BEHAVIOR RESEARCH METHODS

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

Journal:	*Behavior Research Methods*
Manuscript ID	BR-Org-22-042.R1
Manuscript Type:	Original Manuscript
Date Submitted by the Author:	n/a

Complete List of Authors:	Zettersten, Martin; Princeton University, Psychology Yurovsky, Daniel; Carnegie Mellon University, Department of Psychology Xu, Tian; Indiana University, Department of Psychological and Brain Sciences Uner, Sarp; Vanderbilt University, Data Science Institute Tsui, Angeline; Stanford University, Department of Psychology Schneider, Rose; University of California San Diego, Department of Psychology Saleh, Annissa; The University of Texas at Austin, Department of Psychology Meylan, Stephan; Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences; Duke University, Department of Psychology and Neuroscience Marchman, Virginia; Stanford University, Psychology Mankewitz, Jessica; Stanford University, Department of Psychology MacDonald, Kyle; McD Tech Labs Long, Bria; Stanford University, Department of Psychology Lewis, Molly; Carnegie Mellon University, Department of Psychology Kachergis, George; Stanford University, Department of Psychology Handa, Kunal; Brown University deMayo, Benjamin; Princeton University, Department of Psychology Carstensen, Alexandra; University of California San Diego, Department of Psychology Braginsky, Mika; MIT, Department of Brain and Cognitive Sciences Boyce, Veronica; Stanford University, Department of Psychology Bhatt, Naiti; New York University, Department of Psychology Bergey, Claire; The University of Chicago, Department of Psychology Frank, Michael; Stanford University, Psychology

PRINCETON UNIVERSITY

April 29, 2022

Dear Professor Kuperman,

We are writing to resubmit the attached manuscript, 'Peekbank: An open, large-scale repository for developmental eye-tracking data of children's word recognition' (MS ID BR-Org-22-042) for consideration for publication in *Behavior Research Methods*. We have revised the manuscript in response to the helpful suggestions by you and the reviewers, as detailed in our revision letter. Our revisions mainly focused on (1) elaborating on future directions for the current project and how researchers can contribute to expanding the database, (2) discussing in more depth issues that arise when aggregating data stemming from heterogeneous designs, and (3) clarifying several aspects of the database structure.

As with our initial submission, this paper has not been published or accepted for publication, and is not under consideration at another journal. All authors have approved the manuscript and agree with its submission to *BRM*. Thank you for considering this manuscript for publication, we appreciate your time and effort.

Sincerely,

Martin Zettersten (corresponding author), on behalf of the Peekbank team

Email: martincz@princeton.edu, Phone: 210-557-9525, Address: Department of Psychology, Princeton University, Princeton, NJ 08544

Editor

> *Our referees have now considered your paper (see their comments at the bottom of this letter). I thank them for providing insightful and constructive comments. In addition, I have read the ms carefully myself. The reviews are in general favorable and suggest that, subject to appropriate revisions, your paper could be suitable for publication. While this is already in the focus of the paper, I would recommend investing even more in clarifying how future studies can contribute to the resource (see Reviewer 1).*

We thank you and reviewers for your positive and helpful feedback, and have in particular made several revisions to help clarify how researchers can further contribute to Peekbank (see pp. 17; 32-33).

Reviewer 1

> *This manuscript describes a new resource for scholars working to understand word learning. Peekbank, a data repository and associated tools, enables aggregating data across studies that used a looking-while-listening experimental paradigm.*
>
> *This manuscript is clearly written, well organized, and straightforward. Thank you. Specific, illustrative code examples (e.g., the Swingley & Aslin, 2002, vignette, pp. 25-26) will be very helpful to users of this new resource.*
> *Throughout the manuscript, the authors articulate a number of analyses that this new resource will make possible, all of which are important and will advance our understanding of how methodological choices matter for our discoveries when using the looking-while-listening paradigm (e.g., the role of various distractor item schemes, various participant/trial inclusion schemes, interpolation schemes, analysis windows, and more).*

We thank the reviewer for their kind comments about the original submission and their constructive suggestions for improvement.

> *Fulfilling the promise of this new resource requires grappling with two critical issues. The authors have considerable expertise on these issues, and more attention to these in the manuscript would be valuable – it will likely help a broad community of researchers think clearly and practically as they move forward with their work.*
>
> *1. The authors note some limitations about the datasets currently in Peekbank (p. 30). The promise of this tool depends on amassing a much larger and more diverse database. Could the authors please articulate ways to help researchers contribute to this database?*

For example, they could recommend best practices for researchers to implement in consent processes/forms during new looking-while-listening studies in order to enable data-sharing (even provide some example consent form language in supplementary materials and/or OSF site). To the extent that there are known barriers to contributing existing and/or future datasets (e.g., original consent forms didn't allow sharing, various legal frameworks across the world's laboratories, etc.), could the authors comment on needed updates and/or advocacy that would help us collectively make the most of these repositories? I think fellow scholars would benefit considerably from a brief discussion of these issues in this manuscript.

Thank you for raising these important points – Peekbank's utility definitely increases as more data is added. We have revised our subsection on Processing, Validation, and Ingestion to include additional details about the process of data import. We have also added text to the General Discussion highlighting some prospects and potential issues in expanding the database (the number and diversity of datasets, homogeneity of methodologies, and limitations in the breadth of the participant population).

2. The promise of new discoveries about individual word knowledge (as in the apple, book, dog, frog vignette shared in this manuscript) is indeed exciting and a needed (and hard to come by) next step in our understanding of word learning. There are some deep issues about the extent to which aggregation licenses such insights, however. In particular, what is at stake when plucking data (e.g., a trial testing one child's knowledge of the word "apple") from the context in which it arose (e.g., the trial just before that child heard/saw apple, how fatigued the child was at that point in their experiment, etc.) and then aggregating that behavior with other artificially isolated datapoints? The authors nod to this issue (p. 28) and allude to model-based analytic methods (p. 30) that begin to address it. Though a manuscript in BRM is not best suited to a full exploration of these issues, some articulation of essential things for researchers to consider when using this new resource in service of theory development would dramatically improve its uptake and use. Could the authors please articulate some dimensions of data aggregation practices that are especially relevant for theory building? The item analyses could ground such discussion, as could other theoretical questions. These authors' considerable expertise could provide much needed pointers to the field about how to maximize the probability of rigorous and responsible future science when using this new resource.

Thanks so much for this comment. We did not want to expand the manuscript too much, but at the same time the reviewer has hit on precisely the issue that we are grappling with in our ongoing research using the database, namely how to model the correlated variation across trials. We have added some additional discussion of these issues to both the subsection on Item

Analyses (pp. 30-31) and the General Discussion (pp. 31-32) acknowledging the broader sources of variation mentioned by the reviewer.

> *Does the Peekbank common data format easily handle studies with multiple phases, such as a learning phase in which participants are exposed to novel word-object mappings, and a separate test phase? Doing so would considerably expand this repository's utility because it could be used to aggregate studies that tested hypotheses about word learning mechanisms (e.g., by manipulating properties of the learning phase stimuli) in addition to documenting infants' knowledge of familiar words.*

Thanks for this excellent question. In designing Peekbank we have regularly encountered cases where a more flexible database schema would create opportunities for new analyses of a subset of studies, but would require a substantial overhead for the full set of studies. Archiving training data is a good example of this sort of problem. We have typically focused our attention mainly on data from the test phases of looking-while-listening studies as the most critical to studying questions of broad interest about word recognition. The need to focus our attention on a particular data format follows in part from the fact that the import process is already the biggest bottleneck in our system and so adding complexity to the database schema has cascading effects on our ability to include new data.

