Peekbank: Exploring children's word recognition through an open, large-scale repository for developmental eye-tracking data

Peekbank team, Martin Zettersten[1], Claire Bergey[2], Naiti S. Bhatt[3], Veronica Boyce[4], Mika Braginsky[5], Alexandra Carstensen[4], Benny deMayo[1], George Kachergis[4], Molly Lewis[6], Bria Long[4], Kyle MacDonald[7], Jessica Mankewitz[4], Stephan Meylan[5,8], Annissa N. Saleh[9], Rose M. Schneider[10], Angeline Sin Mei Tsui[4], Sarp Uner[8], Tian Linger Xu[11], Daniel Yurovsky[6], & Michael C. Frank[1]

[1] Dept. of Psychology, Princeton University

[2] Dept. of Psychology, University of Chicago

[3] Scripps College

[4] Dept. of Psychology, Stanford University

[5] Dept. of Brain and Cognitive Sciences, MIT

[6] Dept. of Psychology, Carnegie Mellon University

[7] Core Technology, McD Tech Labs

[8] Dept. of Psychology and Neuroscience, Duke University

[9] Dept. of Psychology, UT Austin

[10] Dept. of Psychology, UC San Diego

[11] Dept. of Psychological and Brain Sciences, Indiana University

Abstract

The ability to rapidly recognize words and link them to referents in context is central to children's early language development. This ability, often called word recognition in the developmental literature, is typically studied in the looking-while-listening paradigm, which measures infants' fixation on a target object (vs. a distractor) after hearing a target label. We present a large-scale, open database of infant and toddler eye-tracking data from looking-while-listening tasks. The goal of this effort is to address theoretical and methodological challenges in measuring vocabulary development.

Keywords: word recognition; eye-tracking; vocabulary development; looking-while-listening; visual world paradigm; lexical processing

Word count: X

Peekbank: Exploring children's word recognition through an open, large-scale repository for developmental eye-tracking data

Across their first years of life, children learn words at an accelerating pace (Michael C. Frank, Braginsky, Yurovsky, & Marchman, 2021). While many children will only *produce* their first word at around one year of age, most children show signs of *understanding* many common nouns (e.g., "mommy") and phrases (e.g., "Let's go bye-bye!") much earlier in development (Bergelson & Swingley, 2012). Although early word understanding is an enticing research target, the processes involved are less directly apparent in children's behaviors and are less accessible to observation than developments in speech production (Fernald, Zangl, Portillo, & Marchman, 2008). To understand a spoken word, children must process the incoming auditory signal and link that signal to relevant meanings – a process often referred to as word recognition. A primary means of measuring word recognition in young infants are eye-tracking techniques that use patterns of preferential looking to make inferences about children's word processing (Fernald, Zangl, Portillo, & Marchman, 2008). The key idea of these methods is that if a child preferentially looks at a target referent (rather than a distractor stimulus) upon hearing a word, this indicates that the child is able to recognize the word and activate its meaning during real-time language processing. Measuring early word recognition offers insight into children's early word representations: children's speed of response (i.e., moving their eyes; turning their heads) to the unfolding speech signal can reveal children's level of comprehension (Bergelson, 2020; Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998). Word recognition skills are also thought to build a foundation for children's subsequent language development. Past research has found that early word recognition efficiency is predictive of later linguistic and general cognitive outcomes (Bleses, Makransky, Dale, Højen, & Ari, 2016; Marchman et al., 2018).

While word recognition is a central part of children's language development, mapping the trajectory of word recognition skills has remained elusive. Studies investigating children's word recognition are typically limited in scope to experiments in individual labs involving small samples tested on a handful of items. The limitations of single datasets makes it difficult to understand developmental changes in children's word knowledge at a broad scale. One way to overcome this challenge is to compile existing datasets into a large-scale database in order to expand the scope of research questions that can be asked about the the development word recognition abilities. This strategy capitalizes on the fact that the looking-while-listening paradigm is widely used, and vast amounts of data have been collected across labs on infants' word recognition over the past 35 years (Golinkoff, Ma, Song, & Hirsh-Pasek, 2013). Such datasets have largely remained isolated from one another, but once combined, they have the potential to offer insights into the lexical development at a broad scale. Similar efforts in language development have born fruit in recent years. For example, WordBank aggregated data from the MacArthur-Bates Communicative Development Inventory, a parent-report measure of child vocabulary, to deliver new insights into cross-linguistic patterns and variability in vocabulary development (Michael C. Frank, Braginsky, Yurovsky, & Marchman, 2017, 2021). In this paper, we introduce *Peekbank*, an open database of infant and toddler eye-tracking data aimed at facilitating the study of developmental changes in children's word knowledge and recognition speed.

The "Looking-While-Listening" Paradigm

Word recognition is traditionally studied in the "looking-while-listening" paradigm (Fernald, Zangl, Portillo, & Marchman, 2008; alternatively referred to as the intermodal preferential looking procedure, Hirsh-Pasek, Cauley, Golinkoff, & Gordon, 1987). In such studies, infants listen to a sentence prompting a specific referent (e.g., *Look at the dog!*) while viewing two images on the screen (e.g., an image of a dog – the target image – and an image of a bird – the distractor image). Infants' word recognition is measured in terms of how quickly and accurately they fixate on the correct target image after hearing its label.

Past research has used this same basic method to study a wide range of questions in language development. For example, the looking-while-listening paradigm has been used to investigate early noun knowledge, phonological representations of words, prediction during language processing, and individual differences in language development (Bergelson & Swingley, 2012; Golinkoff, Ma, Song, & Hirsh-Pasek, 2013; Lew-Williams & Fernald, 2007; Marchman et al., 2018; Swingley & Aslin, 2000).

Measuring developmental change in word recognition

While the looking-while-listening paradigm has been fruitful in advancing understanding of early word knowledge, fundamental questions remain. One central question is how to accurately capture developmental change in the speed and accuracy of word recognition. There is ample evidence demonstrating that infants get faster and more accurate in word recognition over the first few years of life (e.g., Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998). However, precisely measuring developmental increases in the speed and accuracy of word recognition remains challenging due to the difficulty of distinguishing developmental changes in word recognition skill from changes in knowledge of specific words. This problem is particularly thorny in studies with young children, since the number of items that can be tested within a single session is limited and items must be selected in an age-appropriate manner (Peter et al., 2019). Another potential challenge are that differences in the design choices and analytic decisions within single studies could obscure changes when comparing individual studies at different developmental time points. One approach to addressing these these challenges is to conduct meta-analyses aggregating effects across studies while testing for heterogeneity due to researcher choices [Lewis et al. (2016); bergmann2018]. However, meta-analyses typically lack the granularity to estimate participant-level and item-level variation or to model behavior beyond coarse-grained effect size estimates. An alternative way to approach this challenge is to aggregate trial-level data from smaller studies measuring word recognition with a wide range of items and design choices into a large-scale dataset that can be analyzed using a unified modeling approach. A sufficiently large dataset would allow researchers to estimate developmental change in word recognition speed and accuracy while generalizing across changes related to specific words or the design features of particular studies.

