Ability})\ + && \text{[ability smooth]} \\
& f_3(\text{Time} * \text{Ability})\ + && \text{[interaction smooth]} \\
& f_i(\text{Time}, \text{Child}_i) && \text{[by-child random smooths]}
\end{aligned}
$$

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.

6.4 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 representation so that partial information can more eagerly activate compatible words.[2]

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.[3] 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.

[2] 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.

[3] 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.

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.

7 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.

7.1 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) used the TRACE model of word recognition (McClelland & Elman, 1986) 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.

7.2 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) 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) 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.

7.3 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). 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.[1] 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.

7.4 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 lack of an explicit selection behavior, such as a mouse click, also means that word recognition accuracy was never directly measured but rather inferred. As a

[1] "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, p. 217).

result, the interpretation of peak looking probability as a measure of word recognition certainty comes with a caveat: It reflects certainty averaging over many familiar words *and maybe a few unfamiliar words.* The idea is as follows: Suppose at age 3 a child does not know four of words well. If they had to click or tap an image, they would have to guess on these trials. We could exclude those trials where they guessed incorrectly, leaving just the trials where the child correctly recognized the word. In this scenario, we would be more justified in interpreting a growth curve peak as a measure of certainty during familiar word recognition because trials involving incorrectly identified words had been excluded. (It bears mentioning that explicit selection behaviors during the experiment are just one way to test a child's knowledge of items; another is a receptive vocabulary test after the experiment which checks whether the child can point to the words from the experiment.)

As it stands here, there is no clear way to tease apart whether the lower growth curve peaks at age 3 reflected greater uncertainty during lexical processing or a greater number of words being unfamiliar (or unknown) to the children. I favor the former interpretation because these were highly familiar words and because children's word recognition improved from age 4 to age 5. We piloted the images/words in two preschool classrooms, using only items that were at least 80% recognizable to children. But even if some words were unfamiliar at age 3, the number of unfamiliar words at age 4 was likely to be very small and therefore unknown words would have exerted a minimal effect on the lexical processing measures. The average peak looking probabilities increased by about .13 at age 4 (from .55 to .68) and by about .09 at age 5 (to .77)—the magnitude of these changes are comparable. Because children also showed improvements at age 5, when the effect of unknown words would be very small, age-related improvements in word recognition certainty likely reflect changes in lexical processing, as opposed to changes in the average mixture of known and unknown words.

The experimental design included semantic and phonological competitors on every trial, so isolating out the semantic and phonological competition effects required some subtlety. As a result, the lexical competition effects are only indirectly observed A more direct design would compare different types of 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 in the analysis of phonological competitors (which focused on just competitors with the same syllable 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 related limitation is that the phonological competitors used here are weak competitors. Adult studies tend to use phonological competitors with substantial overlap between the target and the competitor. For example, the landmark study of adults by Allopenna, Magnuson, and Tanenhaus (1998)—which showed that participants' eyetracking probabilities matched lexical activations from the TRACE model of word recognition (McClelland & Elman, 1986)—used target–cohort pairs that shared a whole syllable: *beaker–beetle*, *candle–candy*, *carrot–carriage*, *castle–casket*, *dollar–dolphin*, *paddle–padlock*, *pickle–picture*, *sandal–sandwich*. With this degree of overlap, there is much more phonological and temporal ambiguity for the cohort to build up activation and compete with the target. In contrast, the words used in this study were all one syllable and the amount of overlap was limited to syllable onset (e.g., *flag–fly*, *pen–pear*). This reduced overlap limits the degree over temporal ambiguity and thus limits the degree to which the competitors can participate in lexical competition. These words were weak phonological competitors, compared to others studies in the adult literature. As a result, the brief advantage of the phonological competitor over the unrelated word may *underestimate* children's sensitivity to phonological competitors: Preschoolers probably will show much more interference from competitors that have a larger degree of overlap. Moreover, with more interference from the competitors, individual differences could emerge more clearly so that child-level measures like speech perception can predict processing differences.

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.

8 Hypothesis check

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 increased 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 design a task and derive a measure from the competitor looking curves that does correlate with child-level measures.

Study 2: Referent selection and mispronunciations

9 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.

9.1 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, p. 148).

Charles-Luce and Luce (1990, 1995) are commonly cited touchstones for this argument. Charles-Luce and Luce compared the expressive (1990) and receptive (1995) 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) rebutted the 1990 study, observing that kids have sparser neighborhoods because they have sparse lexicons. Dollaghan (1994) also showed that young children do indeed have dense neighborhoods in their lexicons. Coady and Aslin (2003) 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. That is, words are more likely to be neighbors if the early lexicon favors more frequent 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) 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 yielded 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 early phonological representations because 14-month-olds could detect the *liff*-*neem* switch (Werker, Cohen, Lloyd, Casasola, & Stager, 1998) but *not* a minimal pair *bih*-*dih* switch (Stager & Werker, 1997).

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), 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) presented 18–23-month-olds two familiar images onscreen, like baby and dog, and the children 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) 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) 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) 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) 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) applied this approach 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.

9.2 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 be 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 (mutual exclusivity principle, Markman & Wachtel, 1988; Novel Name–Nameless Category principle, Mervis & Bertrand, 1994). The nonword trials in studies like Law and Edwards (2015) and White and Morgan (2008) 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), 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) replicated the preference for a super-novel object during a word recognition eyetracking task. In this case, 22-month-old English-learning toddlers were pre-exposed to images of novel objects. Later, 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 listeners' comprehension of familiar words and interpretation of nonwords reflect different processes. McMurray et al. (2012) 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, p. 832).

Bion, Borovsky, and Fernald (2013) 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?*). During later retention trials, 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 (82% versus 68% looks to the target for the 30-month-olds).

For Bion et al. (2013), 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.

9.3 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? Moreover, 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) and Law and Edwards (2015). 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.

In Chapter 11, 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 a 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, 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.

Finally, in Chapter 13, I describe the both sets of analyses together, and Chapter 14 reviews the results of my pre-analysis research hypotheses. In Appendix E, I briefly present the results for specific mispronunciations, although item effects are not formally modeled.

10 Method

Data collection for this experiment occurred during the same longitudinal study as the familiar word recognition experiment of Study 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.

10.1 Mispronunciation task

This experiment is an adaptation of the mispronunciation detection task by White and Morgan (2008) and Law and Edwards (2015). 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., *duck*) or a one-feature mispronunciation of the first consonant of the target word (*guck*). 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., *shann* presented with images of a cup and a novel-looking bassoon reed). Each nonword was constructed to match the phonotactic probability of one of the mispronunciations. Figure 10.1 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).

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

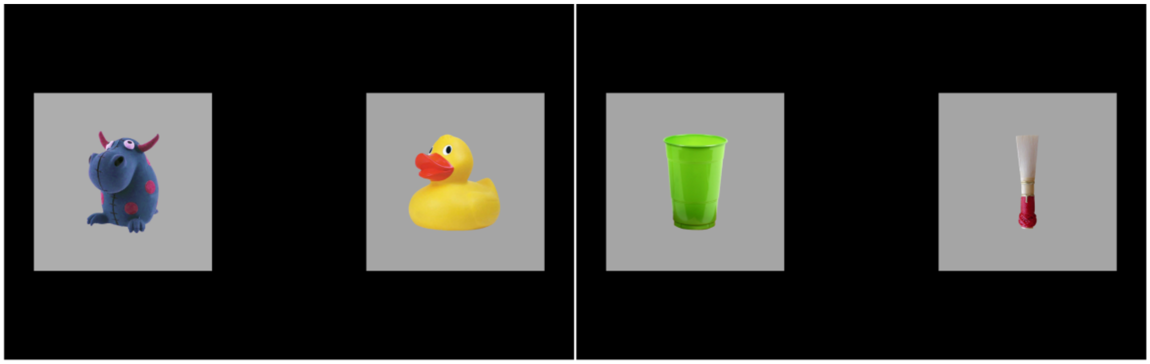


Figure 10.1: Example displays for a trial in which *duck* is mispronounced as "guck" (*left*) and a trial in which the nonword *shann* is presented (*right*).

task. A complete list of the items used in the experiment over the three years of the study is included in Appendix C.

10.2 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.

10.3 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

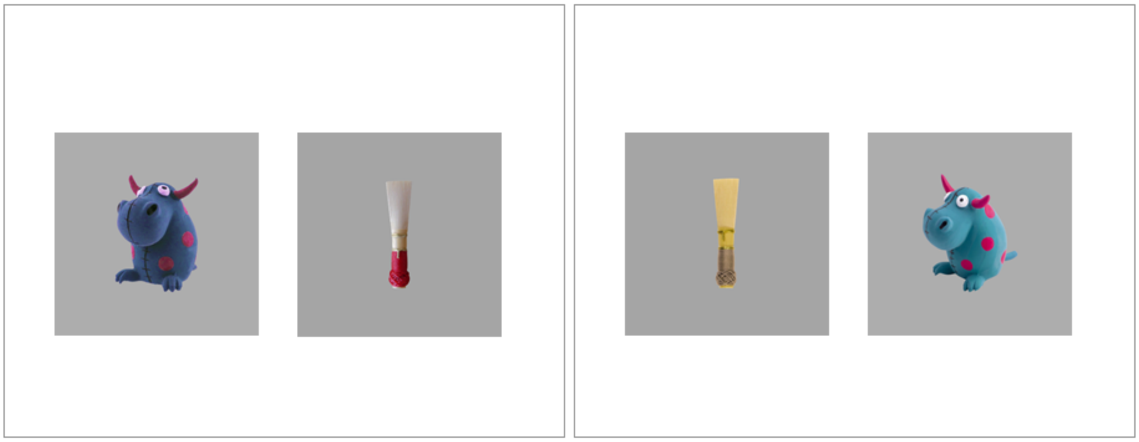


Figure 10.2: Examples of retention trials that tested *guck* and *shann*. During the first retention trial (*left*), children heard one of the unfamiliar words (e.g., *guck*). The correct response was to point to the toy bull creature because it was the unfamiliar object used on the *duck–guck* trials. During a later trial that used different image assets (*right*), children heard the other word (*shann*). The correct response was to point to the bassoon reed because it was the unfamiliar object used on the *shann* trials.

trial and the other was from a nonword trial. For example, the toy creature (*guck*) was paired with the bassoon reed (*shann*) from the nonword condition. The pairs were yoked, as each nonword was designed to match the phonotactic probability of one of the mispronunciations. During the retention test, children saw two images of the novel objects (say, *reed1.jpeg* and *toy1.jpeg*) printed on a letter-size sheet of paper, heard one of novel labels (*shann*), and had to point to the named object. Figure 10.2 shows an example of what the children saw when tested on *guck* and *shann*. In a later trial, the other label (*guck*) was tested but using different image assets for the objects (*toy2.jpeg* and *reed2.jpeg*). In a block of testing, there were 12 trials, 6 for nonwords and 6 for mispronunciations.

10.4 Data screening

Table 10.1 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. After mapping the

Table 10.1: Eyetracking data before and after data screening. For convenience, the number of exclusions is included as Raw − Screened. *Percent Missing*: Percentage of looks offscreen during 0–2000 ms after target onset.

Dataset	Year	Children	Blocks	Trials	Percent Missing
Raw	Age 3	177	341	12245	25.4%
	Age 4	181	349	12600	21.4%
	Age 5	164	325	11736	16.7%
Screened	Age 3	162	305	9062	7.9%
	Age 4	170	320	10031	8.1%
	Age 5	157	306	10113	7.8%
Raw − Screened	Age 3	15	36	3183	17.6%
	Age 4	11	29	2569	13.3%
	Age 5	7	19	1623	8.9%

gaze coordinates onto the onscreen images, I performed data screening following the same set of steps as in Chapter 4. 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 age-level growth curves, I observed an increasing preference for the unfamiliar image for the nonword condition—see Figure 10.3. 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

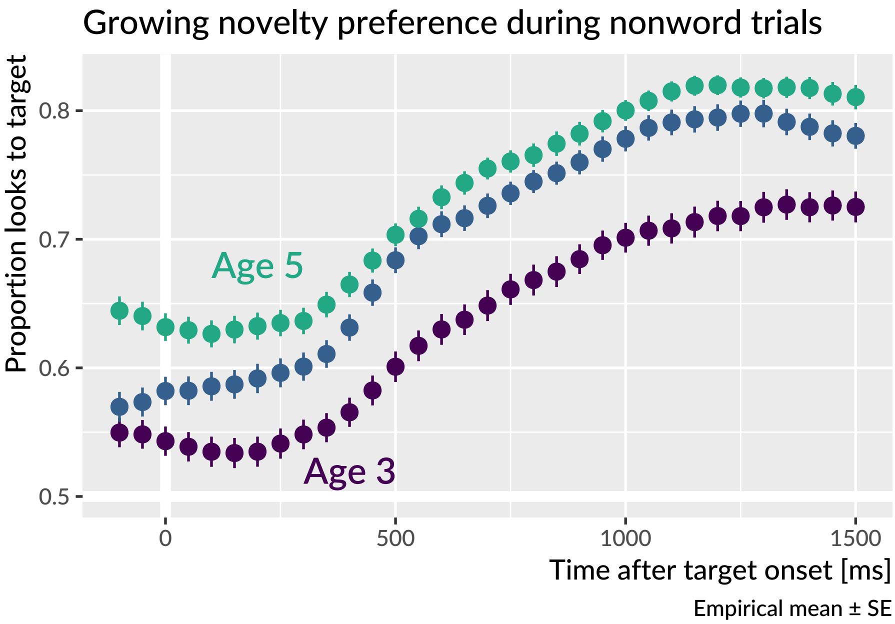


Figure 10.3: Observed average looks to the target on nonword trials. At the onset of the target noun, there is a novelty preference that increases with each year of the study. This novelty preference is the motivation for separating trials based on gaze location at target onset. Points and intervals represent the mean and standard error of children's empirical growth curves.

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 on the assumption that the earlier image better reflected the child's fixation location at the onset of the target word. 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 10.2 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

Table 10.2: Number of trials classified based on initial fixation location.