That said, we already archive several datasets in which children are trained on novel words (and word-object mappings). The problem is that training phases are quite heterogeneous across these experiments. For example, in Yurovsky et al. (2017), the training phase was a movie with dynamic AOIs. In other studies outside of what we have in Peekbank, exposure is done via overhearing (e.g., Messenger, Yuan, & Fisher, 2015), cross-situational training (Smith & Yu, 2008), or sequences of unambiguous exposure trials (Bion, Borovsky, & Fernald, 2013). Given the complexity and heterogeneity in the literature – and the practical fact that often training data are not shared in the same formats – our current compromise is to archive information only from looking-while-listening trials that conform to the general two-alternative structure. This decision dramatically simplifies the import process.

It is still possible to use Peekbank to study word learning. In our current datasets we have used the 'novelty' column associated with each stimulus to encode its general novelty, and the 'condition' column associated with each trial type to encode whether the child learned that novel target and/or distractor in training. Thus, to analyze relations between exposure conditions and learning outcomes, a user would simply need to code this information from the relevant papers and merge it with outcome information from Peekbank.

In sum, while there are trial-level learning analyses that our current schema does not support (for reasons of practicality, primarily), Peekbank still affords substantial opportunities for analyses of novel word learning.

> *The authors report that data are "balanced in terms of gender (47%)" (p. 17). This way of reporting assumes a gender binary. My understanding of recommended practice is to refer to infants' assigned sex and fully report the data (e.g., x% female, x% male, x% unreported, etc., assigned sexes).*

Thanks for this suggestion - we have updated our reporting in the main manuscript.

> *Thank you for your work on this exciting new resource.*

We appreciate the supportive and constructive feedback. Your suggestions have helped us significantly improve the manuscript.

Reviewer 2

> *Thanks for the opportunity to read this interesting report on the PeekBank database, presenting a large aggregation of eye-tracked studies of infant word recognition, and a pipeline for augmenting that database. This is a great contribution that I suspect will be highly prized by the language acquisition and psycholinguistics community.*

Thanks for the supportive comments!

> *The key idea here is a very good one - that we can gain exploratory power and precision by aggregating prior studies of infant word recognition, many of which rely on ~the same paradigm - two pictures on a screen, and an instruction to look at one of the pictures. The authors motivate this idea in terms of other recent attempts to aggregate (developmental) data, such as WordBank, BIDS, and the ChildProject, but of course there's a longer history of this type of work. CHILDES, for example, was motivated in just the same way, defining a clear data standard that diverse researchers could use for building and comparing transcripts. That effort has obviously had a huge impact on the field over decades, and highlighting it may better to the naive reader why this type of exercise is important. Likewise, in cognitive psychology, there's a large history of mega-studies that again are quite comparable to this. Perhaps the best example of this is the English Lexicon Project, which has had the same aim of aggregating huge amounts of data to provide better resolution on questions of interest. So, as a minor change, it may be worth better highlighting the links between this work and related projects in developmental and cognitive psychology.*

We thank the reviewer for this great suggestion. We have edited the introduction to highlight the long history of this type of work in the field and its link to our work (p. 5).

> *My read of the implemented data structure is that it is sensibly minimalist. In terms of the time series, data is recoded to approximately match the sampling rate of hand-coded video-camera-based eye tracking (rather than the higher resolution of a remote eye tracker). Hand-coded data is perhaps not the dominant form any more, although that may have changed with Zoom testing during COVID. Nonetheless, it has been commonplace for a couple of decades, so there is a deep well of this data, and it fits with the relatively low frequency of large saccades. So this decision makes good sense to me.*

We thank the reviewer for their supportive feedback and for considering the tradeoffs in our data structure decisions carefully. We hope to add a large amount of hand-coded data from past work to the database over the next several years, which in part motivated our decision to make the database as compatible with common frame rates for hand-coded data as possible.

> *Downloading and looking at the tables, the data structure is also quite clever in terms of dealing with children's ages, but I don't see it mentioned in the paper. Specifically, the administrations table includes fields for lab_age and lab_age_units, allowing the database to record how the original lab recorded age (e.g., children were 2 years old), and then also translate that into a common format (the age field seems to be uniformly in months). It'd be worth mentioning that in the paper. Also, the transformation between age and lab_age ought to be justified. Looking at the data, the authors have done "number of years * 12", so that children aged 2 years are recorded as being 24 months. But of course the average 2-year-old is 30 months. It might be worth reconsidering this, as it could lead to some artefacts where users treat data from kids as old as 35-months as being equivalent to kids aged 24 months.*

We thank the reviewer for their close attention to the setup of the schema and calculation of age. To address the first point, we have added a sentence to the schema section (p. 14) explaining our encoding of lab age, lab age units, and (standardized format) age.

In our current database, we do not have any datasets that originally expressed age in whole numbers of years–any datasets with a lab age unit of years have decimal points giving month- or day-level precision, e.g., 2.16 to express 2 years and 2 months old. Multiplying by 12 is correct when we have this level of granularity in the raw data. However, this comment has made us aware of the issue with converting age in this way when the raw data is not at least precise to the level of months—in particular, that it could lead to overestimating young children's abilities by coding their age as (up to nearly a year) younger than they actually are. We have updated the guidance in our import templates and codebook to note that lab ages recorded in units of whole years should be converted using the reviewer's suggested conversion, years * 12 + 6, to mitigate this issue.

> *Finally, looking at the Discussion, I think that the limitations of the database to date are well described. But one further point is that - as far as I can tell - the data isn't only homogeneous in terms of participant backgrounds and languages, but also in terms of lab groups where the data was collected from. Nothing wrong with this as a starting point, but it would be great to expand the database to include data from research groups with distinct theoretical and empirical perspectives.*

This is an important point, and one that we now also highlight in our discussion of the limitations of the current database (p. 32).

> *p.s. One further quick Issue - The administrations table that I downloaded had dataset names that did not match those in Table 1 of the paper.*

We thank the reviewer for their close attention to the details of our dataset tables and the table in our paper. We have updated Table 1 of the paper on p. 18 to reflect the matching dataset names.

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

Martin Zettersten[1], Daniel Yurovsky[2], Tian Linger Xu[3], Sarp Uner[4], Angeline Sin Mei Tsui[5], Rose M. Schneider[6], Annissa N. Saleh[7], Stephan Meylan[8,9], Virginia Marchman[5], Jessica Mankewitz[5], Kyle MacDonald[10], Bria Long[5], Molly Lewis[2], George Kachergis[5], Kunal Handa[11], Benjamin deMayo[1], Alexandra Carstensen[6], Mika Braginsky[9], Veronica Boyce[5], Naiti S. Bhatt[12], Claire Augusta Bergey[13], & Michael C. Frank[5]

[1] Department of Psychology, Princeton University

[2] Department of Psychology, Carnegie Mellon University

[3] Department of Psychological and Brain Sciences, Indiana University

[4] Data Science Institute, Vanderbilt University

[5] Department of Psychology, Stanford University

[6] Department of Psychology, University of California, San Diego

[7] Department of Psychology, The University of Texas at Austin

[8] Department of Psychology and Neuroscience, Duke University

[9] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

[10] Core Technology, McD Tech Labs

[11] Brown University

[12] Department of Psychology, New York University

[13] Department of Psychology, University of Chicago

Author Note

Acknowledgements. We would like to thank the labs and researchers that have made their data publicly available in the database. For further information about contributions, see https://langcog.github.io/peekbank-website/docs/contributors/. This work was supported in part by grants from the National Institutes of Health awarded to VM (R01 HD092343, 2R01 HD069150).