A related open theoretical question is understanding changes in children's word recognition at the level of individual items. Looking-while-listening studies have been limited in their ability to assess the development of specific words. One limitation is that studies typically test only a small number of trials for each item, limiting the power the accurately measure the development of word-specific accuracy (DeBolt, Rhemtulla, & Oakes, 2020). A second limitation is that targets are often yoked with a limited set of distractors (often one or two), leaving ambiguous whether accurate looking to a particular target word is largely a function of children's recognition of the target word, their knowledge about the distractor, which allows them to reject the distractor as a response candidate, or both. Aggregating across many looking-while-listening studies has the potential to meet these challenges by increasing the number of observations for specific items at different ages and by increasing the variability in the distractor items co-occurring with a specific target.

Replicability and Reproducibility

A core challenge facing psychology in general, and the study of infant development in particular, are threats to the replicability and reproducibility of core empirical results (M. C. Frank et al., 2017; Nosek et al., 2021). In infant research, many studies are not adequately powered to detect the main effects of interest (Bergmann et al., 2018). This is often compounded by low reliability in infant measures, often due to limits on the number of trials that can be collected from an individual infant in an experimental session (Byers-Heinlein, Bergmann, & Savalei, 2021). One hurdle to improving the power in infant research is that it can often be difficult to develop a priori estimates of effect sizes, and how specific design decisions (e.g., the number of test trials) will impact power and reliability. Large-scale databases of infant behavior can aid researchers' in their decision-making by providing rich datasets that can help constrain expectations about possible effect sizes and can be used to make data-driven design decisions. For example, if a researcher is interested in understanding how the number of test trials could impact the power and reliability of their looking-while-listening design, a large-scale database would allow them to simulate possible outcomes across a range of test trials, based on past eye-tracking data with infants.

In addition to threats to replicability, the field of infant development also faces concerns about analytic reproducibility - the ability for researchers to arrive at the same analytic conclusion reported in the original research article, given the same dataset. A recent estimate based on studies published in a prominent cognitive science journal suggests that analyses can remain difficult to reproduce, even when data is made available to other research teams (Hardwicke et al., 2018). Aggregating data in centralized databases can aid in improving reproducibility in several ways. First, building a large-scale database requires defining a standardized data specification. Recent examples include the brain imaging data structure (BIDS), an effort to specify a unified data format for neuroimaging experiments (Gorgolewski et al., 2016). Defining a data standard - in this case, for infant eye-tracking experiments - supports reproducibility by setting data curation standards that guarantee that critical information will be available in openly shared data and that make it easier for different research teams to understand the data structure. Second, open databases make it easy for researchers to generate open and reproducible analytic pipelines, both for individual studies and for analyses aggregating across datasets. Creating open analytic pipelines across many datasets also serves a pedagogical purpose, providing teaching examples illustrating how to implement analytic techniques using in influential studies and how to conduct reproducible analyses on infant eye-tracking data.

Peekbank: An open database of developmental eye-tracking studies.

What all of these open challenges share is that they are difficult to address at the scale of a single research lab or in a single study. To address this challenge, we developed *Peekbank* a flexible and reproducible interface to an open database of developmental eye-tracking studies. The Peekbank project (a) collects a large set of eye-tracking datasets on children's word recognition, (b) introduces a data format and processing tools for standardizing eye-tracking data across data sources, and (c) provides an interface for accessing and analyzing the database. In the current paper, we introduce the key components of the project and give an overview of the existing database. We then provide worked examples of how researchers can use Peekbank (1) to inform methodological decision-making, (2) to teach through reproducible examples, and (3) ask novel research questions about the development of children's word recognition.

Design and Technical Approach

Database Framework

One of the main challenges in compiling a large-scale eye-tracking database is the lack of a shared data format: both labs and individual experiments can record their results in a wide range of formats. For example, different experiments encode trial-level and subject-level information in many different ways. Therefore, we have developed a common tabular format to support analyses of all studies simultaneously.

As illustrated in Figure 1, the Peekbank framework consists of four main components: (1) a set of tools to *convert* eye-tracking datasets into a unified format, (2) a relational database populated with data in this unified format, (3) a set of tools to *retrieve* data from this database, and (4) a web app (using the Shiny framework) for visualizing the data. These components are supported by three packages. The `peekds` package (for the R language; R Core Team (2020)) helps researchers convert existing datasets to use the standardized format of the database. The `peekbank` module (Python) creates a database with the relational schema and populates it with the standardized datasets produced by `peekds`. The database is served through MySQL, an industry standard relational database server, which may be accessed by a variety of programming languages, and can be hosted on one machine and accessed by many others over the Internet. As is common in relational databases, records of similar types (e.g., participants, trials, experiments, coded looks at each timepoint) are grouped into tables, and records of various types are linked through numeric identifiers. The `peekbankr` package (R) provides an application programming interface, or API, that offers high-level abstractions for accessing the tabular data stored in Peekbank. Most users will access data through this final package, in which case the details of data formatting, processing, and the specifics of connecting to the database are abstracted away from the user.

Database Schema

The Peekbank database contains two major types of data: (1) metadata regarding experiments, participants, and trials, and (2) time course looking data, detailing where on the screen a child is looking at a given point in time (Fig. 2).

Metadata. Metadata can be separated into four parts: (1) participant-level information (e.g., demographics) (2) experiment-level information (e.g., the type of eye tracker used to collect the data) (3) session information (e.g. a participant's age for a specific experimental session) and (4) trial information (e.g., what images or videos were presented onscreen, and paired with which audio).

Participant Information.

