

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

³ Martin Zettersten¹, Daniel Yurovsky², Tian Linger Xu³, Sarp Uner⁴, Angeline Sin Mei Tsui⁵,
⁴ Rose M. Schneider⁶, Annissa N. Saleh⁷, Stephan Meylan^{8,9}, Virginia Marchman⁵, Jessica
⁵ Mankewitz⁵, Kyle MacDonald¹⁰, Bria Long⁵, Molly Lewis², George Kachergis⁵, Kunal
⁶ Handa¹¹, Benjamin deMayo¹, Alexandra Carstensen⁶, Mika Braginsky⁹, Veronica Boyce⁵,
⁷ Naiti S. Bhatt¹², Claire Bergey¹³, & Michael C. Frank⁵

⁸ ¹ Department of Psychology, Princeton University

⁹ ² Department of Psychology, Carnegie Mellon University

¹⁰ ³ Department of Psychological and Brain Sciences, Indiana University

¹¹ ⁴ Data Science Institute, Vanderbilt University

¹² ⁵ Department of Psychology, Stanford University

¹³ ⁶ Department of Psychology, University of California, San Diego

¹⁴ ⁷ Department of Psychology, The University of Texas at Austin

¹⁵ ⁸ Department of Psychology and Neuroscience, Duke University

¹⁶ ⁹ Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

¹⁷ ¹⁰ Core Technology, McD Tech Labs

¹⁸ ¹¹ Brown University

¹⁹ ¹² Department of Psychology, New York University

²⁰ ¹³ Department of Psychology, University of Chicago

21

Author Note

22 **Acknowledgements.** We would like to thank the labs and researchers that have
23 made their data publicly available in the database. For further information about
24 contributions, see <https://langcog.github.io/peekbank-website/docs/contributors/>.

25 **Open Practices Statement.** All code for reproducing the paper is available at
26 <https://github.com/langcog/peekbank-paper>. Raw and standardized datasets are available
27 on the Peekbank OSF repository (<https://osf.io/pr6wu/>) and can be accessed using the
28 peekbankr R package (<https://github.com/langcog/peekbankr>).

29 **CRediT author statement.** Outside of the position of the first and the last author,
30 authorship position was determined by sorting authors' last names in reverse alphabetical
31 order. An overview of authorship contributions following the CRediT taxonomy can be
32 viewed here: https://docs.google.com/spreadsheets/d/e/2PACX-1vRD-LJD_dTAQaAynyBlwXvGpfAVzP-3Pi6JTDG15m3PYZe0c44Y12U2a_hwdmhIstpjyigG2o3na4y/pubhtml.

34 Correspondence concerning this article should be addressed to Martin Zettersten,
35 Department of Psychology, Princeton University, 218 Peretsman Scully Hall, Princeton, NJ
36 08540. E-mail: martincz@princeton.edu

37

Abstract

38 The ability to rapidly recognize words and link them to referents is central to children's
39 early language development. This ability, often called word recognition in the developmental
40 literature, is typically studied in the looking-while-listening paradigm, which measures
41 infants' fixation on a target object (vs. a distractor) after hearing a target label. We present
42 a large-scale, open database of infant and toddler eye-tracking data from
43 looking-while-listening tasks. The goal of this effort is to address theoretical and
44 methodological challenges in measuring vocabulary development. We first present how we
45 created the database, its features and structure, and associated tools for processing and
46 accessing infant eye-tracking datasets. Using these tools, we then work through two
47 illustrative examples to show how researchers can use Peekbank to interrogate theoretical
48 and methodological questions about children's developing word recognition ability.

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

51 Word count: 6605

52 Peekbank: An open, large-scale repository for developmental eye-tracking data of children's
53 word recognition

54 Across their first years of life, children learn words at an accelerating pace (Frank,
55 Braginsky, Yurovsky, & Marchman, 2021). While many children will only produce their first
56 word at around one year of age, most children show signs of understanding many common
57 nouns (e.g., *mommy*) and phrases (e.g., *Let's go bye-bye!*) much earlier in development
58 (Bergelson & Swingley, 2012, 2013; Tincoff & Jusczyk, 1999). Although early word
59 understanding is a critical element of first language learning, the processes involved are less
60 directly apparent in children's behaviors and are less accessible to observation than
61 developments in speech production (Fernald, Zangl, Portillo, & Marchman, 2008;
62 Hirsh-Pasek, Cauley, Golinkoff, & Gordon, 1987). To understand a spoken word, children
63 must process the incoming auditory signal and link that signal to relevant meanings – a
64 process often referred to as word recognition. One of the primary means of measuring word
65 recognition in young infants is using eye-tracking techniques that gauge where children look
66 in response to linguistic stimuli (Fernald, Zangl, Portillo, & Marchman, 2008). The logic of
67 these methods is that if, upon hearing a word, a child preferentially looks at a target
68 stimulus rather than a distractor, the child is able to recognize the word and activate its
69 meaning during real-time language processing. Measuring early word recognition offers
70 insight into children's early word representations: children's speed of response (i.e., moving
71 their eyes; turning their heads) to the unfolding speech signal can reveal children's level of
72 comprehension (Bergelson, 2020; Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998).
73 Word recognition skills are also thought to build a foundation for children's subsequent
74 language development. Past research has found that early word recognition efficiency is
75 predictive of later linguistic and general cognitive outcomes (Bleses, Makransky, Dale, Højlen,
76 & Ari, 2016; Marchman et al., 2018).

77 While word recognition is a central part of children's language development, mapping

78 the trajectory of word recognition skills has remained elusive. Studies investigating children's
79 word recognition are typically limited in scope to experiments in individual labs involving
80 small samples tested on a handful of items. The limitations of single datasets makes it
81 difficult to understand developmental changes in children's word knowledge at a broad scale.

82 One way to overcome this challenge is to compile existing datasets into a large-scale
83 database in order to expand the scope of research questions that can be asked about the
84 development of word recognition abilities. This strategy capitalizes on the fact that the
85 looking-while-listening paradigm is widely used, and vast amounts of data have been
86 collected across labs on infants' word recognition over the past 35 years (Golinkoff, Ma, Song,
87 & Hirsh-Pasek, 2013). Such datasets have largely remained isolated from one another, but
88 once combined, they have the potential to offer general insights into lexical development.

89 There has been a long history of efforts to aggregate data in a unified format in
90 developmental and cognitive psychology, generating projects that have often had a
91 tremendous impact on the field. Prominent examples in language research include the
92 English Lexicon Project, which provides an open repository of psycholinguistic data for over
93 80,000 English words and non-words in order to support large-scale investigations of lexical
94 processing (Balota et al., 2007); the Child Language Data Exchange System (CHILDES),
95 which has played an instrumental role in the study of early language environments by
96 systematizing and aggregating data from naturalistic child-caregiver language interactions
97 (MacWhinney, 2000); and WordBank, which aggregated data from the MacArthur-Bates
98 Communicative Development Inventory, a parent-report measure of child vocabulary, to
99 deliver new insights into cross-linguistic patterns and variability in vocabulary development
100 (Frank, Braginsky, Yurovsky, & Marchman, 2017, 2021).

101 In this paper, we introduce *Peekbank*, an open database of infant and toddler
102 eye-tracking data aimed at facilitating the study of developmental changes in children's word
103 recognition.

¹⁰⁴ **Measuring Word Recognition: The Looking-While-Listening Paradigm**

¹⁰⁵ Word recognition is traditionally studied in the looking-while-listening paradigm
¹⁰⁶ (Fernald, Zangl, Portillo, & Marchman, 2008; alternatively referred to as the intermodal
¹⁰⁷ preferential looking procedure, Hirsh-Pasek, Cauley, Golinkoff, & Gordon, 1987). In these
¹⁰⁸ studies, infants listen to a sentence prompting a specific referent (e.g., *Look at the dog!*)
¹⁰⁹ while viewing two images on the screen (e.g., an image of a dog – the target image – and an
¹¹⁰ image of a bird – the distractor image). Infants' word recognition is evaluated by how
¹¹¹ quickly and accurately they fixate on the target image after hearing its label. Past research
¹¹² has used this basic method to study a wide range of questions in language development. For
¹¹³ example, the looking-while-listening paradigm has been used to investigate early noun
¹¹⁴ knowledge, phonological representations of words, prediction during language processing, and
¹¹⁵ individual differences in language development (Bergelson & Swingley, 2012; Golinkoff, Ma,
¹¹⁶ Song, & Hirsh-Pasek, 2013; Lew-Williams & Fernald, 2007; Marchman et al., 2018; Swingley
¹¹⁷ & Aslin, 2002).

¹¹⁸ While this research has been fruitful in advancing understanding of early word
¹¹⁹ knowledge, fundamental questions remain. One central question is how to accurately capture
¹²⁰ developmental change in the speed and accuracy of word recognition. There is ample
¹²¹ evidence demonstrating that infants become faster and more accurate in word recognition
¹²² over the first few years of life (e.g., Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998).
¹²³ However, precisely measuring developmental increases in the speed and accuracy of word
¹²⁴ recognition remains challenging due to the difficulty of distinguishing developmental changes
¹²⁵ in word recognition skill from changes in knowledge of specific words. This problem is
¹²⁶ particularly thorny in studies with young children, since the number of items that can be
¹²⁷ tested within a single session is limited and items must be selected in an age-appropriate
¹²⁸ manner (Peter et al., 2019). More broadly, key differences in the design choices (e.g., how
¹²⁹ distractor items are selected) and analytic decisions (e.g., how the analysis window is defined)

130 between studies can obscure developmental change if not appropriately taken into account.