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

CRediT author statement. Outside of the position of the first and the last author, authorship position was determined by sorting authors' last names in reverse alphabetical order. An overview of authorship contributions following the CRediT taxonomy can be viewed here: https://docs.google.com/spreadsheets/d/e/2PACX-1vRD-LJD_dTAQaAynyBlwXvGpfAVzP-3Pi6JTDoG15m3PYZe0c44Y12U2a_hwdmhIstpjyigG2o3na4y/pubhtml.

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

Abstract

The ability to rapidly recognize words and link them to referents 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. We first present how we created the database, its features and structure, and associated tools for processing and accessing infant eye-tracking datasets. Using these tools, we then work through two illustrative examples to show how researchers can use Peekbank to interrogate theoretical and methodological questions about children's developing word recognition ability.

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

Word count: 7369

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

Across their first years of life, children learn words at an accelerating pace (Frank, Braginsky, Yurovsky, & Marchman, 2021). While many children will only produce their first word at around one year of age, most children show signs of understanding many common nouns (e.g., *mommy*) and phrases (e.g., *Let's go bye-bye!*) much earlier in development (Bergelson & Swingley, 2012, 2013; Tincoff & Jusczyk, 1999). Although early word understanding is a critical element of first language learning, the processes involved are less directly apparent in children's behaviors and are less accessible to observation than developments in speech production (Fernald, Zangl, Portillo, & Marchman, 2008; Hirsh-Pasek, Cauley, Golinkoff, & Gordon, 1987). To understand a spoken word, children must process the incoming auditory signal and link that signal to relevant meanings – a process often referred to as word recognition. One of the primary means of measuring word recognition in young infants is using eye-tracking techniques that gauge where children look in response to linguistic stimuli (Fernald, Zangl, Portillo, & Marchman, 2008). The logic of these methods is that if, upon hearing a word, a child preferentially looks at a target stimulus rather than a distractor, the child is able to recognize the word and activate its meaning during real-time language processing. Measuring early word recognition offers insight into children's early word representations: children's speed of response (i.e., moving their eyes; turning their heads) to the unfolding speech signal can reveal children's level of comprehension (Bergelson, 2020; Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998). 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 handful of items. The limitations of single datasets makes it difficult to understand developmental changes in children's word knowledge at a broad scale.

One way to overcome this challenge is to compile existing datasets into a large-scale database in order to expand the scope of research questions that can be asked about the development of word recognition abilities. This strategy capitalizes on the fact that the looking-while-listening paradigm is widely used, and vast amounts of data have been collected across labs on infants' word recognition over the past 35 years (Golinkoff, Ma, Song, & Hirsh-Pasek, 2013). Such datasets have largely remained isolated from one another, but once combined, they have the potential to offer general insights into lexical development. There has been a long history of efforts to aggregate data in a unified format in developmental and cognitive psychology, generating projects that have often had a tremendous impact on the field. Prominent examples in language research include the English Lexicon Project, which provides an open repository of psycholinguistic data for over 80,000 English words and non-words in order to support large-scale investigations of lexical processing (Balota et al., 2007); the Child Language Data Exchange System (CHILDES), which has played an instrumental role in the study of early language environments by systematizing and aggregating data from naturalistic child-caregiver language interactions (MacWhinney, 2000); and WordBank, which aggregated data from the MacArthur-Bates Communicative Development Inventory, a parent-report measure of child vocabulary, to deliver new insights into cross-linguistic patterns and variability in vocabulary development (Frank, Braginsky, Yurovsky, & Marchman, 2017, 2021). In this paper, we introduce *Peekbank*, an open database of infant and toddler eye-tracking data aimed at facilitating the study of developmental changes in children's word recognition.

Measuring Word Recognition: The Looking-While-Listening Paradigm

Word recognition is traditionally studied in the looking-while-listening paradigm (Fernald, Zangl, Portillo, & Marchman, 2008; alternatively referred to as the intermodal preferential looking procedure, Hirsh-Pasek, Cauley, Golinkoff, & Gordon, 1987). In these 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 bird – the distractor image). Infants' word recognition is evaluated by how quickly and accurately they fixate on the target image after hearing its label. Past research has used this basic method to study a wide range of questions in language development. For example, the looking-while-listening paradigm has been used to investigate 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, 2002).

While this research has been fruitful in advancing understanding of early word knowledge, fundamental questions remain. One central question is how to accurately capture developmental change in the speed and accuracy of word recognition. There is ample evidence demonstrating that infants become faster and more accurate in word recognition over the first few years of life (e.g., Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998). However, precisely measuring developmental increases in the speed and accuracy of word recognition remains challenging due to the difficulty of distinguishing developmental changes in word recognition skill from changes in knowledge of specific words. This problem is particularly thorny in studies with young children, 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). More broadly, key differences in the design choices (e.g., how distractor items are selected) and analytic decisions (e.g., how the analysis window is defined) between studies can obscure developmental change if not appropriately taken into account.

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

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

Replicability and Reproducibility

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

In addition to threats to replicability, the field of infant development also faces concerns about analytic reproducibility – the ability for researchers to arrive at the same analytic conclusion reported in the original research article, given the same dataset. A recent estimate based on studies published in a prominent cognitive science journal suggests that analyses can remain difficult to reproduce, even when data are made available to other research teams (Hardwicke et al., 2018). Aggregating data in centralized databases can aid in improving reproducibility in several ways. First, building a large-scale database requires defining a standardized data specification. Recent examples include the `brain imaging data structure` (BIDS), an effort to specify a unified data format for neuroimaging experiments (Gorgolewski et al., 2016), and the data formats associated with `ChildProject`, for managing long-form at-home language recordings (Gautheron, Rochat, & Cristia, 2021). Defining a data standard – in this case, for infant eye-tracking experiments – supports reproducibility by guaranteeing that critical information will be available in openly shared data and by making it easier for different research teams to understand the data structure. Second, open databases make it easy for researchers to generate open and reproducible analytic pipelines, both for individual studies and for analyses aggregating across datasets. Creating open analytic pipelines across many datasets also serves a pedagogical purpose, providing teaching examples illustrating how to implement analytic techniques used in influential studies and how to conduct reproducible analyses with infant eye-tracking data.

Peekbank: An open database of developmental eye-tracking studies

What all of these open challenges share is that they are difficult to address at the scale of a single research lab or in a single study. To address this challenge, we developed *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 heterogeneous data sources, and (c) provides an interface for accessing and analyzing the database. In the current paper, we introduce the key components of the project and give an overview of the existing database. We then provide two worked examples of how researchers can use Peekbank. In the first, we examine a classic result in the word recognition literature, and in the second we aggregate data across studies to investigate developmental trends in the recognition of individual words.

