

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 (Frank,
32 Braginsky, Yurovsky, & Marchman, 2021). Although many children will only produce their
33 first word at around one year of age, they show signs of understanding many common nouns
34 (e.g., "mommy") and phrases (e.g., "Let's go bye-bye!") much earlier in development
35 (Bergelson & Swingley, 2012). However, the processes involved in early word understanding
36 are less directly apparent in children's behaviors and are less accessible to observation than
37 developments in speech production (Fernald, Zangl, Portillo, & Marchman, 2008). To
38 understand speech, children must process the incoming auditory signal and link that signal
39 to relevant meanings – a process often referred to as word recognition. Measuring early word
40 recognition offers insight into children's early word representations and as well as the speed
41 and efficiency with which children comprehend language in real time, as the speech signal
42 unfolds (Bergelson, 2020; Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998). Word
43 recognition skills are also thought to build a foundation for children's subsequent language
44 development. Past research has found that early word recognition efficiency is predictive of
45 later linguistic and general cognitive outcomes (Bleses, Makransky, Dale, Højen, & Ari, 2016;
46 Marchman et al., 2018). One explanation for this relationship is that efficiency of word
47 recognition facilitates subsequent word learning: the faster children are at processing speech,
48 the more efficiently they can learn from the input in their environment (Fernald &
49 Marchman, 2012).

50 While word recognition is a central part of children's language development, mapping
51 the trajectory of word recognition skills has remained elusive. Studies investigating children's
52 word recognition are typically limited in scope to experiments in individual labs involving
53 small samples tested on a small set of items. This limitation makes it difficult to understand
54 developmental changes in children's word knowledge at a broad scale. Peekbank provides an

55 openly accessible database of eye-tracking data of children's word recognition, with the
56 primary goal of facilitating the study of developmental changes in children's word knowledge
57 and recognition speed.

58 **The “Looking-While-Listening” Paradigm**

59 Word recognition is traditionally studied in the “looking-while-listening” paradigm
60 (alternatively referred to as the intermodal preferential looking procedure; Fernald et al.,
61 2008; Hirsh-Pasek, Cauley, Golinkoff, & Gordon, 1987). In such studies, infants listen to a
62 sentence prompting a specific referent (e.g., “Look at the dog!”) while viewing two images on
63 the screen (e.g., an image of a dog – the target image – and an image of a bird – the
64 distractor image). Infants' word recognition is measured in terms of how quickly and
65 accurately they fixate on the correct target image after hearing its label. Past research has
66 used this same basic method to study a wide range of questions in language development.
67 For example, the looking-while-listening paradigm has been used to investigate early noun
68 knowledge, phonological representations of words, prediction during language processing, and
69 individual differences in language development (Bergelson & Swingley, 2012; Golinkoff, Ma,
70 Song, & Hirsh-Pasek, 2013; Lew-Williams & Fernald, 2007; Marchman et al., 2018; D.
71 Swingley & Aslin, 2000a).

72 **TO DO: ALIGN CHALLENGES WITH tidybits/ use cases - computational
73 reproducibility and teaching - item-level analyses**

74 **Measuring developmental change in word recognition**

75 While the looking-while-listening paradigm has been highly fruitful in advancing
76 understanding of early word knowledge, fundamental questions remain. One central question
77 is how to accurately capture developmental change in the speed and accuracy of word

78 recognition. There is ample evidence demonstrating that infants get faster and more
79 accurate in word recognition over the first few years of life (e.g., Fernald et al., 1998).
80 However, precisely measuring developmental increases in the speed and accuracy of word
81 recognition remains challenging due to the difficulty of distinguishing developmental changes
82 in word recognition skill from changes in knowledge of specific words. This problem is
83 particularly thorny in studies with young children, since the number of items that can be
84 tested within a single session is limited and items must be selected in an age-appropriate
85 manner (Peter et al., 2019). One way to overcome this challenge is to measure word
86 recognition across development in a large-scale dataset with a wide range of items. A
87 sufficiently large dataset would allow researchers to estimate developmental change in word
88 recognition speed and accuracy while generalizing across changes related to specific words.

89 **Developing methodological best-practices**

90 A second question relates to evaluating methodological best practices. In particular,
91 many fundamental analytic decisions vary substantially across studies, and different decisions
92 may lead to researchers drawing different inferences about children's word recognition. For
93 example, researchers vary in how they select time windows for analysis, transform the
94 dependent measure of target fixations, and model the time course of word recognition
95 (Csibra, Hernik, Mascaro, Tatone, & Lengyel, 2016; Fernald et al., 2008; Huang & Snedeker,
96 2020). This problem is made more complex by the fact that many of these decisions depend
97 on a variety of design-related and participant-related factors (e.g., infant age). Establishing
98 best practices therefore requires a large database of infant word recognition studies varying
99 across such factors, in order to test the potential consequences of methodological decisions
100 on study results.