131 One approach to addressing these challenges is to conduct meta-analyses aggregating
132 effects across studies while testing for heterogeneity due to researcher choices (Bergmann et
133 al., 2018; Lewis et al., 2016). However, meta-analyses typically lack the granularity to
134 estimate participant-level and item-level variation or to model behavior beyond
135 coarse-grained effect size estimates. An alternative way to approach this challenge is to
136 aggregate trial-level data from smaller studies measuring word recognition with a wide range
137 of items and design choices into a large-scale dataset that can be analyzed using a unified
138 modeling approach. A sufficiently large dataset would allow researchers to estimate
139 developmental change in word recognition speed and accuracy while generalizing across
140 changes related to specific words or the design features of particular studies.

141 A related open theoretical question is understanding changes in children's word
142 recognition at the level of individual items. Looking-while-listening studies have been limited
143 in their ability to assess the development of specific words. One limitation is that studies
144 typically test only a small number of trials for each item, reducing power to precisely measure
145 the development of word-specific accuracy (DeBolt, Rhemtulla, & Oakes, 2020). A second
146 limitation is that target stimuli are often yoked with a narrow set of distractor stimuli (i.e., a
147 child sees a target with only one or two distractor stimuli over the course of an experiment),
148 leaving ambiguous whether accurate looking to a particular target word can be attributed to
149 children's recognition of the target word or their knowledge about the distractor.
150 Aggregating across many looking-while-listening studies has the potential to meet these
151 challenges by increasing the number of observations for specific items at different ages and by
152 increasing the size of the inventory of distractor stimuli that co-occur with each target.

153 **Replicability and Reproducibility**

154 A core challenge facing psychology in general, and the study of infant development in
155 particular, are threats to the replicability and reproducibility of core empirical results (Frank
156 et al., 2017; Nosek et al., 2022). In infant research, many studies are not adequately powered
157 to detect the main effects of interest (Bergmann et al., 2018). This issue is compounded by
158 low reliability in infant measures, often due to limits on the number of trials that can be
159 collected from an individual infant in an experimental session (Byers-Heinlein, Bergmann, &
160 Savalei, 2021). One hurdle to improving power in infant research is that it can be difficult to
161 develop a priori estimates of effect sizes and how specific design decisions (e.g., the number
162 of test trials) will impact power and reliability. Large-scale databases of infant behavior can
163 aid researchers in their decision-making by allowing them to directly test how different
164 design decisions affect power and reliability. For example, if a researcher is interested in
165 understanding how the number of test trials could impact the power and reliability of their
166 looking-while-listening design, a large-scale infant eye-tracking database would allow them to
167 simulate possible outcomes across a range of test trials, providing the basis for data-driven
168 design decisions.

169 In addition to threats to replicability, the field of infant development also faces
170 concerns about analytic reproducibility – the ability for researchers to arrive at the same
171 analytic conclusion reported in the original research article, given the same dataset. A recent
172 estimate based on studies published in a prominent cognitive science journal suggests that
173 analyses can remain difficult to reproduce, even when data are made available to other
174 research teams (Hardwicke et al., 2018). Aggregating data in centralized databases can aid
175 in improving reproducibility in several ways. First, building a large-scale database requires
176 defining a standardized data specification. Recent examples include the `brain imaging`
177 `data structure` (BIDS), an effort to specify a unified data format for neuroimaging
178 experiments (Gorgolewski et al., 2016), and the data formats associated with `ChildProject`,

179 for managing long-form at-home language recordings (Gautheron, Rochat, & Cristia, 2021).
180 Defining a data standard – in this case, for infant eye-tracking experiments – supports
181 reproducibility by guaranteeing that critical information will be available in openly shared
182 data and by making it easier for different research teams to understand the data structure.
183 Second, open databases make it easy for researchers to generate open and reproducible
184 analytic pipelines, both for individual studies and for analyses aggregating across datasets.
185 Creating open analytic pipelines across many datasets also serves a pedagogical purpose,
186 providing teaching examples illustrating how to implement analytic techniques used in
187 influential studies and how to conduct reproducible analyses with infant eye-tracking data.

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

189 What all of these open challenges share is that they are difficult to address at the scale
190 of a single research lab or in a single study. To address this challenge, we developed
191 *Peekbank*, a flexible and reproducible interface to an open database of developmental
192 eye-tracking studies. The Peekbank project (a) collects a large set of eye-tracking datasets
193 on children’s word recognition, (b) introduces a data format and processing tools for
194 standardizing eye-tracking data across heterogeneous data sources, and (c) provides an
195 interface for accessing and analyzing the database. In the current paper, we introduce the
196 key components of the project and give an overview of the existing database. We then
197 provide two worked examples of how researchers can use Peekbank. In the first, we examine
198 a classic result in the word recognition literature, and in the second we aggregate data across
199 studies to investigate developmental trends in the recognition of individual words.

200 **Design and Technical Approach**

201 **Database Framework**

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

207 As illustrated in Figure 1, the Peekbank framework consists of four main components:
208 (1) a set of tools to *convert* eye-tracking datasets into a unified format, (2) a relational
209 database populated with data in this unified format, (3) a set of tools to *retrieve* data from
210 this database, and (4) a web app (using the Shiny framework) for visualizing the data. These
211 components are supported by three packages. The `peekds` package (for the R language, R
212 Core Team, 2021) helps researchers convert existing datasets to use the standardized format
213 of the database. The `peekbank` module (Python) creates a database with the relational
214 schema and populates it with the standardized datasets produced by `peekds`. The database
215 is served through MySQL, an industry standard relational database server, which may be
216 accessed by a variety of programming languages, and can be hosted on one machine and
217 accessed by many others over the Internet. As is common in relational databases, records of
218 similar types (e.g., participants, trials, experiments, coded looks at each timepoint) are
219 grouped into tables, and records of various types are linked through numeric identifiers. The
220 `peekbankr` package (R) provides an application programming interface, or API, that offers
221 high-level abstractions for accessing the tabular data stored in Peekbank. Most users will
222 access data through this final package, in which case the details of data formatting,
223 processing, and the specifics of connecting to the database are abstracted away from the user.

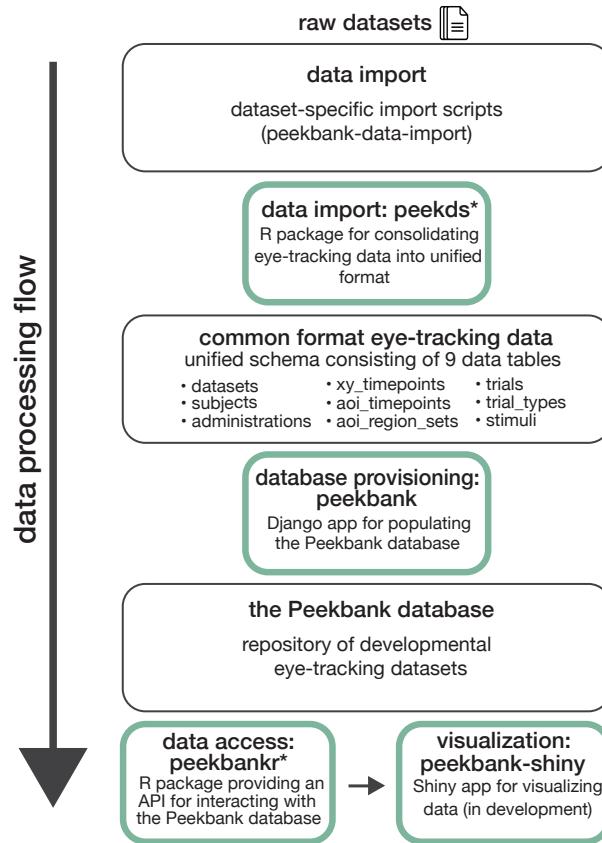


Figure 1. Overview of the Peekbank data ecosystem. Peekbank tools are highlighted in green.
* indicates R packages introduced in this work.

224 Database Schema

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

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

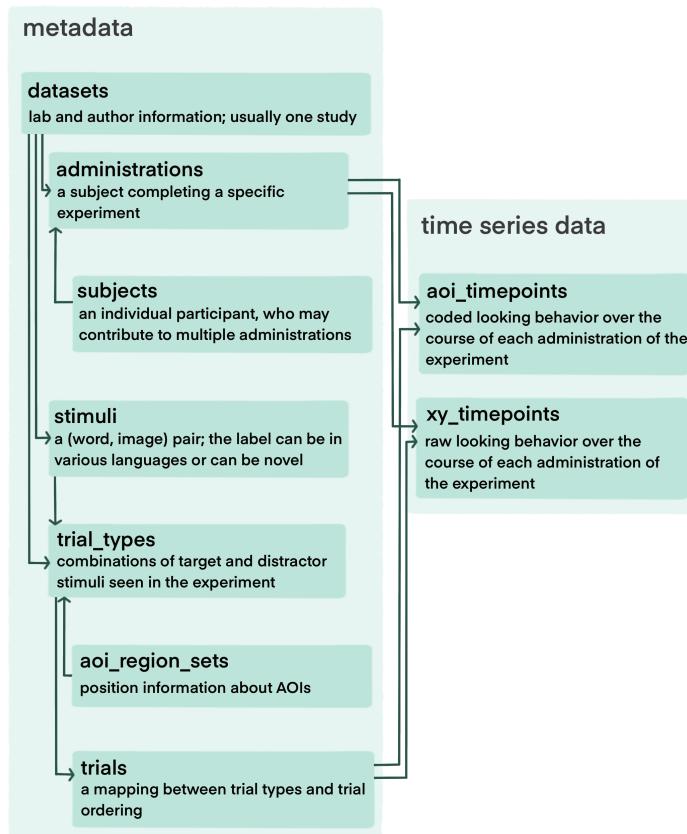


Figure 2. The Peekbank schema. Each darker rectangle represents a table in the relational database.

