

¹ Peekbank: Exploring children's word recognition through an open, large-scale repository for
² developmental eye-tracking data

³ Peekbank team, Martin Zettersten¹, Claire Bergey², Naiti S. Bhatt³, Veronica Boyce⁴, Mika
⁴ Braginsky⁵, Alexandra Carstensen⁴, Benny deMayo¹, George Kachergis⁴, Molly Lewis⁶, Bria
⁵ Long⁴, Kyle MacDonald⁷, Jessica Mankewitz⁴, Stephan Meylan^{5,8}, Annissa N. Saleh⁹, Rose
⁶ M. Schneider¹⁰, Angeline Sin Mei Tsui⁴, Sarp Uner⁸, Tian Linger Xu¹¹, Daniel Yurovsky⁶, &
⁷ Michael C. Frank¹

⁸ ¹ Dept. of Psychology, Princeton University

⁹ ² Dept. of Psychology, University of Chicago

¹⁰ ³ Scripps College

¹¹ ⁴ Dept. of Psychology, Stanford University

¹² ⁵ Dept. of Brain and Cognitive Sciences, MIT

¹³ ⁶ Dept. of Psychology, Carnegie Mellon University

¹⁴ ⁷ Core Technology, McD Tech Labs

¹⁵ ⁸ Dept. of Psychology and Neuroscience, Duke University

¹⁶ ⁹ Dept. of Psychology, UT Austin

¹⁷ ¹⁰ Dept. of Psychology, UC San Diego

¹⁸ ¹¹ Dept. of Psychological and Brain Sciences, Indiana University

19

Abstract

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

27 *Keywords:* tools; processing; analysis / usage examples

28 Word count: X

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

31 Across their first years of life, children learn words at an accelerating pace (Michael C.
32 Frank, Braginsky, Yurovsky, & Marchman, 2021). While many children will only *produce*
33 their first word at around one year of age, most children show signs of *understanding* many
34 common nouns (e.g., “mommy”) and phrases (e.g., “Let’s go bye-bye!”) much earlier in
35 development (Bergelson & Swingley, 2012). Although early word understanding is an
36 enticing research target, the processes involved are less directly apparent in children’s
37 behaviors and are less accessible to observation than developments in speech production
38 (Fernald, Zangl, Portillo, & Marchman, 2008). To understand a spoken word, children must
39 process the incoming auditory signal and link that signal to relevant meanings – a process
40 often referred to as word recognition. A child who looks at a word’s intended referent upon
41 hearing the word is inferred to have recognized and understood the word. Measuring early
42 word recognition offers insight into children’s early word representations and as well as the
43 speed and efficiency with which children comprehend language in real time, as the speech
44 signal unfolds (Bergelson, 2020; Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998).
45 Word recognition skills are also thought to build a foundation for children’s subsequent
46 language development. Past research has found that early word recognition efficiency is
47 predictive of later linguistic and general cognitive outcomes (Bleses, Makransky, Dale, Højen,
48 & Ari, 2016; Marchman et al., 2018).

49 While word recognition is a central part of children’s language development, mapping
50 the trajectory of word recognition skills has remained elusive. Studies investigating children’s
51 word recognition are typically limited in scope to experiments in individual labs involving
52 small samples tested on a handful of items. The limitations of single datasets makes it
53 difficult to understand developmental changes in children’s word knowledge at a broad scale.
54 One way to overcome this challenge is to compile existing datasets into a large-scale

55 database in order to expand the scope of research questions that can be asked about the the
56 development word recognition abilities. This strategy capitalizes on the fact that the
57 looking-while-listening paradigm is widely used, and vast amounts of data have been
58 collected across labs on infants' word recognition over the past 35 years (Golinkoff, Ma, Song,
59 & Hirsh-Pasek, 2013). Such datasets have largely remained isolated from one another, but
60 once combined, they have the potential to offer insights into the lexical development at a
61 broad scale. Similar efforts in language development have born fruit in recent years. For
62 example, WordBank aggregated data from the MacArthur-Bates Communicative
63 Development Inventory, a parent-report measure of child vocabulary, to deliver new insights
64 into cross-linguistic patterns and variability in vocabulary development (Michael C. Frank,
65 Braginsky, Yurovsky, & Marchman, 2017, 2021). In this paper, we introduce *Peekbank*, an
66 open database of infant and toddler eye-tracking data aimed at facilitating the study of
67 developmental changes in children's word knowledge and recognition speed.