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

102 What these two questions share is that they are difficult to answer at the scale of a
103 single study. To address this challenge, we introduce Peekbank, a flexible and reproducible
104 interface to an open database of developmental eye-tracking studies. The Peekbank project
105 (a) collects a large set of eye-tracking datasets on children’s word recognition, (b) introduces
106 a data format and processing tools for standardizing eye-tracking data across data sources,
107 and (c) provides an interface for accessing and analyzing the database. In the current paper,
108 we give an overview of the key components of the project and some initial demonstrations of
109 its utility in advancing theoretical and methodological insights. We report two analyses
110 using the database and associated tools ($N=1,233$): (1) a growth curve analysis modeling
111 age-related changes in infants’ word recognition while generalizing across item-level
112 variability; and (2) a multiverse-style analysis of how a central methodological decision –
113 selecting the time window of analysis – impacts inter-item reliability.

114 Design and Technical Approach**115 Database Framework**

116 One of the main challenges in compiling a large-scale eye-tracking dataset is the lack of
117 a shared data format across individual experiments. Researcher conventions for structuring
118 data vary, as do the technical specifications of different devices (e.g., computer displays and
119 eyetracking cameras), rendering the task of integrating datasets from different labs and data
120 sources difficult. Therefore, our first effort was to develop a common tabular format to
121 support analyses of all studies simultaneously.

122 As illustrated in Figure 1, the Peekbank framework consists of four main components:
123 (1) a set of tools to convert eye-tracking datasets into a unified format, (2) a relational

124 database populated with data in this unified format, (3) a set of tools to retrieve data from
125 this database, and (4) a web app (using the Shiny framework) for visualizing the data. These
126 components are supported by three libraries. The `peekds` library (for the R language; R
127 Core Team (2020)) helps researchers convert existing datasets to use the standardized format
128 of the database. The `peekbank` module (Python) creates a database with the relational
129 schema and populates it with the standardized datasets produced by `peekds`. The database
130 is implemented in MySQL, an industry standard relational database, which may be accessed
131 by a variety of programming languages, and can be hosted on one machine and accessed by
132 many others over the Internet. The `peekbankr` library (R) provides an application
133 programming interface, or API, that offers high-level abstractions for accessing the tabular
134 data stored in Peekbank. Most users will access data through this final library, in which case
135 the details of data formatting and processing are abstracted away from the user.

136 In the following sections, we will begin by providing the details on the database's
137 organization (or *schema*) and the technical implementation on `peekds`. Users who are
138 primarily interested in accessing the database can skip these details and focus on access
139 through the `peekbankr` API and the web apps.

140 Database Schema

141 The `peekbank` database contains two major types of data: (1) metadata regarding the
142 relevant experiment, participant, and trial, and (2) timecourse looking data, detailing where
143 on the screen a child is looking at a given point in time (Fig. 2).

144 Here, we will give an outline of the tables encoding this data. As is common in
145 relational databases, records of similar types (e.g., participants, trials, experiments, coded
146 looks at each timepoint) are grouped into tables, and records of various types are linked
147 through numeric identifiers.

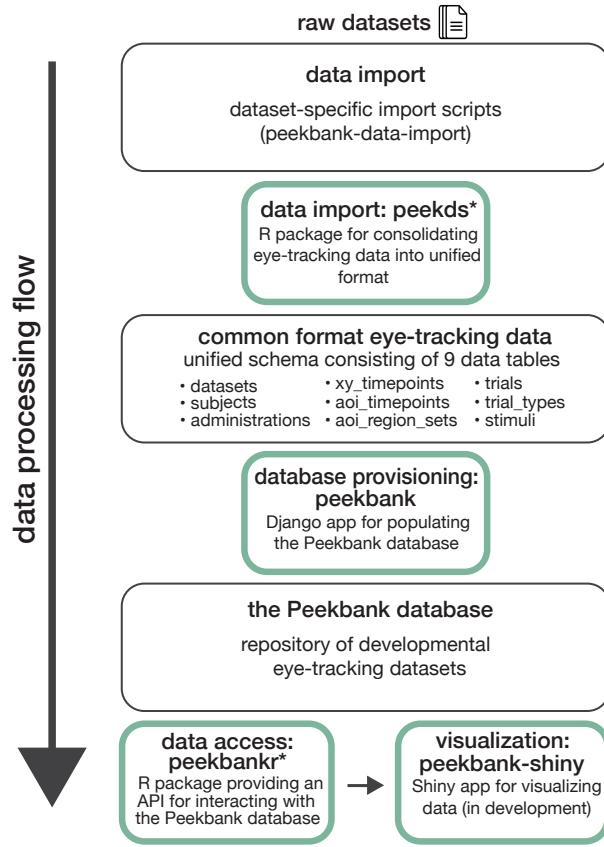


Figure 1. Overview of the Peekbank data ecosystem. Peekbank tools are highlighted in green.
* indicates R packages introduced in this work.