Design and Technical Approach

Database Framework

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

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

Database Schema

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

Metadata. Metadata can be separated into four parts: (1) participant-level information (e.g., demographics), (2) experiment-level information (e.g., the type of eye

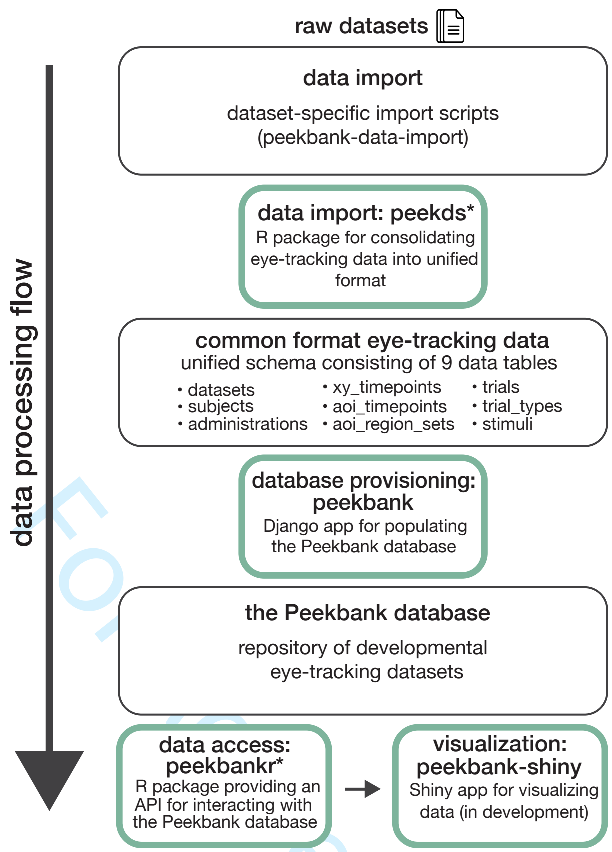

Figure 1. Overview of the Peekbank data ecosystem. Peekbank tools are highlighted in green. * indicates R packages introduced in this work.

tracker used to collect the data), (3) session information (e.g. a participant's age for a specific experimental session), and (4) trial information (e.g., which images or videos were presented onscreen, and paired with which audio).

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`

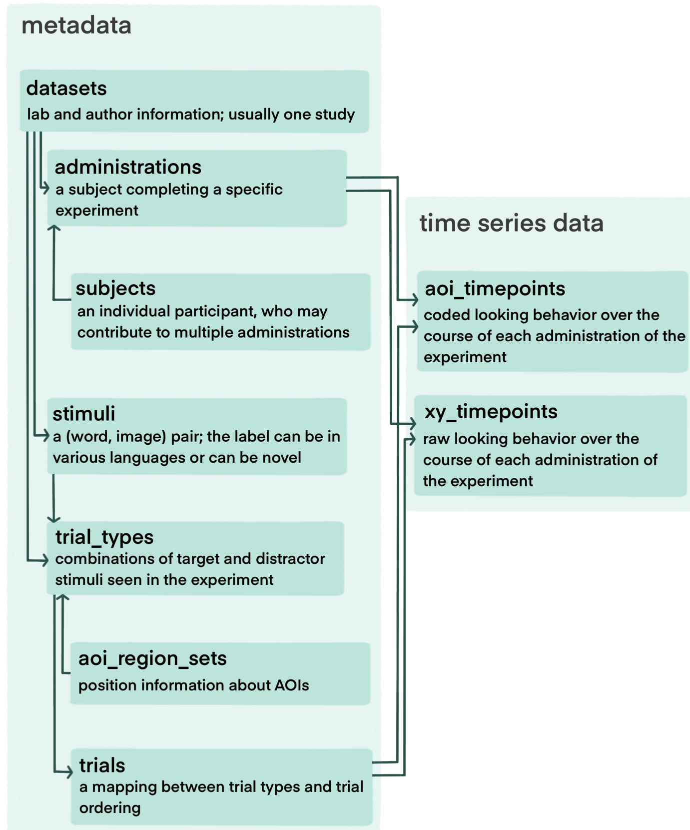

Figure 2. The Peekbank schema. Each darker rectangle represents a table in the relational database. AOIs are areas of interest in an eye-tracking experiment, in this case information about the position of target and distractor stimuli on the screen.

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 administrations.

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

Experiment Information.

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

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

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

[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.

includes the age of the participant, the coding method (eye-tracking vs. hand-coding), and the properties of the monitor that was used. For participant age, we include the fields `lab_age` and `lab_age_units` to record how the original lab encoded age, as well as an additional field, `age`, to encode age in a standardized format across datasets, using standardized months as the common unit of measurement (see the Peekbank codebook for details on how ages are converted into months).

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 corresponding data in the `xy_timepoints` table.

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

To facilitate time course analysis and visualization across datasets, time course data must be resampled to a uniform sampling rate (i.e., such that every trial in every dataset has observations at the same time points). All data in the database is resampled to 40 Hz (observations every 25 ms), which represents a compromise between retaining fine-grained timing information from datasets with dense sampling rates (maximum sampling rate among current datasets: 500 Hz) while minimizing the possibility of introducing artifacts via resampling for datasets with lower sampling rates (minimum sampling rate for current datasets: 30 Hz). Further, 25 ms is a mathematically convenient interval for ensuring consistent resampling; we found that using 33.333 ms (30 Hz) as our interval simply introduced a large number of technical complexities. The resampling operation is accomplished using the `resample_times()` function. During the resampling process, we interpolate using constant interpolation, selecting for each interpolated timepoint the looking location for the earlier-observed time point in the original data for both `aoi_timepoints` and `xy_timepoints` data. Compared to linear interpolation (see e.g., Wass, Smith, & Johnson, 2013), which fills segments of missing or unobserved time points by interpolating between the observed locations of timepoints at the beginning and end of the interpolated segment, constant interpolation has the advantage that it is more conservative, in the sense that it does not introduce new look locations beyond those measured in the original data.

[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.

One possible application of our new dataset is investigating the consequences of other interpolation functions for data analysis.

Processing, Validation, and Ingestion

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

Many of the specific operations involved in importing a dataset can be abstracted across datasets. The `peekds` package offers a library of these functions. Once the data have been extracted in a tabular form, the package also offers a validation function that checks whether all tables have the required fields and data types expected by the database. In an effort to double check the data quality and to make sure that no errors are made in the importing script, we also typically perform a visual check of the import process, creating a time course plot to replicate the results in the paper that first presented each dataset. Once this plot has been created and checked for consistency and all tables pass our validation functions, the processed dataset is ready for reprocessing into the database using the `peekbank` library. This library applies additional data checks, and adds the data to the MySQL database using the Django web framework.

To date, the import process has been carried out by the Peekbank team using data offered by other research teams. Data contributors are also welcome to provide import scripts to facilitate contribution. However, creating these scripts requires familiarity with both R scripting and the specific Peekbank schema, and writing an import script can be somewhat time-consuming in practice. To support future data contributions, import script templates and examples are available for both hand-coded datasets and automatic eye-tracking datasets for research teams to adapt to their data. These import templates walk researchers through each step of data processing using example datasets from Peekbank and include explanations of key decision points, examples of how to use various helper functions available in `peekds`, and further details about the database schema.

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_salientme	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

The database currently includes 20 looking-while-listening datasets comprising N=1594 total participants (Table 1). The current data represents a convenience sample of datasets that were (a) datasets collected by or available to Peekbank team members, (b) made available to Peekbank after informal inquiry or (c) datasets that were openly available. Most datasets (14 out of 20 total) consist of data from monolingual native English speakers. They span a wide age spectrum with participants ranging from 9 to 70 months of age, and are balanced in terms of children's assigned sex (47.30% female; 50.40% male; 2.30%

unreported). The datasets vary across a number of design-related dimensions, and include studies using manually coded video recordings and automated eye-tracking methods (e.g., Tobii, EyeLink) to measure gaze behavior. All studies tested familiar items, but the database also includes 5 datasets that tested novel pseudo-words in addition to familiar words. Users interested in a subset of the data (e.g., only trials testing familiar words) can filter out unwanted trials using columns available in the schema (e.g., using the column `stimulus_novelty` in the `stimuli` table).