68 The “Looking-While-Listening” Paradigm

69 Word recognition is traditionally studied in the “looking-while-listening” paradigm
70 [alternatively referred to as the intermodal preferential looking procedure; Fernald et al.
71 (2008); Hirsh-Pasek, Cauley, Golinkoff, and Gordon (1987)]. In such studies, infants listen to
72 a sentence prompting a specific referent (e.g., *Look at the dog!*) while viewing two images on
73 the screen (e.g., an image of a dog – the target image – and an image of a bird – the
74 distractor image). Infants' word recognition is measured in terms of how quickly and
75 accurately they fixate on the correct target image after hearing its label. Past research has
76 used this same basic method to study a wide range of questions in language development.
77 For example, the looking-while-listening paradigm has been used to investigate early noun
78 knowledge, phonological representations of words, prediction during language processing, and
79 individual differences in language development (Bergelson & Swingley, 2012; Golinkoff et al.,
80 2013; Lew-Williams & Fernald, 2007; Marchman et al., 2018; Swingley & Aslin, 2000).

81 Measuring developmental change in word recognition

82 While the looking-while-listening paradigm has been fruitful in advancing
83 understanding of early word knowledge, fundamental questions remain. One central question
84 is how to accurately capture developmental change in the speed and accuracy of word
85 recognition. There is ample evidence demonstrating that infants get faster and more
86 accurate in word recognition over the first few years of life (e.g., Fernald et al., 1998).
87 However, precisely measuring developmental increases in the speed and accuracy of word
88 recognition remains challenging due to the difficulty of distinguishing developmental changes
89 in word recognition skill from changes in knowledge of specific words. This problem is
90 particularly thorny in studies with young children, since the number of items that can be
91 tested within a single session is limited and items must be selected in an age-appropriate
92 manner (Peter et al., 2019). One way to overcome this challenge is to measure word
93 recognition across development in a large-scale dataset with a wide range of items. A
94 sufficiently large dataset would allow researchers to estimate developmental change in word
95 recognition speed and accuracy while generalizing across changes related to specific words.

96 Understanding the development of word recognition on the item level

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

¹⁰⁷ for specific items at different ages and by increasing the variability in the distractor items
¹⁰⁸ co-occurring with a specific target.

¹⁰⁹ **Replicability and Reproducibility**

¹¹⁰ A core challenge facing psychology in general, and the study of infant development in
¹¹¹ particular, are threats to the replicability and reproducibility of core empirical results (M. C.
¹¹² Frank et al., 2017; Nosek et al., 2021). In infant research, many studies are not adequately
¹¹³ powered to detect the main effects of interest (Bergmann et al., 2018). This is often
¹¹⁴ compounded by low reliability in infant measures, often due to limits on the number of trials
¹¹⁵ that can be collected from an individual infant in an experimental session (Byers-Heinlein,
¹¹⁶ Bergmann, & Savalei, 2021). One hurdle to improving the power in infant research is that it
¹¹⁷ can often be difficult to develop a priori estimates of effect sizes, and how specific design
¹¹⁸ decisions (e.g., the number of test trials) will impact power and reliability. Large-scale
¹¹⁹ databases of infant behavior can aid researchers' in their decision-making by providing rich
¹²⁰ datasets that can help constrain expectations about possible effect sizes and can be used to
¹²¹ make data-driven design decisions. For example, if a researcher is interested in
¹²² understanding how the number of test trials could impact the power and reliability of their
¹²³ looking-while-listening design, a large-scale database would allow them to simulate possible
¹²⁴ outcomes across a range of test trials, based on past eye-tracking data with infants.

¹²⁵ [add paragraph about reproducibility?]

¹²⁶ **Peekbank: An open database of developmental eye-tracking studies.**

¹²⁷ What many of these open challenges share is that they are difficult to address at the
¹²⁸ scale of a single research lab or in a single study. To address this challenge, we developed
¹²⁹ *Peekbank* a flexible and reproducible interface to an open database of developmental
¹³⁰ eye-tracking studies. The Peekbank project (a) collects a large set of eye-tracking datasets
¹³¹ 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 a number of 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

¹⁵⁸ peekbankr package (R) provides an application programming interface, or API, that offers
¹⁵⁹ high-level abstractions for accessing the tabular data stored in Peekbank. Most users will
¹⁶⁰ access data through this final package, in which case the details of data formatting,
¹⁶¹ processing, and the specifics of connecting to the database are abstracted away from the user.