Year	Condition	Familiar initial	Unfamiliar initial	Neither/excluded
Age 3	Real word	1629 (53.7%)	1250 (41.2%)	154 (5.10%)
	Nonword	1284 (43.8%)	1453 (49.6%)	194 (6.60%)
	Mispronunciation	1608 (51.9%)	1305 (42.1%)	185 (6.00%)
Age 4	Real word	1561 (45.9%)	1693 (49.8%)	145 (4.30%)
	Nonword	1280 (39.2%)	1799 (55.2%)	183 (5.60%)
	Mispronunciation	1686 (50.0%)	1552 (46.1%)	132 (3.90%)
Age 5	Real word	1718 (50.5%)	1558 (45.8%)	125 (3.70%)
	Nonword	1172 (35.2%)	1959 (58.9%)	194 (5.80%)
	Mispronunciation	1752 (51.7%)	1487 (43.9%)	148 (4.40%)

for classifying trials based on the initial fixation location.

10.5 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, year, condition, initial fixation location, and time, and I created orthogonal polynomials to use as time features for the model. Figure 10.4, Figure 10.5, and Figure 10.6 shows the empirical growth curves for each condition following the above-described data screening and preparation steps.

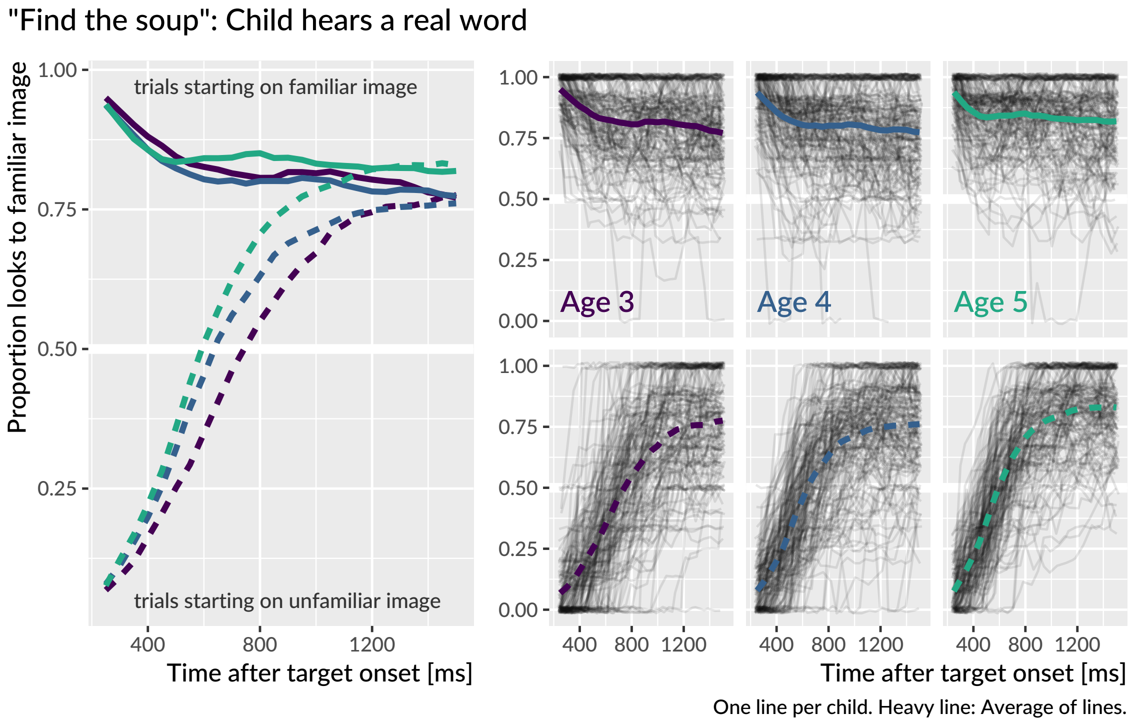


Figure 10.4: Empirical word recognition growth curves for the real words. Each line represents an individual child's proportion of looks to the target image over time. The heavy lines are the averages of the lines for each year. Only the steep, upward growth curves from unfamiliar-initial trials are analyzed.

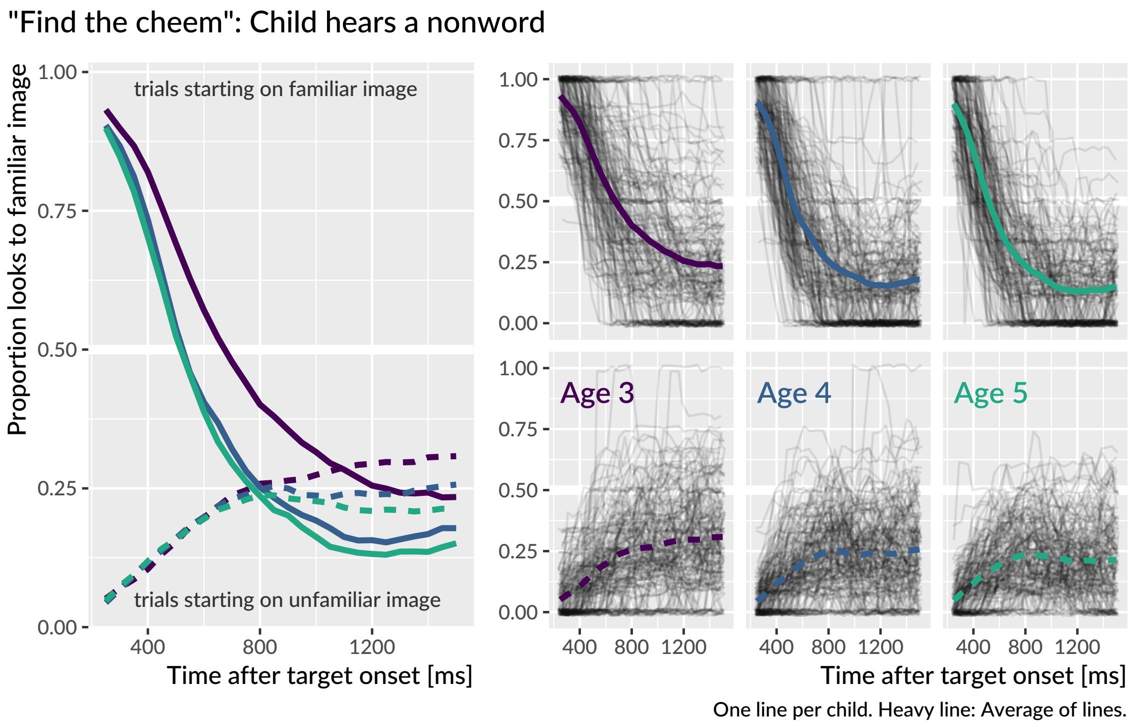


Figure 10.5: Empirical word recognition growth curves for the nonwords. Only the steep, downward growth curves from familiar-initial trials are analyzed.

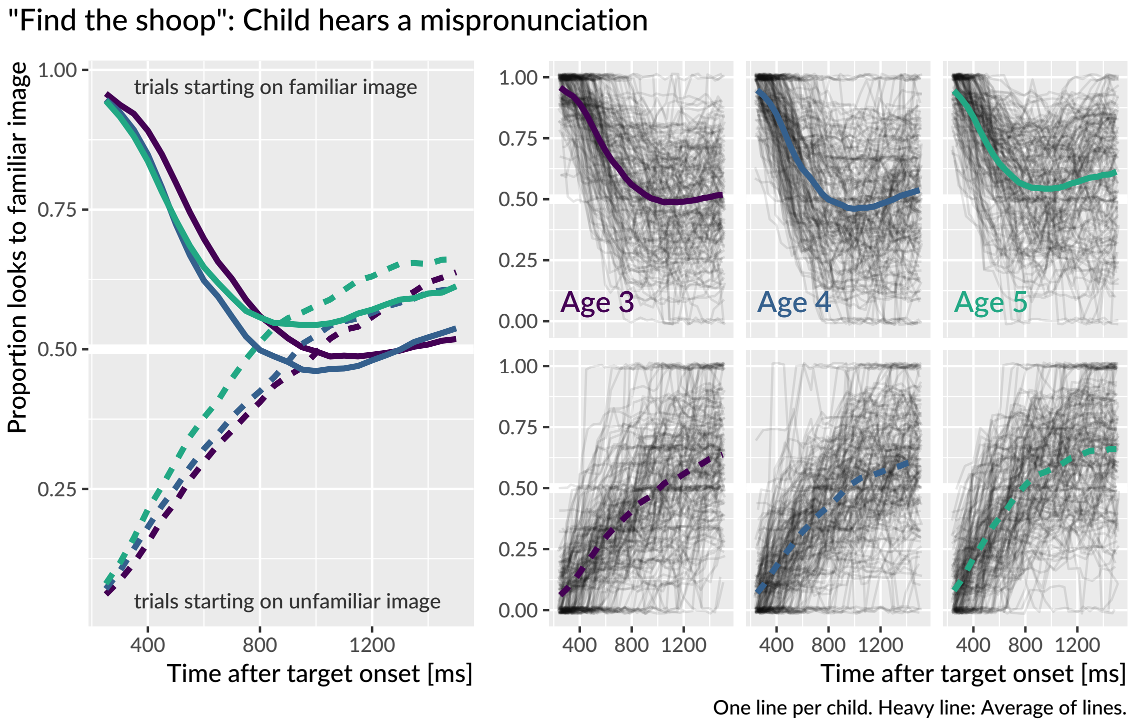


Figure 10.6: Empirical word recognition growth curves for the mispronunciations. Both types of curves are analyzed.

11 Development of referent selection

11.1 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. 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 with linear, quadratic and cubic time features but it adds a condition effect which interacts with these features. The linear model was:

$$\begin{aligned}\log \text{odds}(\text{looking}) = \beta_0 + \beta_1\text{Time}^1 + \beta_2\text{Time}^2 + \beta_3\text{Time}^3 + & \qquad \text{[nonword curve]} \\ (\gamma_0 + \gamma_1\text{Time}^1 + \gamma_2\text{Time}^2 + \gamma_3\text{Time}^3) * \text{Condition} & \qquad \text{[real words]}\end{aligned}$$

I fit a separate model for each year of the study. Appendix D 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

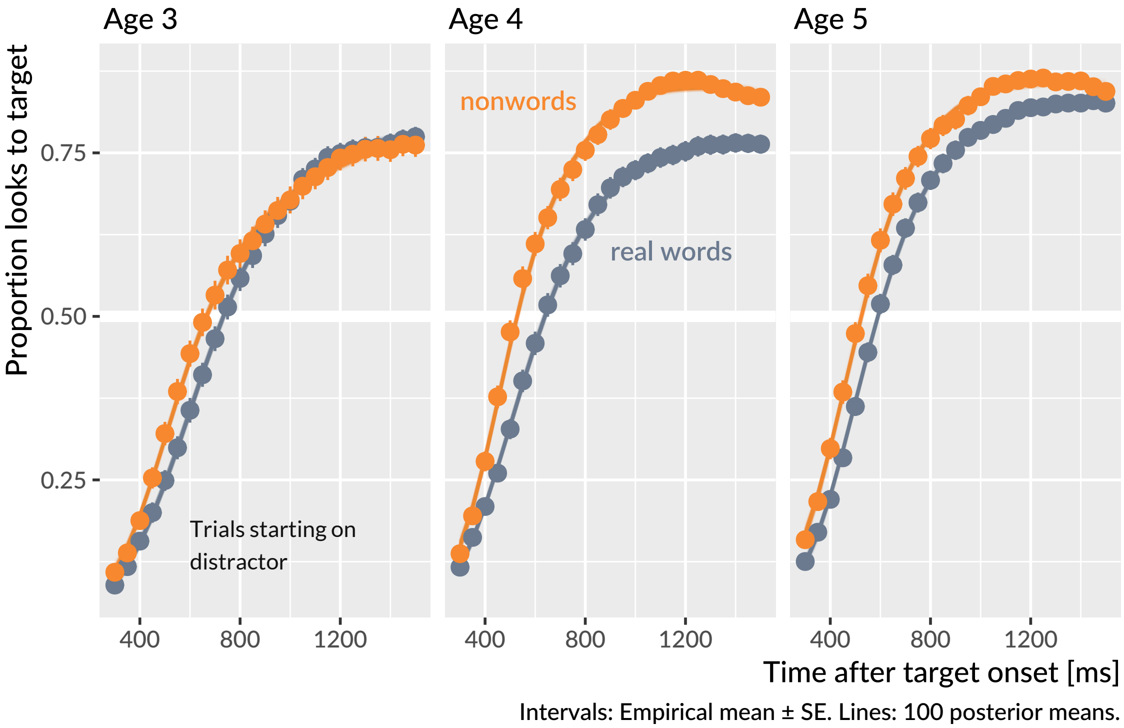


Figure 11.1: Averages of participants' growth curves in each condition and age. The lines represent 100 posterior predictions of the group average.

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 11.1 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.

In Chapter 5, 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 11.1. 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 9.00% [−1.00%, 18.0%] 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 linear time effect where the nonword curve was 13.0% [1.00%, 25.0%] 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 11.2 shows the posterior averages of the

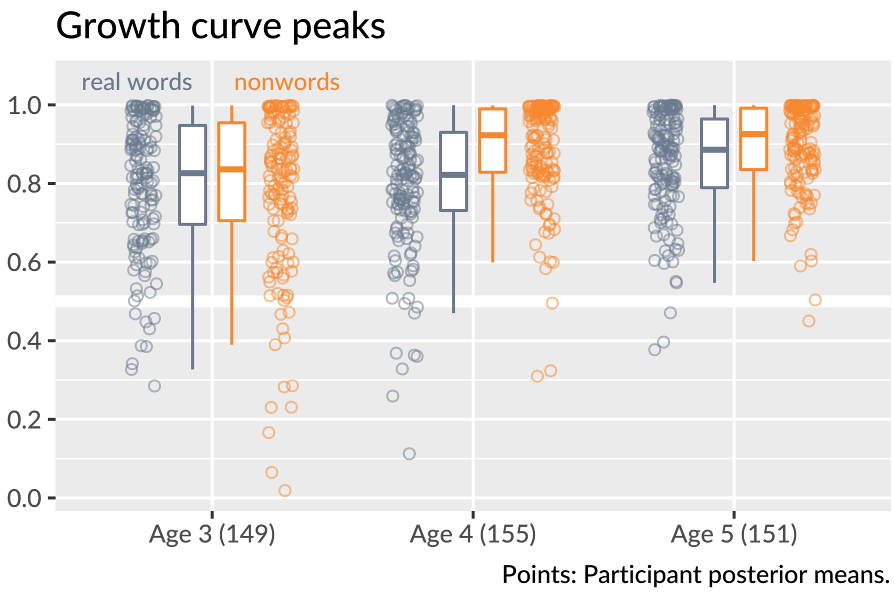


Figure 11.2: Growth curve peaks by child, condition and year of the study. The movement of the medians conveys how the nonword peaks effect increased from age 3 to age 4 and the real word peaks increased from age 4 to age 5. The piling of points near the 1.0 line depicts how children reached ceiling performance on this task.

growth curves peaks for each participant.