Invariant information about individuals who participate in one or more studies (e.g, a

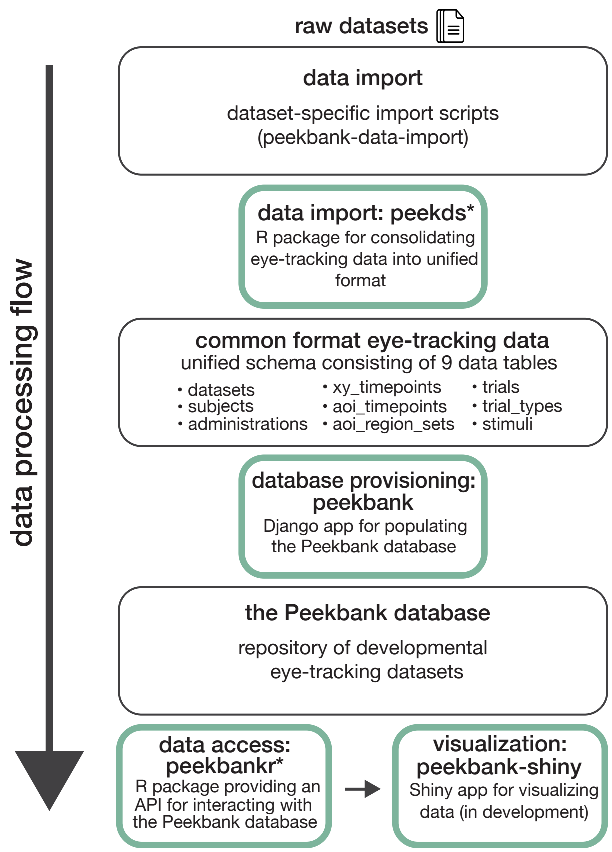

Figure 1. Overview of the Peekbank data ecosystem. Peekbank tools are highlighted in green. * indicates R packages introduced in this work.

subject's first language) is recorded in the `subjects` table, while the `administrations` table contains information about a subject's participation in a single session of a study (see Session Information, below). This division allows Peekbank to gracefully handle longitudinal designs: a single subject can be associated with many administrations.

Subject-level data includes all participants who have experiment data. In general, we include as many participants as possible in the database and leave it to end-users to apply the appropriate exclusion criteria for their analysis.

Experiment Information.

The `datasets` table includes information about the lab conducting the study and the relevant publications to cite regarding the data.

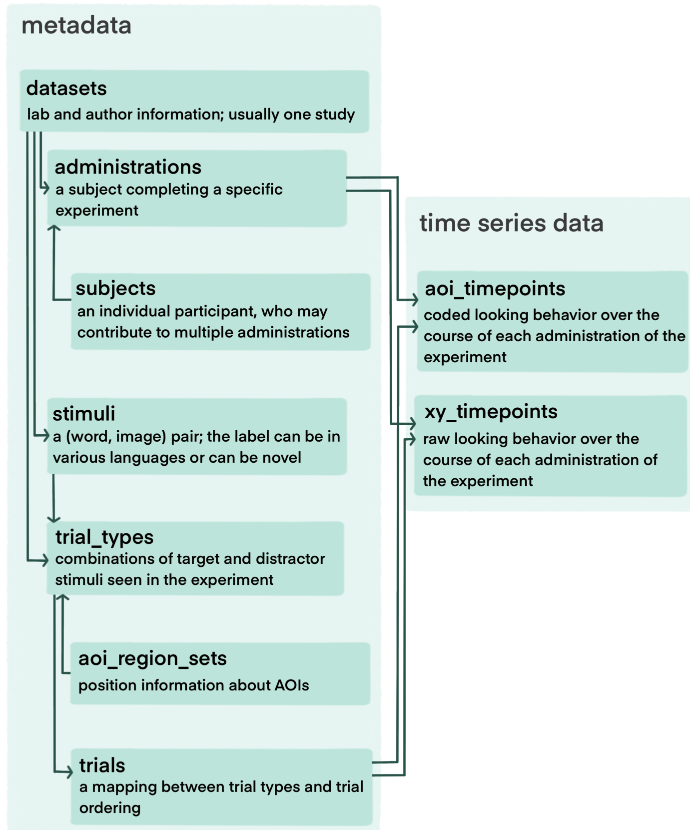

Figure 2. The Peekbank schema. Each square represents a table in the relational database.

In most cases, a dataset corresponds to a single study.

Information about the experimental design is split across the `trial_types` and `stimuli` tables. The `trial_types` table encodes information about each trial *in the design of the experiment*,[1] including the target stimulus and location (left vs. right), the distractor stimulus and location, and the point of disambiguation for that trial. If a dataset used automatic eye-tracking rather than manual coding, each trial type is additionally linked to a set of area of interest (x, y) coordinates, encoded in the `aoi_region_sets` table. The `trial_types` table links trial types to the `aoi_region_sets` table and the `trials` table. Each trial_type record links to two records in the `stimuli` table, identified by the

[1] We note that the term *trial* is often overloaded, to refer to a particular combination of stimuli seen by many participants, vs. a participant seeing that particular combination at a paraticular point in the experiment. We track the latter in the ‘trials‘ table.

distractor_id and the target_id fields.

Each record in the stimuli table is a (word, image) pair. In most experiments, there is a one-to-one mapping between images and labels (e.g., each time an image of a dog appears it is referred to as "dog"). For studies in which there are multiple potential labels per image (e.g., "dog" and "chien" are both used to refer to an image of a dog), images can have multiple rows in the stimuli table with unique labels as well as a row with no label to be used when the image appears solely as a distractor (and thus its label is ambiguous). This structure is useful for studies on synonymy or using multiple languages. For studies in which the same label refers to multiple images (e.g., the word "dog" refers to an image of a dalmatian and a poodle), the same label can have multiple rows in the stimuli table with unique images.

Session Information.

The administrations table includes information about the participant or experiment that may change between sessions of the same study, even for the same participant. This includes the age of the participant, the coding method (eye-tracking vs. hand-coding), and the properties of the monitor that was used.

Trial Information.

The trials table includes information about a specific participant completing a specific instance of a trial type. This table links each record in the raw data (described below) to the trial type and specifies the order of the trials seen by a specific participant.

Time course data. Raw looking data is a series of looks to AOIs or to (x, y) coordinates on the experiment screen, linked to points in time. For data generated by eye-trackers, we typically have (x, y) coordinates at each time point, which will be encoded in the xy_timepoints table. These looks will also be recoded into AOIs according to the

AOI coordinates in the `aoi_region_sets` table using the `add_aois()` function in `peekds`, which will be encoded in the `aoi_timepoints` table. For hand-coded data, we typically have a series of AOIs; these will be recoded into the categories in the Peekbank schema (target, distractor, other, and missing) and encoded in the `aoi_timepoints` table, and these datasets will not have an `xy_timepoints` table.

Typically, timepoints in the `xy_timepoints` table and `aoi_timepoints` table need to be regularized to center each trial's time around the point of disambiguation–such that 0 is the time of target word onset in the trial (i.e., the beginning of *dog* in "Can you find the *dog*?"). If time values run throughout the experiment rather than resetting to zero at the beginning of each trial, `rezero_times()` is used to reset the time at each trial. After this, each trial's times are centered around the point of disambiguation using `normalize_times()`. When these steps are complete, the time course is ready for resampling.