Versioning and Reproducibility

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

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

Interfacing with Peekbank

Peekbankr

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

- `connect_to_peekbank()` opens a connection with the Peekbank database to allow tables to be downloaded with the following functions
- `get_datasets()` gives each dataset name and its citation information
- `get_subjects()` gives information about persistent participant identifiers (e.g., native languages, sex)
- `get_administrations()` gives information about specific experimental administrations (e.g., participant age, monitor size, gaze coding method)
- `get_stimuli()` gives information about word–image pairings that appeared in experiments
- `get_trial_types()` gives information about pairings of stimuli that appeared in the experiment (e.g., point of disambiguation, target and distractor stimuli, condition, language)
- `get_trials()` gives the trial orderings for each administration, linking trial types to the trial IDs used in time course data
- `get_aoi_region_sets()` gives coordinate regions for each area of interest (AOI) linked to trial type IDs
- `get_xy_timepoints()` gives time course data for each participant's looking behavior in each trial, as (x, y) coordinates on the experiment monitor
- `get_aoi_timepoints()` gives time course data for each participant's looking behavior in each trial, coded into areas of interest

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

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

Shiny App

One goal of the Peekbank project is to allow a wide range of users to easily explore and learn from the database. We therefore have created an interactive web application – `peekbank-shiny` – that allows users to quickly and easily create informative visualizations of individual datasets and aggregated data (https://peekbank-shiny.com/). `peekbank-shiny` is built using Shiny, a software package for creating web apps for data exploration with R, as well as the `peekbankr` package. All code for the Shiny app is publicly available (https://github.com/langcog/peekbank-shiny). The Shiny app allows users to create commonly used visualizations of looking-while-listening data, based on data from the Peekbank database. Specifically, users can visualize:

1. the *time course of looking data* in a profile plot depicting infant target looking across trial time
2. *overall accuracy*, defined as the proportion target looking within a specified analysis window
3. *reaction times* in response to a target label, defined as how quickly participants shift

fixation to the target image on trials in which they were fixating on the distractor image at onset of the target label

4. an *onset-contingent plot*, which shows the time course of participant looking as a function of their look location at the onset of the target label

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

OSF site

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

Peekbank in Action

In the following section, we provide examples of how users can access and analyze the data in Peekbank. First, we provide an overview of some general properties of the datasets in the database. We then demonstrate two potential use-cases for Peekbank data. In each case, we provide sample code to demonstrate the ease of doing simple analyses using the database. Our first example shows how we can investigate the findings of a classic study.

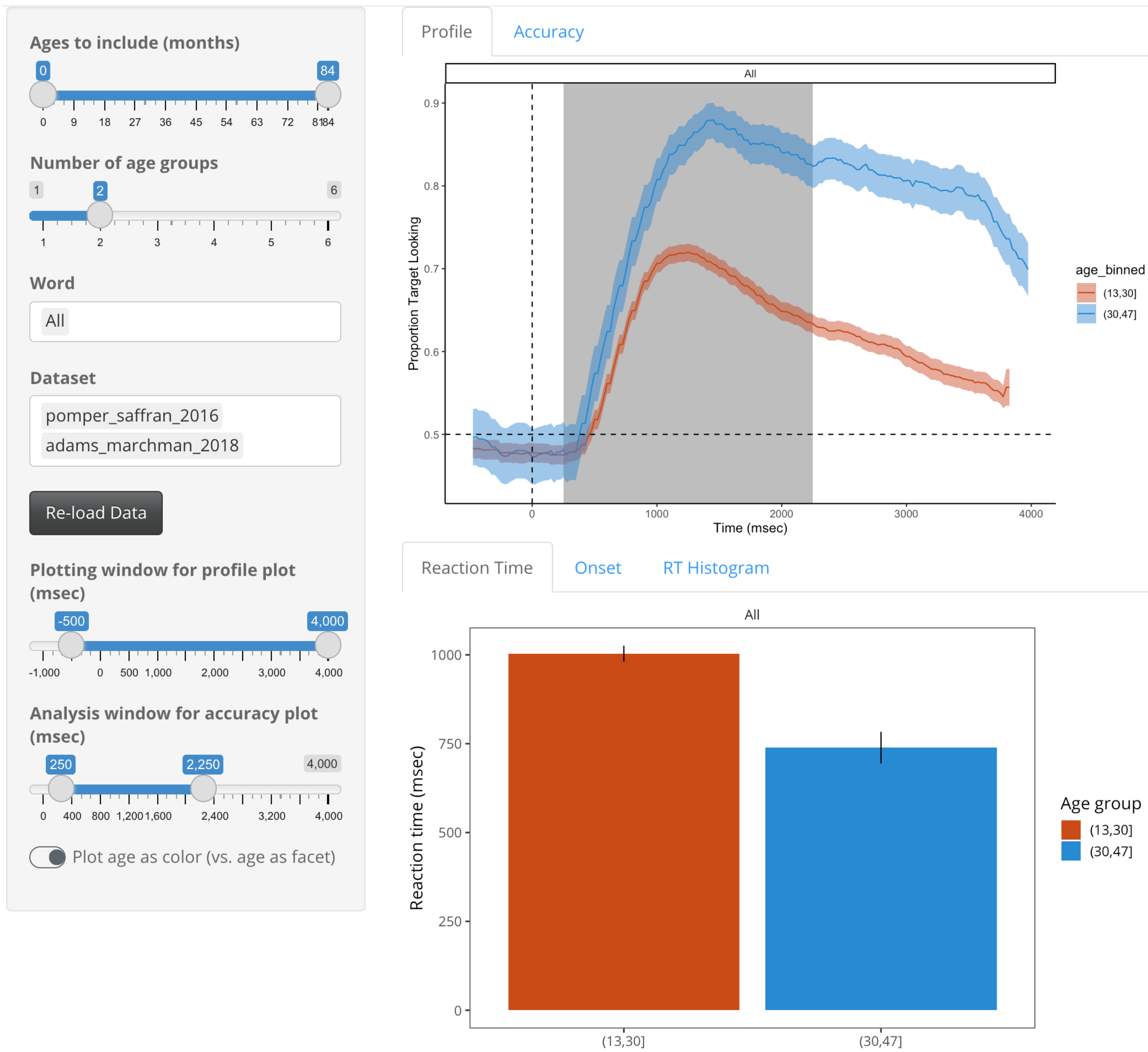

Figure 3. Screenshot of the Peekbank Shiny app, which shows a variety of standard analysis plots as a function of user-selected datasets, words, age ranges, and analysis windows. Shown here are mean reaction time and proportion target looking over time by age group for two selected datasets.

This type of investigation can be a very useful exercise for teaching students about best practices for data analysis (e.g., Hardwicke et al., 2018) and also provides an easy way to explore looking-while-listening time course data in a standardized format. Our second example shows an 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 could in principle be used for optimizing the stimuli for new experiments, especially as the Peekbank dataset grows and gains coverage over a greater number of items. All analyses are conducted using R [Version 4.1.1; R Core Team (2021)][3]

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.

[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)], *tidyr* [Version 1.1.3; Wickham (2021b)], *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)].