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. 1.1.18.1; Bates, Mächler, Bolker, & Walker, 2015). 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 & Hill, 2007 use this approach with mixed models.)

At age 3, the two conditions did not significantly differ, $B_{\text{real}-\text{nonword}} = .01$, $t = 0.95$. 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 11.1. 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 11.3 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.

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.

Table 11.1: Correlations between curve peaks and vocabulary measures. Vocbulary measures are standard scores for receptive vocabulary (PPVT-4) and expressive vocabulary (EVT-2). Significance conventions: $p \leqslant .05$*, $p \leqslant .01$**, $p \leqslant .001$***.

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

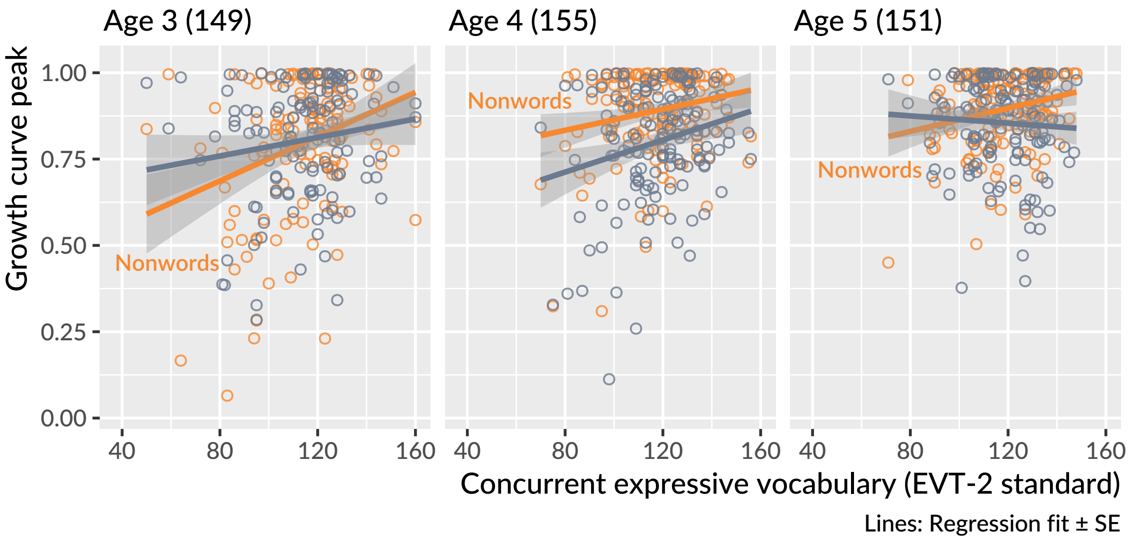


Figure 11.3: Relationships between expressive vocabulary and growth curve peaks at each age.

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.

11.2 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, 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, $p = .005$, $\Delta R^2 = .03$, 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 11.4 depicts the difference between the two conditions with a flat line for the real condition and small slope for the nonword condition.

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

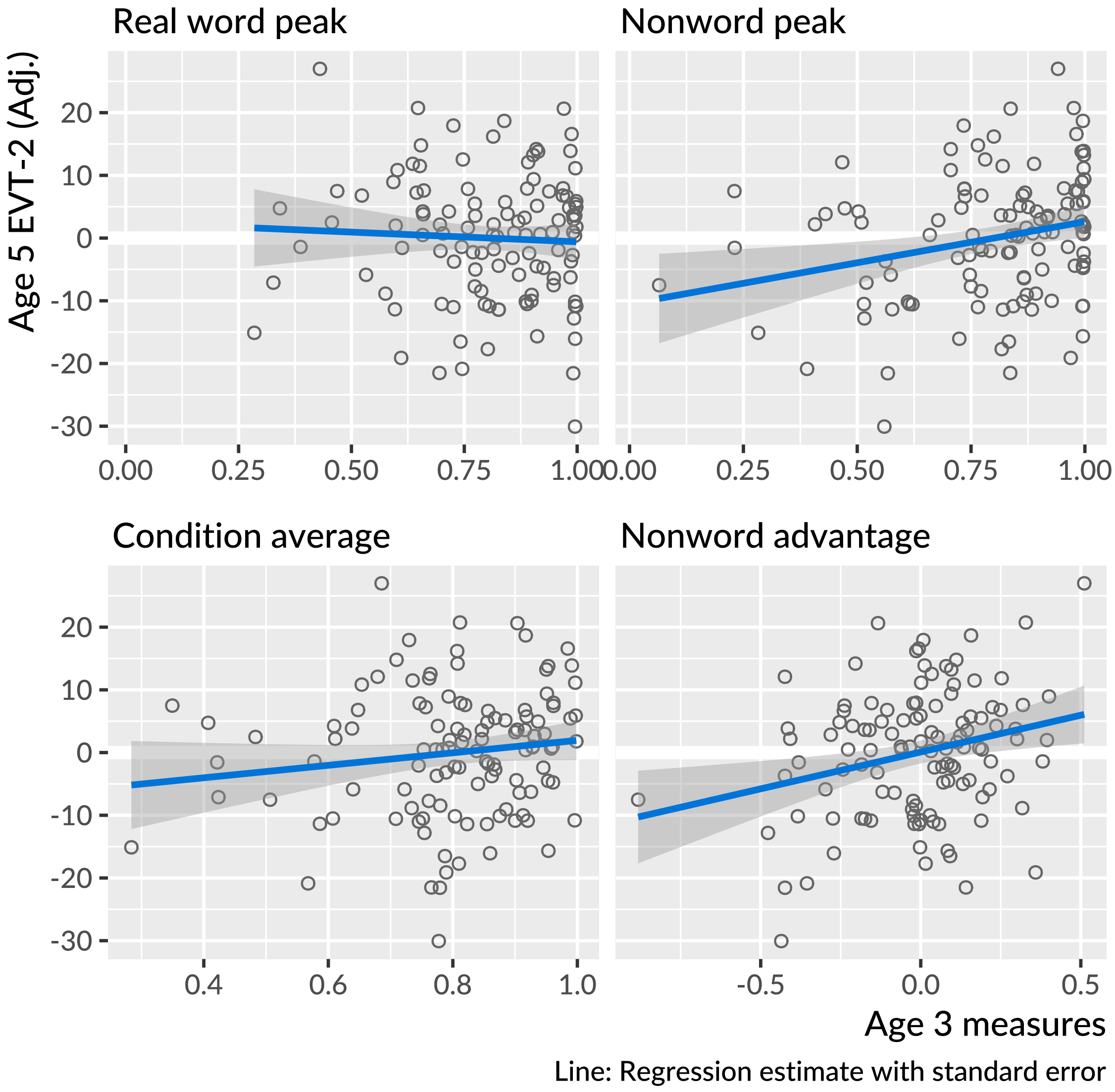


Figure 11.4: Marginal effects of age-3 referent selection measures on age-5 expressive vocabulary standard scores. The vocabulary scores were adjusted (residualized) to control for age-3 vocabulary, so these regression lines show the effects of the predictors over and above age-3 vocabulary.

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.

11.3 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 presence 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) and Law and Edwards (2015). 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 predictive effect 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 have an effect over and above the difference of the peaks. This result was unexpected, given how lexical processing can predict future language outcomes as in Chapter 5. On the other hand, familiar word recognition with a familiar object and novel object is probably not demanding enough for individual differences to predict future vocabulary size

For these two conditions, I hypothesized that 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 therefore an 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; Mather & Plunkett, 2012). An additional factor may be the presence of the mispronunciation trials—reported in Chapter 12. 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 may cause children to downweight the syllable-initial sounds. Such an adaptation would lead to lower overall activation of the real words. This possibility is a limitation of this study: A design with just real words and nonwords would provide a better comparison of the two kinds of words. Alternatively, the novelty bias could interfere with processing of familiar words. For some trials, children could have ignored the familiar word and focused attention on the novel object due to a novelty bias.

12 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. The linear model was therefore:

$$\begin{aligned}
\log \text{odds}(\text{looking}) = {} & \beta_0 + \beta_1 \text{Time}^1 + \beta_2 \text{Time}^2 + \beta_3 \text{Time}^3 + && \text{[age 3 growth curve]} \\
& (\gamma_0 + \gamma_1 \text{Time}^1 + \gamma_2 \text{Time}^2 + \gamma_3 \text{Time}^3) * \text{Age}\,4 + && \text{[age 4 adjustments]} \\
& (\delta_0 + \delta_1 \text{Time}^1 + \delta_2 \text{Time}^2 + \delta_3 \text{Time}^3) * \text{Age}\,5 && \text{[age 5 adjustments]}
\end{aligned}$$

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 D 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.

12.1 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 12.1 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.

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 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 12.2 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

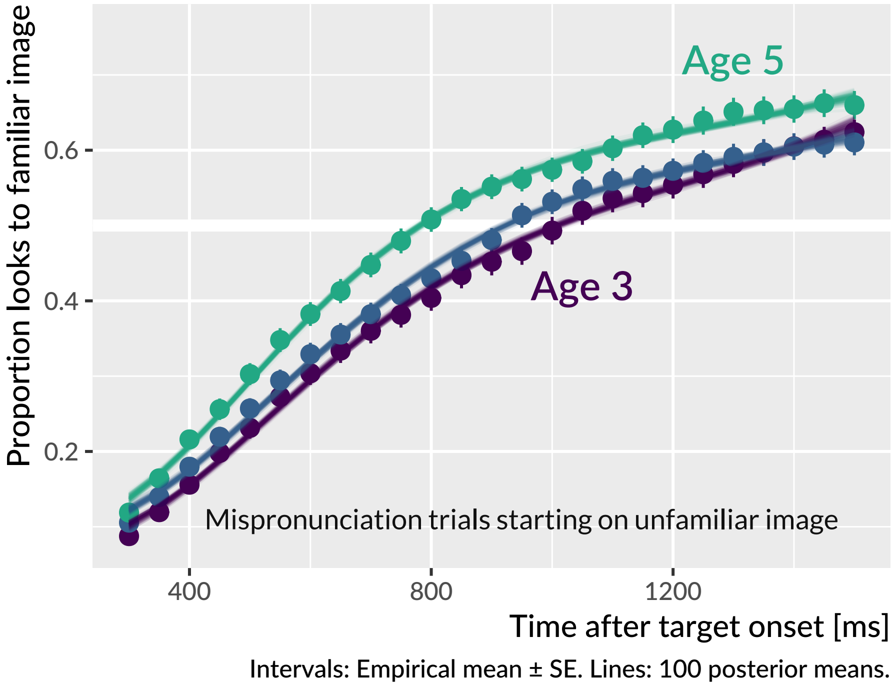


Figure 12.1: Average looks to the familiar image for mispronunciation trials starting on the unfamiliar image at each age. Lines represent 100 posterior predictions of the group average (the average of participants' individually predicted growth curves).

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 12.2 also indicates how most of the children at each age favored the familiar object over the unfamiliar 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 12.1 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

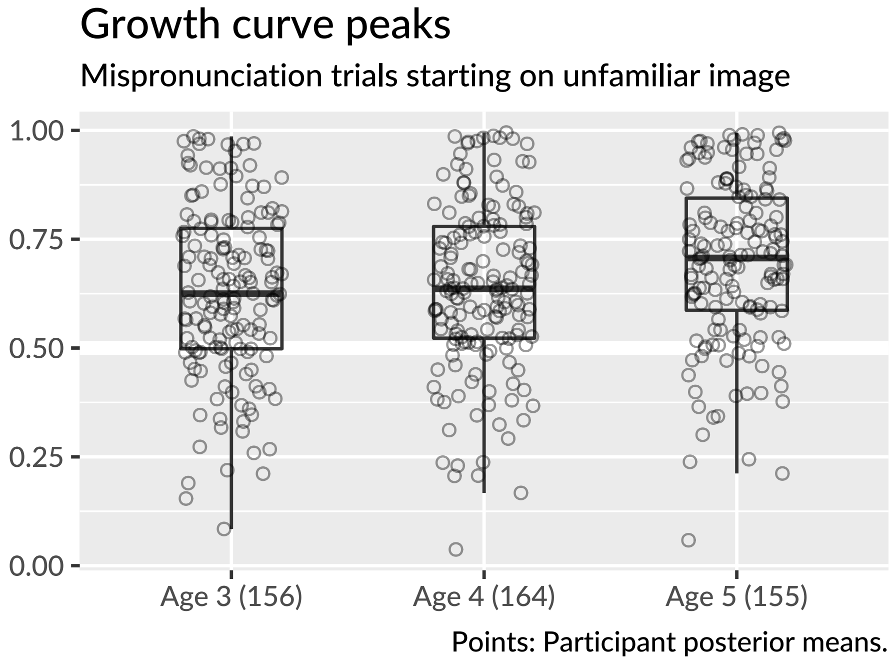


Figure 12.2: Growth curve peaks by age for mispronunciation trials starting on the unfamiliar image.

