

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) timecourse looking data,
142 detailing where on the screen a child is looking at a given point in time, and (2) metadata
143 regarding the relevant experiment, participant, and trial (Fig XX). Here, we will give an
144 outline of the tables encoding this data. As is common in relational databases, records of
145 similar types (e.g., participants, trials, experiments, coded looks at each timepoint) are
146 grouped into tables, and records of various types are linked through numeric identifiers.

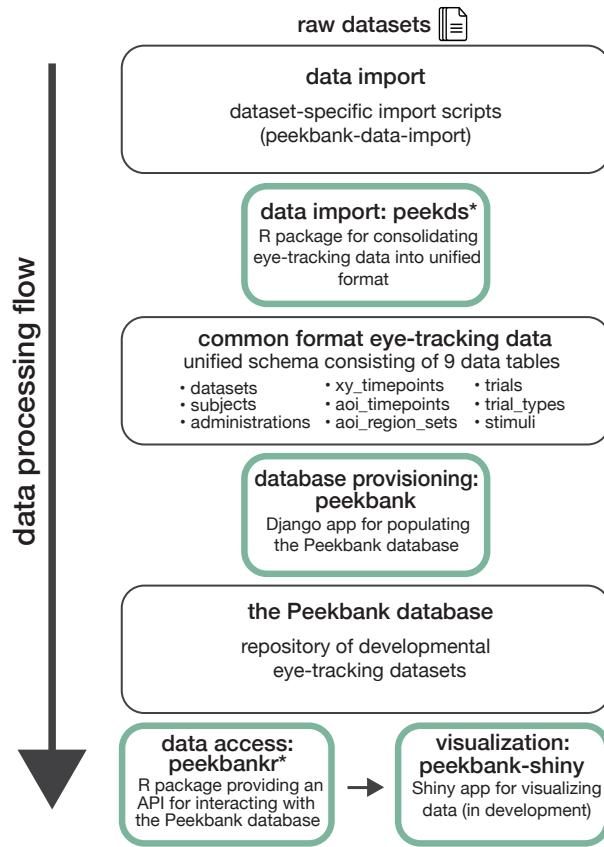


Figure 1. Overview of the Peekbank data ecosystem. Peekbank tools are highlighted in green.
* indicates R packages introduced in this work.

147

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

148 `aoi_timepoints` and `xy_timepoints`. The `aoi_timepoints` table encodes where a child is
 149 looking at each point in time, by specifying the coded area of interest (AOI): looks to the
 150 target, looks to the distractor, looks on the screen but away from target and distractor, and
 151 missing looks. All datasets must include this timecourse data, as it represents the main
 152 record of children's looking behavior. For eyetracking experiments that are automatically
 153 rather than manually coded, the `xy_timepoints` table additionally encodes the inferred (x,
 154 y) coordinates of fixations on the screen over the course of each trial. Both the
 155 `aoi_timepoints` and `xy_timepoints` tables are resampled to a consistent sampling rate, as
 156 described in the Import section below. To normalize across trials and across experiments, all
 157 timecourses are computed so that the time of 0 ms represents the onset of disambiguating

158 material (i.e., the beginning of *dog* in “Can you find the *dog*?”).

159 **Metadata.** Each record in the timecourse data is linked to several metadata records.

160 This metadata can be separated into three parts: (1) subject-level information (e.g.,
161 demographics) (2) experiment-level information (e.g., a subject’s age for a specific
162 experiment, or the particular eyetracker used to collect the data) and (3) trial information
163 and experimental design (what images or videos were presented onscreen, and paired with
164 which audio). Information about individuals who participate in one or more studies, for
165 example a subject’s sex and first language, is recorded in the `subjects` table, while the
166 `administrations` table contains information about a specific subject participating in a
167 specific experiment. This division allows Peekbank to gracefully handle longitudinal designs:
168 a single subject can be associated with many administrations.

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

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

180 Import

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

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

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

205 **trials** table.

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

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

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

227 To facilitate time course analysis and visualization across datasets, timecourse data
 228 must be resampled to a uniform sampling rate. To do this, we use the **resample()** function.
 229 During the resampling process, we interpolate using constant interpolation, selecting the

²³⁰ looking location for the nearest time point in the original data for both `aoi_timepoints`
²³¹ and `xy_timepoints` data. Compared to linear interpolation (see e.g. Wass et al., 2014),
²³² constant interpolation has the advantage that it does not introduce new look locations, so it
²³³ is a more conservative method of resampling.