Participant Information.

All information about individual participants in Peekbank is completely de-identified under United States law, containing none of the key identifiers listed under the “Safe Harbor” standard for data de-identification. All participant-level linkages are made using anonymous participant identifiers.

Invariant information about individuals who participate in one or more studies (e.g., a participant’s first language) is recorded in the `subjects` table, while the `administrations` table contains information about each individual session in a given study (see Session Information, below). This division allows Peekbank to gracefully handle longitudinal designs: a single participant can complete multiple sessions and thus be associated with multiple

243 administrations.

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

247 ***Experiment Information.***

248 The **datasets** table includes information about the lab conducting the study and the
249 relevant publications to cite regarding the data. In most cases, a dataset corresponds to a
250 single study.

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

260 Each record in the **stimuli** table is a (word, image) pair. In most experiments, there
261 is a one-to-one mapping between images and labels (e.g., each time an image of a dog
262 appears it is referred to as *dog*). For studies in which there are multiple potential labels per
263 image (e.g., *dog* and *chien* are both used to refer to an image of a dog), images can have
264 multiple rows in the **stimuli** table with unique labels. This structure is useful for studies on

¹ We note that the term *trial* is ambiguous and could be used to refer to both a particular combination of stimuli seen by many participants and a participant seeing that particular combination at a particular point in the experiment. We track the former in the **trial_types** table and the latter in the **trials** table.

synonymy or using multiple languages. It is also possible for an image to be associated with a row with no label, if the image appears solely as a distractor (and thus its label is ambiguous). For studies in which the same label refers to multiple images (e.g., the word *dog* refers to an image of a dalmatian and a poodle), the same label can have multiple rows in the `stimuli` table with unique images.

Session Information.

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

Trial Information.

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

Time course data. Raw looking data is a series of looks to areas of interest (AOIs), such as looks to the left or right of the screen, or to (x, y) coordinates on the experiment screen, linked to points in time. For data generated by eye-trackers, we typically have (x, y) coordinates at each time point, which we encode in the `xy_timepoints` table. These looks are also recoded into AOIs according to the AOI coordinates in the `aoi_region_sets` table using the `add_aois()` function in `peekds`, and encoded in the `aoi_timepoints` table. For hand-coded data, we typically have a series of AOIs (i.e., looks to the left vs. right of the screen), but lack information about exact gaze positions on-screen; in these cases the AOIs are recoded into the categories in the Peekbank schema (target, distractor, other, and missing) and encoded in the `aoi_timepoints` table; however, these datasets do not have any

290 corresponding data in the `xy_timepoints` table.

291 Typically, timepoints in the `xy_timepoints` table and `aoi_timepoints` table need to
292 be regularized to center each trial's time around the point of disambiguation – such that 0 is
293 the time of target word onset in the trial (i.e., the beginning of *dog* in *Can you find the*
294 *dog?*). We re-centered timing information to the onset of the target label to facilitate
295 comparison of target label processing across all datasets.² If time values run throughout the
296 experiment rather than resetting to zero at the beginning of each trial, `rezero_times()` is
297 used to reset the time at each trial. After this, each trial's times are centered around the
298 point of disambiguation using `normalize_times()`. When these steps are complete, the
299 time course is ready for resampling.

300 To facilitate time course analysis and visualization across datasets, time course data
301 must be resampled to a uniform sampling rate (i.e., such that every trial in every dataset has
302 observations at the same time points). All data in the database is resampled to 40 Hz
303 (observations every 25 ms), which represents a compromise between retaining fine-grained
304 timing information from datasets with dense sampling rates (maximum sampling rate among
305 current datasets: 500 Hz) while minimizing the possibility of introducing artifacts via
306 resampling for datasets with lower sampling rates (minimum sampling rate for current
307 datasets: 30 Hz). Further, 25 ms is a mathematically convenient interval for ensuring
308 consistent resampling; we found that using 33.333 ms (30 Hz) as our interval simply
309 introduced a large number of technical complexities. The resampling operation is
310 accomplished using the `resample_times()` function. During the resampling process, we
311 interpolate using constant interpolation, selecting for each interpolated timepoint the looking
312 location for the earlier-observed time point in the original data for both `aoi_timepoints`

² While information preceding the onset of the target label in some datasets such as co-articulation cues (Mahr, McMillan, Saffran, Ellis Weismer, & Edwards, 2015) or adjectives (Fernald, Marchman, & Weisleder, 2013) can in principle disambiguate the target referent, we use a standardized point of disambiguation based on the onset of the label for the target referent. Onset times for other potentially disambiguating information (such as adjectives) can typically be recovered from the raw data provided on OSF.

313 and `xy_timepoints` data. Compared to linear interpolation (see e.g., Wass, Smith, &
314 Johnson, 2013) – which fills segments of missing or unobserved time points by interpolating
315 between the observed locations of timepoints at the beginning and end of the interpolated
316 segment –, constant interpolation has the advantage that it is more conservative, in the sense
317 that it does not introduce new look locations beyond those measured in the original data.
318 One possible application of our new dataset is investigating the consequences of other
319 interpolation functions for data analysis.

320 Processing, Validation, and Ingestion

321 Although Peekbank provides a common data format, the crux issue of populating the
322 database is the conversion of existing datasets to this format. Each dataset is imported via a
323 custom import script, which documents the process of conversion. Often various decisions
324 must be made in this import process (for example, how to characterize a particular trial type
325 within the options available in the Peekbank schema); these scripts provide a reproducible
326 record of this decision-making process. Our data import repository (available on GitHub at
327 <https://github.com/langcog/peekbank-data-import>) contains all of these scripts, links to
328 internal documentation on data import, and a set of generic import templates for different
329 formats.

330 Many of the specific operations involved in importing a dataset can be abstracted
331 across datasets. The `peekds` package offers a library of these functions. Once the data have
332 been extracted in a tabular form, the package also offers a validation function that checks
333 whether all tables have the required fields and data types expected by the database. In an
334 effort to double check the data quality and to make sure that no errors are made in the
335 importing script, we also typically perform a visual check of the import process, creating a
336 time course plot to replicate the results in the paper that first presented each dataset. Once
337 this plot has been created and checked for consistency and all tables pass our validation

³³⁸ functions, the processed dataset is ready for reprocessing into the database using the
³³⁹ `peekbank` library. This library applies additional data checks, and adds the data to the
³⁴⁰ MySQL database using the Django web framework.

³⁴¹ To date, the import process has been carried out by the Peekbank team using data
³⁴² offered by other research teams. There is no technical obstacle to data contributors also
³⁴³ providing an import script to facilitate contribution, though in practice creating these scripts
³⁴⁴ requires familiarity with both R scripting and the specific Peekbank schema; writing a first
³⁴⁵ import script can be somewhat time-consuming.

³⁴⁶ Current Data Sources

Table 1
Overview of the datasets in the current database.

Study Citation	Dataset name	N	Mean age (mos.)	Age range (mos.)	Method	Language
Adams et al., 2018	adams_marchman_2018	69	17.1	13–20	manual coding	English
Byers-Heinlein et al., 2017	byers-heinlein_2017	48	20.1	19–21	eye-tracking	English, French
Casillas et al., 2017	casillas_tseltal_2015	23	31.3	9–48	manual coding	Tseltal
Fernald et al., 2013	fmw_2013	80	20.0	17–26	manual coding	English
Frank et al., 2016	frank_tablet_2016	69	35.5	12–60	eye-tracking	English
Garrison et al., 2020	garrison_bergelson_2020	35	14.5	12–18	eye-tracking	English
Hurtado et al., 2007	xsectional_2007	49	23.8	15–37	manual coding	Spanish
Hurtado et al., 2008	hurtado_2008	76	21.0	17–27	manual coding	Spanish
Mahr et al., 2015	mahr_coartic	29	20.8	18–24	eye-tracking	English
Perry et al., 2017	perry_cowpig	45	20.5	19–22	manual coding	English
Pomper & Saffran, 2016	pomper_saffran_2016	60	44.3	41–47	manual coding	English
Pomper & Saffran, 2019	pomper_saliencyme	44	40.1	38–43	manual coding	English
Potter & Lew-Williams, unpub.	potter_canine	36	23.8	21–27	manual coding	English
Potter et al., 2019	potter_remix	44	22.6	18–29	manual coding	Spanish, English
Ronfard et al., 2021	ronfard_2021	40	20.0	18–24	manual coding	English
Swingley & Aslin, 2002	swingley_aslin_2002	50	15.1	14–16	manual coding	English
Weisleder & Fernald, 2013	weisleder_stl	29	21.6	18–27	manual coding	Spanish
Yurovsky & Frank, 2017	attword_processed	288	25.5	13–59	eye-tracking	English
Yurovsky et al., 2013	reflook_socword	435	33.6	12–70	eye-tracking	English
Yurovsky et al., unpub.	reflook_v4	45	34.2	11–60	eye-tracking	English