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

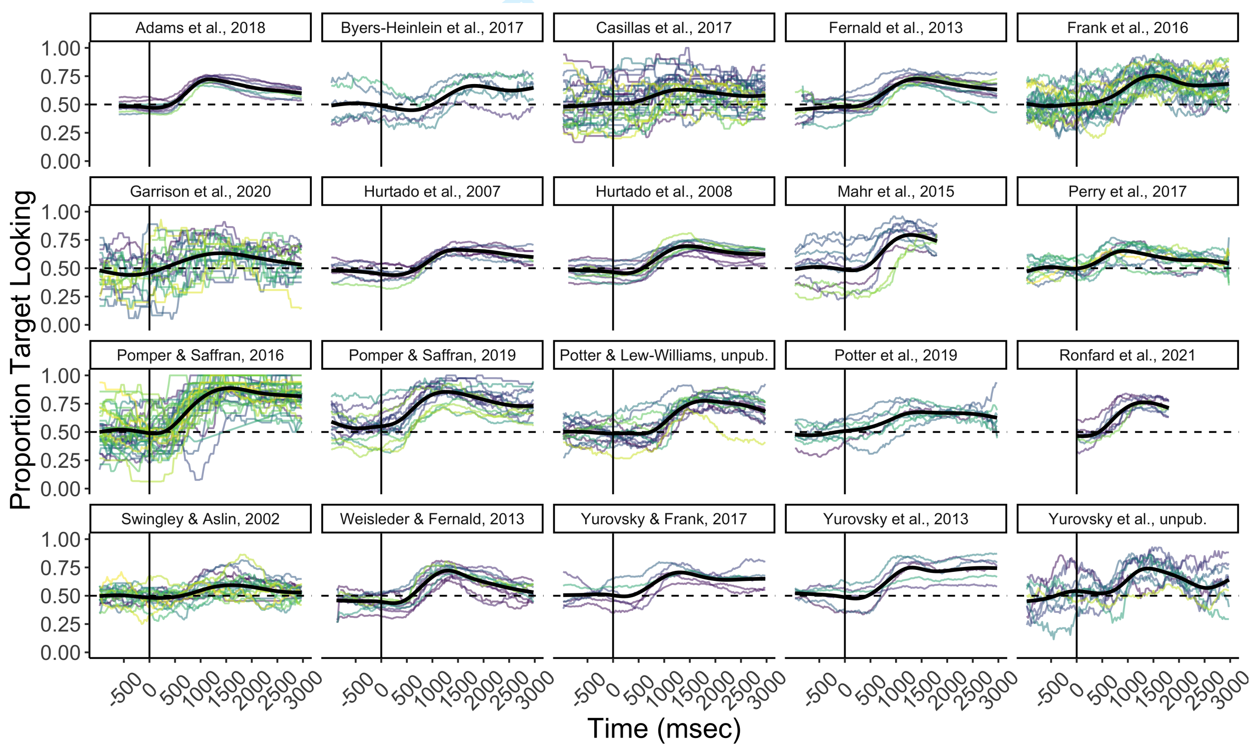

Figure 4. Item-level variability in proportion target looking within each dataset (chance=0.5). Time is centered on the onset of the target label (vertical line). Colored lines represent specific target labels. Black lines represent smoothed average fits based on a general additive model using cubic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

A second question of interest is about the variability across items (i.e., target labels) within specific studies. Some studies use a smaller set of items (e.g., 8 nouns, Adams et al.,

2018) while others use dozens of different items (e.g., Garrison, Baudet, Breitfeld, Aberman, & Bergelson, 2020). Figure 4 gives an overview of the variability in proportion looking to the target item for individual words in each dataset. Although all datasets show a gradual rise in average proportion target looking over chance performance, the number of unique target labels and their associated accuracy vary widely across datasets.

Investigating prior findings: Swingley and Aslin (2002)

Swingley and Aslin (2002) investigated the specificity of 14-16-month-olds' word representations using the looking-while-listening paradigm, asking whether recognition would be slower and less accurate for mispronunciations, e.g. *opal* (mispronunciation) instead of *apple* (correct pronunciation).[4] In this short vignette, we show how easily the data in Peekbank can be used to visualize this result. Our goal here is not to provide a precise analytical reproduction of the analyses reported in the original paper, but rather to demonstrate the use of the Peekbank framework to analyze datasets of this type. In particular, because Peekbank uses a uniform data import standard, it is likely that there will be minor numerical discrepancies between analyses on Peekbank data and analyses that use another processing pipeline.


```r
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")
```

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

[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.

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"))
```

As the code above shows, once the data are joined, condition information for each timepoint is present and so we can easily filter out filler trials and set up the conditions for 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()))
```

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

Figure 5 shows the average time course of looking for the two conditions, as produced by the code above. Looks after the correctly pronounced noun appeared both faster (deviating from chance earlier) and more accurate (showing a higher asymptote). Overall, this example demonstrates the ability to produce this visualization in just a few lines of code.

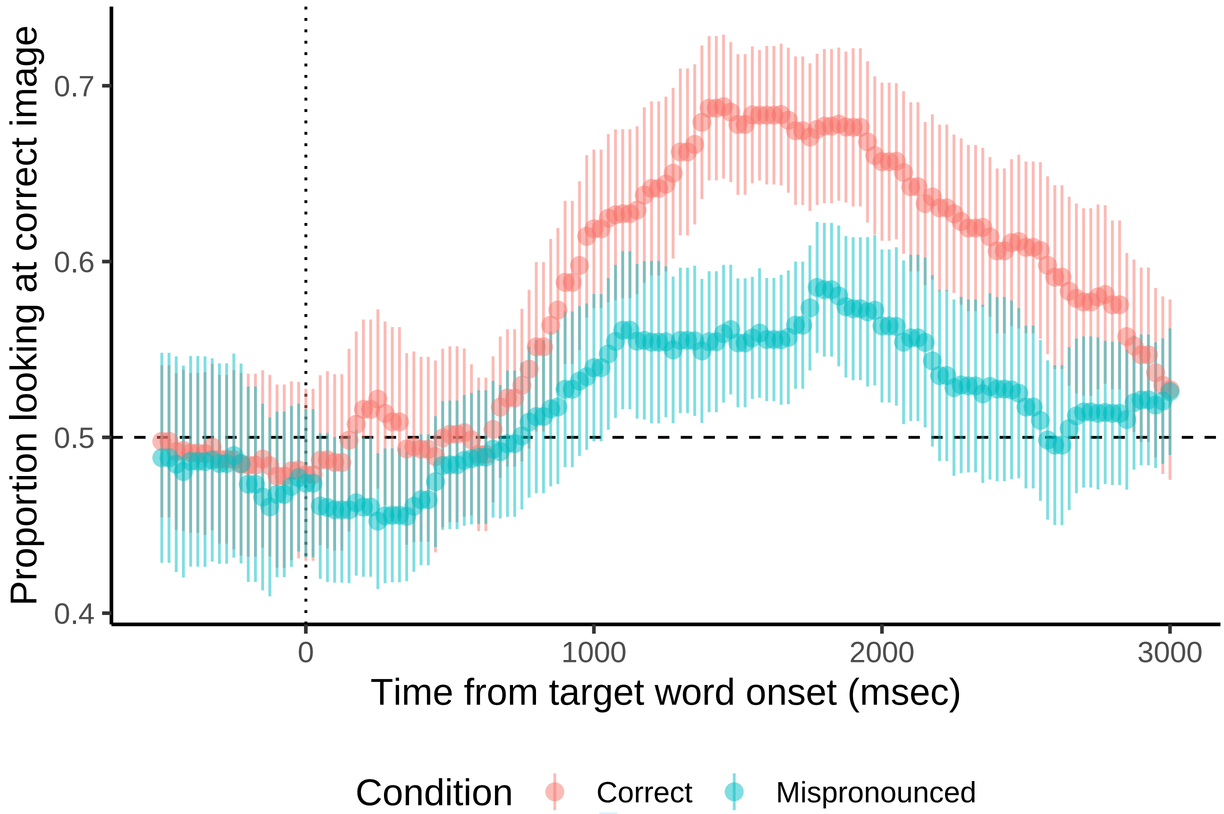

Figure 5. Proportion looking at the correct referent by time from the point of disambiguation (the onset of the target noun) in Swingley & Aslin (2002). Colors show the two pronunciation conditions; points give means and ranges show 95% confidence intervals. The dotted line shows the point of disambiguation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

