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

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

[1] Dept. of Psychology, Princeton University

[2] Dept. of Psychology, University of Chicago

[3] Scripps College

[4] Dept. of Psychology, Stanford University

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

[6] Dept. of Psychology, Carnegie Mellon University

[7] Core Technology, McD Tech Labs

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

[9] Dept. of Psychology, UT Austin

[10] Dept. of Psychology, UC San Diego

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

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

The authors made the following contributions. Peekbank team: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Peekbank team, Postal address. E-mail: my@email.com

Abstract

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

Keywords: keywords

Word count: X

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

Introduction

Across their first years of life, children learn words in their native tongues at a rapid pace (Frank, Braginsky, Yurovsky, & Marchman, 2021). [notes about the size/ pace] A key part of the word learning process is children's emerging ability to rapidly process words and link them to relevant meanings – often referred to as word recognition. Measuring early word recognition offers insight into children's early word representations and the processes supporting early language comprehension (Bergelson, 2020). Word recognition skills are also thought to build a foundation for children's subsequent language development. Past research has found that early word recognition efficiency is predictive of later linguistic and general cognitive outcomes (Bleses, Makransky, Dale, Højen, & Ari, 2016; Marchman et al., 2018). While word recognition is a central part of children's language development, mapping the trajectory of word recognition skills has remained elusive. Studies investigating children's word recognition are typically limited in scope to experiments in individual labs involving small samples tested on a limited set of items. This limitation in scale makes it difficult to understand developmental changes in children's word knowledge at a broad scale. Peekbank provides an openly accessible database of eye-tracking data of children's word recognition, with the primary goal of facilitating the study of developmental changes in children's word knowledge and recognition speed.

The "Looking-While-Listening" Paradigm

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

Measuring developmental change in word recognition

While the looking-while-listening paradigm has been highly fruitful in advancing understanding of early word knowledge, fundamental questions remain both about the trajectory of children's word recognition ability and the nature of the method itself. One central question is how to measure developmental change in word recognition. A key idea in the language learning literature is that processing speed - the ability to quickly link a word with its referent - supports language learning. Age-related changes in speed of processing are thought to accelerate infants' subsequent language learning: the faster infants are able to process incoming speech input, the better able they become to learn from their language environment. Similarly, longitudinal analyses have found that individual differences in word recognition speed predict linguistic and cognitive outcomes later in childhood (e.g., Marchman & Fernald, 2008). However, measuring increases in the speed and accuracy of word recognition faces the challenge of distinguishing developmental changes in word recognition skill from changes in knowledge of specific words. This problem is particularly thorny in child development, since the number of items that can be tested within a single session is limited and items must be selected in an age-appropriate manner (Peter et al., 2019). Measuring developmental change therefore requires large-scale datasets with a range of items, in order to generalize age-related changes across words.

Developing methodological best-practices

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

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

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

Design and Technical Approach

Database Framework.

The Peekbank data framework consists of three components: (1) processing raw experimental datasets; (2) populating a relational database; and (3) providing an interface to the database (Fig XX). The peekds library (for the R language; R Development Core Team, 2020) helps researchers convert and validate existing datasets to use the relational format of the database. The peekbank module (Python) creates a database with the relational schema and populates it with the standardized datasets produced by peekds. The database is implemented in MySQL, an industry standard relational database, which may be accessed by a variety of programming languages over the internet. The peekbankr library (R) provides an application programming interface, or API, that offers high-level abstractions for accessing data in Peekbank.

Data Format and Processing.

One of the main challenges in compiling a large-scale eye-tracking dataset is the lack of a shared re-usable data format across individual experiments. Researcher conventions for structuring data vary, as do the technical specifications of different devices, rendering the task of integrating datasets from different labs and data sources difficult. We developed a common, tidy format for the eye-tracking data in Peekbank to ease the process of conducting cross-dataset analyses (Wickham et al., 2019). The schema of the database is sufficiently