148 **Metadata.** We encode the metadata available for each study and participation in
 149 the database. This metadata can be separated into three parts: (1) subject-level information
 150 (e.g., demographics) (2) experiment-level information (e.g., a subject's age for a specific
 151 experiment, or the particular eyetracker used to collect the data) and (3) trial information
 152 and experimental design (e.g., what images or videos were presented onscreen, and paired
 153 with which audio). Information about individuals who participate in one or more studies
 154 (e.g., a subject's sex and first language), is recorded in the `subjects` table, while the
 155 `administrations` table contains information about a subject's participation in a single
 156 administration of a study (e.g., a subject's age of participation or the eyetracker that was
 157 used). This division allows Peekbank to gracefully handle longitudinal designs: a single
 158 subject can be associated with many administrations.

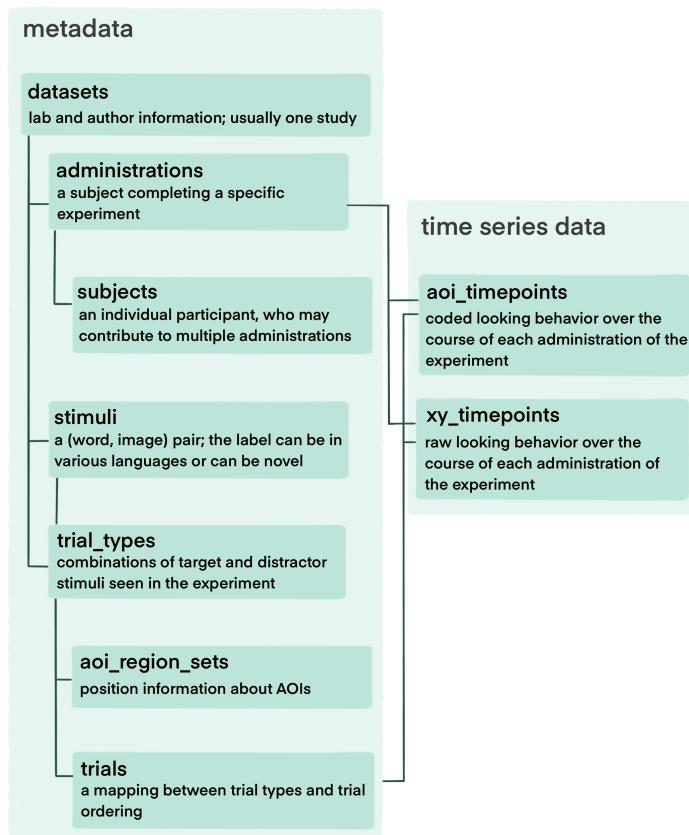


Figure 2. The Peekbank schema. Each square represents a table in the relational database.

159 The `stimuli` and `trial_types` tables store information about trials, which in turn
 160 may reflect specifics of the experiment design. Stimuli are (label, image) mappings that are
 161 seen in the experiment. The `trial_types` table encodes information about each trial of the
 162 experiment, including the target stimulus and location, the distractor stimulus and location,
 163 and the point of disambiguation for that trial. If this dataset used automatic eyetracking
 164 rather than manual coding, each trial type is additionally linked to a set of area of interest
 165 (x, y) coordinates, encoded in the `aoi_region_sets` table.

166 Because individual trial types can be repeated multiple times within an administration,
 167 the order of the trials is encoded in the `trials` table. Each unique ordering that occurred in
 168 the experiment is encoded in this table. The `trial_id`, which links a trial type to the order
 169 it was presented in an administration, is attached to the timecourse looking data.

170 **Timecourse data.** Timecourse looking data is encoded in two tables:

171 `aoi_timepoints` and `xy_timepoints`. The `aoi_timepoints` table encodes where a child is
172 looking at each point in time, by specifying the coded area of interest (AOI): looks to the
173 target, looks to the distractor, looks on the screen but away from target and distractor, and
174 missing looks. All datasets must include this timecourse data, as it represents the main
175 record of children’s looking behavior. For eyetracking experiments that are automatically
176 rather than manually coded, the `xy_timepoints` table additionally encodes the inferred (x,
177 y) coordinates of fixations on the screen over the course of each trial. Both the
178 `aoi_timepoints` and `xy_timepoints` tables are resampled to a consistent sampling rate, as
179 described in the Import section below. To normalize across trials and across experiments, all
180 timecourses are computed so that the time of 0 ms represents the onset of disambiguating
181 material (i.e., the beginning of *dog* in “Can you find the *dog*?”).