Item analyses

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

Our first step is to collect and join the data from the relevant tables including timepoint data, trial and stimulus data, and administration data (for participant ages). We join these into a single dataframe for easy manipulation; this dataframe is a common starting point for analyses of item-level data.

```r
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)
```

Next we select a set of four target words (chosen based on having more than 100 children contributing data for each word across several one-year age groups). We create age groups, aggregate, and compute timepoint-by-timepoint confidence intervals using the z approximation.

```r
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())
```

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

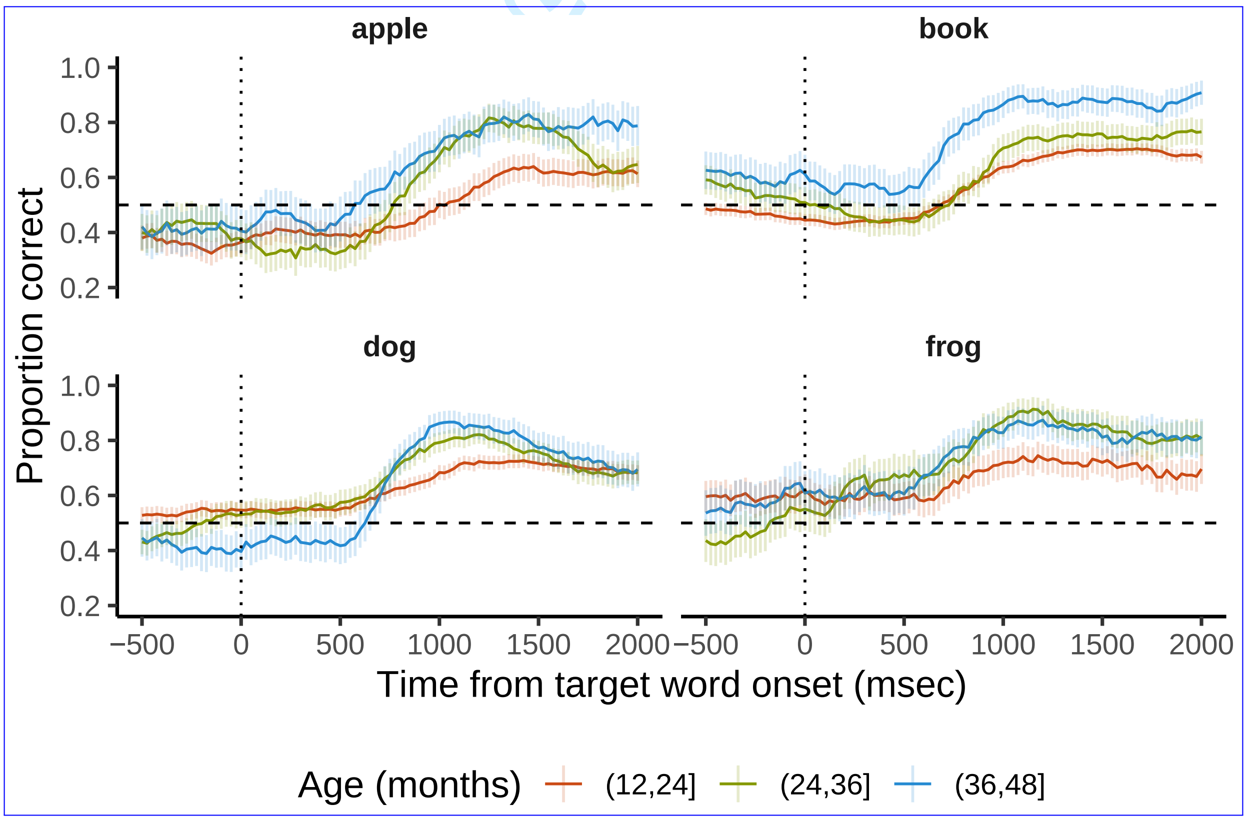

Figure 6. Time course plot for four well-represented target items in the Peekbank dataset, split by three age groups. Each line represents children's average looking to the target image after the onset of the target label (dashed vertical line). Error bars represent 95% CIs.

This simple averaging approach is a proof-of-concept to demonstrate some of the potential of the Peekbank dataset. An eye-movement trajectory on an individual trial reflects myriad factors, including the age and ability of the child, the target and distractor stimuli on that trial, the position of the trial within the experiment, and the general parameters of the experiment (for example, stimulus timing, eye-tracker type and calibration, etc.). Although we often neglect these statistically in the analysis of individual experiments – for example, averaging across items and trial orders – they may lead to imprecision when we average across multiple studies in Peekbank. For example, studies with older children may use more difficult items or faster trial timing, leading to the impression that children's abilities overall increase more slowly than they in fact do. Even in our example in Figure 6, we see hints of this confounding – for example, the low baseline looks to *apple* may be an artifact of an attractive distractor being paired with this item in one or two studies. In future work, we hope to introduce model-based analytic methods that use mixed effects regression to factor out study-level and individual-level variance in order to recover developmental effects more appropriately (see e.g., Zettersten et al., 2021 for a prototype of such an analysis).

Discussion

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; Sanchez et al., 2019; 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 overview of the structure of the database, shows how users can access the database, and demonstrates how it can be used both to investigate prior experiments and to synthesize data across studies.

The current database has a number of limitations, particularly in the number and diversity of datasets it contains. With 20 datasets currently available in the database, idiosyncrasies of particular designs and condition manipulations still have a substantial influence on the results of particular analyses, as discussed above in our item analysis example. Expanding the set of distinct datasets will allow us to increase the number of datasets that contain specific items, leading to more robust generalizations across the many sources of variation that are confounded within studies (e.g., item set, participant age range, and specific experimental parameters). A critical next step will be the development of analytic models that take this structure into account in making generalizations across datasets.

A second limitation stems from the fact that the database represents a convenience sample of data readily available to the Peekbank team, which leads the database to be relatively homogeneous in a number of key respects. First, the datasets primarily come from labs that share similar theoretical perspectives and implement the looking-while-listening method in similar ways. 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 (Henrich, Heine, & Norenzayan, 2010; Muthukrishna et al., 2020). Increasing the diversity of lab sources, participant backgrounds, and languages will expand the scope of the generalizations we can form about child word recognition, while also providing new opportunities for describing cross-lab, cross-cultural, and cross-linguistic variation.

Towards the goal of expanding our database, we invite researchers to contribute their data. On the Peekbank website, we provide technical documentation for potential contributors. Although we anticipate being involved in most new data imports, as discussed above, our import process is transparently documented and the repository contains examples for most commonly-used eye-trackers. Contributing data to an open repository also can raise questions about participant privacy. Potential contributors should consult with their local institutional review boards for guidance on any challenges, but we do not foresee obstacles because of the de-identified nature of the data. Under United States regulation, all data contributed to Peekbank are considered de-identified and hence not considered "human subjects data"; hence, institutional review boards should not regulate this contribution process. Under the European Union's Generalized Data Protection Regulation (GDPR), labs may need to take special care to provide a separate set of participant identifiers that can never be re-linked to their own internal records.

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 key advances in our understanding of infant cognition (Aslin, 2007; Baillargeon, Spelke, & Wasserman, 1985; Bergelson & Swingley, 2012; Fantz, 1963; Liu, Ullman, Tenenbaum, & Spelke, 2017; Quinn, Eimas, & Rosenkrantz, 1993). 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.