³⁴⁷ The database currently includes 20 looking-while-listening datasets comprising $N=1594$
³⁴⁸ total participants (Table 1). The current data represents a convenience sample of datasets
³⁴⁹ that were (a) datasets collected by or available to Peekbank team members, (b) made
³⁵⁰ available to Peekbank after informal inquiry or (c) datasets that were openly available. Most
³⁵¹ datasets (14 out of 20 total) consist of data from monolingual native English speakers. They

352 span a wide age spectrum with participants ranging from 9 to 70 months of age, and are
353 balanced in terms of gender (47.30% female; 50.40% male; 2.30% unreported). The datasets
354 vary across a number of design-related dimensions, and include studies using manually coded
355 video recordings and automated eye-tracking methods (e.g., Tobii, EyeLink) to measure gaze
356 behavior. All studies tested familiar items, but the database also includes 5 datasets that
357 tested novel pseudo-words in addition to familiar words. Users interested in a subset of the
358 data (e.g., only trials testing familiar words) can filter out unwanted trials using columns
359 available in the schema (e.g., using the column `stimulus_novelty` in the `stimuli` table).

360 **Versioning and Reproducibility**

361 The content of Peekbank will change as we add additional datasets and revise previous
362 ones. To facilitate reproducibility of analyses, we use a versioning system by which
363 successive releases are assigned a name reflecting the year and version, e.g., 2022.1. By
364 default, users will interact with the most recent version of the database available, though the
365 `peekbankr` API allows researchers to run analyses against any previous version of the
366 database. For users with intensive use-cases, each version of the database may be
367 downloaded as a compressed .sql file and installed on a local MySQL server.

368 Peekbank allows for fully reproducible analyses using our source data, but the goal is
369 not to reproduce precisely the analyses – or even the datasets – in the publications whose
370 data we archive. Because of our emphasis on a standardized data importing and formatting
371 pipeline, there may be minor discrepancies in the time course data that we archive compared
372 with those reported in original publications. Further, we archive all of the data that are
373 provided to us – including participants that might have been excluded in the original studies,
374 if these data are available – rather than attempting to reproduce specific exclusion criteria.
375 We hope that Peekbank can be used as a basis for comparing different exclusion and filtering
376 criteria – as such, an inclusive policy regarding importing all available data helps us provide

377 a broad base of data for investigating these decisions.

378

Interfacing with Peekbank

379 **Peekbankr**

380 The `peekbankr` API offers a way for users to access data from the database and
381 flexibly analyze it in R. The majority of API calls simply allow users to download tables (or
382 subsets of tables) from the database. In particular, the package offers the following functions:

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

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

403 Once users have downloaded tables, they can be merged using `join` commands via their
404 linked IDs. A set of standard merges are shown below in the “Peekbank in Action” section;
405 these allow the common use-case of examining time course data and metadata jointly.

406 Because of the size of the XY and AOI data tables, downloading data across multiple
407 studies can be time-consuming. Many of the most common analyses of the Peekbank data
408 require downloading the `aoi_timepoints` table, thus we have put substantial work into
409 optimizing transfer times. In particular, `connect_to_peekbank` offers a data compression
410 option, and `get_aoi_timepoints` by default downloads time courses via a compressed
411 (run-length encoded) representation, which is then uncompressed on the client side. More
412 information about these options (including how to modify them) can be found in the
413 package documentation.

414 **Shiny App**

415 One goal of the Peekbank project is to allow a wide range of users to easily explore and

416 learn from the database. We therefore have created an interactive web application –

417 `peekbank-shiny` – that allows users to quickly and easily create informative visualizations
418 of individual datasets and aggregated data (<https://peekbank-shiny.com/>).

419 `peekbank-shiny` is built using Shiny, a software package for creating web apps for data
420 exploration with R, as well as the `peekbankr` package. All code for the Shiny app is publicly
421 available (<https://github.com/langcog/peekbank-shiny>). The Shiny app allows users to
422 create commonly used visualizations of looking-while-listening data, based on data from the
423 Peekbank database. Specifically, users can visualize:

- 424 1. the *time course of looking data* in a profile plot depicting infant target looking across

425 trial time
426 2. *overall accuracy*, defined as the proportion target looking within a specified analysis
427 window
428 3. *reaction times* in response to a target label, defined as how quickly participants shift
429 fixation to the target image on trials in which they were fixating on the distractor
430 image at onset of the target label
431 4. an *onset-contingent plot*, which shows the time course of participant looking as a
432 function of their look location at the onset of the target label

433 Users are given various customization options for each of these visualizations, e.g.,
434 choosing which datasets to include in the plots, controlling the age range of participants,
435 splitting the visualizations by age bins, and controlling the analysis window for time course
436 analyses. Plots are then updated in real time to reflect users' customization choices. A
437 screenshot of the app is shown in Figure 3. The Shiny app thus allows users to quickly
438 inspect basic properties of Peekbanks datasets and create reproducible visualizations without
439 incurring any of the technical overhead required to access the database through R.

440 OSF site

441 In addition to the Peekbank database proper, all data is openly available on the
442 Peekbank OSF webpage (<https://osf.io/pr6wu/>). The OSF site also includes the original raw
443 data (both time series data and metadata, such as trial lists and participant logs) that was
444 obtained for each study and subsequently processed into the standardized Peekbank format.
445 Where available, the OSF page also includes additional information about the stimuli used in
446 each dataset, including in some instances the original stimulus sets (e.g., image and audio
447 files).

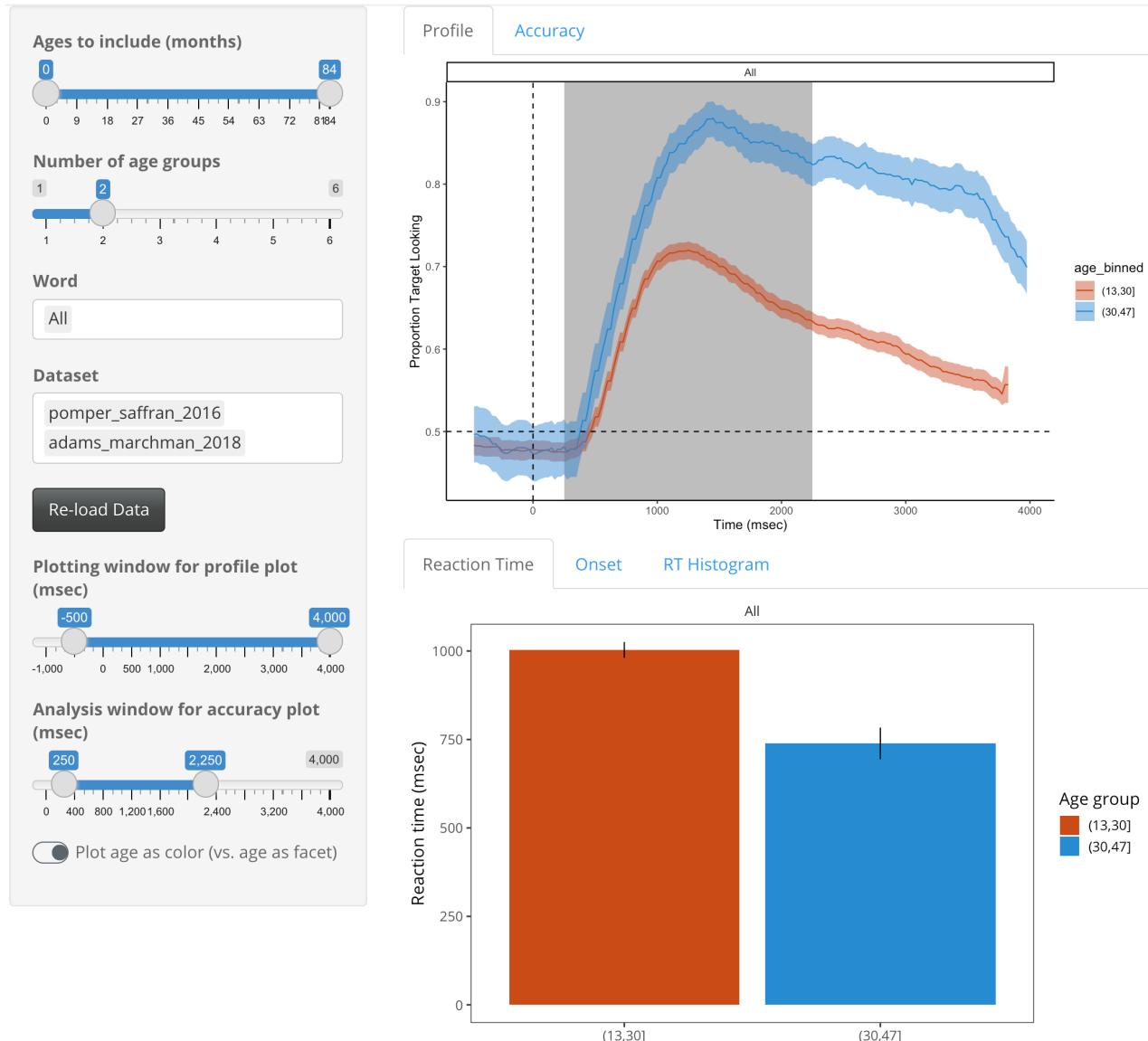


Figure 3. Screenshot of the Peekbank Shiny app, which shows a variety of standard analysis plots as a function of user-selected datasets, words, age ranges, and analysis windows. Shown here are mean reaction time and proportion target looking over time by age group for two selected datasets.