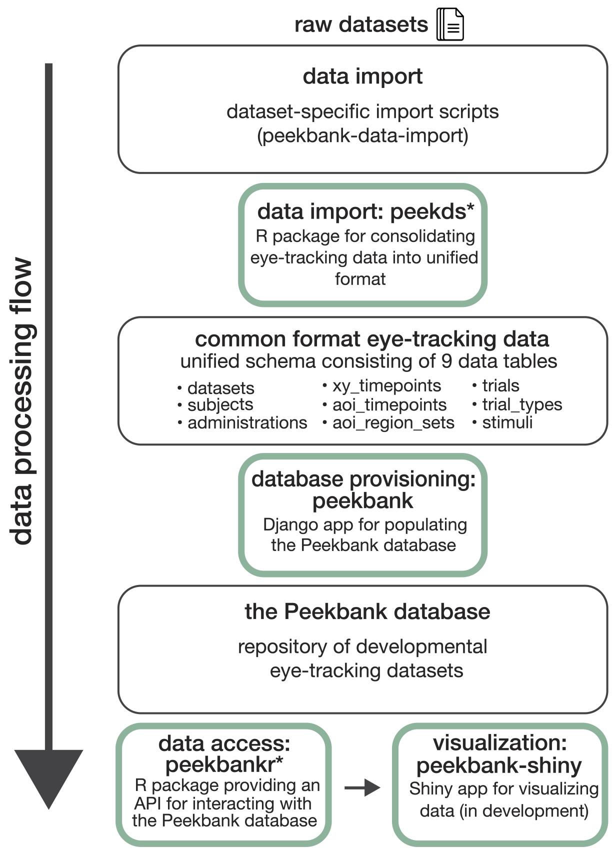


Figure 1. Overview of the Peekbank data ecosystem. Peekbank tools are highlighted in green. *custom R packages.

general to handle heterogeneous datasets, including both manually coded and automated eye-tracking data.

During data import, raw eye-tracking datasets are processed to conform to the Peekbank data schema. The centerpiece of the schema is the aoi_timepoints table (Fig XX), which records whether participants looked to the target or the distractor stimulus at each timepoint of a given trial. Additional tables track information about data sources, participant characteristics, trial characteristics, stimuli, and raw eye-tracking data. In addition to unifying the data format, we conduct several additional pre-processing steps to facilitate analyses across datasets, including resampling observations to a common sampling rate (40 Hz) and normalizing time relative to the onset of the target label.