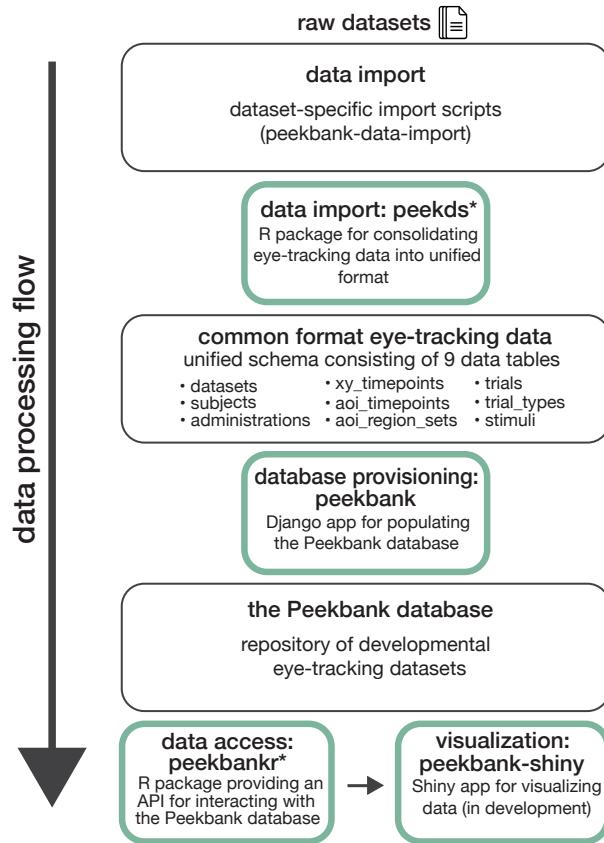


Figure 1. Overview of the Peekbank data ecosystem. Peekbank tools are highlighted in green.
 * indicates R packages introduced in this work.

¹⁶² Database Schema

¹⁶³ The Peekbank database contains two major types of data: (1) metadata regarding
¹⁶⁴ experiments, participants, and trials, and (2) time course looking data, detailing where on
¹⁶⁵ the screen a child is looking at a given point in time (Fig. 2).

¹⁶⁶ **Metadata.** Metadata can be separated into four parts: (1) participant-level
¹⁶⁷ information (e.g., demographics) (2) experiment-level information (e.g., the type of eye
¹⁶⁸ tracker used to collect the data) (3) session information (e.g. a participant's age for a specific

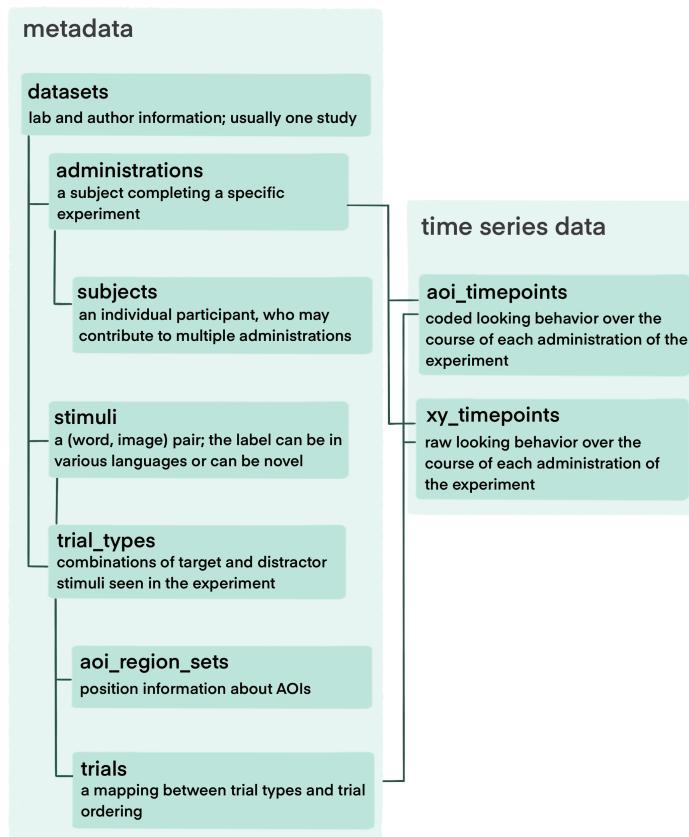


Figure 2. The Peekbank schema. Each square represents a table in the relational database.

169 experimental session) and (4) trial information (e.g., what images or videos were presented
170 onscreen, and paired with which audio).

171 ***Participant Information.***

172 Invariant information about individuals who participate in one or more studies (e.g., a
173 subject's first language) is recorded in the **subjects** table, while the **administrations**
174 table contains information about a subject's participation in a single session of a study (see
175 Session Information, below). This division allows Peekbank to gracefully handle longitudinal
176 designs: a single subject can be associated with many administrations.

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

180 ***Experiment Information.***

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

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

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

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

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 particular point in the experiment. We track the latter in the ‘trials’ table.

203 ***Session Information.***

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

208 ***Trial Information.***

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

212 **Time course data.** Raw looking data is a series of looks to AOIs or to (x, y)
213 coordinates on the experiment screen, linked to points in time. For data generated by
214 eye-trackers, we typically have (x, y) coordinates at each time point, which will be encoded
215 in the `xy_timepoints` table. These looks will also be recoded into AOIs according to the
216 AOI coordinates in the `aoi_region_sets` table using the `add_aois()` function in `peekds`,
217 which will be encoded in the `aoi_timepoints` table. For hand-coded data, we typically have
218 a series of AOIs; these will be recoded into the categories in the Peekbank schema (target,
219 distractor, other, and missing) and encoded in the `aoi_timepoints` table, and these
220 datasets will not have an `xy_timepoints` table.