448

Peekbank in Action

449 In the following section, we provide examples of how users can access and analyze the
450 data in Peekbank. First, we provide an overview of some general properties of the datasets
451 in the database. We then demonstrate two potential use-cases for Peekbank data. In each
452 case, we provide sample code to demonstrate the ease of doing simple analyses using the
453 database. Our first example shows how we can investigate the findings of a classic study.
454 This type of investigation can be a very useful exercise for teaching students about best
455 practices for data analysis (e.g., Hardwicke et al., 2018) and also provides an easy way to
456 explore looking-while-listening time course data in a standardized format. Our second
457 example shows an exploration of developmental changes in the recognition of particular
458 words. Besides its theoretical interest (which we will explore more fully in subsequent work),
459 this type of analysis could in principle be used for optimizing the stimuli for new
460 experiments, especially as the Peekbank dataset grows and gains coverage over a greater
461 number of items. All analyses are conducted using R [Version 4.1.1; R Core Team (2021)]³

³ We, furthermore, used the R-packages *dplyr* [Version 1.0.7; Wickham, François, Henry, and Müller (2021)], *forcats* [Version 0.5.1; Wickham (2021a)], *ggplot2* [Version 3.3.5; Wickham (2016)], *ggthemes* [Version 4.2.4; Arnold (2021)], *here* [Version 1.0.1; Müller (2020)], *papaja* [Version 0.1.0.9997; Aust and Barth (2020)], *peekbankr* [Version 0.1.1.9002; Braginsky, MacDonald, and Frank (2021)], *purrr* [Version 0.3.4; Henry and Wickham (2020)], *readr* [Version 2.0.1; Wickham and Hester (2021)], *stringr* [Version 1.4.0; Wickham (2019)], *tibble* [Version 3.1.4; Müller and Wickham (2021)], *tidyverse* [Version 1.3.1; Wickham et al. (2019)], *tinylabels* (Barth, 2021), *viridis* [Version 0.6.1; Garnier et al. (2021a); Garnier et al. (2021b)], *viridisLite* [Version 0.4.0; Garnier et al. (2021b)], and *xtable* [Version 1.8.4; Dahl, Scott, Roosen, Magnusson, and Swinton (2019)].

⁴⁶² **General Descriptives**

Study Citation	Unique Items	Prop. Target	95% CI
Adams et al., 2018	8	0.65	[0.63, 0.67]
Byers-Heinlein et al., 2017	6	0.55	[0.52, 0.58]
Casillas et al., 2017	30	0.59	[0.54, 0.63]
Fernald et al., 2013	12	0.65	[0.63, 0.67]
Frank et al., 2016	24	0.64	[0.6, 0.68]
Garrison et al., 2020	87	0.60	[0.56, 0.64]
Hurtado et al., 2007	8	0.59	[0.55, 0.63]
Hurtado et al., 2008	12	0.61	[0.59, 0.63]
Mahr et al., 2015	10	0.71	[0.68, 0.74]
Perry et al., 2017	12	0.61	[0.58, 0.63]
Pomper & Saffran, 2016	40	0.77	[0.75, 0.8]
Pomper & Saffran, 2019	16	0.74	[0.72, 0.75]
Potter & Lew-Williams, unpub.	16	0.65	[0.61, 0.68]
Potter et al., 2019	8	0.63	[0.58, 0.67]
Ronfard et al., 2021	8	0.69	[0.65, 0.73]
Swingley & Aslin, 2002	22	0.57	[0.55, 0.59]
Weisleder & Fernald, 2013	12	0.63	[0.6, 0.66]
Yurovsky & Frank, 2017	6	0.63	[0.62, 0.65]
Yurovsky et al., 2013	6	0.61	[0.6, 0.63]
Yurovsky et al., unpub.	10	0.61	[0.57, 0.65]

Table 2

Average proportion target looking in each dataset.

⁴⁶³ One of the values of the uniform data format we use in Peekbank is the ease of
⁴⁶⁴ providing cross-dataset descriptions that can give an overview of some of the general
⁴⁶⁵ patterns found in our data. A first broad question is about the degree of accuracy in word
⁴⁶⁶ recognition found across studies. In general, participants demonstrated robust, above-chance
⁴⁶⁷ word recognition in each dataset (chance=0.5). Table 2 shows the average proportion of
⁴⁶⁸ target looking within a standard critical window of 367-2000ms after the onset of the label
⁴⁶⁹ for each dataset (Swingley & Aslin, 2002). Proportion target looking was generally higher for
⁴⁷⁰ familiar words ($M = 0.66$, 95% CI = [0.65, 0.67], $n = 1543$) than for novel words learned
⁴⁷¹ during the experiment ($M = 0.59$, 95% CI = [0.58, 0.61], $n = 822$).

⁴⁷² A second question of interest is about the variability across items (i.e., target labels)
⁴⁷³ within specific studies. Some studies use a smaller set of items (e.g., 8 nouns, Adams et al.,
⁴⁷⁴ 2018) while others use dozens of different items (e.g., Garrison, Baudet, Breitfeld, Aberman,
⁴⁷⁵ & Bergelson, 2020). Figure 4 gives an overview of the variability in proportion looking to the

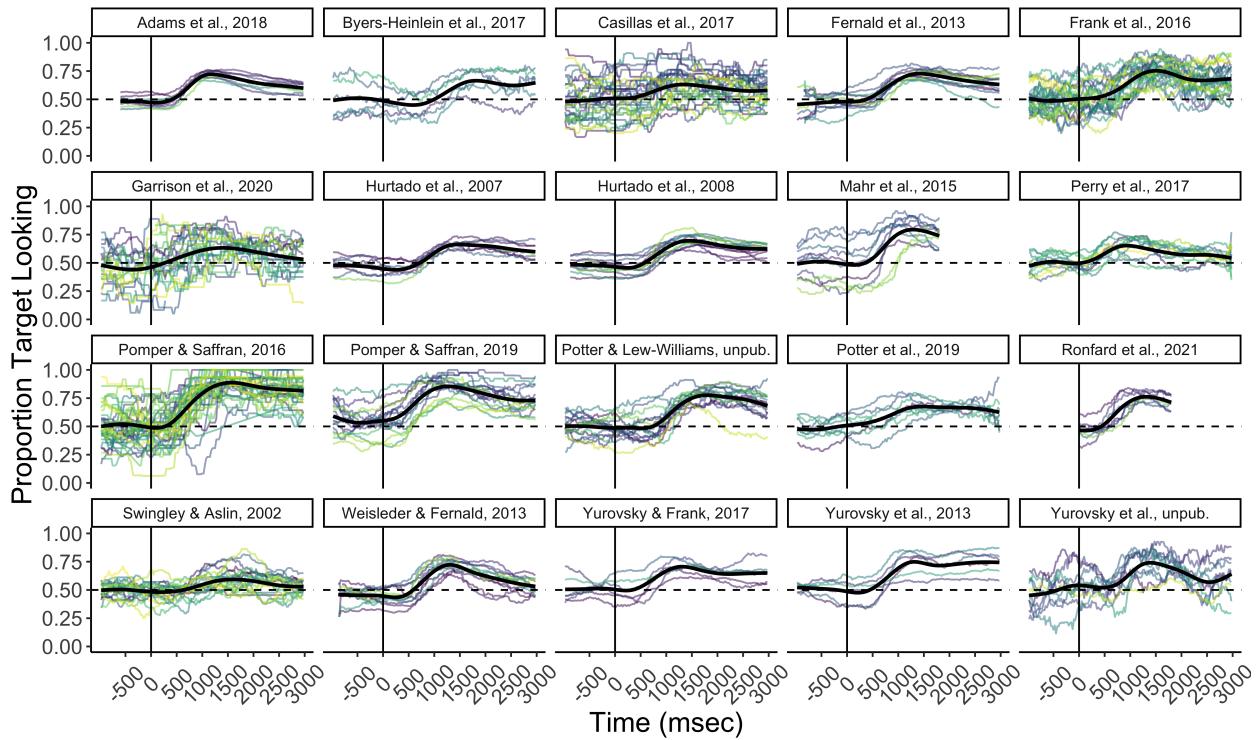


Figure 4. Item-level variability in proportion target looking within each dataset (chance=0.5). Time is centered on the onset of the target label (vertical line). Colored lines represent specific target labels. Black lines represent smoothed average fits based on a general additive model using cubic splines.

476 target item for individual words in each dataset. Although all datasets show a gradual rise in
 477 average proportion target looking over chance performance, the number of unique target
 478 labels and their associated accuracy vary widely across datasets.