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 hig-

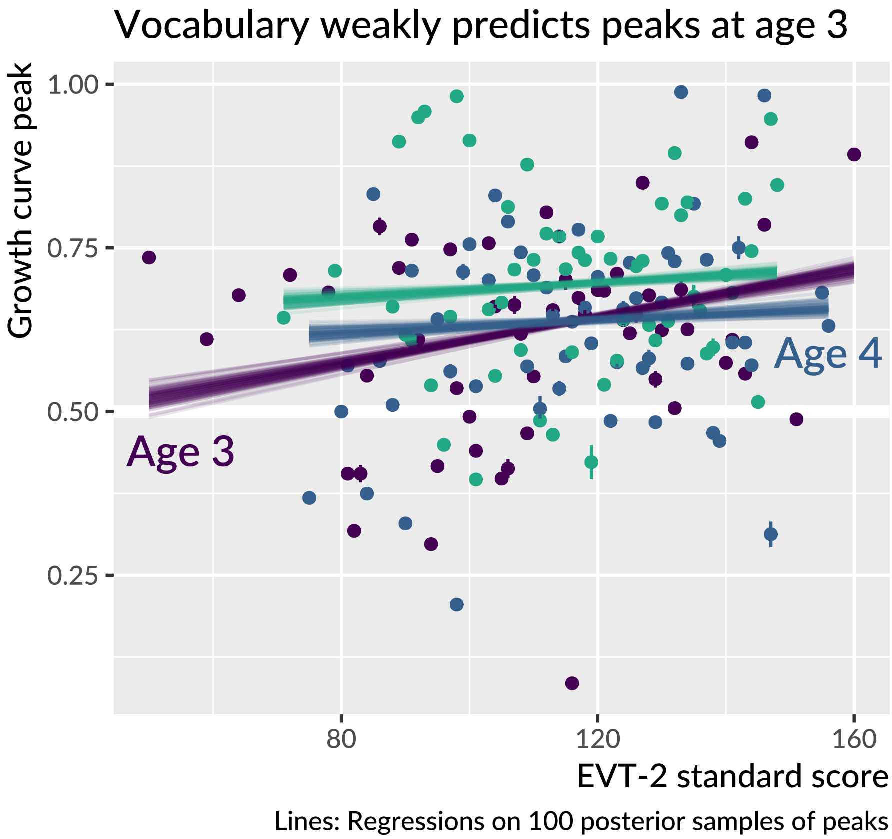


Figure 12.3: Relationship between expressive vocabulary and growth curve peaks for mispronunciation trials starting on the unfamiliar image. I took 100 draws from the posterior distribution and computed participants' growth curve peaks for each draw. Points represent the mean and standard error of 100 peaks. Lines represent regressions fit for each draw.

her average probabilities. These effects, however, were very small. As shown in Figure 12.3, 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.

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.

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

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 unfamiliar 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 12.4 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.

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 12.5 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,

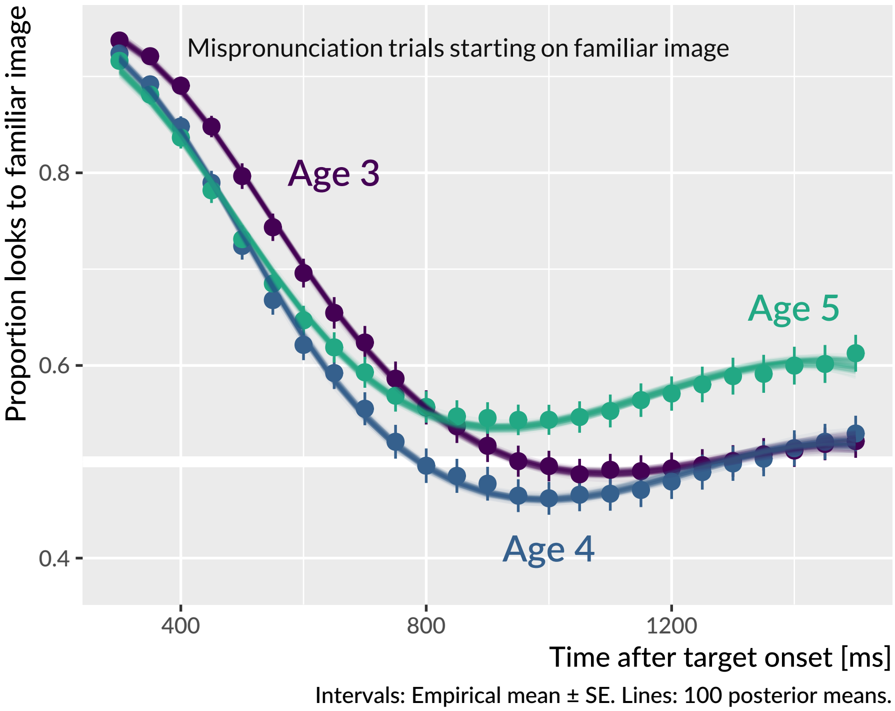


Figure 12.4: Average looks to familiar image for mispronunciation trials starting on the familiar image at each age. Lines represent 100 posterior predictions of the group average (the average of participants' individual growth curves).

.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.

For the linear time feature, the slope becomes flatter year over year, decreasing by 19.0% [7.00%, 29.0%] from age 3 to age 4 and decreasing by 30.0% [17.0%, 42.0%] 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 12.5, 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

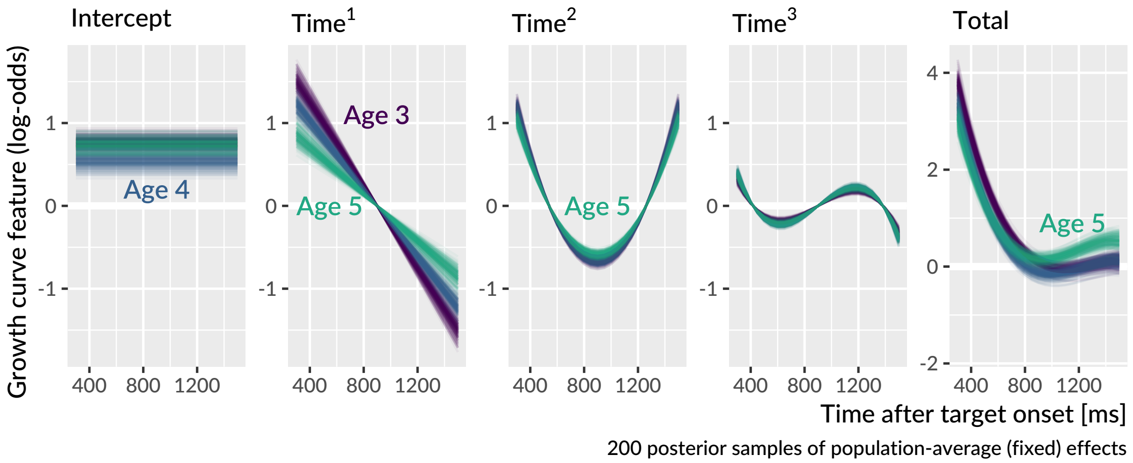


Figure 12.5: Weighted growth curve features. For the first four panels, the y axes are scaled to the same range. This scaling highlights how the cubic time component contributes less to the overall shape than the other features.

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 12.0% [1.00%, 22.0%] 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 contributes only a small amount to the overall shape of the curves.

The combination of these effects shows in the final panel of Figure 12.5. 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 interpretation 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 to stay looking at the familiar object than look at both images equally.

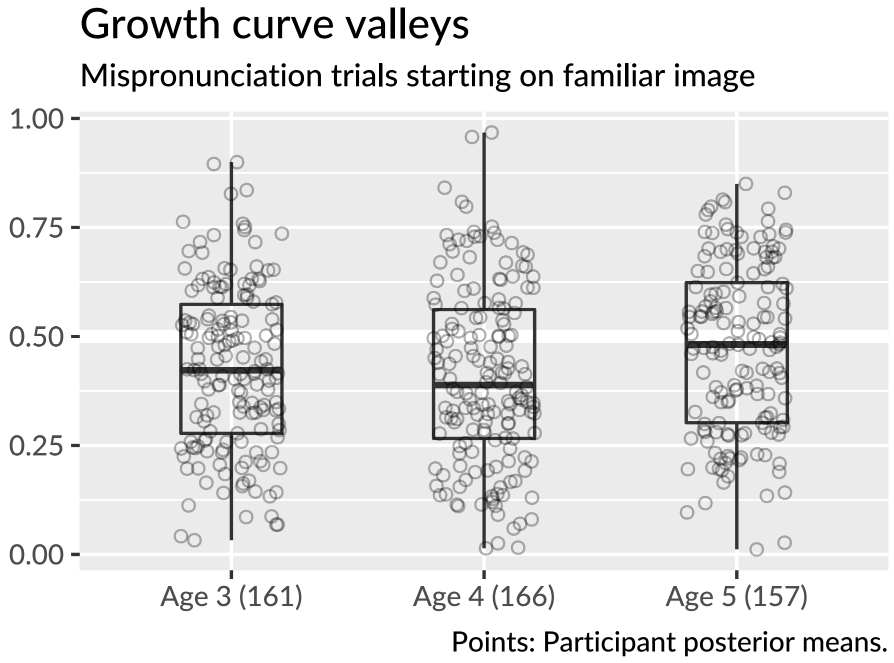


Figure 12.6: Growth curve valleys by age for mispronunciation trials starting on the familiar image.

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 12.6 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 did not dip as low as the other curves.

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 their valleys and plotted their individual growth curves. Figure 12.7 shows the results from age 3. The final two bins show children who stayed with the familiar image throughout the mispronunciation trials. The first two bins mostly contain children who switched to the unfamiliar 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 highlight 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.

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 negligibly small, as shown in Figure 12.8.

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 unfamiliar image. For an average participant, a 15-point increase in expressive vocabulary predicted a 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 simultaneously 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

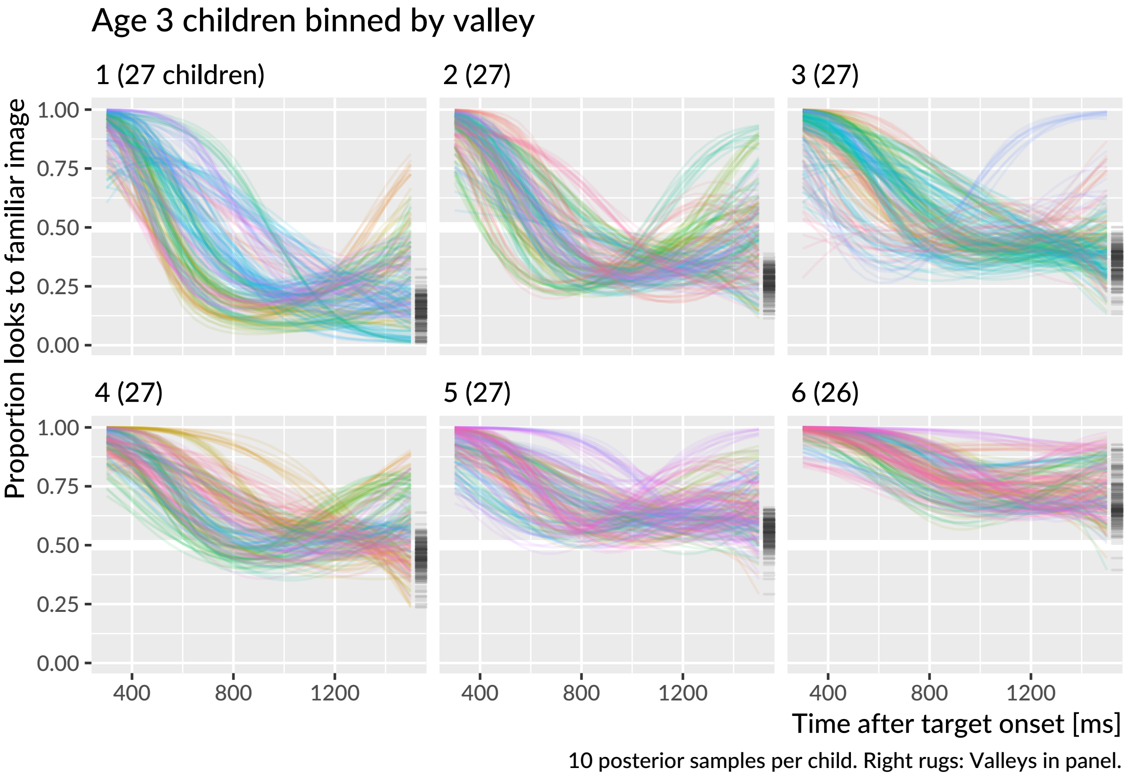


Figure 12.7: Growth curves for mispronunciation trials starting on the familiar image at age 3. Children were grouped into sextiles based on the posterior mean of their growth curve valleys—that is, the lowest point on the growth curve. Ten lines are drawn per child to visualize uncertainty. Children were assigned colors arbitrarily. On the right side of each panel are "rugs" which mark the valleys in that panel.

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 also interacted so that a large change 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 12.7 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

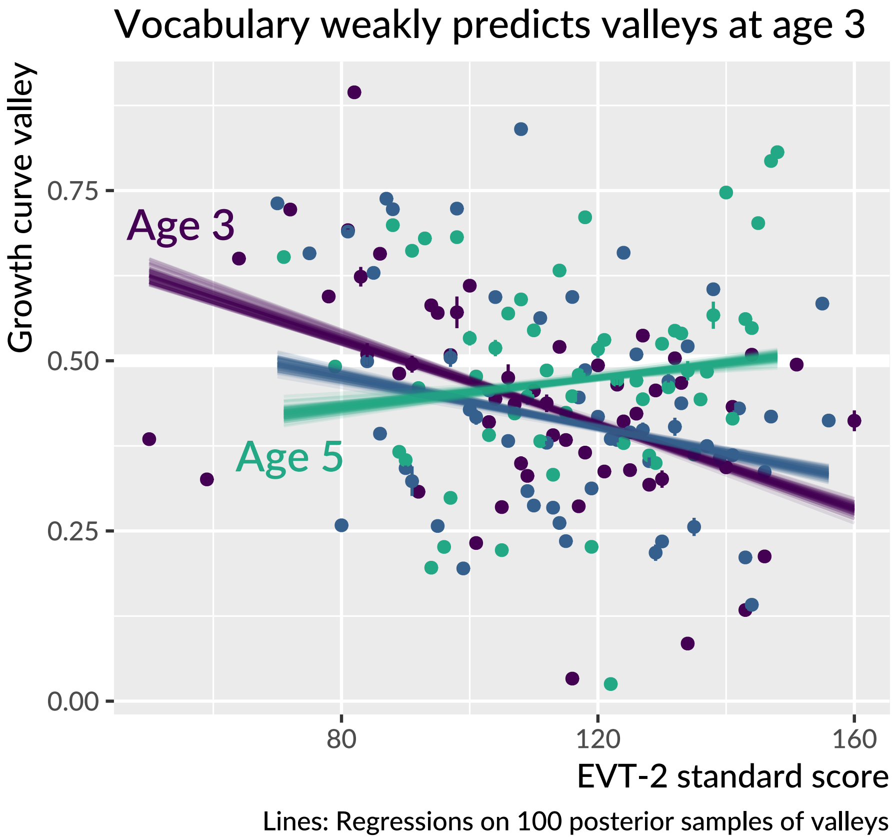


Figure 12.8: Relationship between expressive vocabulary and growth curve valleys for mispronunciation trials starting on the familiar image.