²³⁴ After resampling, the final step of dataset import is validation. The `peekds` package
²³⁵ offers functions to check the now processed data tables against the most updated database
²³⁶ schema to ensure that all tables have the required fields and correct data types for database
²³⁷ ingestion. In an effort to double check the data quality and to make sure that no errors are
²³⁸ made in the importing script, we also create a time course plot based on our processed tables
²³⁹ to replicate the results in the original paper in the validation step.

²⁴⁰ **CHECK and edit resampling section for ties and for maximum time over which**
²⁴¹ **we interpolate**

²⁴² 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

²⁴³ The database currently includes 15 looking-while-listening datasets comprising
²⁴⁴ $N=1320$ total participants (Table 1). Most datasets (12 out of 15 total) consist of data from

245 monolingual native English speakers. They span a wide age spectrum with participants
246 ranging from 9 to 70 months of age, and are balanced in terms of gender (46% female). The
247 datasets vary across a number of design-related dimensions, and include studies using
248 manually coded video recordings and automated eye-tracking methods (e.g., Tobii, EyeLink)
249 to measure gaze behavior. All studies tested familiar items, but the database also includes 5
250 datasets that tested novel pseudo-words in addition to familiar words. All data are openly
251 available on the Open Science Framework (<https://osf.io/pr6wu/>).

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

253 **Versioning + Expanding the database**

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

261 **Interfacing with peekbank**

262 **Shiny App**

263 One goal of the Peekbank project is to allow a wide range of users to easily explore and
264 learn from the database. We therefore have created an interactive web application –
265 `peekbank-shiny` – that allows users to quickly and easily create informative visualizations
266 of individual datasets and aggregated data. `peekbank-shiny` is built using Shiny, a software

267 package for creating web apps using R. The Shiny app allows users to create commonly used
268 visualizations of looking-while-listening data, based on data from the Peekbank database.

269 Specifically, users can visualize

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

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

284 **Peekbankr**

285 Functions:

- 286 • `connect_to_peekbank()`
- 287 • `get_datasets()`
- 288 • `get_subjects()`
- 289 • `get_administrations()`
- 290 • `get_stimuli()`

- 291 • `get_aoi_timepoints()`
- 292 • `get_trials()`
- 293 • `get_trial_types()`
- 294 • `get_xy_timepoints()`
- 295 • `get_aoi_region_sets()`

296 **OSF site**

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

298 **Peekbank: General Descriptives**

299 [Accuracy, Reaction Times, Item variability?]

300 **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.

301 In general, participants demonstrated robust, above-chance word recognition in each
 302 dataset (chance=0.5). Table 2 shows the average proportion of target looking within a

303 standard critical window of 367-2000ms after the onset of the label for each dataset (D.
304 Swingley & Aslin, 2000b). Proportion target looking was generally higher for familiar words
305 ($M = 0.66$, 95% CI = [0.65, 0.67], $n = 1269$) than for novel words learned during the
306 experiment ($M = 0.59$, 95% CI = [0.58, 0.61], $n = 822$).

307 **Item-level variability**

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

311 **Peekbank in Action**

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

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

324 D. Swingley and Aslin (2000a) investigated the specificity of 14-16 month-olds' word
 325 representations using the looking-while-listening paradigm, asking whether recognition would
 326 be slower and less accurate for mispronunciations, e.g. "oppel" (close mispronunciation) or
 327 "opel" (distant mispronunciation) instead of "apple" (correct condition). In this short
 328 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")
```

329 We begin by retrieving the relevant tables from the database, `aoi_timepoints`,
 330 `administrations`, `trial_types`, and `trials`. As discussed above, each of these can be
 331 downloaded using a simple API call through `peekbankr`, which returns dataframes that
 332 include ID fields. These ID fields allow for easy joining of the data into a single dataframe
 333 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"))
```

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

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

338 The final step in our analysis is to create a summary dataframe using `dplyr`
 339 commands. We first group the data by timestep, participant, and condition and compute the
 340 proportion looking at the correct image. We then summarize again, averaging across
 341 participants, computing both means and 95% confidence intervals (via the approximation of
 342 1.96 times the standard error of the mean). The resulting dataframe can be used for
 343 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))
```

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

348 Item analyses

349 A second use case for Peekbank is to examine item-level variation in word recognition.
 350 While individual datasets rarely have enough statistical power to show reliable

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

354 Our first step is to collect and join the data from the relevant tables including
 355 timepoint data, trial and stimulus data, and administration data (for participant ages). We
 356 join these into a single dataframe for easy access; this dataframe is a common starting point
 357 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)
```