479 Investigating prior findings: Swingley and Aslin (2002)

480 Swingley and Aslin (2002) investigated the specificity of 14-16-month-olds' word
 481 representations using the looking-while-listening paradigm, asking whether recognition would
 482 be slower and less accurate for mispronunciations, e.g. *opal* (mispronunciation) instead of
 483 *apple* (correct pronunciation).⁴ In this short vignette, we show how easily the data in

⁴ The original paper investigated both close (e.g., *opple*, /apl/) and distant (e.g., *opal*, /opl/) mispronunciations. For simplicity, here we combine both mispronunciation conditions since the close vs. distant mispronunciation manipulation showed no effect in the original paper.

484 Peekbank can be used to visualize this result. Our goal here is not to provide a precise
 485 analytical reproduction of the analyses reported in the original paper, but rather to
 486 demonstrate the use of the Peekbank framework to analyze datasets of this type. In
 487 particular, because Peekbank uses a uniform data import standard, it is likely that there will
 488 be minor numerical discrepancies between analyses on Peekbank data and analyses that use
 489 another processing pipeline.

```
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")
```

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

495 As the code above shows, once the data are joined, condition information for each
 496 timepoint is present and so we can easily filter out filler trials and set up the conditions for
 497 further analysis.

```
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()))
```

498 The final step in our analysis is to create a summary dataframe using `dplyr`

499 commands. We first group the data by timestep, participant, and condition and compute the
 500 proportion looking at the correct image. We then summarize again, averaging across
 501 participants, computing both means and 95% confidence intervals (via the approximation of
 502 1.96 times the standard error of the mean). The resulting dataframe can be used for
 503 visualization of the time course of looking.

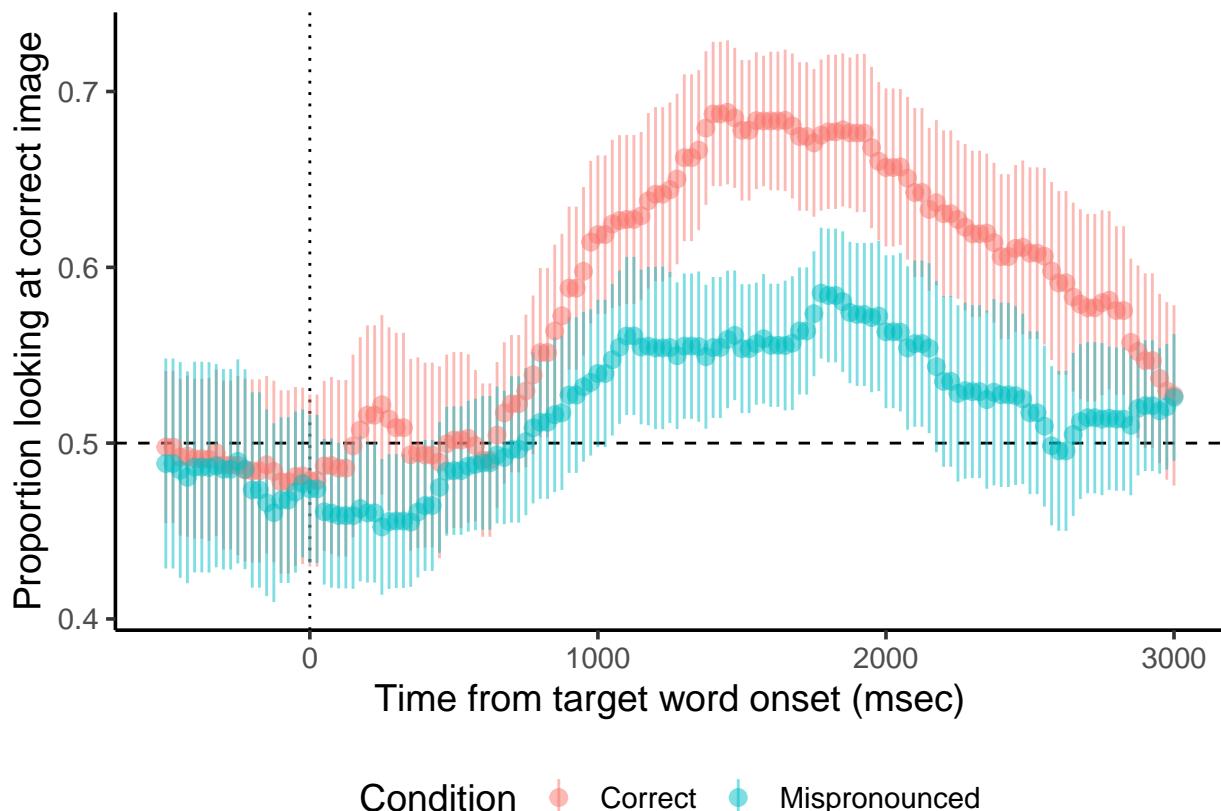


Figure 5. Proportion looking at the correct referent by time from the point of disambiguation (the onset of the target noun) in Swingley & Aslin (2002). Colors show the two pronunciation conditions; points give means and ranges show 95% confidence intervals. The dotted line shows the point of disambiguation and the dashed line shows chance performance.

504 Figure 5 shows the average time course of looking for the two conditions, as produced

505 by the code above. Looks after the correctly pronounced noun appeared both faster

506 (deviating from chance earlier) and more accurate (showing a higher asymptote). Overall,
507 this example demonstrates the ability to produce this visualization in just a few lines of code.

508 **Item analyses**

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

514 Our first step is to collect and join the data from the relevant tables including
515 timepoint data, trial and stimulus data, and administration data (for participant ages). We
516 join these into a single dataframe for easy manipulation; this dataframe is a common
517 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)
```

518 Next we select a set of four target words (chosen based on having more than 100
 519 children contributing data for each word across several one-year age groups). We create age
 520 groups, aggregate, and compute timepoint-by-timepoint confidence intervals using the z
 521 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 = sum(aoi == "target") /
    sum(aoi %in% c("target", "distractor"))) |>
  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())
```

522 Finally, we plot the data as time courses split by age. Our plotting code is shown below
 523 (with styling commands removed for clarity). Figure 6 shows the resulting plot, with time
 524 courses for each of three (rather coarse) age bins. Although some baseline effects are visible
 525 across items, we still see clear and consistent increases in looking to the target, with the
 526 increase appearing earlier and in many cases asymptoting at a higher level for older children.

527 This simple averaging approach is a proof-of-concept to demonstrate some of the
 528 potential of the Peekbank dataset. An eye-movement trajectory on an individual trial reflects