182 **Import**

183 During data import, raw eye-tracking datasets are processed to conform to the
184 Peekbank data schema. The following section is a description of the import process for
185 Peekbank. It serves as both a description of our method in importing the datasets already in
186 the database, as well as a high-level overview of the import process for researchers looking to
187 import their data in the future. First, we will describe the import of metadata, and second,
188 we will describe import of the timecourse looking data, including processing functions in
189 `peekds` for normalizing and resampling looking behavior.

190 **Metadata.** Subject-level data is imported for all participants who have experiment
191 data. In general, we import data without particular exclusions, including as many
192 participants as possible in the database. The `subjects` and `administrations` tables
193 separate information at the subject level from information about runs of the experiment,
194 such that longitudinal studies have multiple administrations linked to each subject.

The `stimuli` table has a row for each (word, image) pair, and thus is used slightly differently across different experiment designs. 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.

The `trial_types` table contains each pair of stimuli, a target and distractor, seen in the experiment. The `trial_types` table links trial types to the `aoi_region_sets` table and the `trials` table.

The `trials` table encodes each unique ordering of trial types seen in all runs of an experiment. For example, for experiments with a fixed trial order, the `trials` table will have as many rows as there are stimuli in the experiment; for experiments with a randomized trial order, there will be many rows linking the trial orderings to the trial types. The `trials` table links all experiment design information to the timecourse data.

Timecourse 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 eyetrackers, 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

221 datasets will not have an `xy_timepoints` table.

222 Typically, timepoints in the `xy_timepoints` table and `aoi_timepoints` table need to
223 be regularized to center each trial's time around the point of disambiguation—the time of
224 target word onset in the trial. If time values run throughout the experiment rather than
225 resetting to zero at the beginning of each trial, `rezero_times()` is used to reset the time at
226 each trial. After this, each trial's times are centered around the point of disambiguation
227 using `normalize_times()`. When these steps are complete, the time course is ready for
228 resampling.

229 To facilitate time course analysis and visualization across datasets, timecourse data
230 must be resampled to a uniform sampling rate (i.e., such that every trial in every dataset has
231 observations at the same time points). To do this, we use the `resample()` function. During
232 the resampling process, we interpolate using constant interpolation, selecting for each
233 interpolated timepoint the looking location for the nearest observed time point in the
234 original data for both `aoi_timepoints` and `xy_timepoints` data. Compared to linear
235 interpolation (see e.g. Wass et al., 2014), constant interpolation has the advantage that it
236 does not introduce new look locations, so it is a more conservative method of resampling.

237 Validation and ingestion into the database

238 After resampling, the final step of dataset import is validation. The `peekds` package
239 offers functions to check the now processed data tables against the database schema to
240 ensure that all tables have the required fields and correct data types for database ingestion.
241 In an effort to double check the data quality and to make sure that no errors are made in the
242 importing script, as part of the import procedure we create a timecourse plot based on our
243 processed tables to replicate the results in the original paper. Once this plot has been
244 created and checked for consistency and all tables pass our validation functions, the

²⁴⁵ processed dataset is ready for ingestion into the database.

²⁴⁶ Currently, the import process is carried out by the Peekbank team using data offered
²⁴⁷ by other research teams. In the future, we hope to allow research teams to carry out their
²⁴⁸ own import processes with checks from the Peekbank team before ingestion. To this end,
²⁴⁹ import script templates are available for both hand-coded datasets and automatic
²⁵⁰ eyetracking datasets for research teams to adapt to their data.

²⁵¹ **CHECK and edit resampling section for ties, interpolating forward/back in
²⁵² time, and for maximum time over which we interpolate**

²⁵³ 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

²⁵⁴ The database currently includes 15 looking-while-listening datasets comprising
²⁵⁵ $N=1320$ total participants (Table 1). Most datasets (12 out of 15 total) consist of data from
²⁵⁶ monolingual native English speakers. They span a wide age spectrum with participants
²⁵⁷ ranging from 9 to 70 months of age, and are balanced in terms of gender (46% female). The
²⁵⁸ datasets vary across a number of design-related dimensions, and include studies using
²⁵⁹ manually coded video recordings and automated eye-tracking methods (e.g., Tobii, EyeLink)

260 to measure gaze behavior. All studies tested familiar items, but the database also includes 5
261 datasets that tested novel pseudo-words in addition to familiar words. All data are openly
262 available on the Open Science Framework (<https://osf.io/pr6wu/>).

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

264 **Versioning + Expanding the database**

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

272 **Interfacing with peekbank**

273 **Shiny App**

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

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

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

295 **Peekbankr**

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

301 Functions:

- 302 • `connect_to_peekbank()`
- 303 • `get_datasets()`
- 304 • `get_subjects()`

- 305 • get_administrations()
- 306 • get_stimuli()
- 307 • get_aoi_timepoints()
- 308 • get_trials()
- 309 • get_trial_types()
- 310 • get_xy_timepoints()
- 311 • get_aoi_region_sets()

312 **OSF site**

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

314 **Peekbank: General Descriptives**

315 [Accuracy, Reaction Times, Item variability?]