221 Typically, timepoints in the `xy_timepoints` table and `aoi_timepoints` table need to
222 be regularized to center each trial's time around the point of disambiguation—such that 0 is
223 the time of target word onset in the trial (i.e., the beginning of *dog* in “Can you find the
224 *dog*?”). If time values run throughout the experiment rather than resetting to zero at the
225 beginning of each trial, `rezero_times()` is used to reset the time at each trial. After this,
226 each trial's times are centered around the point of disambiguation using `normalize_times()`.
227 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()` 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. Compared to linear interpolation (see e.g. Wass et al., 2014), constant interpolation has the advantage that it does not introduce new look locations, so it is a more conservative method of resampling.

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.

²⁵² **CHECK and edit resampling section for ties, interpolating forward/back in
²⁵³ time, and for maximum time over which we interpolate**

²⁵⁴ **Current Data Sources**

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
ft_pt	Adams et al., 2018	69	17.1	13–20	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
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
yoursmy	Garrison et al., 2020	35	14.5	12–18	eye-tracking	English

²⁵⁵ The database currently includes 15 looking-while-listening datasets comprising
²⁵⁶ $N=1320$ total participants (Table 1). Most datasets (12 out of 15 total) consist of data from
²⁵⁷ monolingual native English speakers. They span a wide age spectrum with participants
²⁵⁸ ranging from 9 to 70 months of age, and are balanced in terms of gender (46% female). The
²⁵⁹ datasets vary across a number of design-related dimensions, and include studies using
²⁶⁰ manually coded video recordings and automated eye-tracking methods (e.g., Tobii, EyeLink)
²⁶¹ to measure gaze behavior. All studies tested familiar items, but the database also includes 5
²⁶² datasets that tested novel pseudo-words in addition to familiar words. All data are openly
²⁶³ available on the Open Science Framework (<https://osf.io/pr6wu/>).

²⁶⁴ How selected? Language coverage? More details about lab and design variation?

²⁶⁵ **Versioning + Expanding the database**

²⁶⁶ The content of Peekbank will change as we add additional datasets and revise previous
²⁶⁷ ones. To facilitate reproducibility of analyses, we use a versioning system where successive

268 releases are assigned a name reflecting the year and version, e.g., 2021.1. By default, users
269 will interact with the most recent version of the database available, though `peekbankr` API
270 allows researchers to run analyses against any previous version of the database. For users
271 with intensive use-cases, each version of the database may be downloaded as a compressed
272 .sql file and installed on a local MySQL server.

273 **Interfacing with peekbank**

274 **Shiny App**

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

281 Specifically, users can visualize

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

288 Users are given various customization options for each of these visualizations, e.g.,
289 choosing which datasets to include in the plots, controlling the age range of participants,
290 splitting the visualizations by age bins, and controlling the analysis window for time course
291 analyses. Plots are then updated in real time to reflect users' customization choices, and
292 users are given options to share the visualizations they created. The Shiny app thus allows

293 users to quickly inspect basic properties of Peekbanks datasets and create reproducible
294 visualizations without incurring any of the technical overhead required to access the
295 database through R.

296 **Peekbankr**

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

302 Functions:

- 303 • `connect_to_peekbank()` opens a connection with the Peekbank database to allow
304 tables to be downloaded with the following functions
- 305 • `get_datasets()` gives each dataset name and its citation information
- 306 • `get_subjects()` gives information about persistent subject identifiers (e.g., native
307 languages, sex)
- 308 • `get_administrations()` gives information about specific experimental
309 administrations (e.g., subject age, monitor size, gaze coding method)
- 310 • `get_stimuli()` gives information about word–image pairings that appeared in
311 experiments
- 312 • `get_trial_types()` gives information about pairings of stimuli that appeared in the
313 experiment (e.g., point of disambiguation, target and distractor stimuli, condition,
314 language)
- 315 • `get_trials()` gives the trial orderings for each administration, linking trial types to
316 the trial IDs used in time course data
- 317 • `get_aoi_region_sets()` gives coordinate regions for each area of interest (AOI)

318 linked to trial type IDs

- 319 • `get_xy_timepoints()` gives time course data for each subject's looking behavior in
- 320 each trial, as (x, y) coordinates on the experiment monitor
- 321 • `get_aoi_timepoints()` gives time course data for each subject's looking behavior in
- 322 each trial, coded into areas of interest

323 **OSF site**

324 Stimuli Data in raw format (if some additional datum needed, e.g. pupil size?)

325 **Peekbank: General Descriptives**

326 [Accuracy, Reaction Times, Item variability?]