References

Adams, K. A., Marchman, V. A., Loi, E. C., Ashland, M. D., Fernald, A., & Feldman, H. M. (2018). Caregiver talk and medical risk as predictors of language outcomes in full term and preterm toddlers. *Child Development*, *89*(5), 1674–1690.

Arnold, J. B. (2021). *Ggthemes: Extra themes, scales and geoms for 'ggplot2'.* Retrieved from https://CRAN.R-project.org/package=ggthemes

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48–53.

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

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, *20*(3), 191–208. https://doi.org/10.1016/0010-0277(85)90008-3

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Barth, M. (2021). *tinylabels: Lightweight variable labels.* Retrieved from https://github.com/mariusbarth/tinylabels

Bergelson, E. (2020). The comprehension boost in early word learning: Older infants are better learners. *Child Development Perspectives*, *14*(3), 142–149.

Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258.

Bergelson, E., & Swingley, D. (2013). The acquisition of abstract words by young infants. *Cognition*, *127*(3), 391–397.

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development*, *89*(6), 1996–2009.

Bleses, D., Makransky, G., Dale, P. S., Højen, A., & Ari, B. A. (2016). Early productive vocabulary predicts academic achievement 10 years later. *Applied Psycholinguistics*, *37*(6), 1461–1476.

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

Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2021). Six solutions for more reliable infant research. *Infant and Child Development*. https://doi.org/https://doi.org/10.1002/icd.2296

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

DeBolt, M. C., Rhemtulla, M., & Oakes, L. M. (2020). Robust data and power in infant research: A case study of the effect of number of infants and number of trials in visual preference procedures. *Infancy*, *25*(4), 393–419. https://doi.org/10.1111/infa.12337

Fantz, R. L. (1963). Pattern vision in newborn infants. *Science*, *140*(3564), 296–297.

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, *16*(2), 234–248. https://doi.org/10.1111/desc.12019

Fernald, A., Pinto, J. P., Swingley, D., Weinberg, A., & McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological Science*, *9*(3), 228–231.

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In I. A. Sekerina, E. M. Fernandez, & H. Clahsen (Eds.), *Developmental psycholinguistics: On-line methods in children's language processing* (pp. 97–135). Amsterdam: John Benjamins.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., … Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. https://doi.org/10.1111/infa.12182

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and Consistency in Early Language Learning: The Wordbank Project*. Cambridge, MA: MIT Press.

Garnier, Simon, Ross, Noam, Rudis, Robert, … Cédric. (2021a). *viridis - colorblind-friendly color maps for r*. https://doi.org/10.5281/zenodo.4679424

Garnier, Simon, Ross, Noam, Rudis, Robert, … Cédric. (2021b). *viridis - colorblind-friendly color maps for r*. https://doi.org/10.5281/zenodo.4679424

Garrison, H., Baudet, G., Breitfeld, E., Aberman, A., & Bergelson, E. (2020). Familiarity plays a small role in noun comprehension at 12–18 months. *Infancy*, *25*(4), 458–477.

Gautheron, L., Rochat, N., & Cristia, A. (2021). Managing, storing, and sharing long-form recordings and their annotations. *PsyArXiv*. Retrieved from https://doi.org/10.31234/osf.io/w8trm

Golinkoff, R. M., Ma, W., Song, L., & Hirsh-Pasek, K. (2013). Twenty-five years using the intermodal preferential looking paradigm to study language acquisition: What have we learned? *Perspectives on Psychological Science*, *8*(3), 316–339.

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., … Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*(1), 160044. https://doi.org/10.1038/sdata.2016.44

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., … Frank, M. C. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science*, *5*(8). https://doi.org/10.1098/rsos.180448

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. https://doi.org/10.1017/S0140525X0999152X

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

Hirsh-Pasek, K., Cauley, K. M., Golinkoff, R. M., & Gordon, L. (1987). The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of Child Language*, *14*(1), 23–45.

Hurtado, N., Marchman, V. A., & Fernald, A. (2007). Spoken word recognition by Latino children learning Spanish as their first language. *Journal of Child Language*, *34*(2), 227–249. https://doi.org/10.1017/S0305000906007896

Hurtado, N., Marchman, V. A., & Fernald, A. (2008). Does input influence uptake? Links between maternal talk, processing speed and vocabulary size in Spanish-learning children. *Developmental Science*, *11*(6), 31–39. https://doi.org/10.1111/j.1467-7687.2008.00768.x

Lewis, M., Braginsky, M., Tsuji, S., Bergmann, C., Piccinini, P. E., Cristia, A., & Frank, M. C. (2016). *A Quantitative Synthesis of Early Language Acquisition Using Meta-Analysis. PsyArXiv.* https://doi.org/10.31234/osf.io/htsjm

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science*, *18*(3), 193–198.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041. https://doi.org/10.1126/science.aag2132

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk.* Mahwah, NJ: Lawrence Erlbaum Associates.

Mahr, T., McMillan, B. T. M., Saffran, J. R., Ellis Weismer, S., & Edwards, J. (2015). Anticipatory coarticulation facilitates word recognition in toddlers. *Cognition*, *142*, 345–350. https://doi.org/10.1016/j.cognition.2015.05.009

Marchman, V. A., Loi, E. C., Adams, K. A., Ashland, M., Fernald, A., & Feldman, H. M. (2018). Speed of language comprehension at 18 months old predicts school-relevant outcomes at 54 months old in children born preterm. *Journal of Developmental & Behavioral Pediatrics*, *39*(3), 246–253.

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological Science*, *31*(6), 678–701.

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

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., … Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748. https://doi.org/https://doi.org/10.31234/osf.io/ksfvq

Peter, M. S., Durrant, S., Jessop, A., Bidgood, A., Pine, J. M., & Rowland, C. F. (2019). Does speed of processing or vocabulary size predict later language growth in toddlers? *Cognitive Psychology*, *115*, 101238.

Potter, C., & Lew-Williams, C. (unpublished). Behold the canine!: How does toddlers' knowledge of typical frames and familiar words interact to influence their sentence processing?

Quinn, P. C., Eimas, P. D., & Rosenkrantz, S. L. (1993). Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants.

Perception, *22*(4), 463–475. https://doi.org/10.1068/p220463

R Core Team. (2021). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ronfard, S., Wei, R., & Rowe, M. L. (2021). Exploring the linguistic, cognitive, and social skills underlying lexical processing efficiency as measured by the looking-while-listening paradigm. *Journal of Child Language*, 1–24. https://doi.org/10.1017/S0305000921000106

Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2019). childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods*, *51*(4), 1928–1941. https://doi.org/10.3758/s13428-018-1176-7

Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science*, *13*(5), 480–484. https://doi.org/10.1111/1467-9280.00485

The ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed speech preference. *Advances in Methods and Practices in Psychological Science*, *3*(1), 24–52.

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, *10*(2), 172–175. https://doi.org/10.1111/1467-9280.00127

Wass, S. V., Smith, T. J., & Johnson, M. H. (2013). Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behavior Research Methods*, *45*(1), 229–250.

https://doi.org/10.3758/s13428-012-0245-6

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143–2152. https://doi.org/10.1177/0956797613488145

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. (2021). *Readr: Read rectangular text data*. Retrieved from https://CRAN.R-project.org/package=readr

Yurovsky, D., Wade, A., Kraus, A. M., Gengoux, G. W., Hardan, A. Y., & Frank, M. C. (unpublished). Developmental changes in the speed of social attention in early word learning.

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