316 **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.

317 In general, participants demonstrated robust, above-chance word recognition in each
318 dataset (chance=0.5). Table 2 shows the average proportion of target looking within a
319 standard critical window of 367-2000ms after the onset of the label for each dataset (D.
320 Swingley & Aslin, 2000b). Proportion target looking was generally higher for familiar words
321 ($M = 0.66$, 95% CI = [0.65, 0.67], $n = 1269$) than for novel words learned during the
322 experiment ($M = 0.59$, 95% CI = [0.58, 0.61], $n = 822$).

323 **Item-level variability**

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

327 **Peekbank in Action**

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

339 Computational reproducibility example: D. Swingley and Aslin (2000a)

340 D. Swingley and Aslin (2000a) investigated the specificity of 14-16 month-olds' word
 341 representations using the looking-while-listening paradigm, asking whether recognition would
 342 be slower and less accurate for mispronunciations, e.g. "oppel" (close mispronunciation) or
 343 "opel" (distant mispronunciation) instead of "apple" (correct condition). In this short
 344 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")
```

345 We begin by retrieving the relevant tables from the database, `aoi_timepoints`,
 346 `administrations`, `trial_types`, and `trials`. As discussed above, each of these can be
 347 downloaded using a simple API call through `peekbankr`, which returns dataframes that
 348 include ID fields. These ID fields allow for easy joining of the data into a single dataframe
 349 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"))
```

350 As the code above shows, once the data are joined, condition information for each
 351 timepoint is present and so we can easily filter out filler trials and set up the conditions for
 352 further analysis. For simplicity, here we combine both mispronunciation conditions since this
 353 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()))
```

354 The final step in our analysis is to create a summary dataframe using `dplyr`
 355 commands. We first group the data by timestep, participant, and condition and compute the
 356 proportion looking at the correct image. We then summarize again, averaging across
 357 participants, computing both means and 95% confidence intervals (via the approximation of
 358 1.96 times the standard error of the mean). The resulting dataframe can be used for
 359 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))
```

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

364 Item analyses

365 A second use case for Peekbank is to examine item-level variation in word recognition.
 366 Individual datasets rarely have enough statistical power to show reliable developmental

367 differences within items. To illustrate the power of aggregating data across multiple datasets,
368 we select the four words with the most data available across studies and ages (apple, book,
369 dog, and frog) and show average recognition trajectories.