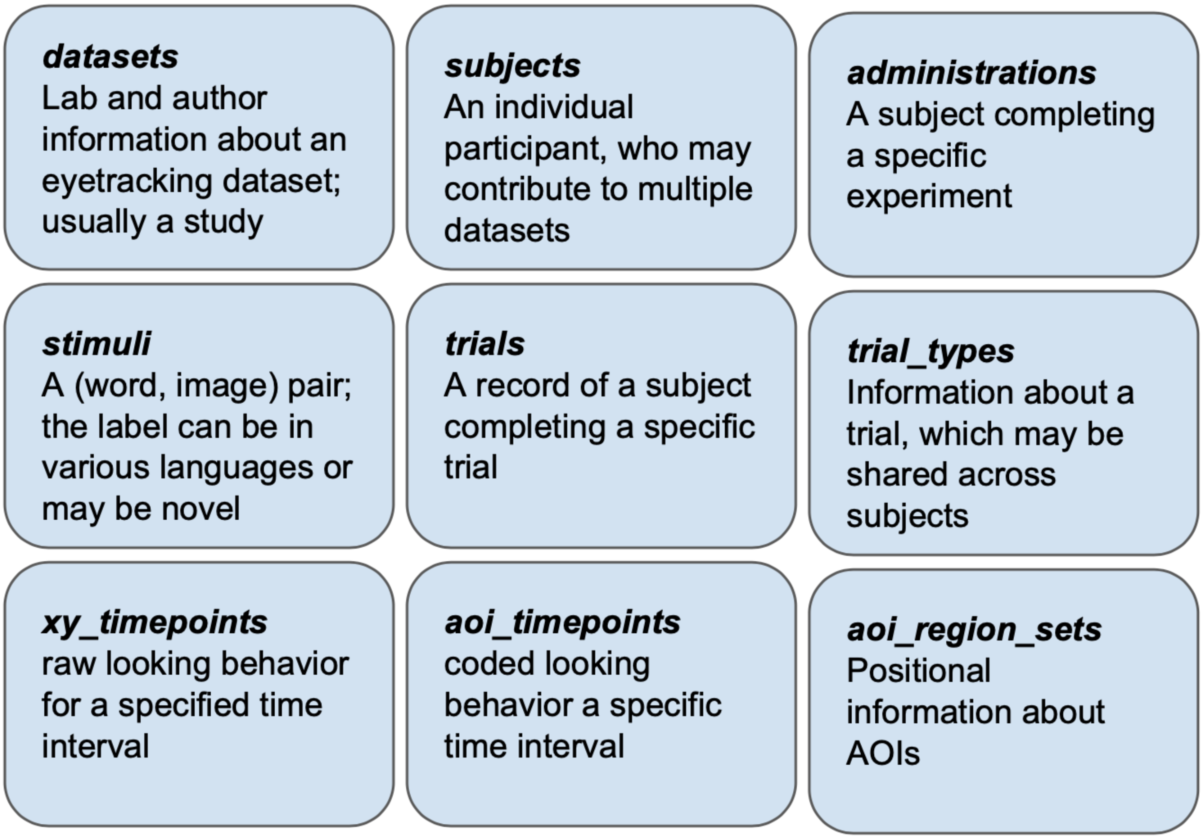


Figure 2. The Peekbank schema. Each square represents a table in the relational database.

Current Data Sources.

The database currently includes 11 looking-while-listening datasets comprising *N*=1320 total participants (Table XX). Most datasets (10 out of 11 total) consist of data from monolingual native English speakers. They span a wide age spectrum with participants ranging from 8 to 84 months of age, and are balanced in terms of gender (48% female). The datasets vary across a number of dimensions related to design and methodology, and include studies using manually coded video recordings and automated eye-tracking methods (e.g., Tobii, EyeLink) to measure gaze behavior. Most studies focused on testing familiar items, but the database also includes studies with novel pseudowords. All data (and accompanying references) are openly available on the Open Science Framework (https://osf.io/pr6wu/?view_only=07a3887eb7a24643bdc1b2612f2729de).

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

Table 1

Overview over the datasets in the current database.

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

Versioning + Expanding the database

Information about versioning approach/ regularity of updates Steps for extending the database?

Interfacing with peekbank

Shiny App

Peekbankr

Functions: connect_to_peekbank() get_datasets() get_subjects() get_administrations() get_stimuli() get_aoi_timepoints() get_trials() get_trial_types() get_xy_timepoints() get_aoi_region_sets()