327 **Overall Word Recognition Accuracy**

Dataset Name	Unique Items	Prop.	Target	95% CI
attword	6	0.63	[0.61, 0.64]	
canine	16	0.64	[0.61, 0.67]	
coartic	10	0.70	[0.67, 0.73]	
cowpig	12	0.60	[0.58, 0.63]	
ft_pt	8	0.64	[0.63, 0.66]	
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	0.62	[0.58, 0.66]	
salientme	16	0.73	[0.71, 0.75]	
switchingCues	40	0.77	[0.75, 0.79]	
tablet	24	0.63	[0.6, 0.67]	
tseltal	30	0.59	[0.54, 0.63]	
yoursmy	87	0.60	[0.56, 0.64]	

Table 2

Average proportion target looking in each dataset.

328 In general, participants demonstrated robust, above-chance word recognition in each
 329 dataset (chance=0.5). Table 2 shows the average proportion of target looking within a
 330 standard critical window of 367-2000ms after the onset of the label for each dataset
 331 (**Swingley2000?**). Proportion target looking was generally higher for familiar words ($M =$

³³² 0.66, 95% CI = [0.65, 0.67], $n = 1269$) than for novel words learned during the experiment

³³³ ($M = 0.59$, 95% CI = [0.58, 0.61], $n = 822$).

³³⁴ **Item-level variability**

³³⁵ Figure 3 gives an overview of the variability in accuracy for individual words in each
³³⁶ dataset. The number of unique target labels and their associated accuracy vary widely
³³⁷ across datasets.

³³⁸ **Peekbank in Action**

³³⁹ We provide two potential use-cases for Peekbank data. In each case, we provide sample
³⁴⁰ code so as to model how easy it is to do simple analyses using data from the database. Our
³⁴¹ first example shows how we can replicate the analysis for a classic study. This type of
³⁴² computational reproducibility can be a very useful exercise for teaching students about best
³⁴³ practices for data analysis (e.g., Hardwicke et al., 2018) and also provides an easy way to
³⁴⁴ explore looking-while-listening time course data in a standardized format. Our second
³⁴⁵ example shows an in-depth exploration of developmental changes in the recognition of
³⁴⁶ particular words. Besides its theoretical interest (which we will explore more fully in
³⁴⁷ subsequent work), this type of analysis could in principle be used for optimizing the stimuli
³⁴⁸ for new experiments, especially as the Peekbank dataset grows and gains coverage over a
³⁴⁹ greater number of items.

³⁵⁰ **Computational reproducibility example: Swingley and Aslin (2000)**

³⁵¹ Swingley and Aslin (2000) investigated the specificity of 14-16 month-olds' word
³⁵² representations using the looking-while-listening paradigm, asking whether recognition would
³⁵³ be slower and less accurate for mispronunciations, e.g. "oppel" (close mispronunciation) or
³⁵⁴ "opel" (distant mispronunciation) instead of "apple" (correct pronunciation). In this short
³⁵⁵ vignette, we show how easily the data in Peekbank can be used to visualize this result.

```
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")
```

356 We begin by retrieving the relevant tables from the database, `aoi_timepoints`,
 357 `administrations`, `trial_types`, and `trials`. As discussed above, each of these can be
 358 downloaded using a simple API call through `peekbankr`, which returns dataframes that
 359 include ID fields. These ID fields allow for easy joining of the data into a single dataframe
 360 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"))
```

361 As the code above shows, once the data are joined, condition information for each
 362 timepoint is present and so we can easily filter out filler trials and set up the conditions for
 363 further analysis. For simplicity, here we combine both mispronunciation conditions since the
 364 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()))
```

365 The final step in our analysis is to create a summary dataframe using `dplyr`
 366 commands. We first group the data by timestep, participant, and condition and compute the
 367 proportion looking at the correct image. We then summarize again, averaging across

368 participants, computing both means and 95% confidence intervals (via the approximation of
 369 1.96 times the standard error of the mean). The resulting dataframe can be used for
 370 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))
```

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

375 Item analyses

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

381 Our first step is to collect and join the data from the relevant tables including
 382 timepoint data, trial and stimulus data, and administration data (for participant ages). We
 383 join these into a single dataframe for easy manipulation; this dataframe is a common
 384 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)

```

385 Next we select a set of four target words (chosen based on having more than XXX
 386 children contributing data for each across several one-year age groups). We create age
 387 groups, aggregate, and compute timepoint-by-timepoint confidence intervals using the z
 388 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

404 building a cumulative developmental science (Bergmann et al. (2018); Michael C. Frank et
405 al. (2017); The ManyBabies Consortium (2020)]. The Peekbank project expands on these
406 efforts by building an infrastructure for aggregating eye-tracking data across studies, with a
407 specific focus on the looking-while-listening paradigm. This paper presents an illustration of
408 some of the key theoretical and methodological questions that can be addressed using
409 Peekbank: generalizing across item-level variability in children’s word recognition and
410 providing data-driven guidance on methodological choices.