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 12.9 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.

I regressed the u-shaped curve feature onto expressive vocabulary standard score

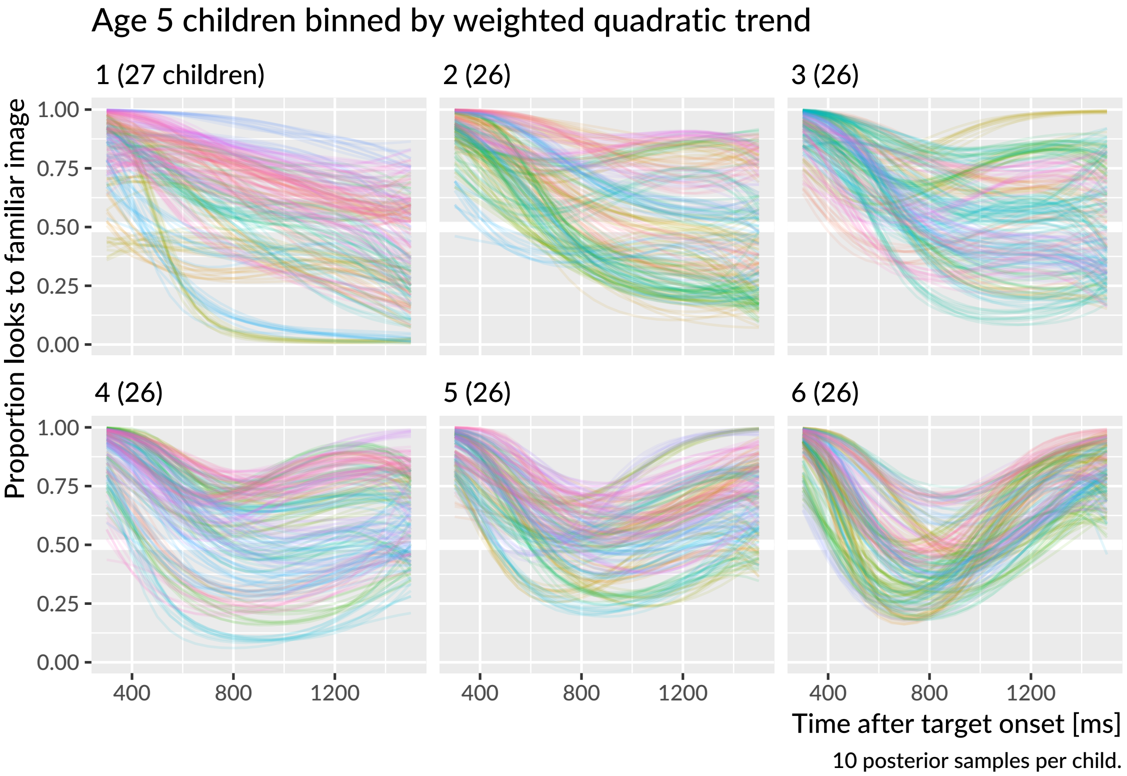


Figure 12.9: Growth curves for mispronunciation trials starting on the familiar image at age 5. Children were grouped into sextiles based on the posterior mean of their curves quadratic trend weighted by the height of the curve in the final time bins. Ten lines are drawn per child to visualize uncertainty. Children were assigned colors arbitrarily.

at each age and onto minimal pair discrimination for age 3. There was a tiny yet significant effect 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.

Table 12.1: Results for the item retention tests.

Word Group	Type	Item	Trials Correct	Percent Correct ± SE
cake	mispronunciation	gake	61 / 107	57.0% ± 4.8
	nonword	pumm	56 / 107	52.0% ± 4.8
duck	mispronunciation	guck	71 / 107	66.0% ± 4.6
	nonword	shann	97 / 107	91.0% ± 2.8
girl	mispronunciation	dirl	71 / 107	66.0% ± 4.6
	nonword	naydge	98 / 107	92.0% ± 2.7
rice	mispronunciation	wice	79 / 107	74.0% ± 4.2
	nonword	bape	80 / 107	75.0% ± 4.2
shoes	mispronunciation	suze	60 / 107	56.0% ± 4.8
	nonword	geeve	90 / 107	84.0% ± 3.5
soup	mispronunciation	shoup	63 / 107	59.0% ± 4.8
	nonword	cheem	93 / 107	87.0% ± 3.3
(all)	mispronunciation		405 / 642	63.0% ± 1.9
	nonword		514 / 642	80.0% ± 1.6

12.3 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 12.1 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 the pairs and performed about equally well on the remaining two (*gake–pumm*, *wice–bape*). I performed an item-response analysis using a mixed-effects logistic regression model. Appendix D 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 12.10 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 mispronunciation 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.

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

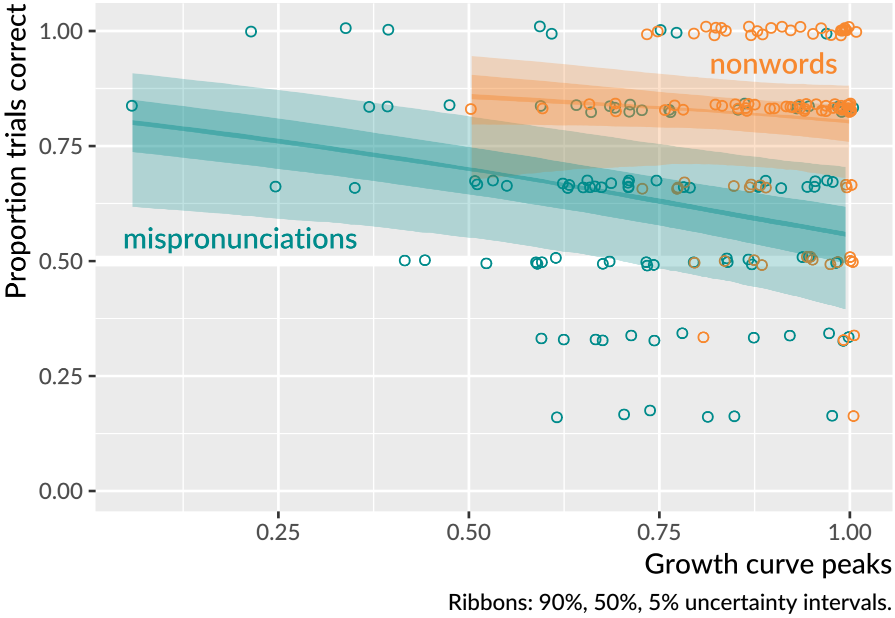


Figure 12.10: Effect of growth curve peaks on children's accuracy on retention trials. For the mispronunciations, I used the peak looks to the familiar image on trials where the child started on the unfamiliar image, so it represents, say, how much a child looked at *shoes* given "suze". Thus, more permissive listeners looked performed more poorly on the retention trials. For the nonwords, I used the peak look to the unfamiliar image on trials where the child started on the familiar image. Points were jittered by 1% to avoid overplotting. There were six trials per condition which is why the points fall into 6 bands.

12.4 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 12.11—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.

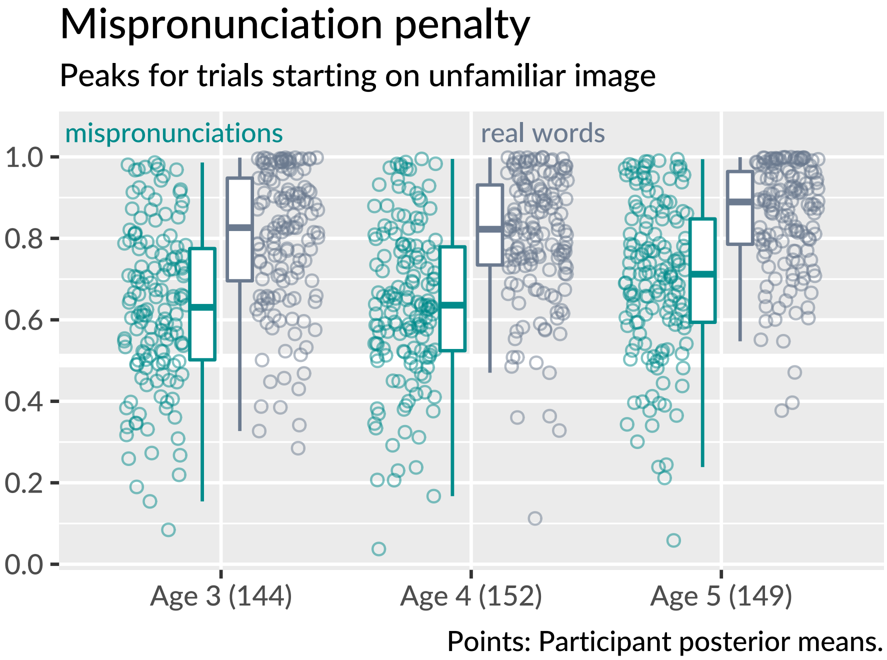


Figure 12.11: Comparison of growth curve peaks for the real words and mispronunciations.

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. Although 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 demonstrate 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 behavior. The u-shaped curve features were not reliably related to any child-level factors, outside of negligibly 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, notices that they have already found the image, and then looks to 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 10.4 and Figure 10.5 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 who fixated on the familiar image might have internally named the object and 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: Children who looked less to the unfamiliar object on mispronunciation trials were less likely to recall that unfamiliar object during retention testing.

13 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.

13.1 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"-initial 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 explanation is that this is a two-image task. Children have ample time to view each image before the noun onset and build up 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 Study 1, which do not provide any evidence for changes in 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.

13.2 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). 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) used the TRACE model of word recognition (McClelland & Elman, 1986) to simulate these kinds of situations. In one set of simulations, the novel object receives 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 an alternative set of simulations, the novel word is added as to the lexicon as a low-frequency word, and the absence 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 processing.

The results here probably support 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. Thus, there is a novelty preference that for these trials gets stronger with development. Plus, children are also learning during this task: At age 5, children were on average were able to recall the unfamiliar image paired with 5/6 nonwords. This learning is consistent with the strategy of simulating a nonword as a low-frequency lexical item.

The findings of Swingley and Aslin (2007) 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 fast-referent selection involves a probability calculation in which children "evaluat[e] the likelihood that an utterance conveys a new word". In this experiment (that is, Study 2), we presented an image of the familiar (mispronounced) object during the mispronunciation trials, which further reduces the likelihood that the mispronunciation indeed reflects a new, as-yet unlearned word. Thus, the children in the retention task averaged around 3–4 mispronunciations correct because those words had a low likelihood of conveying a new word. In fact, the more children discounted that probability—that is, the more they looked to the familiar word on the mispronunciation trials—the less likely they were to retain the mispronunciation items.

The interference from the familiar word on encoding and retaining the mispronunciation and image pairing is beneficial for word learning. Suppose that wordform-object associations are encoded in proportion to the activation of the wordforms and objects. For the unambiguous nonwords, the strong activations allow the associations to be built up quickly, whereas for the mispronunciations, most of activation is spent on the familiar word, meaning that the association is built up more slowly. The use of the word *likelihood* in Swingley and Aslin (2007) also suggests an analogous framing using Bayesian terminology: Known words are like priors which influence (regularize) how unfamiliar words are interpreted, so it takes more exposure to overcome these priors and encode a mispronunciation as a new word. In both framings (lexical activation or Bayesian inference), children showed less retention of the mispronunciations because the words the children know made the mispronunciations ambiguous and that ambiguity made each instance of the mispronunciations less informative. The slowed encoding is a good thing, because spurious associations from mispronunciations should *not* be learned. McMurray et al. (2012) argues the same point: "Slow [word] learning may be more optimal in that it prevents children from committing too strongly to a single (perhaps erroneous) mapping before they have enough data" [p. 870].

13.3 Limitations and implications

The primary limitation for this study is that it applied a procedure designed for toddlers (White & Morgan, 2008) on preschoolers. That is to say, the two-image task was too easy for there to be large year-over-year developmental changes in children's performance. Children were successful at recognizing real words and fast-selecting referents for nonwords at age 3, and by age 4, a quarter of the children performed at ceiling on the nonword condition. The most difficult condition, based on the absence of ceiling effects, was the mispronunciation condition in which children showed much more uncertainty on how to process these words. For this condition, a developmental trend was also observed where preschoolers at age 5 had a larger preference for the familiar object on the trials.

Another limitation here is that the mispronunciations are a particular kind of mispronunciation: One-feature onset mismatches. That initial segment sets the stage for the processing of a word as it activates a word's onset neighbors. Just as it takes longer for a rime to influence processing of word—as evidence needs to pile up from multiple compatible sound in order to overcome an initial mismatch—it should also take longer for a child to recover from an onset-mispronunciation. Conversely, we might also expect vowel or rime mispronunciation to be less disruptive for word recognition because of the useful information on the starting segment. Indeed, Swingley (2009) demonstrates that this is probably the case. That study tested onset and coda mispronunciations in an eyetracking task that used two familiar objects. Both adults and toddlers responded to coda mispronunciations (e.g., "dut" for *duck*) by showing delayed looks *away* from the (mispronounced) target image.