529 myriad factors, including the age and ability of the child, the target and distractor stimuli on
 530 that trial, the position of the trial within the experiment, and the general parameters of the
 531 experiment (for example, stimulus timing, eye-tracker type and calibration, etc.). Although
 532 we often neglect these statistically in the analysis of individual experiments – for example,
 533 averaging across items and trial orders – they may lead to imprecision when we average
 534 across multiple studies in Peekbank. For example, studies with older children may use more
 535 difficult items or faster trial timing, leading to the impression that children’s abilities overall
 536 increase more slowly than they do in fact. Even in our example in Figure 6, we see hints of
 537 this confounding – for example, the low baseline looks to *apple* may be an artifact of an
 538 attractive distractor being paired with this item in one or two studies. In future work, we
 539 hope to introduce model-based analytic methods that use mixed effects regression to factor
 540 out study-level and individual-level variance in order to recover developmental effects more
 541 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)
```

542

Discussion

543 Theoretical progress in understanding child development requires rich datasets, but
 544 collecting child data is expensive, difficult, and time-intensive. Recent years have seen a
 545 growing effort to build open source tools and pool research efforts to meet the challenge of
 546 building a cumulative developmental science (Bergmann et al., 2018; Frank, Braginsky,
 547 Yurovsky, & Marchman, 2017; Sanchez et al., 2019; The ManyBabies Consortium, 2020).

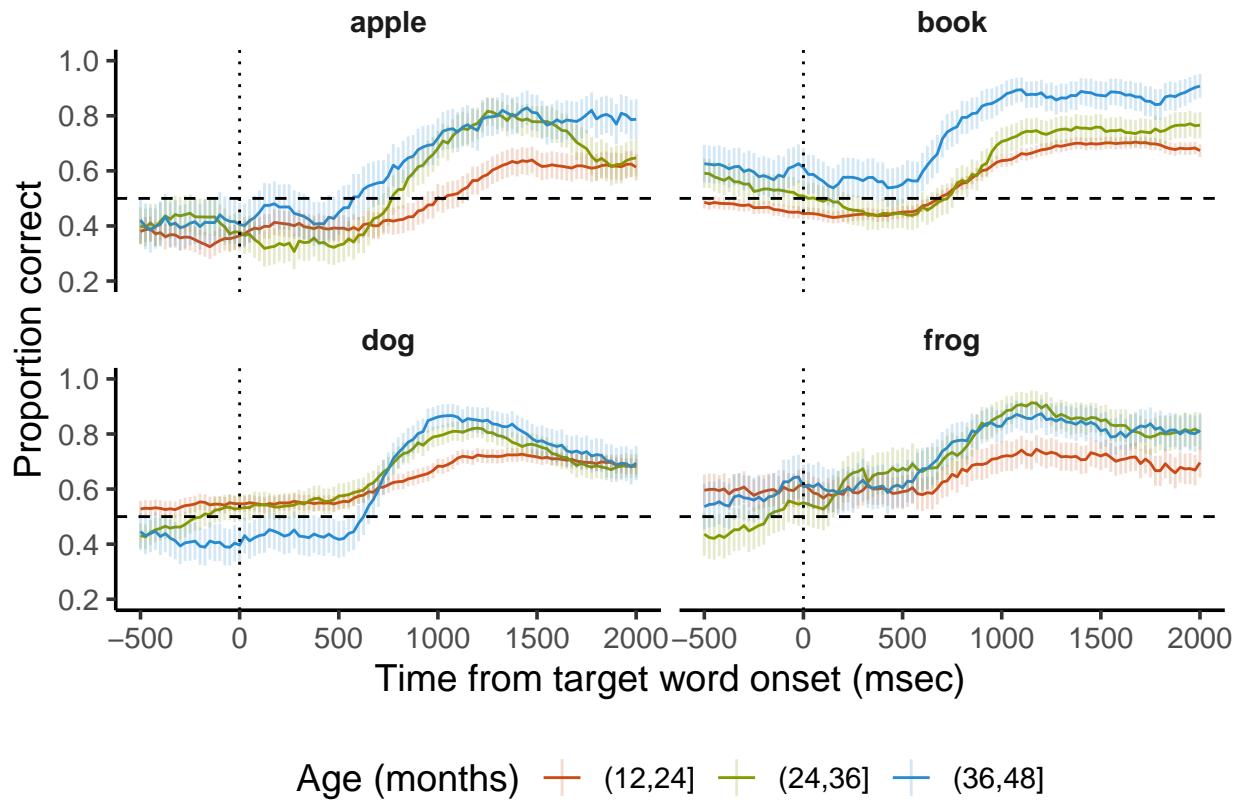


Figure 6. Time course plot for four well-represented target items in the Peekbank dataset, split by three age groups. Each line represents children's average looking to the target image after the onset of the target label (dashed vertical line). Error bars represent 95% CIs.

548 The Peekbank project expands on these efforts by building an infrastructure for aggregating
 549 eye-tracking data across studies, with a specific focus on the looking-while-listening
 550 paradigm. This paper presents an overview of the structure of the database, shows how users
 551 can access the database, and demonstrates how it can be used both to investigate prior
 552 experiments and to synthesize data across studies.

553 The current database has a number of limitations, particularly in the number and
 554 diversity of datasets it contains. With 20 datasets currently available in the database,
 555 idiosyncrasies of particular designs and condition manipulations still have a substantial
 556 influence on the results of particular analyses, as discussed above in our item analysis
 557 example. Expanding the set of distinct datasets will allow us to increase the number of
 558 datasets that contain specific items, leading to more robust generalizations across the many

559 sources of variation that are confounded within studies (e.g., item set, participant age range,
560 and specific experimental parameters). A critical next step will be the development of
561 analytic models that take this structure into account in making generalizations across
562 datasets.

563 A second limitation stems from the fact that the database represents a convenience
564 sample of data readily available to the Peekbank team, which leads the database to be
565 relatively homogeneous in a number of key respects. First, the datasets primarily come from
566 labs that share similar theoretical perspectives and implement the looking-while-listening
567 method in similar ways. The current database is also limited by the relatively homogeneous
568 background of its participants, both with respect to language (almost entirely monolingual
569 native English speakers) and cultural background (Henrich, Heine, & Norenzayan, 2010;
570 Muthukrishna et al., 2020). Increasing the diversity of lab sources, participant backgrounds,
571 and languages will expand the scope of the generalizations we can form about child word
572 recognition, while also providing new opportunities for describing cross-lab, cross-cultural,
573 and cross-linguistic variation.

574 Towards the goal of expanding our database, we invite researchers to contribute their
575 data. On the Peekbank website we provide technical documentation for potential
576 contributors. Although we anticipate being involved in most new data imports, as discussed
577 above, our import process is transparently documented and the repository contains examples
578 for most commonly-used eye-trackers.

579 Contributing data to an open repository also can raise questions about participant
580 privacy. Potential contributors should consult with their local institutional review boards for
581 guidance on any challenges, but we do not foresee obstacles because of the de-identified
582 nature of the data. Under United States regulation, all data contributed to Peekbank are
583 considered de-identified and hence not considered “human subjects data”; hence, institutional
584 review boards should not regulate this contribution process. Under the European Union’s

585 Generalized Data Protection Regulation (GDPR), labs may need to take special care to
586 provide a separate set of participant identifiers that can never be re-linked to their own
587 internal records.

588 While the current database is focused on studies of word recognition, the tools and
589 infrastructure developed in the project can in principle be used to accommodate any
590 eye-tracking paradigm, opening up new avenues for insights into cognitive development.
591 Gaze behavior has been at the core of many key advances in our understanding of infant
592 cognition (Aslin, 2007; Baillargeon, Spelke, & Wasserman, 1985; Bergelson & Swingley, 2012;
593 Fantz, 1963; Liu, Ullman, Tenenbaum, & Spelke, 2017; Quinn, Eimas, & Rosenkrantz, 1993).
594 Aggregating large datasets of infant looking behavior in a single, openly-accessible format
595 promises to bring a fuller picture of infant cognitive development into view.

596

References

- 597 Adams, K. A., Marchman, V. A., Loi, E. C., Ashland, M. D., Fernald, A., & Feldman,
598 H. M. (2018). Caregiver talk and medical risk as predictors of language outcomes
599 in full term and preterm toddlers. *Child Development*, 89(5), 1674–1690.
- 600 Arnold, J. B. (2021). *Ggthemes: Extra themes, scales and geoms for 'ggplot2'*.
601 Retrieved from <https://CRAN.R-project.org/package=ggthemes>
- 602 Aslin, R. N. (2007). What's in a look? *Developmental Science*, 10(1), 48–53.
- 603 Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown*.
604 Retrieved from <https://github.com/crsh/papaja>
- 605 Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in
606 five-month-old infants. *Cognition*, 20(3), 191–208.
607 [https://doi.org/10.1016/0010-0277\(85\)90008-3](https://doi.org/10.1016/0010-0277(85)90008-3)
- 608 Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., ...
609 Treiman, R. (2007). The English Lexicon project. *Behavior Research Methods*,
610 39(3), 445–459. <https://doi.org/10.3758/BF03193014>
- 611 Barth, M. (2021). *tinylabes: Lightweight variable labels*. Retrieved from
612 <https://github.com/mariusbarth/tinylabes>
- 613 Bergelson, E. (2020). The comprehension boost in early word learning: Older infants
614 are better learners. *Child Development Perspectives*, 14(3), 142–149.
- 615 Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the
616 meanings of many common nouns. *Proceedings of the National Academy of
617 Sciences*, 109(9), 3253–3258.

- 618 Bergelson, E., & Swingley, D. (2013). The acquisition of abstract words by young
619 infants. *Cognition*, 127(3), 391–397.
- 620 Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C.,
621 & Cristia, A. (2018). Promoting replicability in developmental research through
622 meta-analyses: Insights from language acquisition research. *Child Development*,
623 89(6), 1996–2009.
- 624 Bleses, D., Makransky, G., Dale, P. S., Højen, A., & Ari, B. A. (2016). Early
625 productive vocabulary predicts academic achievement 10 years later. *Applied
626 Psycholinguistics*, 37(6), 1461–1476.
- 627 Braginsky, M., MacDonald, K., & Frank, M. (2021). *Peekbankr: Accessing the
628 peekbank database*. Retrieved from <http://github.com/langcog/peekbankr>
- 629 Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2021). Six solutions for more
630 reliable infant research. *Infant and Child Development*.
631 <https://doi.org/https://doi.org/10.1002/icd.2296>
- 632 Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., & Swinton, J. (2019). *Xtable:
633 Export tables to LaTeX or HTML*. Retrieved from
634 <https://CRAN.R-project.org/package=xtable>
- 635 DeBolt, M. C., Rhemtulla, M., & Oakes, L. M. (2020). Robust data and power in
636 infant research: A case study of the effect of number of infants and number of
637 trials in visual preference procedures. *Infancy*, 25(4), 393–419.
638 <https://doi.org/10.1111/infa.12337>
- 639 Fantz, R. L. (1963). Pattern vision in newborn infants. *Science*, 140(3564), 296–297.
- 640 Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language

- 641 processing skill and vocabulary are evident at 18 months. *Developmental Science*,
642 16(2), 234–248. <https://doi.org/10.1111/desc.12019>
- 643 Fernald, A., Pinto, J. P., Swingley, D., Weinberg, A., & McRoberts, G. W. (1998).
644 Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological
645 Science*, 9(3), 228–231.
- 646 Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while