358 Next we select a set of four target words (chosen based on having more than children
 359 contributing data for each across several one-year age groups). We create age groups,
 360 aggregate, and compute timepoint-by-timepoint confidence intervals using the z
 361 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)
```

373

Discussion/ Conclusion

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

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

391 (almost entirely monolingual native English speakers) and cultural background (all but one
392 dataset comes from WEIRD populations; (Muthukrishna et al., 2020). Increasing the
393 diversity of participant backgrounds and languages will expand the scope of the
394 generalizations we can form about child word recognition.

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

401 **Acknowledgements**

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

404

References

- 405 Bergelson, E. (2020). The comprehension boost in early word learning: Older infants are
406 better learners. *Child Development Perspectives*, 14(3), 142–149.
- 407 Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of
408 many common nouns. *PNAS*, 109(9), 3253–3258.
- 409 Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., &
410 Cristia, A. (2018). Promoting replicability in developmental research through
411 meta-analyses: Insights from language acquisition research. *Child Development*,
412 89(6), 1996–2009.
- 413 Bleses, D., Makransky, G., Dale, P. S., Højen, A., & Ari, B. A. (2016). Early productive
414 vocabulary predicts academic achievement 10 years later. *Applied Psycholinguistics*,
415 37(6), 1461–1476.
- 416 Csibra, G., Hernik, M., Mascaro, O., Tatone, D., & Lengyel, M. (2016). Statistical treatment
417 of looking-time data. *Developmental Psychology*, 52(4), 521–536.
- 418 Fernald, A., & Marchman, V. A. (2012). Individual differences in lexical processing at 18
419 months predict vocabulary growth in typically developing and late-talking toddlers.
420 *Child Development*, 83(1), 203–222.
- 421 Fernald, A., Pinto, J. P., Swingley, D., Weinberg, A., & McRoberts, G. W. (1998). Rapid
422 gains in speed of verbal processing by infants in the 2nd year. *Psychological Science*,
423 9(3), 228–231.
- 424 Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening:
425 Using eye movements to monitor spoken language comprehension by infants and
426 young children. In I. A. Sekerina, E. M. Fernandez, & H. Clahsen (Eds.),

- 427 *Developmental psycholinguistics: On-line methods in children's language processing*
428 (pp. 97–135). Amsterdam: John Benjamins.
- 429 Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open
430 repository for developmental vocabulary data. *Journal of Child Language*, 44(3),
431 677–694.
- 432 Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and*
433 *Consistency in Early Language Learning: The Wordbank Project*. Cambridge, MA:
434 MIT Press.
- 435 Golinkoff, R. M., Ma, W., Song, L., & Hirsh-Pasek, K. (2013). Twenty-five years using the
436 intermodal preferential looking paradigm to study language acquisition: What have
437 we learned? *Perspectives on Psychol. Science*, 8(3), 316–339.
- 438 Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M.
439 C., . . . Frank, M. C. (2018). Data availability, reusability, and analytic