370 Our first step is to collect and join the data from the relevant tables including
371 timepoint data, trial and stimulus data, and administration data (for participant ages). We
372 join these into a single dataframe for easy manipulation; this dataframe is a common
373 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)
```

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

389

Discussion and Conclusion

390 Theoretical progress in understanding child development requires rich datasets, but
 391 collecting child data is expensive, difficult, and time-intensive. Recent years have seen a
 392 growing effort to build open source tools and pool research efforts to meet the challenge of
 393 building a cumulative developmental science (Bergmann et al. (2018); Frank, Braginsky,
 394 Yurovsky, and Marchman (2017); The ManyBabies Consortium (2020)]. The Peekbank
 395 project expands on these efforts by building an infrastructure for aggregating eye-tracking
 396 data across studies, with a specific focus on the looking-while-listening paradigm. This paper
 397 presents an illustration of some of the key theoretical and methodological questions that can
 398 be addressed using Peekbank: generalizing across item-level variability in children’s word
 399 recognition and providing data-driven guidance on methodological choices.

400 There are a number of limitations surrounding the current scope of the database. A
 401 priority in future work will be to expand the size of the database. With 11 datasets currently
 402 available in the database, idiosyncrasies of particular designs and condition manipulations
 403 still have substantial influence on modeling results. Expanding the set of distinct datasets
 404 will allow us to increase the number of observations per item across datasets, leading to more
 405 robust generalizations across item-level variability. The current database is also limited by
 406 the relatively homogeneous background of its participants, both with respect to language

407 (almost entirely monolingual native English speakers) and cultural background (all but one
408 dataset come from WEIRD populations, potentially limiting generalizability; see
409 Muthukrishna et al. (2020)). Increasing the diversity of participant backgrounds and
410 languages will expand the scope of the generalizations we can form about child word
411 recognition.

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

418 **Acknowledgements**

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

421

References

- 422 Bergelson, E. (2020). The comprehension boost in early word learning: Older infants are
423 better learners. *Child Development Perspectives*, 14(3), 142–149.
- 424 Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of