As I note in Appendix E, the small repertoire of mispronunciations is another limitation for this study. It is conceivable that specific mispronunciations change in severity with development, even though the canonical form of the word is a familiar and well known word. In this study, for example, the distance between the *girl* and *dirl* increased each year, even though *girl* was well known to three-year-olds, but this distance was driven by children becoming faster and more efficient at recognizing *girl*. Put another way, a mispronunciation penalty also reflects children's knowledge of the canonical word. In this data, the age-5 *rice* receives as many looks as the age-4 *dirl*. Do children know *rice* less or accept *dirl* more? With so few items, it is unclear whether these differences are accidental or systematic.

One implication for this research is that children's recognition of familiar words and referent selection for nonwords improved a modest amount over the preschool years. Although these tasks can be done by toddlers, full mastery does not begin to emerge until age 4. This finding agrees with the main finding from my analysis of familiar word recognition: Although children can ostensibly know a word very well, their recognition and processing of that word can still improve during preschool.

The finding that mispronunciations were harder to retain than nonwords also has implications for teaching or intervention. Namely, teaching words that are confusable with known or relevant words is more difficult because a child’s known words can influence whether the child accepts a taught word as a novel word.

14 Hypothesis check

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 a 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 a 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 predictor 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. The size of this effect was rather small, however.

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 E. 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.

Overall discussion

15 General discussion of both studies

In this chapter, I integrate results from the two studies. First, I describe the mechanisms underlying children's word recognition. I then briefly discuss some clinical implications of this research, and I outline the main contributions of the research.

15.1 Mechanisms of word recognition

What cognitive or word-recognition mechanisms can explain the data observed from these two studies? These are the essential findings that the model of word recognition needs to account for:

- Developmental improvements in familiar word recognition
- Early advantage of phonologically similar words (over unrelated words)
- Late advantage of semantically similar words (over unrelated words)
- Developmental changes in the advantage of these similar words
- Disrupted processing of onset-mispronunciations
- Effortless processing of unambiguous nonwords
- Individual differences in familiar word recognition

As a baseline for word recognition mechanisms, I will start with 1) a continuous activation model 2) that uses different levels of representation—in other words, TRACE (McClelland & Elman, 1986). In my preceding interpretations of the data from Study 1 and Study 2, I assumed a TRACE-like architecture, so it is helpful to briefly review what this model does.

TRACE interprets an input pattern by spreading activation (energy) through a network of processing units. The pattern of activation over the network is its interpretation of the input signal, so that more active units represent more likely interpretations. Over many processing cycles, the network propagates energy among its connections until it settles into a stable pattern of activation. (Activation also decays over cycles so that the model can start from and return to a resting state.) This activation process is *continuous*; the model's interpretation evolves continuously. We can ask at any point during (or after) presentation of a word what the model's interpretation of that word is. Thus, the listener does not need to hear a whole word to generate a plausible guess for that word (e.g., Fernald et al., 2001).

The model involves three levels of representation: perceptual/phonetic features, phoneme units and lexical units. The input for TRACE is a mock-speech signal that activates the perceptual feature-detectors. These units respond to phonetic features like voicing or vocalic resonance. The perceptual units activate phoneme units, and the phoneme units activate lexical word units. For example, the bundle of features representing /b/ would activate /b/ but to a lesser extent also activate the phonetically similar /d/ (different place), /p/ (voice), /v/ (manner), or /m/ (nasality). The initial /b/ sound activates a neighborhood of words containing /b/, and the phonetically similar phonemes like /d/ or /p/ also activate compatible similar words, albeit to a weaker extent.

The combination of continuous processing and these levels of representation means that ambiguities can arise during word recognition. Suppose that after /b/, the sound /i/ arrives, activating a set of phoneme units and in turn activating words containing /i/. The sequence of /bi/ favors a particular neighborhood of cohorts: *be*, *bee*, *beam*, *beak*, *beat*, *beetle*, etc. At this point, however, the signal is ambiguous. Any of the words in the cohort are plausible interpretations, and more information is needed to refine the interpretation. In Swingley et al. (1999), 24-month-olds were slower to respond to trials of *doggie* versus *doll*, compared to *doggie–tree* or *doll–truck* trials, where the delay reflected the momentary ambiguity from the words sharing an onset consonant and vowel.

The mechanisms described thus far can account for the advantage of the phonological competitors over unrelated words from the first study. The initial phoneme in a word activates a cohort of words that share that sound, so the cohorts briefly represent more plausible interpretations of the target than words that are not phonologically related. A child acts on that early information and shifts their gaze to the phonological competitor.

Words in TRACE compete with each other through lateral inhibition, so that an active word will dampen the activation of other competitors. Inhibition allows the model to reinforce or revise an interpretation. In the earlier example, the arrival of /m/ after /bi/ would strongly favor *beam* as the most plausible interpretation of the word, and *beam* will inhibit the other candidates like *beak* or *beat* so that it can be the decisive interpretation of the word. The transient effect of the phonological competitor suggests lateral inhibition: The advantage of the phonological competitor over the unrelated word is short lived because the target word builds up activation and inhibits the phonological competitor.

To account for the effect of the semantic competitor, we need to make a few more assumptions. Semantic information is not explicitly included as a part of TRACE, but we can stipulate that semantic information is part of a word's lexical representation. We also need a way for semantically related words to coactivate, so that hearing *bee* will generate some spurious looks to *fly*. In this case, we can assume that there are excitatory connections between semantically related words so that hearing a word also activates its semantic relatives. In my earlier discussions, I used the term *cascading activation* to describe this arrangement. For children to generate looks to the semantic competitor, they first need to build up activation of the word and that activation would cascade over to semantic relatives. The time course of cascading activation here is consistent with the late effects of the semantic competitor. The semantic competitor exerts an advantage over the unrelated word *after* semantic information comes online.

The relative advantage of the phonological and semantic competitors increased each year, as did children's overall recognition of the familiar word. In other words, children became better at activating the target *and* the words related to the target. In Chapter 7, I argue that these developmental changes in the first study reflected stronger bottom-up phoneme–word connections (for greater activation of the target and phonological competitors) and stronger semantic connections between words. Alternatively, one might assume that phonologically similar words coactivate in a similar way as the semantically related words activate each other. The problem with this interpretation is that it would not resolve lexical ambiguity to have similar sounding words supporting each other. The phonological similarity between words lives not in the connections between them but in the phonemes that the words share and that mutually activate them. The phonologically related words compete with each other, and they may inhibit each other so that the most plausible interpretation can quickly suppress competing interpretations. For these data, I did not observe any developmental changes in inhibition, so I favored an interpretation that focused on stronger bottom-up connections. (I discuss inhibition more below when I discuss open questions.)

One prediction of TRACE is that rhymes and rimes (one-syllable rhymes) can affect word recognition. But these rhymes are at a disadvantage. Early in the processing of a word, all the action is in the bottom-up connections from the phonetic features to the phonological units onto the words. Cohorts show an early advantage in word recognition because they receive activation before lexical units start to inhibit each other. A rhyme mismatches the input from the start of the word, so it undergoes inhibition early on. But as the word unfolds, subsequent phonemes can build up activation of the rhyme word, and the word can overcome the initial disadvantage. Allopenna et al. (1998) found a strong similarity between TRACE's activation patterns and adult listeners' looking patterns. Namely, adults can hear *beaker* and look to the word, but they also might generate spurious early looks to a cohort (*beetle*) and late looks to a rhyme (*speaker*). (Anecdotally, my name is Tristan, but in grade school, I always snapped to attention whenever Kristen's name was called.)

The mechanisms that predict how rhymes can engage in lexical competition also explain the disruptive mispronunciation effects observed in Study 2. The initial /s/ in *suze* sends the listener down a lexical garden path, activating /s/-initial words. The arrival of the rest of the word—plus the presentation of an image shoes onscreen—supports *shoes* as an interpretation of the word. But there is much less certainty in this situation. At age 3, I observed 80% looking to the image of the shoes for the real word *shoes* compared to 50% looking (to the shoes) for *suze*.[1] There was a small developmental improvement for the mispronunciation and real word conditions. For example, at age 5, *suze* reached 60% looking to the familiar image and *shoes* reached 87% looking. Developmentally, children became more likely to activate the familiar word when given a mispronunciation, and this change likely reflects general improvements in activation efficiency. Gains in activation efficiency are consistent with the results for familiar word recognition in Study 1 where children showed increases in overall looking probability and in how quickly looking probabilities changed during

[1] It should be noted that these mispronunciations were all one-syllable words, so they did not have much phonological substance that could overlap with the target. If the mispronunciation-target pairs were longer, as in a *beaker–speaker* rhyme, more segments would overlap, leading to greater activation of the mispronounced target.

a trial.

What about the effortless processing of the unambiguous nonwords? Surely, children do not have a lexical item *geeve* to activate the first time they hear the word. On these trials, however, the children did know *sock* and know that *geeve* was not the name for the sock, so they looked to the trolley instead. For McMurray et al. (2012), the problem facing a child is reference selection: Children have to select a visual referent for a spoken word. In their model, all words can refer to all visual referents initially, so the model has to prune away unnecessary connections to build up selective word recognition. Development of the *sock*-sock pairing pruned away other visual referents or words from activating *sock*. Thus, *geeve* is not likely to activate *sock* but the viability of a *geeve*-trolley pairing allows the child to select the correct referent for the nonword. In TRACE simulations, Mayor and Plunkett (2014) handled this situation by treating the nonword as a low-frequency word. In both situations, a novel nonword is recognized despite not being well known to the child. This recognition is possible because the new word is not affected by lexical competition from any other plausible alternatives.

This framework also allows us to account for the differences in retention for the nonwords and mispronunciations at age 5. In McMurray et al. (2012), learning was associative. The model developed connections between spoken words, lexical items, and visual referents when spoken words and visual referents occurred together, and each co-occurrence built up the connections. On the mispronunciation trials from Study 2, a child heard a mispronunciation of a familiar word and also saw an image of the familiar (mispronounced) word. On average, they tended to interpret the mispronunciation as the familiar word. Thus, the familiar word competed with the mispronunciation, leading the child to develop a weaker association between the novel object and the mispronunciation. The effect of looking behavior, where children who looked more to the familiar image on mispronunciation trials showed poorer retention, helps explain how the familiar image could impede the association of the mispronunciation and the novel object. In contrast, for the unambiguous nonword trials, children could associate the novel object and novel word more strongly. This difference in lexical competition manifested in the retention performance where children were better able to retain nonwords than mispronunciations.

So far, I have described a general framework of word recognition, and I claimed children's developmental changes in word recognition reflect more efficient representations and activation pathways. I now describe task differences and individual differences under this framework.

Word learning is a matter of degree. I like to draw a distinction between “shallow” receptive knowledge and “deeper” expressive knowledge, based on the idea that recognition is easier than generation. But we can imagine a finer continuum with degrees of recognition ability. For example, a word can be recognized in one situation but may not be recognized in a more challenging situation. For example, McMurray et al. (2012) tested a word-learning model’s comprehension by simulating alternative-forced choice (AFC) tasks where a named target was displayed and pitted against visual competitors. The model showed graded performance, with better comprehension on 3-AFC (2 competitor) tests than 5-AFC tests, and better performance on 5-AFC tests than 10-AFC tests. Thus, the 4-AFC task in my first study provided a more challenging word-recognition environment than the 2-AFC task in the second study. For example, children demonstrated ceiling performance on the nonword condition at age 4, whereas children had room to develop each year in the 4-AFC task.

Individual differences in word recognition reflect differences in children’s lexicons and their lexical representations. Although all the words on the 4-AFC were familiar to preschoolers, children differed in their peak looking probabilities and rate of fixating on the target. In lexical processing terms, children differed in peak activation and the rate at which activation reached the target word. Differences in word recognition were stable from year to year. Even though all the children became faster, more reliable and more certain during word recognition with age, the children who were faster and more reliable at age 3 were also faster and more reliable at age 5. The children who performed better at age 3 had more familiarity with the words and more reliable representations of them—thus, these children had a head start and they built on top of that advantage as they grew older. This interpretation can also account for how word recognition performance at age 3 correlated with vocabulary scores at later ages.

Open questions about word recognition mechanisms

There are three immediate open questions from this research. First, how does lexical inhibition change over this developmental window? The results from Study 1 show that phonologically and semantically similar words become more relevant during word recognition as children grow older. (The words became more active, compared to the unrelated word.) I did not observe any clear changes in how quickly those words were *rejected* as possible interpretations of the input, and thus, I could not make any claims about the development of inhibition.

In principle, developmental changes could have been observed in this experiment. Lexical inhibition would affect how quickly the phonological competitor's advantage decays as the target word becomes the favored interpretation. A developmental change in lexical inhibition would cause the competitor's activation to decay more quickly (or more slowly) at older ages. But the growth curves observed here were parallel; they decayed at the same rate. Changes in lexical inhibition are detectable by an experiment paradigm like this one (with a target, competitor and an unrelated word), but in the present case, I did not observe these changes. Thus, developmental change in lexical inhibition during the preschool years remains an open question.

Based on other work, I expect older children to show greater inhibition. Rigler et al. (2015) showed that 9-year-old children were more sensitive to phonological cohorts and rimes than 16-year-old listeners, suggesting children need to develop inhibitory connections that suppress the interference from these words. Blomquist and McMurray (2017) used a cross-splicing paradigm to test lexical inhibition in 7–8-year-old versus 12–13-year-old children. In this paradigm, a target like *cap* is created by splicing an initial *ca* onset with a different token (*ca(p)p*), with a cohort competitor (*ca(t)p*) and a nonword (*ca(k)p*), the idea being that the sublexical information in the cohort splice will favor *cat* and therefore inhibit *cap* whereas a nonword splice cannot inhibit *cap*. This manipulation held in both groups, but the older children were more disrupted by the cohort splice. It would be revealing to see both paradigms applied to this age range. For the preschool years, however, the development trajectory seems to be the strengthening of connections so that the phonological competitors can participate in word recognition with later childhood being a time to develop inhibitory connections. In other words, a child has to develop sensitivity to cohorts first in order to demonstrate the ability to quickly inhibit them.