To facilitate time course analysis and visualization across datasets, time course data must be resampled to a uniform sampling rate (i.e., such that every trial in every dataset has observations at the same time points). To do this, we use the `resample_times()` function. During the resampling process, we interpolate using constant interpolation, selecting for each interpolated timepoint the looking location for the nearest observed time point in the original data for both `aoi_timepoints` and `xy_timepoints` data. In the case of ties, the look location observed at the earlier timepoint in the original data is chosen for the resampled timepoint. Currently, all data is resampled to 40 Hz (observations every 25 ms) by default, which represents a compromise between retaining fine-grained timing information from datasets with dense sampling rates (maximum sampling rate among current datasets: 500 Hz) while minimizing the possibility of introducing artifacts via resampling for datasets with lower sampling rates (minimum sampling rate for current datasets: 30 Hz). Compared to linear interpolation (see e.g. Wass et al., 2014), constant interpolation has the advantage that it is more conservative, in the sense that it does not introduce new look locations beyond those measured in the original data.

Processing, Validation and Ingestion

The `peekds` package offers functions to extract the above data. Once this data has been extracted in a tabular form, the package also offers a function to check whether all tables have the required fields and data types expected by the database. In an effort to double check the data quality and to make sure that no errors are made in the importing script, as part of the import procedure we create a time course plot based on our processed tables to replicate the results in the paper that first presented each dataset. Once this plot has been created and checked for consistency and all tables pass our validation functions, the processed dataset is ready for ingestion into the database using the `peekbank` library. This library applies additional data checks, and adds the data to the MySQL database using the Django web framework.

Currently, the import process is carried out by the Peekbank team using data offered by other research teams. In the future, we hope to allow research teams to carry out their own import processes with checks from the Peekbank team before ingestion. To this end, import script templates are available for both hand-coded datasets and automatic eye-tracking datasets for research teams to adapt to their data.

Current Data Sources

The database currently includes 20 looking-while-listening datasets comprising N=1594 total participants (Table 1). The current data represents a convenience sample of datasets that were (a) datasets collected by or available to Peekbank team members, (b) made available to Peekbank after informal inquiry or (c) datasets that were openly available. Most datasets (14 out of 20 total) consist of data from monolingual native English speakers. They

Table 1
Overview of the datasets in the current database.

Dataset name	Citation	N	Mean age (mos.)	Age range (mos.)	Method	Language
attword	Yurovsky & Frank, 2017	288	25.5	13–59	eye-tracking	English
canine	unpublished	36	23.8	21–27	manual coding	English
coartic	Mahr et al., 2015	29	20.8	18–24	eye-tracking	English
cowpig	Perry et al., 2017	45	20.5	19–22	manual coding	English
fmw	Fernald et al., 2013	80	20.0	17–26	manual coding	English
ft_pt	Adams et al., 2018	69	17.1	13–20	manual coding	English
input_uptake	Hurtado et al., 2008	76	21.0	17–27	manual coding	Spanish
lsc	Ronfard et al., 2021	40	20.0	18–24	manual coding	English
mispron	Swingley & Aslin, 2002	50	15.1	14–16	manual coding	English
mix	Byers-Heinlein et al., 2017	48	20.1	19–21	eye-tracking	English, French
reflook_socword	Yurovsky et al., 2013	435	33.6	12–70	eye-tracking	English
reflook_v4	unpublished	45	34.2	11–60	eye-tracking	English
remix	Potter et al., 2019	44	22.6	18–29	manual coding	Spanish, English
salientme	Pomper & Saffran, 2019	44	40.1	38–43	manual coding	English
stl	Weisleder & Fernald, 2013	29	21.6	18–27	manual coding	Spanish
switchingCues	Pomper & Saffran, 2016	60	44.3	41–47	manual coding	English
tablet	Frank et al., 2016	69	35.5	12–60	eye-tracking	English
tseltal	Casillas et al., 2017	23	31.3	9–48	manual coding	Tseltal
xsectional	Hurtado et al., 2007	49	23.8	15–37	manual coding	Spanish
yoursmy	Garrison et al., 2020	35	14.5	12–18	eye-tracking	English

span a wide age spectrum with participants ranging from 9 to 70 months of age, and are balanced in terms of gender (47% female). The datasets vary across a number of design-related dimensions, and include studies using manually coded video recordings and automated eye-tracking methods (e.g., Tobii, EyeLink) to measure gaze behavior. All studies tested familiar items, but the database also includes 5 datasets that tested novel pseudo-words in addition to familiar words.

Versioning + Expanding the database

The content of Peekbank will change as we add additional datasets and revise previous ones. To facilitate reproducibility of analyses, we use a versioning system where successive releases are assigned a name reflecting the year and version, e.g., `2021.1`. By default, users will interact with the most recent version of the database available, though `peekbankr` API allows researchers to run analyses against any previous version of the database. For users with intensive use-cases, each version of the database may be downloaded as a compressed .sql file and installed on a local MySQL server.

Interfacing with peekbank

Shiny App

One goal of the Peekbank project is to allow a wide range of users to easily explore and learn from the database. We therefore have created an interactive web application – `peekbank-shiny` – that allows users to quickly and easily create informative visualizations of individual datasets and aggregated data. `peekbank-shiny` is built using Shiny, a software package for creating web apps using R. The Shiny app allows users to create commonly used visualizations of looking-while-listening data, based on data from the Peekbank database. Specifically, users can visualize

1. the time course of looking data in a profile plot depicting infant target looking across trial time
2. overall accuracy (proportion target looking) within a specified analysis window
3. reaction times (speed of fixating the target image) in response to a target label
4. an onset-contingent plot, which shows the time course of participant looking as a function of their look location at the onset of the target label

Users are given various customization options for each of these visualizations, e.g., choosing which datasets to include in the plots, controlling the age range of participants, splitting the visualizations by age bins, and controlling the analysis window for time course analyses. Plots are then updated in real time to reflect users' customization choices, and users are given options to share the visualizations they created. The Shiny app thus allows users to quickly inspect basic properties of Peekbanks datasets and create reproducible visualizations without incurring any of the technical overhead required to access the database through R.

Peekbankr

The `peekbankr` API offers a way for users to access data from the database and flexibly analyze it in `R`. Users can download tables from the database, as specified in the Schema section above, and merge them using their linked IDs to examine time course data and metadata jointly. In the sections below, we work through some examples to outline the possibilities for analyzing data downloaded using `peekbankr`.