440 reproducibility: Evaluating the impact of a mandatory open data policy at the journal
441 Cognition. *Royal Society Open Science*, 5(8). <https://doi.org/10.1098/rsos.180448>
- 442 Hirsh-Pasek, K., Cauley, K. M., Golinkoff, R. M., & Gordon, L. (1987). The eyes have it:
443 Lexical and syntactic comprehension in a new paradigm. *Journal of Child Language*,
444 14(1), 23–45.
- 445 Huang, Y., & Snedeker, J. (2020). Evidence from the visual world paradigm raises questions
446 about unaccusativity and growth curve analyses. *Cognition*, 200, 104251.
- 447 Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of
448 grammatical gender in spoken word recognition. *Psychological Science*, 18(3),
449 193–198.

- 450 Marchman, V. A., Loi, E. C., Adams, K. A., Ashland, M., Fernald, A., & Feldman, H. M.
451 (2018). Speed of language comprehension at 18 months old predicts school-relevant
452 outcomes at 54 months old in children born preterm. *Journal of Dev. & Behav.*
453 *Pediatrics*, 39(3), 246–253.
- 454 Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J.,
455 & Thue, B. (2020). Beyond Western, Educated, Industrial, Rich, and Democratic
456 (WEIRD) Psychology: Measuring and Mapping Scales of Cultural and Psychological
457 Distance. *Psychological Science*, 31(6), 678–701.
- 458 Peter, M. S., Durrant, S., Jessop, A., Bidgood, A., Pine, J. M., & Rowland, C. F. (2019).
459 Does speed of processing or vocabulary size predict later language growth in toddlers?
460 *Cognitive Psychology*, 115, 101238.
- 461 R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna,
462 Austria: R Foundation for Statistical Computing. Retrieved from
463 <https://www.R-project.org/>
- 464 Swingley, D., & Aslin, R. N. (2000a). Spoken word recognition and lexical representation in