A second open question is when does a nonword engage in lexical competition and interfere with word recognition in children at this age. For adults, nonwords can affect processing very quickly. Kapnoula et al. (2015) used a cross-splicing paradigm with adults and observed that newly learned words compete with familiar ones immediately. Magnuson, Tanenhaus, Aslin, and Dahan (2003) trained adult participants with artificial lexicons and observed that after one day of training, cohort and rimes effects *within the artificial lexicon* were comparable, but after a second day, the cohort showed an early advantage. For preschoolers, I would expect them to show cohort and rime effects with enough training. The prediction is based on the competitor effects observed in the first study where age-5 children showed the early advantage of the phonological competitor over the unrelated word. Sensitivity to lexical inhibition via cross-splicing is an open question for preschoolers in general, even for familiar words. If lexical inhibition develops over later childhood, it is conceivable that preschoolers could show equal sensitivity to cross-splicing from cohorts and nonwords.

A third open question, given the previous discussion of models and mechanisms, is whether a word recognition model like TRACE can replicate the developmental changes observed here. There is no reason to assume that it would not be able to simulate the results from each year, given that it has been used to simulate word-recognition data from adults (Allopenna et al., 1998), adults with aphasia (Mirman, Yee, Blumstein, & Magnuson, 2011), toddlers (Mayor & Plunkett, 2014), and adolescents with specific language impairment (McMurray et al., 2010). These simulations have shed light on their respective listener populations. For the toddler data, Mayor and Plunkett (2014) had to use reduced lexical inhibition parameters in order to replicate graded mispronunciation effects of White and Morgan (2008), suggesting that lexical inhibition is not a crucial feature of toddler word recognition. McMurray et al. (2010) used TRACE simulations with different modeling parameters to test different theories of specific language impairment. Ultimately, they found that lexical decay—"the ability to maintain words in memory" (p. 23)—was the most important model parameter, implying that individual differences in word recognition for listeners with specific language impairment are rooted in lexical processes (and not perceptual or phonological ones).

For the current data, the goal of the simulations would be the developmental story: Which parameters would need to change each year to have the model match the empirical data? I would posit that the changes would involve some manipulation of the rate of lexical activation so that bottom-up information can activate relevant words the more quickly. I would also expect changes in the degree of lexical inhibition to play a role, based on the simulations in McMurray et al. (2010) in which cohort effects increased as lexical inhibition decreased. Even though I did not observe any changes in lexical inhibition in terms of how quickly the competitor advantages decayed in the first study, it is still plausible that lexical inhibition changes are needed to accommodate increased bottom-up activation.

15.2 Clinical implications

The results of Study 1 remind us that words are not simply acquired—they are recognized, learned, and *integrated.* In the first study, children's recognition of highly familiar words improved each year. Children's representations of familiar words will continue to develop, even when they ostensibly *know* the word. One might attribute this development change to improvements in visual processing, sensory processing, or some other nonlinguistic factor. This study cannot rule out those explanations. The increasing effect of the phonological and semantic competitors, however, suggests that changes in lexical representations are needed to explain these results.

Part of the promise of eyetracking-based research is that word recognition can predict later outcomes. One common conclusion in this research is that word recognition may provide an early screening tool: "[t]ime-course measures of comprehension in very young language learners could ultimately prove useful in improving early identification of children at risk for persistent language disorders" (Fernald & Marchman, 2012, p. 219). The developmental results here stress that such a tool has to be developmentally appropriate. Children's processing of real words on the two-image task did not predict language outcomes, but the slightly more challenging nonword condition did yield a small predictive effect. For the more difficult four-image task, individual differences were greatest and most predictive at age 3 and the range of variability decreased with age. Thus, recognition of familiar words is perhaps best understood as a lexical measure that needs to be scaled with children's vocabulary norms.

Finally, the observed difficulty of retaining mispronunciations emphasizes that children will err on the side of known words during nonword referent selection when the known words are plausible interpretations. In particular, it does not seem like a contrastive method of teaching, say, *pear* by having it compete against a known word *bear* would be effective because the known word would interfere with the encoding of the nonword.

15.3 Contributions

The most important contribution of this research is that children became more sensitive to phonological and semantic competitors as they grew older. When they erred, they were more likely to look to a relevant word. This result indicates that children improve in word recognition by being able to activate phonologically plausible words quickly, from partial information, and activate semantically related words on the basis of cascading activation. The developmental trend is that of more *engagement* (Leach & Samuel, 2007), with children developing the connections among related words and harnessing those similarities to their advantage.

Another contribution is the description of individual differences in word recognition: Namely, differences are stable over time but diminish in magnitude, so that early differences are more predictive than later ones. Although there has been ample evidence of how word recognition in toddlers predicted later outcomes, it was not clear whether those differences held over the preschool years. The results here indicate that the differences are task-specific: A more age-appropriate four-image task can better differentiate preschoolers than a simpler two-image one.

A final contribution comes from Study 2. This study was limited by how apparently easy it was for preschoolers. After all, it was a 2-AFC task where one of the images was always a familiar object and the other was an unfamiliar object. That limitation was revealing, showing that this design does not scale up for preschoolers developmentally. Children had mastered mutual-exclusivity-type referent selection on this task by age 4. Children could effortlessly associate nonwords to novel objects, provided that the nonword is not under competition from any known words, as was observed for the mispronunciations.

A Items used in Study 1

Table A.1 lists the items used for the Visual World experiment in Study 1. 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. That is, *pear* had *ring* as its unrelated competitor in one block and *vase* in the other block. This happened due to a design oversight. For the analysis of phonological competitors, I only used trials where the target and the phonological foil shared the same syllable onset (Table A.2). For the analysis of semantic competitors, I only used trials where the target and the semantic foil belonged to the same category (Table A.3).

Table A.1: 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

Table A.2: 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 A.3: 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

B Computational details for Study 1

B.1 Growth curve analyses

These models were fit in R (vers. 3.4.3; R Core Team, 2018) with the RStanARM package (vers. 2.16.3; Gabry & Goodrich, 2018).

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 Time^1 was equal to change from the start to the end of the analysis window. Table B.1 shows the ranges of the time features. It took approximately 24 hours to run the model on four Monte Carlo sampling chains with 1,000 warm-up iterations and 1,000 sampling iterations. Warm-up iterations are discarded, so the model comprises 4,000 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 longitudinal project. Mnemonically, `ot` stands for *orthogonal time* and the number is the degree of the polynomial. This convention is used by Mirman (2014). `Study` refers to the timepoint (year) of the larger longitudinal investigation. Conceptually, I use *study* to mean a data-collection

Table B.1: Ranges of the polynomial time features.

Feature	Min	Max	Range
Time^1	–0.50	0.50	1.00
Time^2	–0.33	0.60	0.93
Time^3	–0.63	0.63	1.26
Trial window (ms)	250	1500	1250

unit, and I think of each wave of testing with their somewhat different tasks and protocols as separate studies. `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.

```r
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),
  control = list(
    adapt_delta = .95,
    max_treedepth = 15),
  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 age. This code 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`. Put another way, `... | ResearchID/Study` expands into `... | ResearchID` and `... | ResearchID:Study`. Thus, for each child, we have general `ResearchID` effects for the intercept, Time^1, Time^2, and Time^3. 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

Time[1] 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, 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).

These priors are very conservative, including information about the size of an effect but not its direction. Gelman and Carlin (2014) describe two types of errors that can arise when estimating an effect or model parameter: Type S errors where the *sign* of the estimated effect is wrong and Type M errors where the magnitude of the estimated effect is wrong. From this perspective, the priors here are uninformative in terms of the sign: Both positive and negative effects are equally likely before seeing the data. Future work on these models should incorporate sign information into the priors: For example, it is a safe bet that the linear time effect will be positive—the curves goes up—so that prior can be adjusted to have a positive, non-zero mean. For this model, I incorporated weak information regarding the magnitude of the effects (an SD of 3 for all effects). On the basis of the estimates here, I employed more

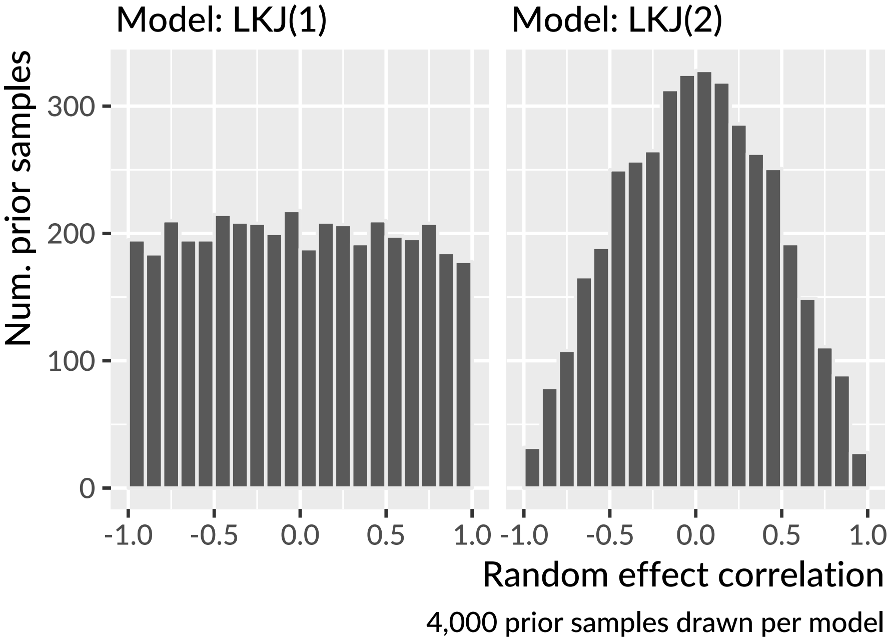


Figure B.1: Samples of correlation effects drawn from LKJ(1) and LKJ(2) priors.

informative priors in Study 2 with an SD of 2 for the linear time effect and an SD of 1 for other effects.

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 B.1 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.

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).
```

Convergence diagnostics for Bayesian models

For the Bayesian models estimated in Study 1 and Study 2, I assessed model convergence by checking software warnings and checking sampling diagnostics. Stan programs emit warnings when the Hamiltonian Monte Carlo sampler runs into problems like divergent transitions or a low Bayesian Fraction of Missing Information statistic. When I encountered these warnings, I handled them by adjusting the sampling controls, as documented in <`http://mc-stan.org/misc/warnings.html`>, or incorporating more information into the model's priors. In the above model, for example, the `adapt_delta` and `max_treedepth` controls were increased to help the model more carefully explore the posterior distribution.

```
rstan::check_hmc_diagnostics(m$stanfit)
#>
#> Divergences:
#> 3 of 4000 iterations ended with a divergence (0.075%).
#> Try increasing 'adapt_delta' to remove the divergences.
#>
#> Tree depth:
#> 0 of 4000 iterations saturated the maximum tree depth of 15.
#>
#> Energy:
#> E-BFMI indicated no pathological behavior.
```

Additionally, I also checked for convergence by using Markov Chain Monte Carlo diagnostics. The models consisted of four sampling chains which explore the posterior distribution from random starting locations. The split $\hat{\mathsf{R}}$ ("*R*-hat") diagnostic checks how well the sampling chains mix together (Gelman et al., 2014, p. 285; Stan Development Team, 2017, p. 371). If the chains are stuck in their own neighborhoods of the parameter space, then the values sampled in each chain will not mix very well. The *split* designation means that each chain is first split in half so that the also diagnostic checks for within-chain mixing. At convergence $\hat{\mathsf{R}}$ equals 1, so a rule of thumb is that $\hat{\mathsf{R}}$ should be less than 1.1 (e.g., Gelman et al., 2014, p. 285). In the bayesplot package (Gabry & Mahr, 2018), we use the convention that values below 1.05 are *good* and values above 1.05 but below than 1.1 are *okay*.

The other diagnostic I monitored was the number of effective samples. If we think of a sampling chain as exploring a parameter space, then the samples form a random walk with each step being a movement from an earlier location. This situation raises the risk of *autocorrelation* where neighboring sampling steps within a chain are correlated with each other. The number of effective samples ("*n* eff.") diagnostic estimates the number of posterior samples, taking into account sampling autocorrelation (Gelman et al., 2014, p. 285; Stan Development Team, 2017, p. 373). The square root of this number is used to calculate Monte Carlo standard error statistics (e.g., Gelman et al., 2014, p. 267), so I think of the number of effective samples as the amount of precision available for a parameter estimate. Interpreting this statistic depends on the quantity being estimated and the amount of precision desired. As a rule of thumb, Gelman et al. (2014) mentions 10 effective samples per chain as a baseline for diagnosing non-convergence: "Having an effective sample size of 10 per sequence should typically correspond to stability of all the simulated sequences. For some purposes, more precision will be desired, and then a higher effective sample size threshold can be used. [p. 287]"

B.2 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) with support from tools in the itsadug R package (vers. 2.3; van Rij et al., 2017).

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.

```r
# Create a Study dummy variabe with Age 4 as the reference level
phon_d$S <- factor(
  phon_d$Study,
  levels = 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.