411 There are a number of limitations surrounding the current scope of the database. A
412 priority in future work will be to expand the size of the database. With 11 datasets currently
413 available in the database, idiosyncrasies of particular designs and condition manipulations
414 still have substantial influence on modeling results. Expanding the set of distinct datasets
415 will allow us to increase the number of observations per item across datasets, leading to more
416 robust generalizations across item-level variability. The current database is also limited by
417 the relatively homogeneous background of its participants, both with respect to language
418 (almost entirely monolingual native English speakers) and cultural background (all but one
419 dataset come from WEIRD populations, potentially limiting generalizability; see
420 Muthukrishna et al. (2020)). Increasing the diversity of participant backgrounds and
421 languages will expand the scope of the generalizations we can form about child word
422 recognition.

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

429

Acknowledgements

430 We would like to thank the labs and researchers that have made their data publicly

431 available in the database.

432

References

- 433 Bergelson, E. (2020). The comprehension boost in early word learning: Older infants
 434 are better learners. *Child Development Perspectives*, 14(3), 142–149.
- 435 Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the
 436 meanings of many common nouns. *PNAS*, 109(9), 3253–3258.
- 437 Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C.,
 438 & Cristia, A. (2018). Promoting replicability in developmental research through
 439 meta-analyses: Insights from language acquisition research. *Child Development*,
 440 89(6), 1996–2009.
- 441 Bleses, D., Makransky, G., Dale, P. S., Højen, A., & Ari, B. A. (2016). Early
 442 productive vocabulary predicts academic achievement 10 years later. *Applied
 443 Psycholinguistics*, 37(6), 1461–1476.
- 444 Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2021). Six solutions for more reliable
 445 infant research. *PsyArXiv*. <https://doi.org/https://doi.org/10.31234/osf.io/ksfvq>
- 446 Fernald, A., Pinto, J. P., Swingley, D., Weinberg, A., & McRoberts, G. W. (1998).
 447 Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological
 448 Science*, 9(3), 228–231.
- 449 Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while
 450 listening: Using eye movements to monitor spoken language comprehension by
 451 infants and young children. In I. A. Sekerina, E. M. Fernandez, & H. Clahsen
 452 (Eds.), *Developmental psycholinguistics: On-line methods in children's language
 453 processing* (pp. 97–135). Amsterdam: John Benjamins.
- 454 Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ...
 455 Yurovsky, D. (2017). A Collaborative Approach to Infant Research: Promoting
 456 Reproducibility, Best Practices, and Theory-Building. *Infancy*, 22(4), 421–435.
 457 <https://doi.org/10.1111/infa.12182>
- 458 Frank, Michael C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017).

- 459 Wordbank: An open repository for developmental vocabulary data. *Journal of*
460 *Child Language*, 44(3), 677–694.
- 461 Frank, Michael C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021).
462 *Variability and Consistency in Early Language Learning: The Wordbank Project*.
463 Cambridge, MA: MIT Press.
- 464 Golinkoff, R. M., Ma, W., Song, L., & Hirsh-Pasek, K. (2013). Twenty-five years
465 using the intermodal preferential looking paradigm to study language acquisition:
466 What have we learned? *Perspectives on Psychol. Science*, 8(3), 316–339.
- 467 Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C.,
468 Kidwell, M. C., ... Frank, M. C. (2018). Data availability, reusability, and
469 analytic reproducibility: Evaluating the impact of a mandatory open data policy
470 at the journal *Cognition*. *Royal Society Open Science*, 5(8).
471 <https://doi.org/10.1098/rsos.180448>
- 472 Hirsh-Pasek, K., Cauley, K. M., Golinkoff, R. M., & Gordon, L. (1987). The eyes
473 have it: Lexical and syntactic comprehension in a new paradigm. *Journal of Child*
474 *Language*, 14(1), 23–45.
- 475 Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid
476 use of grammatical gender in spoken word recognition. *Psychological Science*,
477 18(3), 193–198.
- 478 Marchman, V. A., Loi, E. C., Adams, K. A., Ashland, M., Fernald, A., & Feldman, H.
479 M. (2018). Speed of language comprehension at 18 months old predicts
480 school-relevant outcomes at 54 months old in children born preterm. *Journal of*
481 *Dev. & Behav. Pediatrics*, 39(3), 246–253.
- 482 Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A.,
483 McInerney, J., & Thue, B. (2020). Beyond Western, Educated, Industrial, Rich,
484 and Democratic (WEIRD) Psychology: Measuring and Mapping Scales of
485 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/>
- 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.
- 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.