425 many common nouns. *PNAS*, 109(9), 3253–3258.
- 426 Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., &
427 Cristia, A. (2018). Promoting replicability in developmental research through
428 meta-analyses: Insights from language acquisition research. *Child Development*,
429 89(6), 1996–2009.
- 430 Bleses, D., Makransky, G., Dale, P. S., Højen, A., & Ari, B. A. (2016). Early productive
431 vocabulary predicts academic achievement 10 years later. *Applied Psycholinguistics*,
432 37(6), 1461–1476.
- 433 Csibra, G., Hernik, M., Mascaro, O., Tatone, D., & Lengyel, M. (2016). Statistical treatment
434 of looking-time data. *Developmental Psychology*, 52(4), 521–536.
- 435 Fernald, A., & Marchman, V. A. (2012). Individual differences in lexical processing at 18
436 months predict vocabulary growth in typically developing and late-talking toddlers.
437 *Child Development*, 83(1), 203–222.
- 438 Fernald, A., Pinto, J. P., Swingley, D., Weinberg, A., & McRoberts, G. W. (1998). Rapid
439 gains in speed of verbal processing by infants in the 2nd year. *Psychological Science*,
440 9(3), 228–231.
- 441 Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening:
442 Using eye movements to monitor spoken language comprehension by infants and
443 young children. In I. A. Sekerina, E. M. Fernandez, & H. Clahsen (Eds.),

- 444 *Developmental psycholinguistics: On-line methods in children's language processing*
445 (pp. 97–135). Amsterdam: John Benjamins.
- 446 Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open
447 repository for developmental vocabulary data. *Journal of Child Language*, 44(3),
448 677–694.
- 449 Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and*
450 *Consistency in Early Language Learning: The Wordbank Project*. Cambridge, MA:
451 MIT Press.
- 452 Golinkoff, R. M., Ma, W., Song, L., & Hirsh-Pasek, K. (2013). Twenty-five years using the
453 intermodal preferential looking paradigm to study language acquisition: What have
454 we learned? *Perspectives on Psychol. Science*, 8(3), 316–339.
- 455 Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M.
456 C., . . . Frank, M. C. (2018). Data availability, reusability, and analytic
457 reproducibility: Evaluating the impact of a mandatory open data policy at the journal
458 Cognition. *Royal Society Open Science*, 5(8). <https://doi.org/10.1098/rsos.180448>
- 459 Hirsh-Pasek, K., Cauley, K. M., Golinkoff, R. M., & Gordon, L. (1987). The eyes have it:
460 Lexical and syntactic comprehension in a new paradigm. *Journal of Child Language*,
461 14(1), 23–45.
- 462 Huang, Y., & Snedeker, J. (2020). Evidence from the visual world paradigm raises questions
463 about unaccusativity and growth curve analyses. *Cognition*, 200, 104251.