Functions:

- `connect_to_peekbank()` opens a connection with the Peekbank database to allow tables to be downloaded with the following functions
- `get_datasets()` gives each dataset name and its citation information
- `get_subjects()` gives information about persistent subject identifiers (e.g., native languages, sex)
- `get_administrations()` gives information about specific experimental administrations (e.g., subject age, monitor size, gaze coding method)
- `get_stimuli()` gives information about word–image pairings that appeared in experiments
- `get_trial_types()` gives information about pairings of stimuli that appeared in the experiment (e.g., point of disambiguation, target and distractor stimuli, condition, language)
- `get_trials()` gives the trial orderings for each administration, linking trial types to the trial IDs used in time course data
- `get_aoi_region_sets()` gives coordinate regions for each area of interest (AOI) linked to trial type IDs
- `get_xy_timepoints()` gives time course data for each subject's looking behavior in each trial, as (x, y) coordinates on the experiment monitor

- `get_aoi_timepoints()` gives time course data for each subject's looking behavior in each trial, coded into areas of interest

OSF site

In addition to the Peekbank database proper, all data is openly available on the Peekbank OSF webpage (https://osf.io/pr6wu/). The OSF site also includes the original raw data (both time series data and metadata, such as trial lists and participant logs) that was obtained for each study and subsequently processed into the standardized Peekbank format. Users who are interested in inspecting or reproducing the processing pipeline for a given dataset can use the respective import script (openly available on GitHub, https://github.com/langcog/peekbank-data-import) to download and process the raw data from OSF into its final standardized format. Where available, the OSF page also includes additional information about the stimuli used in each dataset, including in some instances the original stimulus sets (e.g., image and audio files).

Peekbank: General Descriptives

[Accuracy, Reaction Times, Item variability?]

Overall Word Recognition Accuracy

Dataset Name	Unique Items	Prop. Target	95% CI
attword	6	0.63	[0.62, 0.65]
canine	16	0.65	[0.61, 0.68]
coartic	10	0.71	[0.68, 0.74]
cowpig	12	0.61	[0.58, 0.63]
fmw	12	0.65	[0.63, 0.67]
ft_pt	8	0.65	[0.63, 0.67]
input_uptake	12	0.61	[0.59, 0.63]
lsc	8	0.69	[0.65, 0.73]
mispron	22	0.57	[0.55, 0.59]
mix	6	0.55	[0.52, 0.58]
reflook_socword	6	0.61	[0.6, 0.63]
reflook_v4	10	0.61	[0.57, 0.65]
remix	8	0.63	[0.58, 0.67]
salientme	16	0.74	[0.72, 0.75]
stl	12	0.63	[0.6, 0.66]
switchingCues	40	0.77	[0.75, 0.8]
tablet	24	0.64	[0.6, 0.68]
tseltal	30	0.59	[0.54, 0.63]
xsectional	8	0.59	[0.55, 0.63]
yoursmy	87	0.60	[0.56, 0.64]

Table 2
Average proportion target looking in each dataset.

In general, participants demonstrated robust, above-chance word recognition in each dataset (chance=0.5). Table 2 shows the average proportion of target looking within a standard critical window of 367-2000ms after the onset of the label for each dataset (Swingley & Aslin, 2000). Proportion target looking was generally higher for familiar words ($M = 0.66$, 95% CI = [0.65, 0.67], $n = 1543$) than for novel words learned during the experiment ($M = 0.59$, 95% CI = [0.58, 0.61], $n = 822$).

Item-level variability

Figure 3 gives an overview of the variability in accuracy for individual words in each dataset. The number of unique target labels and their associated accuracy vary widely across datasets.

Peekbank in Action

We provide two potential use-cases for Peekbank data. In each case, we provide sample code so as to model how easy it is to do simple analyses using data from the database. Our first example shows how we can replicate the analysis for a classic study. This type of computational reproducibility can be a very useful exercise for teaching students about best practices for data analysis (e.g., Hardwicke et al., 2018) and also provides an easy way to explore looking-while-listening time course data in a standardized format. Our second example shows an in-depth exploration of developmental changes in the recognition of particular words. Besides its theoretical interest (which we will explore more fully in subsequent work), this type of analysis could in principle be used for optimizing the stimuli for new experiments, especially as the Peekbank dataset grows and gains coverage over a greater number of items.

Computational reproducibility example: Swingley and Aslin (2000)

Swingley and Aslin (2000) investigated the specificity of 14-16 month-olds' word representations using the looking-while-listening paradigm, asking whether recognition would be slower and less accurate for mispronunciations, e.g. "oppel" (close mispronunciation) or "opel" (distant mispronunciation) instead of "apple" (correct pronunciation). In this short vignette, we show how easily the data in Peekbank can be used to visualize this result.

```r
library(peekbankr)
aoi_timepoints <- get_aoi_timepoints(dataset_name = "swingley_aslin_2002")
administrations <- get_administrations(dataset_name = "swingley_aslin_2002")
trial_types <- get_trial_types(dataset_name = "swingley_aslin_2002")
trials <- get_trials(dataset_name = "swingley_aslin_2002")
```

We begin by retrieving the relevant tables from the database, `aoi_timepoints`, `administrations`, `trial_types`, and `trials`. As discussed above, each of these can be downloaded using a simple API call through `peekbankr`, which returns dataframes that include ID fields. These ID fields allow for easy joining of the data into a single dataframe containing all the information necessary for the analysis.

```
swingley_data <- aoi_timepoints %>%
  left_join(administrations) %>%
  left_join(trials) %>%
  left_join(trial_types) %>%
  filter(condition != "filler") %>%
  mutate(condition = if_else(condition == "cp", "Correct", "Mispronounced"))
```

As the code above shows, once the data are joined, condition information for each timepoint is present and so we can easily filter out filler trials and set up the conditions for further analysis. For simplicity, here we combine both mispronunciation conditions since the close vs. distant mispronunciation manipulation showed no effect in the original paper.

```
accuracies <- swingley_data  %>%
  group_by(condition, t_norm, administration_id) %>%
  summarize(correct = sum(aoi == "target") /
              sum(aoi %in% c("target","distractor"))) %>%
  group_by(condition, t_norm) %>%
  summarize(mean_correct = mean(correct),
            ci = 1.96 * sd(correct) / sqrt(n()))
```

The final step in our analysis is to create a summary dataframe using `dplyr` commands. We first group the data by timestep, participant, and condition and compute the proportion looking at the correct image. We then summarize again, averaging across participants, computing both means and 95% confidence intervals (via the approximation of 1.96 times the standard error of the mean). The resulting dataframe can be used for visualization of the time course of looking.

```
ggplot(accuracies, aes(x = t_norm, y = mean_correct, color = condition)) +
  geom_hline(yintercept = 0.5, linetype = "dashed", color = "black") +
  geom_vline(xintercept = 0, linetype = "dotted", color = "black") +
  geom_pointrange(aes(ymin = mean_correct - ci,
```

```
                          ymax = mean_correct + ci)) +
  labs(x = "Time from target word onset (msec)",
       y = "Proportion looking at correct image",
       color = "Condition") +
  lims(x = c(-500, 3000))
```

Figure 4 shows the average time course of looking for the two conditions, as produced by the code above. Looks after the correctly pronounced noun appeared both faster (deviating from chance earlier) and more accurate (showing a higher asymptote). Overall, this example demonstrates the ability to produce this visualization in just a few lines of code.