647 listening: Using eye movements to monitor spoken language comprehension by
648 infants and young children. In I. A. Sekerina, E. M. Fernandez, & H. Clahsen
649 (Eds.), *Developmental psycholinguistics: On-line methods in children's language
650 processing* (pp. 97–135). Amsterdam: John Benjamins.
- 651 Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ...
652 Yurovsky, D. (2017). A collaborative approach to infant research: Promoting
653 reproducibility, best practices, and theory-building. *Infancy*, 22(4), 421–435.
654 <https://doi.org/10.1111/infa.12182>
- 655 Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank:
656 An open repository for developmental vocabulary data. *Journal of Child
657 Language*, 44(3), 677–694.
- 658 Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability
659 and Consistency in Early Language Learning: The Wordbank Project*. Cambridge,
660 MA: MIT Press.
- 661 Garnier, Simon, Ross, Noam, Rudis, Robert, ... Cédric. (2021a). *viridis -
662 colorblind-friendly color maps for r*. <https://doi.org/10.5281/zenodo.4679424>
- 663 Garnier, Simon, Ross, Noam, Rudis, Robert, ... Cédric. (2021b). *viridis -
664 colorblind-friendly color maps for r*. <https://doi.org/10.5281/zenodo.4679424>

- 665 Garrison, H., Baudet, G., Breitfeld, E., Aberman, A., & Bergelson, E. (2020).
666 Familiarity plays a small role in noun comprehension at 12–18 months. *Infancy*,
667 25(4), 458–477.
- 668 Gautheron, L., Rochat, N., & Cristia, A. (2021). Managing, storing, and sharing
669 long-form recordings and their annotations. *PsyArXiv*. Retrieved from
670 <https://doi.org/10.31234/osf.io/w8trm>
- 671 Golinkoff, R. M., Ma, W., Song, L., & Hirsh-Pasek, K. (2013). Twenty-five years
672 using the intermodal preferential looking paradigm to study language acquisition:
673 What have we learned? *Perspectives on Psychological Science*, 8(3), 316–339.
- 674 Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P.,
675 ... Poldrack, R. A. (2016). The brain imaging data structure, a format for
676 organizing and describing outputs of neuroimaging experiments. *Scientific Data*,
677 3(1), 160044. <https://doi.org/10.1038/sdata.2016.44>
- 678 Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C.,
679 Kidwell, M. C., ... Frank, M. C. (2018). Data availability, reusability, and
680 analytic reproducibility: Evaluating the impact of a mandatory open data policy
681 at the journal *Cognition*. *Royal Society Open Science*, 5(8).
682 <https://doi.org/10.1098/rsos.180448>
- 683 Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?
684 *Behavioral and Brain Sciences*, 33(2-3), 61–83.
685 <https://doi.org/10.1017/S0140525X0999152X>
- 686 Henry, L., & Wickham, H. (2020). *Purrrr: Functional programming tools*. Retrieved
687 from <https://CRAN.R-project.org/package=purrr>
- 688 Hirsh-Pasek, K., Cauley, K. M., Golinkoff, R. M., & Gordon, L. (1987). The eyes

- 689 have it: Lexical and syntactic comprehension in a new paradigm. *Journal of Child*
690 *Language*, 14(1), 23–45.
- 691 Hurtado, N., Marchman, V. A., & Fernald, A. (2007). Spoken word recognition by
692 Latino children learning Spanish as their first language. *Journal of Child*
693 *Language*, 34(2), 227–249. <https://doi.org/10.1017/S0305000906007896>
- 694 Hurtado, N., Marchman, V. A., & Fernald, A. (2008). Does input influence uptake?
695 Links between maternal talk, processing speed and vocabulary size in
696 Spanish-learning children. *Developmental Science*, 11(6), 31–39.
697 <https://doi.org/10.1111/j.1467-7687.2008.00768.x>
- 698 Lewis, M., Braginsky, M., Tsuji, S., Bergmann, C., Piccinini, P. E., Cristia, A., &
699 Frank, M. C. (2016). *A Quantitative Synthesis of Early Language Acquisition*
700 *Using Meta-Analysis*. *PsyArXiv*. <https://doi.org/10.31234/osf.io/htsjm>
- 701 Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid
702 use of grammatical gender in spoken word recognition. *Psychological Science*,
703 18(3), 193–198.
- 704 Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old
705 infants infer the value of goals from the costs of actions. *Science*, 358(6366),
706 1038–1041. <https://doi.org/10.1126/science.aag2132>
- 707 MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk*. Mahwah,
708 NJ: Lawrence Erlbaum Associates.
- 709 Mahr, T., McMillan, B. T. M., Saffran, J. R., Ellis Weismer, S., & Edwards, J. (2015).
710 Anticipatory coarticulation facilitates word recognition in toddlers. *Cognition*,
711 142, 345–350. <https://doi.org/10.1016/j.cognition.2015.05.009>

- 712 Marchman, V. A., Loi, E. C., Adams, K. A., Ashland, M., Fernald, A., & Feldman, H.
713 M. (2018). Speed of language comprehension at 18 months old predicts
714 school-relevant outcomes at 54 months old in children born preterm. *Journal of*
715 *Developmental & Behavioral Pediatrics*, 39(3), 246–253.
- 716 Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A.,
717 McInerney, J., & Thue, B. (2020). Beyond Western, Educated, Industrial, Rich,
718 and Democratic (WEIRD) psychology: Measuring and mapping scales of cultural
719 and psychological distance. *Psychological Science*, 31(6), 678–701.
- 720 Müller, K. (2020). *Here: A simpler way to find your files*. Retrieved from
721 <https://CRAN.R-project.org/package=here>
- 722 Müller, K., & Wickham, H. (2021). *Tibble: Simple data frames*. Retrieved from
723 <https://CRAN.R-project.org/package=tibble>
- 724 Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A.,
725 ... Vazire, S. (2022). Replicability, robustness, and reproducibility in
726 psychological science. *Annual Review of Psychology*, 73, 719–748.
727 <https://doi.org/https://doi.org/10.31234/osf.io/ksfvq>
- 728 Peter, M. S., Durrant, S., Jessop, A., Bidgood, A., Pine, J. M., & Rowland, C. F.
729 (2019). Does speed of processing or vocabulary size predict later language growth
730 in toddlers? *Cognitive Psychology*, 115, 101238.
- 731 Potter, C., & Lew-Williams, C. (unpublished). Behold the canine!: How does
732 toddlers' knowledge of typical frames and familiar words interact to influence their
733 sentence processing?
- 734 Quinn, P. C., Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for representations
735 of perceptually similar natural categories by 3-month-old and 4-month-old infants.

- 736 *Perception*, 22(4), 463–475. <https://doi.org/10.1068/p220463>
- 737 R Core Team. (2021). *R: A language and environment for statistical computing*.
- 738 Vienna, Austria: R Foundation for Statistical Computing. Retrieved from
- 739 <https://www.R-project.org/>
- 740 Ronfard, S., Wei, R., & Rowe, M. L. (2021). Exploring the linguistic, cognitive, and
- 741 social skills underlying lexical processing efficiency as measured by the
- 742 looking-while-listening paradigm. *Journal of Child Language*, 1–24.
- 743 <https://doi.org/10.1017/S0305000921000106>
- 744 Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank,
- 745 M. C. (2019). childe-db: A flexible and reproducible interface to the child
- 746 language data exchange system. *Behavior Research Methods*, 51(4), 1928–1941.
- 747 <https://doi.org/10.3758/s13428-018-1176-7>
- 748 Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form
- 749 representations of 14-month-olds. *Psychological Science*, 13(5), 480–484.
- 750 <https://doi.org/10.1111/1467-9280.00485>
- 751 The ManyBabies Consortium. (2020). Quantifying sources of variability in infancy
- 752 research using the infant-directed speech preference. *Advances in Methods and*
- 753 *Practices in Psychological Science*, 3(1), 24–52.
- 754 Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in
- 755 6-month-olds. *Psychological Science*, 10(2), 172–175.
- 756 <https://doi.org/10.1111/1467-9280.00127>
- 757 Wass, S. V., Smith, T. J., & Johnson, M. H. (2013). Parsing eye-tracking data of
- 758 variable quality to provide accurate fixation duration estimates in infants and
- 759 adults. *Behavior Research Methods*, 45(1), 229–250.

- 760 https://doi.org/10.3758/s13428-012-0245-6
- 761 Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language
762 experience strengthens processing and builds vocabulary. *Psychological Science*,
763 24(11), 2143–2152. https://doi.org/10.1177/0956797613488145
- 764 Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag
765 New York. Retrieved from https://ggplot2.tidyverse.org
- 766 Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string*
767 operations. Retrieved from https://CRAN.R-project.org/package=stringr
- 768 Wickham, H. (2021a). *Forcats: Tools for working with categorical variables (factors)*.
769 Retrieved from https://CRAN.R-project.org/package=forcats
- 770 Wickham, H. (2021b). *Tidyr: Tidy messy data*. Retrieved from
771 https://CRAN.R-project.org/package=tidyr
- 772 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ...
773 Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*,
774 4(43), 1686. https://doi.org/10.21105/joss.01686
- 775 Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of*
776 *data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr
- 777 Wickham, H., & Hester, J. (2021). *Readr: Read rectangular text data*. Retrieved from
778 https://CRAN.R-project.org/package=readr
- 779 Yurovsky, D., Wade, A., Kraus, A. M., Gengoux, G. W., Hardan, A. Y., & Frank, M.
780 C. (unpublished). Developmental changes in the speed of social attention in early
781 word learning.

- 782 Zettersten, M., Bergey, C., Bhatt, N., Boyce, V., Braginsky, M., Carstensen, A., ...
- 783 Frank, M. C. (2021). Peekbank: Exploring children's word recognition through an
- 784 open, large-scale repository for developmental eye-tracking data. *Proceedings of*
- 785 *the 43rd Annual Conference of the Cognitive Science Society.*