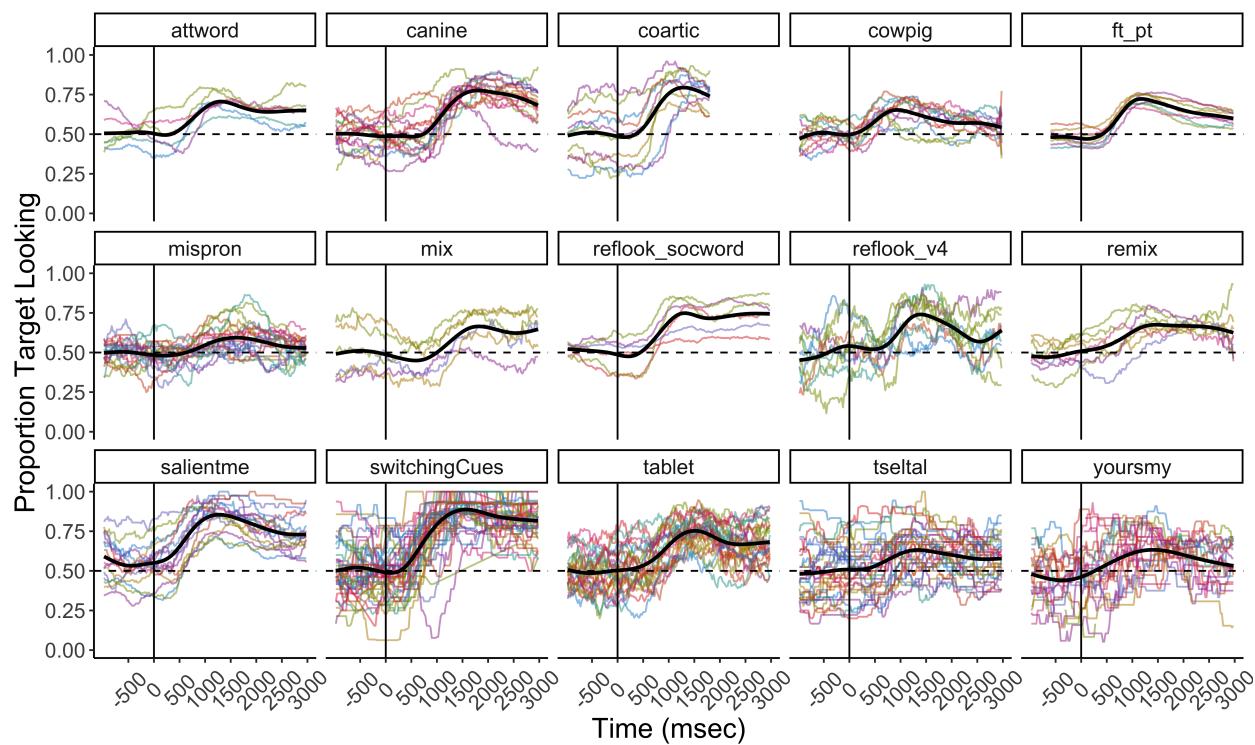


Figure 3. Item-level variability in proportion target looking within each dataset (chance=0.5). Time is centered on the onset of the target label (vertical line). Colored lines represent specific target labels. Black lines represent smoothed average fits based on a general additive model using cubic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