465 very young children. *Cognition*, 76(2), 147–166.
- 466 Swingley, D., & Aslin, R. N. (2000b). Spoken word recognition and lexical representation in
467 very young children. *Cognition*, 76(2), 147–166.
- 468 The ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research
469 using the infant-directed speech preference. *Advances in Methods and Practices in*
470 *Psychological Science*, 3(1), 24–52.
- 471 Zettersten, M., Bergey, C., Bhatt, N., Boyce, V., Braginsky, M., Carstensen, A., ... others.
472 (2021). Peekbank: Exploring children's word recognition through an open, large-scale
473 repository for developmental eye-tracking data.

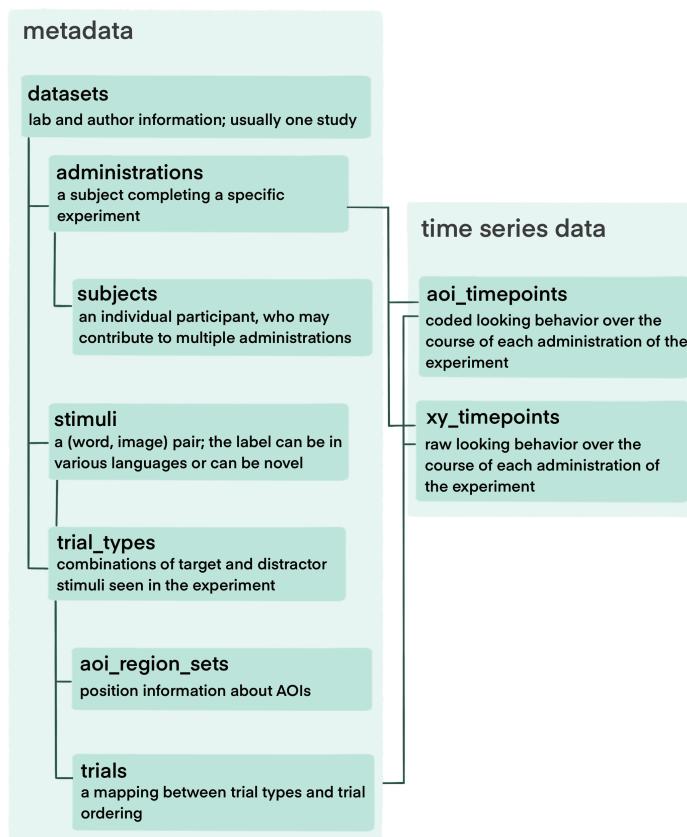


Figure 2. The Peekbank schema. Each square represents a table in the relational database.

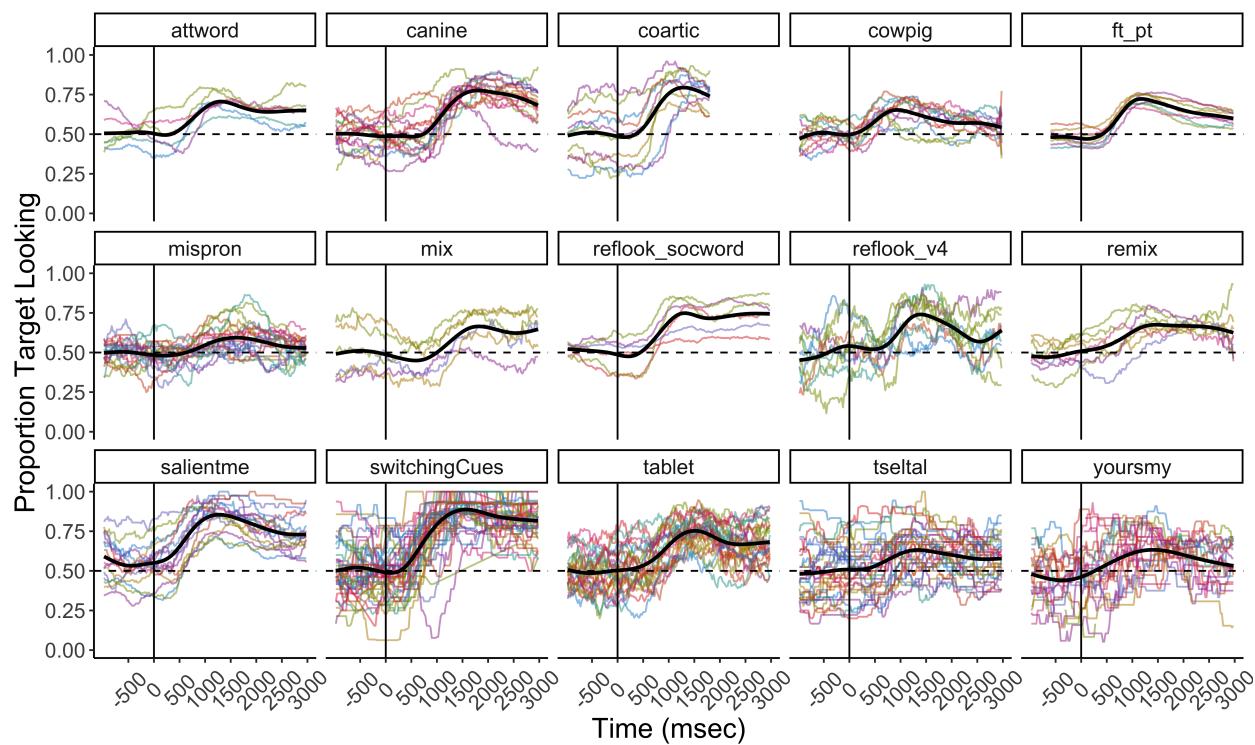


Figure 3. Item-level variability in proportion target looking within each dataset (chance=0.5). Time is centered on the onset of the target label (vertical line). Colored lines represent specific target labels. Black lines represent smoothed average fits based on a general additive model using cubic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

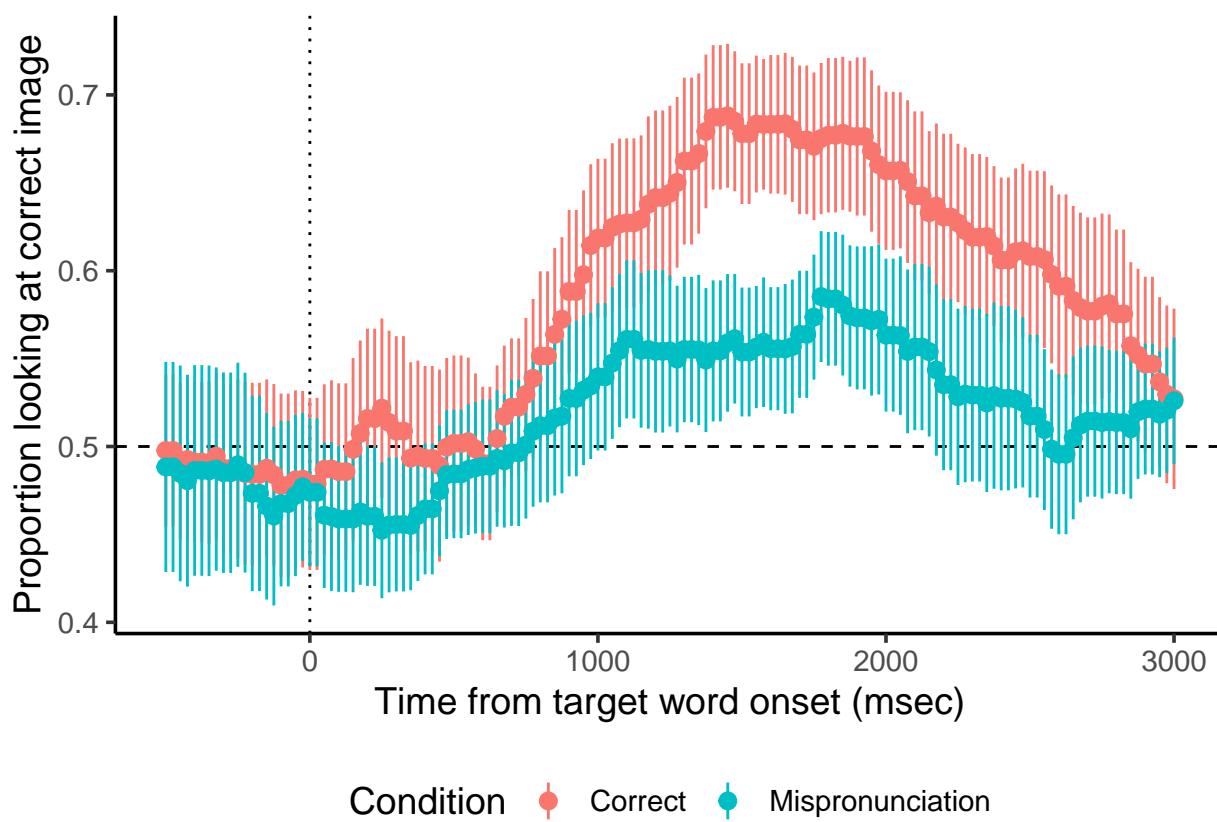


Figure 4. Proportion looking at the correct referent by time from the point of disambiguation (the onset of the target noun). Colors show the two pronunciation conditions; points give means and ranges show 95% confidence intervals. The dotted line shows the point of disambiguation and the dashed line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

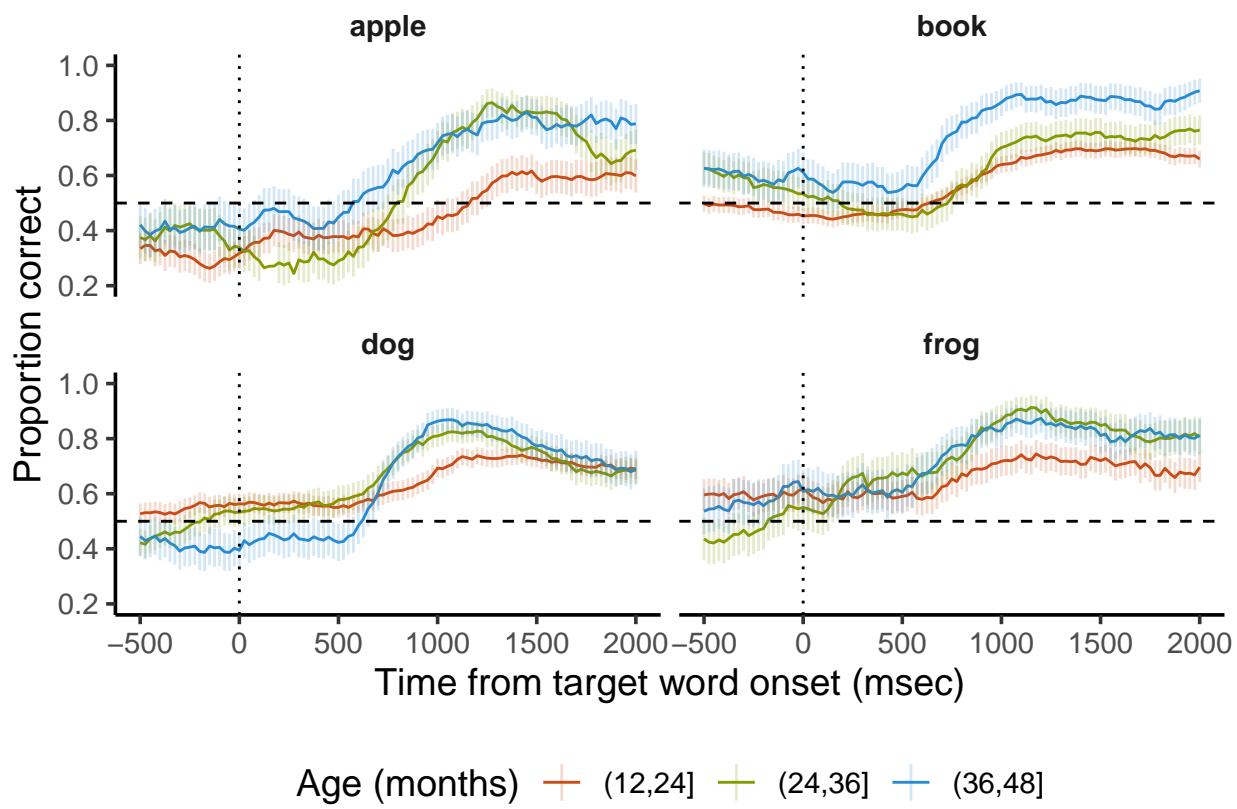


Figure 5. Add caption here.