Item analyses

A second use case for Peekbank is to examine item-level variation in word recognition. Individual datasets rarely have enough statistical power to show reliable developmental differences within items. To illustrate the power of aggregating data across multiple datasets, we select the four words with the most data available across studies and ages (apple, book, dog, and frog) and show average recognition trajectories.

Our first step is to collect and join the data from the relevant tables including timepoint data, trial and stimulus data, and administration data (for participant ages). We join these into a single dataframe for easy manipulation; this dataframe is a common starting point for analyses of item-level data.

```
all_aoi_timepoints <- get_aoi_timepoints()

all_stimuli <- get_stimuli()

all_administrations <- get_administrations()

all_trial_types <- get_trial_types()

all_trials <- get_trials()
```

```
aoi_data_joined <- all_aoi_timepoints %>%
  right_join(all_administrations) %>%
  right_join(all_trials) %>%
  right_join(all_trial_types) %>%
  mutate(stimulus_id = target_id) %>%
  right_join(all_stimuli) %>%
  select(administration_id, english_stimulus_label, age, t_norm, aoi)
```

Next we select a set of four target words (chosen based on having more than XXX children contributing data for each across several one-year age groups). We create age groups, aggregate, and compute timepoint-by-timepoint confidence intervals using the z approximation.

```
target_words <- c("book","dog","frog","apple")

target_word_data <- aoi_data_joined %>%
  filter(english_stimulus_label %in% target_words) %>%
  mutate(age_group = cut(age, breaks = seq(12,48,12))) %>%
  filter(!is.na(age_group)) %>%
  group_by(t_norm, administration_id, age_group, english_stimulus_label) %>%
  summarise(correct = mean(aoi == "target") /
              mean(aoi %in% c("target","distractor"), na.rm=TRUE)) %>%
  group_by(t_norm, age_group, english_stimulus_label) %>%
  summarise(ci = 1.96 * sd(correct, na.rm=TRUE) / sqrt(length(correct)),
          correct = mean(correct, na.rm=TRUE),
          n = n())
```

Finally, we plot the data as time courses split by age. Our plotting code is shown below (with styling commands again removed for clarity). Figure 5 shows the resulting plot, with time courses for each of three (rather coarse) age bins. Although some baseline effects are visible across items, we still see clear and consistent increases in looking to the target, with the increase appearing earlier and in many cases asymptoting at a higher level for older children. On the other hand, this simple averaging approach ignores study-to-study variation (perhaps responsible for the baseline effects we see in the "apple" and "frog" items especially). In future work, we hope to introduce model-based analytic methods that use mixed effects regression to factor out study-level and individual-level variance in order to recover developmental effects more appropriately (see e.g. Zettersten et al. (2021) for a prototype of such an analysis).

```
ggplot(target_word_data,
       aes(x = t_norm, y = correct, col = age_group)) +
  geom_line() +
  geom_linerange(aes(ymin = correct - ci, ymax = correct + ci),
                 alpha = .2) +
  facet_wrap(~english_stimulus_label)
```

Discussion and Conclusion

Theoretical progress in understanding child development requires rich datasets, but collecting child data is expensive, difficult, and time-intensive. Recent years have seen a growing effort to build open source tools and pool research efforts to meet the challenge of building a cumulative developmental science (Bergmann et al., 2018; Michael C. Frank, Braginsky, Yurovsky, & Marchman, 2017; The ManyBabies Consortium, 2020). The Peekbank project expands on these efforts by building an infrastructure for aggregating eye-tracking data across studies, with a specific focus on the looking-while-listening paradigm. This paper presents an overview of the structure of the database, as well as how users can access the database and some initial demonstrations of how it can be used both to facilitate reproducibility, for teaching and for exploring theoretical questions beyond on the scope of an individual study.

There are a number of limitations surrounding the current scope of the database. A priority in future work will be to expand the size of the database. With 20 datasets currently available in the database, idiosyncrasies of particular designs and condition manipulations still have substantial influence on modeling results. Expanding the set of distinct datasets will allow us to increase the number of observations per item across datasets, leading to more robust generalizations across item-level variability. The current database is also limited by the relatively homogeneous background of its participants, both with respect to language (almost entirely monolingual native English speakers) and cultural background (all but one dataset come from WEIRD populations, potentially limiting generalizability; see Muthukrishna et al. (2020)). Increasing the diversity of participant backgrounds and languages will expand the scope of the generalizations we can form about child word recognition.

Finally, while the current database is focused on studies of word recognition, the tools and infrastructure developed in the project can in principle be used to accommodate any eye-tracking paradigm, opening up new avenues for insights into cognitive development. Gaze behavior has been at the core of many of the key advances in our understanding of infant cognition. Aggregating large datasets of infant looking behavior in a single, openly-accessible format promises to bring a fuller picture of infant cognitive development into view.

Acknowledgements

We would like to thank the labs and researchers that have made their data publicly available in the database.

References

Bergelson, E. (2020). The comprehension boost in early word learning: Older infants are better learners. *Child Development Perspectives*, *14*(3), 142–149.

Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. *PNAS*, *109*(9), 3253–3258.

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development*, *89*(6), 1996–2009.

Bleses, D., Makransky, G., Dale, P. S., Højen, A., & Ari, B. A. (2016). Early productive vocabulary predicts academic achievement 10 years later. *Applied Psycholinguistics*, *37*(6), 1461–1476.

Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2021). Six solutions for more reliable infant research. *PsyArXiv*. https://doi.org/https://doi.org/10.31234/osf.io/ksfvq

DeBolt, M. C., Rhemtulla, M., & Oakes, L. M. (2020). Robust data and power in infant research: A case study of the effect of number of infants and number of trials in visual preference procedures. *Infancy*, *25*(4), 393–419. https://doi.org/10.1111/infa.12337

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, *16*(2), 234–248. https://doi.org/10.1111/desc.12019

Fernald, A., Pinto, J. P., Swingley, D., Weinberg, A., & McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological*

Science, *9*(3), 228–231.

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In I. A. Sekerina, E. M. Fernandez, & H. Clahsen (Eds.), *Developmental psycholinguistics: On-line methods in children's language processing* (pp. 97–135). Amsterdam: John Benjamins.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., … Yurovsky, D. (2017). A Collaborative Approach to Infant Research: Promoting Reproducibility, Best Practices, and Theory-Building. *Infancy*, *22*(4), 421–435. https://doi.org/10.1111/infa.12182

Frank, Michael C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694.