- 464 Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of
465 grammatical gender in spoken word recognition. *Psychological Science*, 18(3),
466 193–198.

- 467 Marchman, V. A., Loi, E. C., Adams, K. A., Ashland, M., Fernald, A., & Feldman, H. M.
468 (2018). Speed of language comprehension at 18 months old predicts school-relevant
469 outcomes at 54 months old in children born preterm. *Journal of Dev. & Behav.*
470 *Pediatrics*, 39(3), 246–253.
- 471 Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J.,
472 & Thue, B. (2020). Beyond Western, Educated, Industrial, Rich, and Democratic
473 (WEIRD) Psychology: Measuring and Mapping Scales of Cultural and Psychological
474 Distance. *Psychological Science*, 31(6), 678–701.
- 475 Peter, M. S., Durrant, S., Jessop, A., Bidgood, A., Pine, J. M., & Rowland, C. F. (2019).
476 Does speed of processing or vocabulary size predict later language growth in toddlers?
477 *Cognitive Psychology*, 115, 101238.
- 478 R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna,
479 Austria: R Foundation for Statistical Computing. Retrieved from
480 <https://www.R-project.org/>
- 481 Swingley, D., & Aslin, R. N. (2000a). Spoken word recognition and lexical representation in
482 very young children. *Cognition*, 76(2), 147–166.
- 483 Swingley, D., & Aslin, R. N. (2000b). Spoken word recognition and lexical representation in
484 very young children. *Cognition*, 76(2), 147–166.
- 485 The ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research
486 using the infant-directed speech preference. *Advances in Methods and Practices in*
487 *Psychological Science*, 3(1), 24–52.
- 488 Zettersten, M., Bergey, C., Bhatt, N., Boyce, V., Braginsky, M., Carstensen, A., ... others.
489 (2021). Peekbank: Exploring children's word recognition through an open, large-scale
490 repository for developmental eye-tracking data.

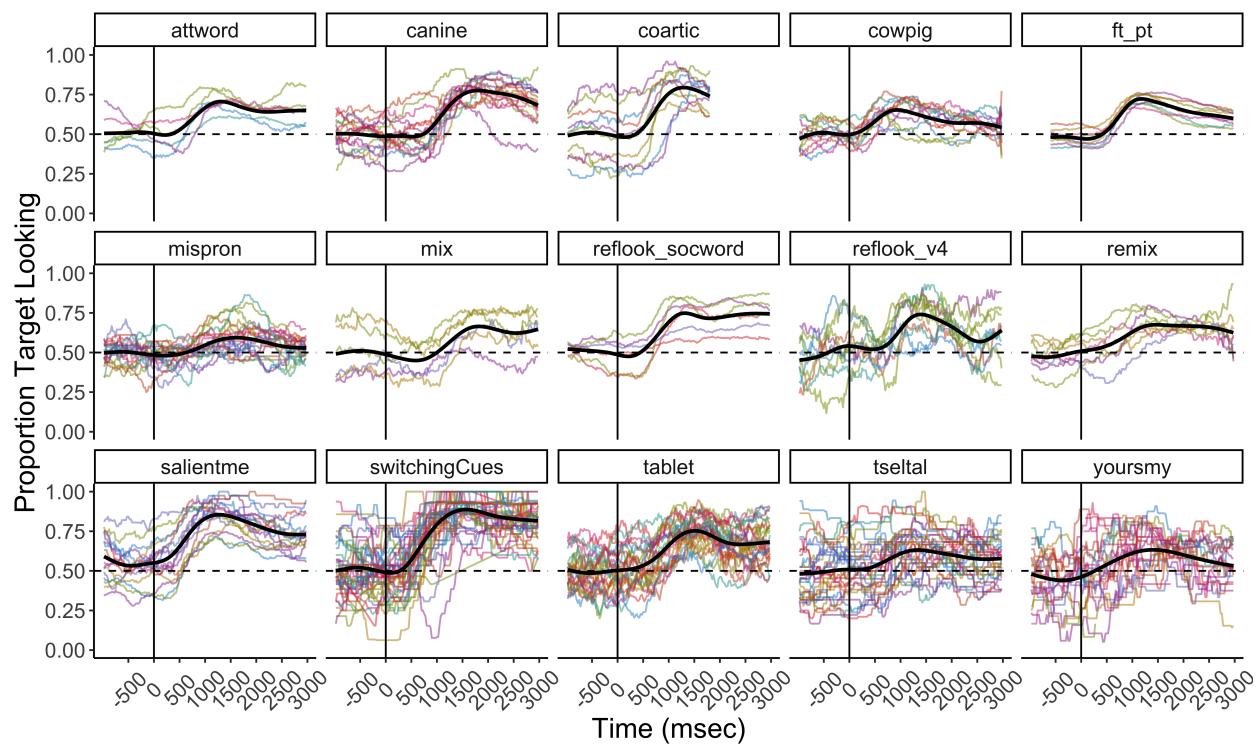


Figure 3. Item-level variability in proportion target looking within each dataset (chance=0.5). Time is centered on the onset of the target label (vertical line). Colored lines represent specific target labels. Black lines represent smoothed average fits based on a general additive model using cubic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

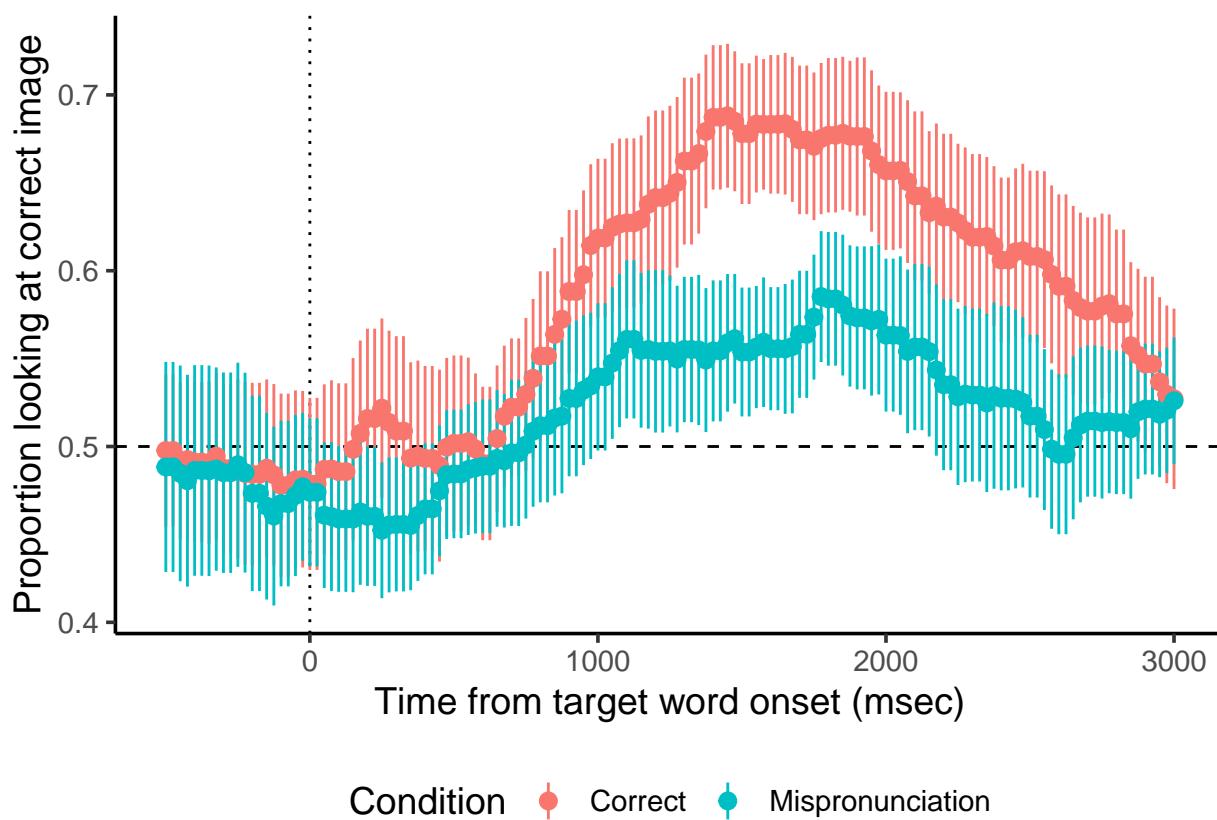


Figure 4. Proportion looking at the correct referent by time from the point of disambiguation (the onset of the target noun). Colors show the two pronunciation conditions; points give means and ranges show 95% confidence intervals. The dotted line shows the point of disambiguation and the dashed line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

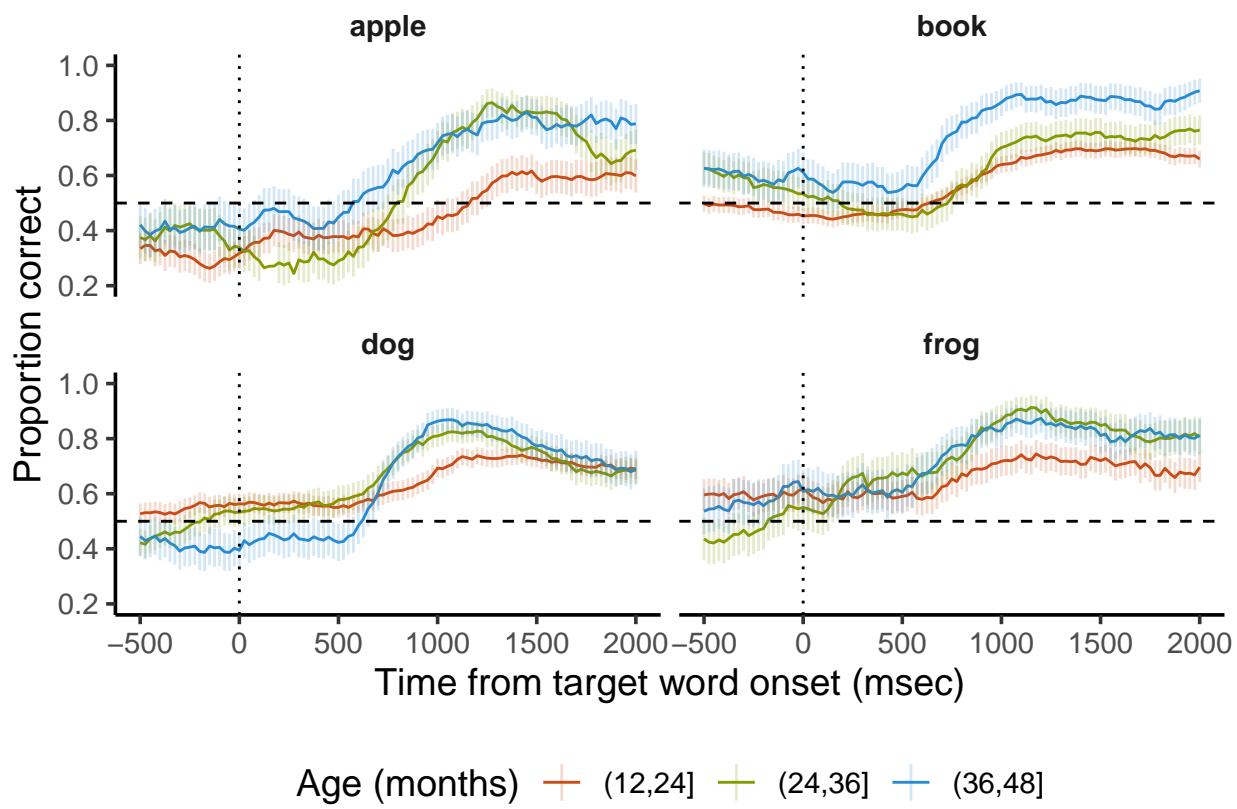


Figure 5. Add caption here.