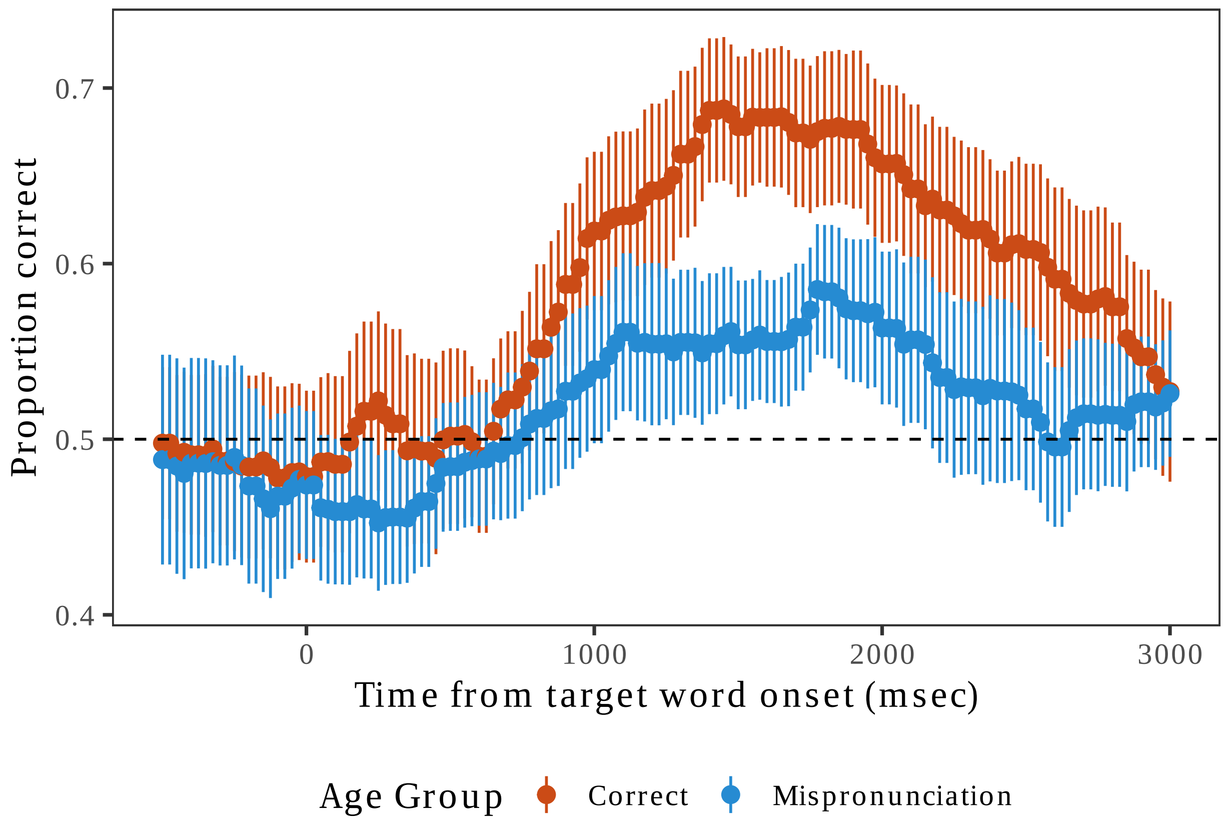
OSF site

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

Peekbank in Action

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

Computational reproducibility example: Swingley & Aslin (2002)



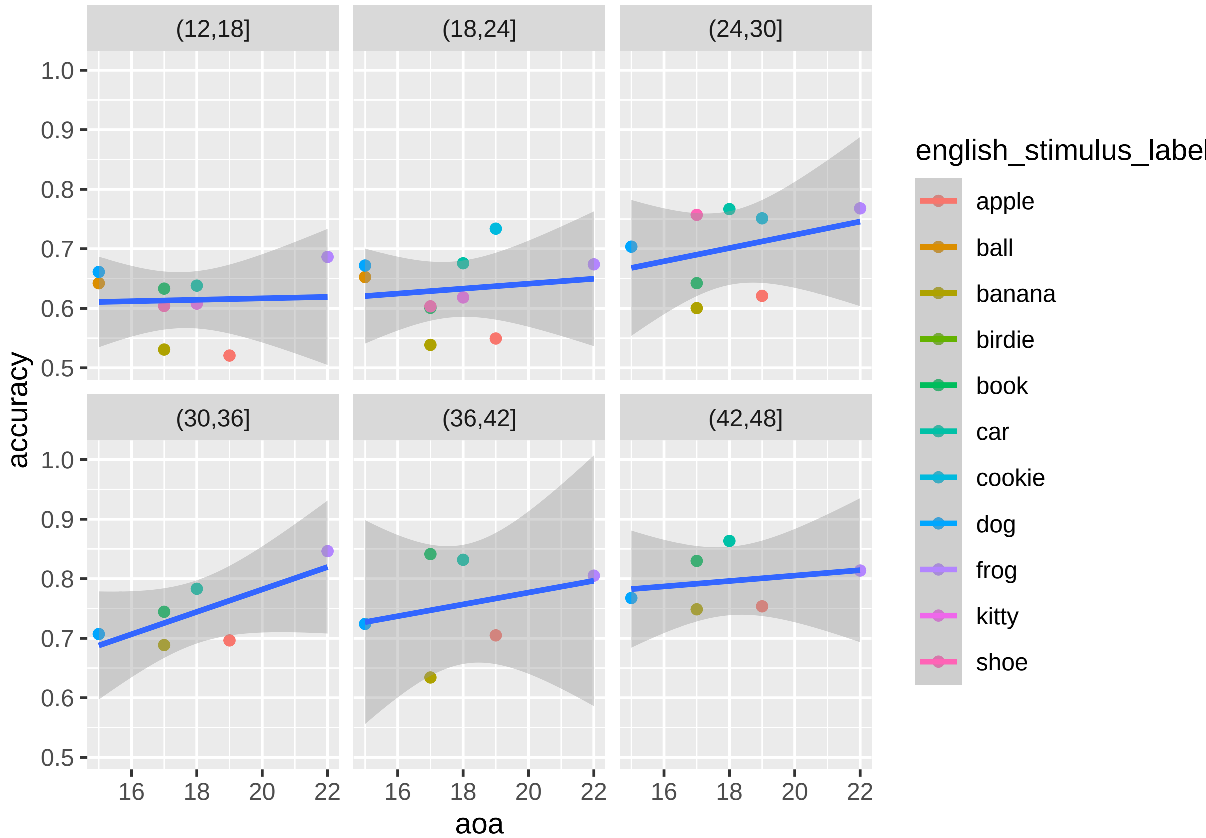
Item analyses

To illustrate the power of aggregating data across multiple datasets, we,

aspirational goal -

ALso, item selection but maybe not yet?