Frank, Michael C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and Consistency in Early Language Learning: The Wordbank Project.* Cambridge, MA: MIT Press.

Golinkoff, R. M., Ma, W., Song, L., & Hirsh-Pasek, K. (2013). Twenty-five years using the intermodal preferential looking paradigm to study language acquisition: What have we learned? *Perspectives on Psychol. Science*, *8*(3), 316–339.

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., … Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*(1), 160044. https://doi.org/10.1038/sdata.2016.44

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., … Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, *5*(8). https://doi.org/10.1098/rsos.180448

Hirsh-Pasek, K., Cauley, K. M., Golinkoff, R. M., & Gordon, L. (1987). The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of Child Language*, *14*(1), 23–45.

Hurtado, N., Marchman, V. A., & Fernald, A. (2007). Spoken word recognition by Latino children learning Spanish as their first language. *Journal of Child Language*, *34*(2), 227–249. https://doi.org/10.1017/S0305000906007896

Hurtado, N., Marchman, V. A., & Fernald, A. (2008). Does input influence uptake? Links between maternal talk, processing speed and vocabulary size in Spanish-learning children. *Developmental Science*, *11*(6), 31–39. https://doi.org/10.1111/j.1467-7687.2008.00768.x

Lewis, M., Braginsky, M., Tsuji, S., Bergmann, C., Piccinini, P. E., Cristia, A., & Frank, M. (2016). *A Quantitative Synthesis of Early Language Acquisition Using Meta-Analysis*. https://doi.org/10.31234/osf.io/htsjm

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science*, *18*(3), 193–198.

Marchman, V. A., Loi, E. C., Adams, K. A., Ashland, M., Fernald, A., & Feldman, H. M. (2018). Speed of language comprehension at 18 months old predicts school-relevant outcomes at 54 months old in children born preterm. *Journal of*

Dev. & Behav. Pediatrics, *39*(3), 246–253.

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) Psychology: Measuring and Mapping Scales of Cultural and Psychological Distance. *Psychological Science*, *31*(6), 678–701.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., … Vazire, S. (2021). Replicability, Robustness, and Reproducibility in Psychological Science. *PsyArXiv*. https://doi.org/https://doi.org/10.31234/osf.io/ksfvq

Peter, M. S., Durrant, S., Jessop, A., Bidgood, A., Pine, J. M., & Rowland, C. F. (2019). Does speed of processing or vocabulary size predict later language growth in toddlers? *Cognitive Psychology*, *115*, 101238.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ronfard, S., Wei, R., & Rowe, M. L. (2021). Exploring the linguistic, cognitive, and social skills underlying lexical processing efficiency as measured by the looking-while-listening paradigm. *Journal of Child Language*, 1–24. https://doi.org/10.1017/S0305000921000106

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*(2), 147–166.

The ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed speech preference. *Advances in Methods and Practices in Psychological Science*, *3*(1), 24–52.

Weisleder, A., & Fernald, A. (2013). Talking to Children Matters: Early Language Experience Strengthens Processing and Builds Vocabulary. *Psychological Science*, *24*(11), 2143–2152. https://doi.org/10.1177/0956797613488145

Zettersten, M., Bergey, C., Bhatt, N., Boyce, V., Braginsky, M., Carstensen, A., … others. (2021). Peekbank: Exploring children's word recognition through an open, large-scale repository for developmental eye-tracking data.

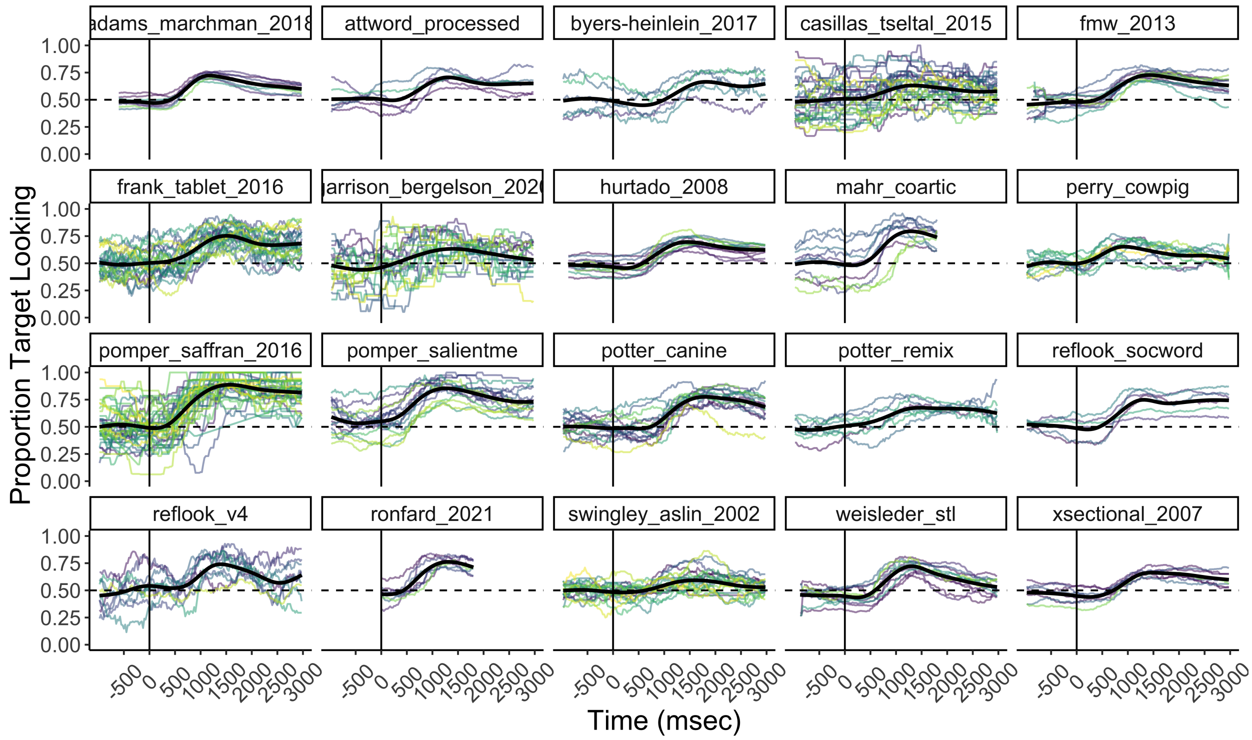

Figure 3. Item-level variability in proportion target looking within each dataset (chance=0.5). Time is centered on the onset of the target label (vertical line). Colored lines represent specific target labels. Black lines represent smoothed average fits based on a general additive model using cubic splines.