```r
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) and Baayen, Rij, Cat, and Wood (2016) 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
```

C Items used in Study 2

Table C.1 shows the stimuli design of each year for the mispronunciation experiment in Study 2. The word list changed between Age 3 and Age 4 where *dog/tog* was replaced with *rice/wice*.

Table C.1: Items used for the mispronunciation experiment.

Age	Word Group	Condition	Word	/IPA/	Familiar	Unfamiliar
3	dog	Real word	dog	/dɔɡ/	dog	wombat
		Mispronunciation	tog	/tɔɡ/	dog	wombat
		Nonword	vafe	/vef/	ball	sextant
3, 4, 5	cake	Real word	cake	/kek/	cake	horned melon
		Mispronunciation	gake	/ɡek/	cake	horned melon
		Nonword	pumm	/pʌm/	book	churn
3, 4, 5	duck	Real word	duck	/dʌk/	duck	toy creature
		Mispronunciation	guck	/ɡʌk/	duck	toy creature
		Nonword	shann	/ʃæn/	cup	reed
3, 4, 5	girl	Real word	girl	/ɡɝl/	girl	marmoset
		Mispronunciation	dirl	/dɝl/	girl	marmoset
		Nonword	naydge	/neʤ/	car	work holder
3, 4, 5	shoes	Real word	shoes	/ʃuz/	shoes	flasks
		Mispronunciation	suze	/suz/	shoes	flasks
		Nonword	geeve	/ɡiv/	sock	trolley
3, 4, 5	soup	Real word	soup	/sup/	soup	steamer
		Mispronunciation	shoup	/ʃup/	soup	steamer
		Nonword	cheem	/ʧim/	bed	pastry mixer
4, 5	rice	Real word	rice	/ɹaɪs/	rice	anise
		Mispronunciation	wice	/waɪs/	rice	anise
		Nonword	bape	/bep/	ball	sextant

D Computational details for Study 2

D.1 Real words versus nonwords growth curves

These models were fit in R (vers. 3.5.0; R Core Team, 2018) with the brms package (vers. 2.3.1; Bürkner, 2017).

The orthogonal polynomial features for Time, they were scaled as in Study 1, so that the linear feature ranged from $-.5$ to $.5$. Under this scaling a unit change in Time^1 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`.

```r
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 specified below. The regression effects (`class = "b"`) have a 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). These priors are uninformative in terms of direction–both positive and negative effects are equally likely–but they are informative in terms of magnitude. A weakly informative LKJ(2) prior is put on the random-effect correlations. I review the role of the LKJ prior in Appendix B. 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.

```r
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.

```r
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).
```

D.2 Mispronunciation 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 unfamiliar-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 D.1 illustrates the diffe-

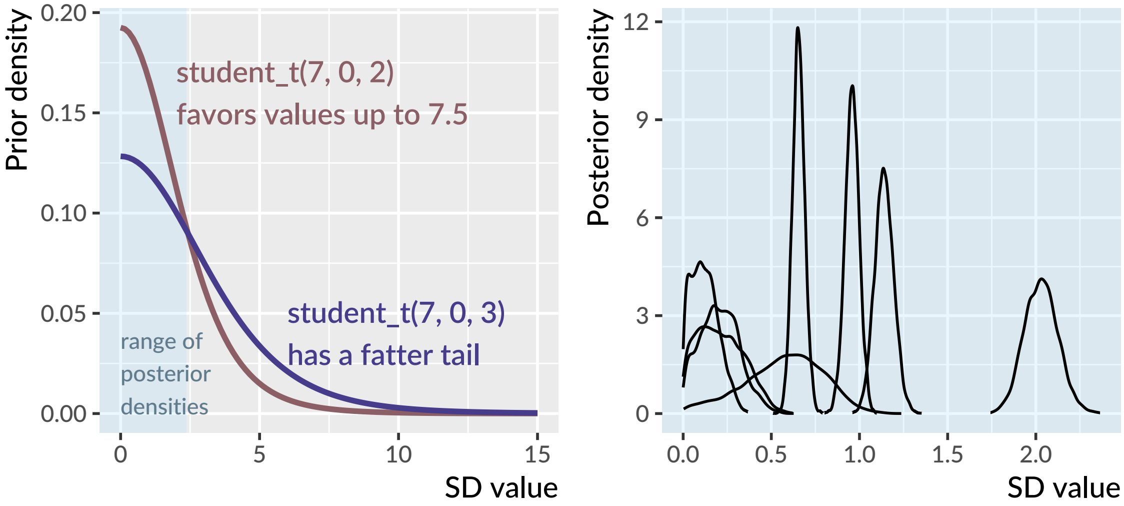


Figure D.1: Prior densities (*left*) versus posterior densities (*right*) for the random-effect standard deviations. I changed the prior to be tighter, so that it favor values up to 7.5. This prior still turned out to be very conservative, given that the posterior samples for these values are all less than 3.

rences 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.

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).
```

D.3 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) with the brms package (vers. 2.3.1; Bürkner, 2017). 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
#> Intercept                     0.57      0.37    -0.03     1.16
#> ItemTypenonword               0.90      0.40     0.22     1.54
#> c_peak_10                    -0.12      0.06    -0.22    -0.02
#> ItemTypenonword:c_peak_10     0.04      0.14    -0.20     0.28
#>                           Eff.Sample Rhat
#> Intercept                       1992 1.00
#> ItemTypenonword                 2180 1.00
#> c_peak_10                       4000 1.00
#> ItemTypenonword:c_peak_10       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).
```

E Effects of specific mispronunciations in Study 2

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 E.1 shows the growth curves for real word and mispronunciations.

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 E.2. 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

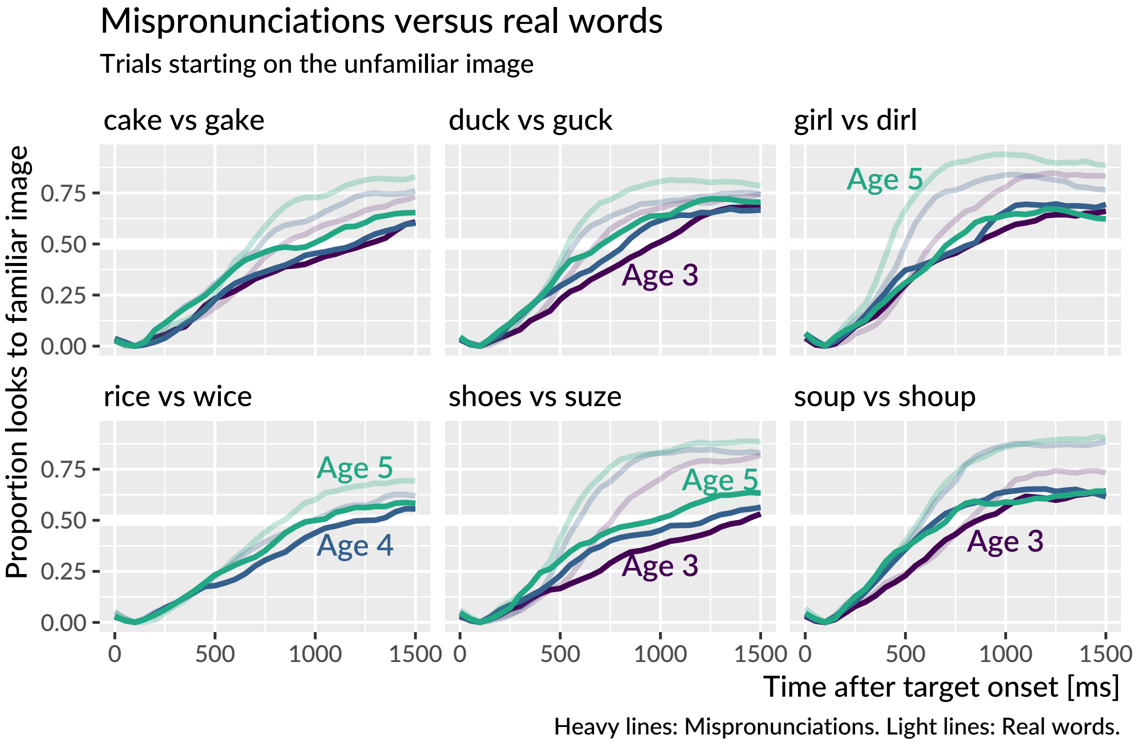


Figure E.1: Average proportion of looks to the familiar object for real words and mispronunciations. A *dog–tog* pair was administered at age 3 but it was replaced by *rice–wice* at age 4.

worse realization of *girl* as children grow older. A similar change happens with *shoup* for *soup*.

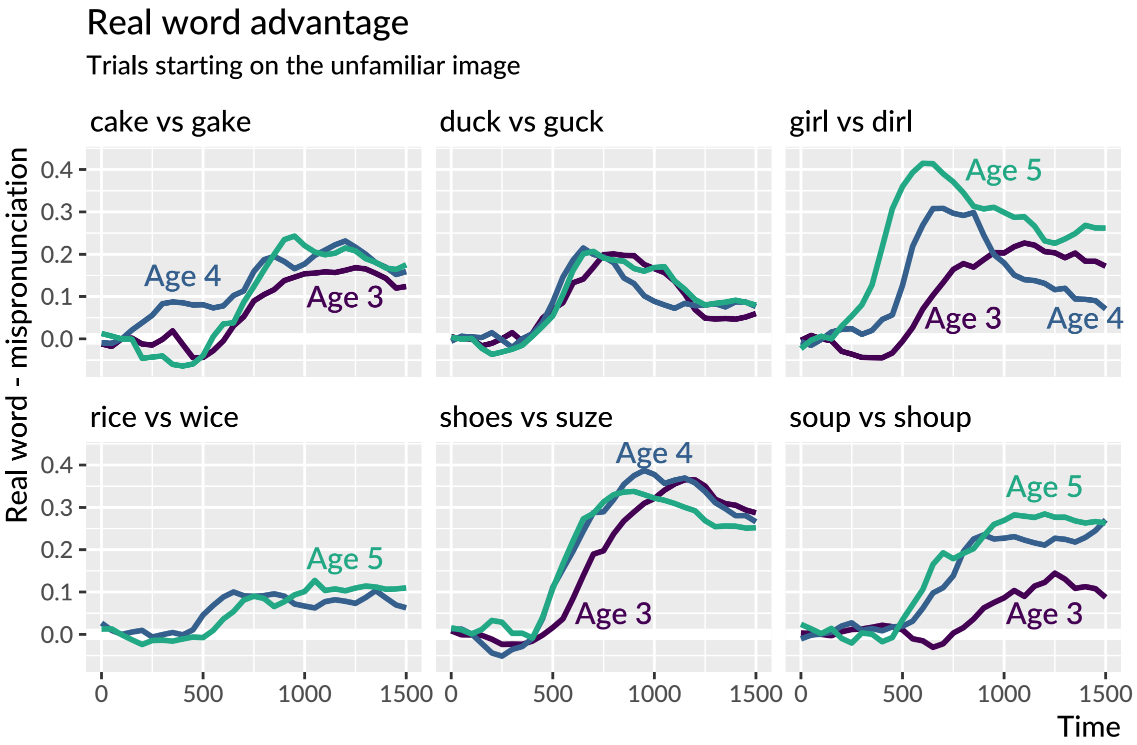


Figure E.2: Differences between the average proportion of looks to real words and mispronunciations.

F 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) analyzed data from the four-image experiment in Study 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) 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 Study 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 (2018) was the manuscript I originally authored for my preliminary examinations. The paper analyzes the same kinds of relations as Weisleder and Fernald (2013) 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 Study 1, vocabulary size, and home language input data from age 3 predicted vocabulary size at 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

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439. doi:10.1006/jmla.1997.2558

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

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

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

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

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

Baylis, A. L., Munson, B., & Moller, K. T. (2008). Factors affecting articulation skills in children with velocardiofacial syndrome and children with cleft palate or velopharyngeal dysfunction: A preliminary report. *The Cleft Palate-Craniofacial Journal*, *45*(2), 193–207. doi:10.1597/06-012.1

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

Blomquist, C., & McMurray, B. (2017). *War of the words: Development of inter-lexical inhibition in typical children.* The University of Iowa. Retrieved from `https://ir.uiowa.edu/honors_theses/31`

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

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

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

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

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

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

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

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

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

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

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

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

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

Gabry, J., & Mahr, T. (2018). *bayesplot: Plotting for Bayesian models.* Retrieved from `https://CRAN.R-project.org/package=bayesplot`

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., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. doi:10.1177/1745691614551642

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

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press/Taylor & Francis Group.

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

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

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica*, *137*(2), 151–171. doi:10.1016/j.actpsy.2010.11.003

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

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 1–29. doi:10.3758/s13423-016-1221-4

Lany, J. (2017). Lexical-processing efficiency leverages novel word learning in infants and toddlers. *Developmental Science*. doi: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

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

Leach, L., & Samuel, A. G. (2007). Lexical configuration and lexical engagement: When adults learn new words. *Cognitive Psychology*, *55*(4), 306–353. doi:10.1016/j.cogpsych.2007.01.001

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

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

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The time course of spoken word learning and recognition: Studies with artificial lexicons. *Journal of Experimental Psychology: General*, *132*(2), 202–227. doi:10.1037/0096-3445.132.2.202

Mahr, T., & Edwards, J. R. (2018). Using language input and lexical processing to predict vocabulary size. *Developmental Science*, e12685. doi:10.1111/desc.12685

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

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

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

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

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

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

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

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

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

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

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

Mervis, C. B., & Bertrand, J. (1994). Acquisition of the novel name–nameless category (N3C) principle. *Child Development*, *65*(6), 1646–1662. doi:10.1111/j.1467-8624.1994.tb00840.x

Mirman, D. (2014). *Growth curve analysis and visualization using R*. Boca Raton, FL: CRC Press/Taylor & Francis Group.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494. doi:10.1016/j.jml.2007.11.006

Mirman, D., Yee, E., Blumstein, S. E., & Magnuson, J. S. (2011). Theories of spoken word recognition deficits in Aphasia: Evidence from eye-tracking and computational modeling. *Brain and Language*, *117*(2), 53–68. doi:10.1016/j.bandl.2011.01.004

Oleson, J. J., Cavanaugh, J. E., McMurray, B., & Brown, G. (2017). Detecting time-specific differences between temporal nonlinear curves: Analyzing data from the visual world paradigm. *Statistical Methods in Medical Research*, *26*(6), 2708–2725. doi:10.1177/0962280215607411

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

Seedorff, M., Oleson, J. J., & McMurray, B. (2018). Detecting when timeseries differ: Using the Bootstrapped Differences of Timeseries (BDOTS) to analyze Visual World Paradigm data (and more). *Journal of Memory and Language*, *102*, 55–67. doi:https://doi.org/10.1016/j.jml.2018.05.004

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

Stan Development Team. (2017). *Stan modeling language: User's guide and reference manual.*

Swingley, D. (2009). Onsets and codas in 1.5-year-olds' word recognition. *Journal of Memory and Language*, *60*(2), 252–269. doi:10.1016/j.jml.2008.11.003

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

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

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

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

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

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

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

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

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Boca Raton, FL: CRC Press/Taylor & Francis Group.