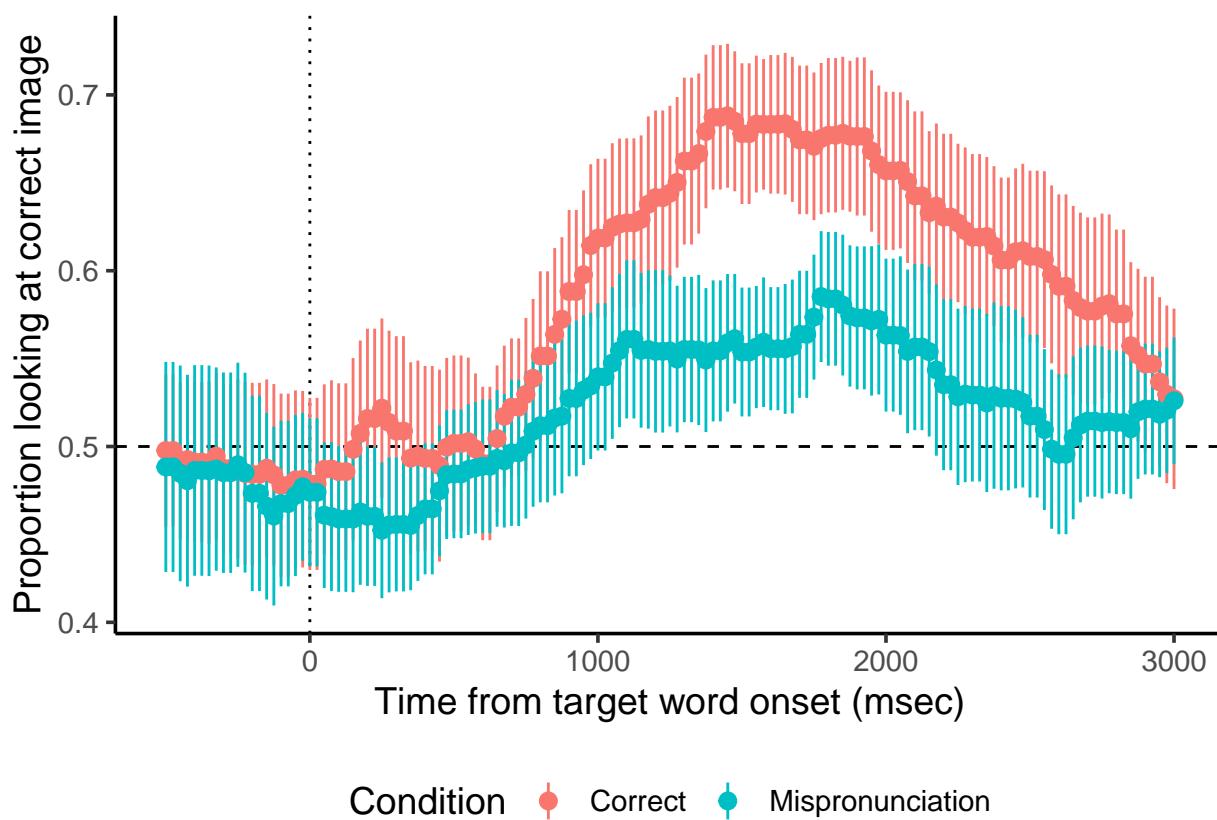


Figure 4. Proportion looking at the correct referent by time from the point of disambiguation (the onset of the target noun). Colors show the two pronunciation conditions; points give means and ranges show 95% confidence intervals. The dotted line shows the point of disambiguation and the dashed line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

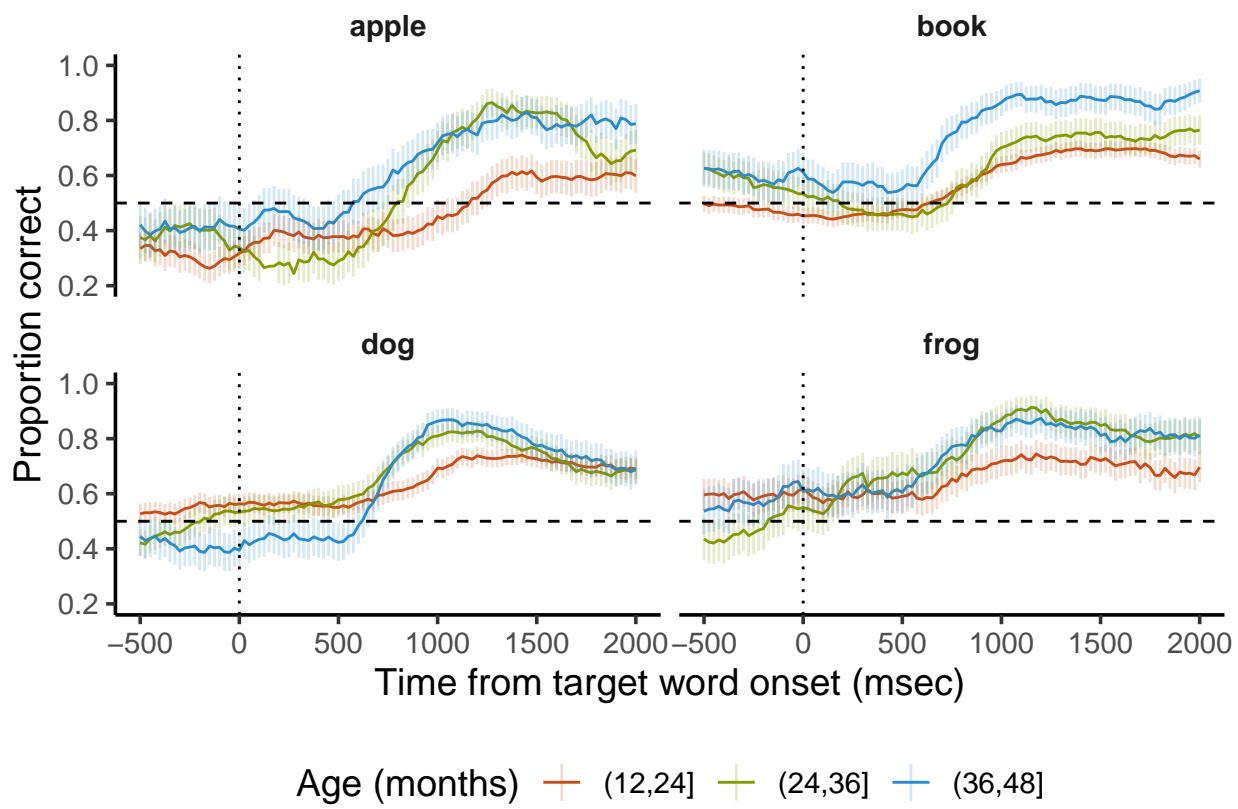


Figure 5. Add caption here.