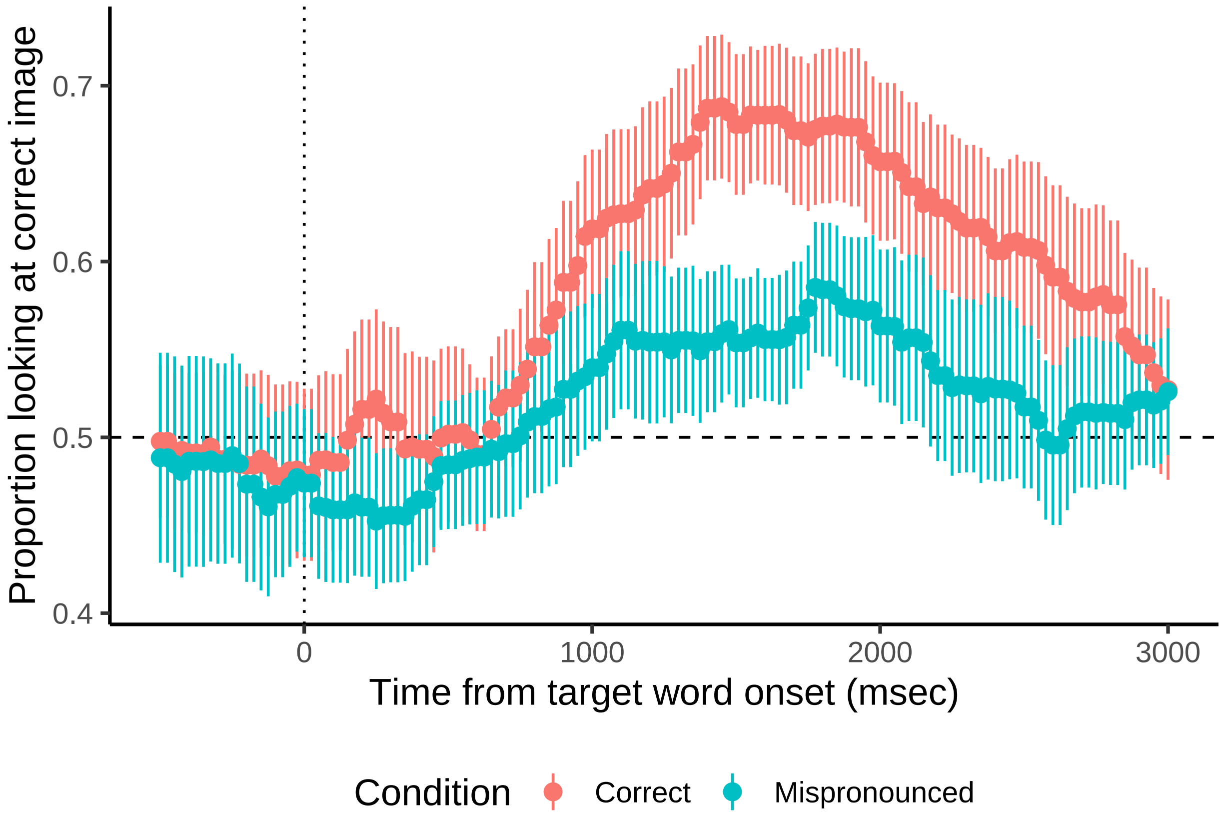

Figure 4. Proportion looking at the correct referent by time from the point of disambiguation (the onset of the target noun). Colors show the two pronunciation conditions; points give means and ranges show 95% confidence intervals. The dotted line shows the point of disambiguation and the dashed line shows chance performance.

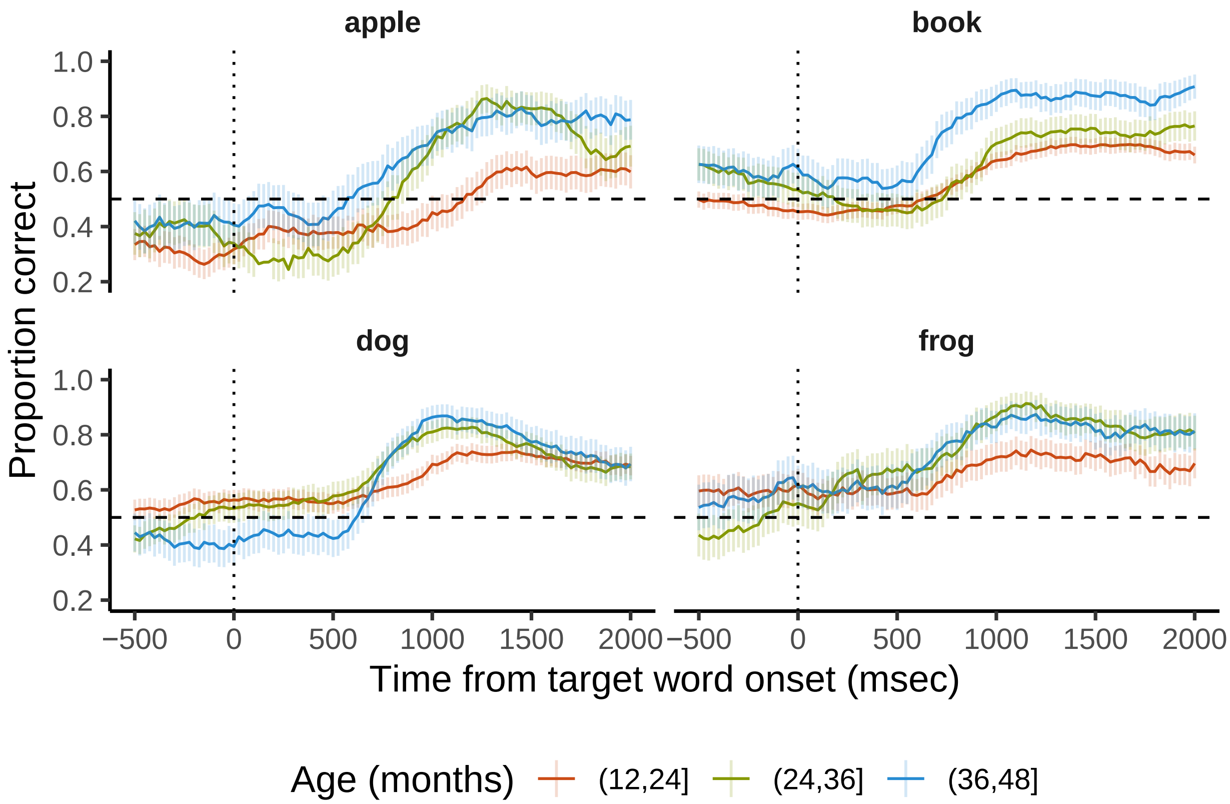

Figure 5. Add caption here.