Links to parent report vocabulary data



Discussion/ Conclusion

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

There are a number of limitations surrounding the current scope of the database. A priority in future work will be to expand the size of the database. With 11 datasets currently available in the database, idiosyncrasies of particular designs and condition manipulations still have substantial influence on modeling results. Expanding the set of distinct datasets will allow us to increase the number of observations per item across datasets, leading to more robust generalizations across item-level variability. The current database is also limited by the relatively homogeneous background of its participants, both with respect to language (almost entirely monolingual native English speakers) and cultural background (all but one dataset comes from WEIRD populations; Muthukrishna et al., 2020). Increasing the diversity of participant backgrounds and languages will expand the scope of the generalizations we can form about child word recognition. Finally, while the current database is focused on studies of word recognition, the tools and infrastructure developed in the project can in principle be used to accommodate any eye-tracking paradigm, opening up new avenues for insights into cognitive development. Gaze behavior has been at the core of many of the key advances in our understanding of infant cognition. Aggregating large datasets of infant looking behavior in a single, openly-accessible format promises to bring a fuller picture of infant cognitive development into view.

Acknowledgements

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

We used R (Version 4.0.3; R Core Team, 2020) and the R-packages *dplyr* (Version 1.0.3; Wickham et al., 2021), *extrafont* (Version 0.17; Winston Chang, 2014), *forcats* (Version 0.5.0; Wickham, 2021a), *ggplot2* (Version 3.3.3; Wickham, 2016), *here* (Version 1.0.1; Müller, 2020), *papaja* (Version 0.1.0.9997; Aust & Barth, 2020), *peekbankr* (Version 0.1.1.9001; Braginsky, MacDonald, & Frank, 2021), *plyr* (Version 1.8.6; Wickham et al., 2021; Wickham, 2011), *png* (Version 0.1.7; Urbanek, 2013), *purrr* (Version 0.3.4; Henry & Wickham, 2020), *readr* (Version 1.4.0; Wickham & Hester, 2020), *stringr* (Version 1.4.0; Wickham, 2019), *tibble* (Version 3.0.5; Müller & Wickham, 2021), *tidyr* (Version 1.1.2; Wickham, 2021b), *tidyverse* (Version 1.3.0; Wickham, Averick, et al., 2019), and *xtable* (Version 1.8.4; Dahl, Scott, Roosen, Magnusson, & Swinton, 2019) for all our analyses.

References

Aust, F., & Barth, M. (2020). *papaja: Create APA manuscripts with R Markdown.* Retrieved from https://github.com/crsh/papaja

Braginsky, M., MacDonald, K., & Frank, M. (2021). *Peekbankr: Accessing the peekbank database.* Retrieved from http://github.com/langcog/peekbankr

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., & Swinton, J. (2019). *Xtable: Export tables to latex or html.* Retrieved from https://CRAN.R-project.org/package=xtable

Henry, L., & Wickham, H. (2020). *Purrr: Functional programming tools.* Retrieved from https://CRAN.R-project.org/package=purrr

Müller, K. (2020). *Here: A simpler way to find your files.* Retrieved from https://CRAN.R-project.org/package=here

Müller, K., & Wickham, H. (2021). *Tibble: Simple data frames.* Retrieved from https://CRAN.R-project.org/package=tibble

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

Urbanek, S. (2013). *Png: Read and write png images.* Retrieved from https://CRAN.R-project.org/package=png

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*(1), 1–29. Retrieved from http://www.jstatsoft.org/v40/i01/

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis.* Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string operations.* Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H. (2021a). *Forcats: Tools for working with categorical variables (factors).* Retrieved from https://CRAN.R-project.org/package=forcats

Wickham, H. (2021b). *Tidyr: Tidy messy data.* Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of data manipulation.* Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Hester, J. (2020). *Readr: Read rectangular text data.* Retrieved from https://CRAN.R-project.org/package=readr

Winston Chang. (2014). *Extrafont: Tools for using fonts.* Retrieved from https://CRAN.R-project.org/package=extrafont