The Art of Wrangling: Working with Web-based Visual World Paradigm Eye-tracking Data in Language Research

Authors: Adam A. Bramlett[1], Seth Wiener[2]

Affiliation: [1, 2] Carnegie Mellon University

Abstract

Web-based eye-tracking is more accessible than ever. Researchers can now carry out visual world paradigm studies remotely and access never before tested, multilingual populations via the internet all without the need for an expensive eye-tracker. Web-based eye-tracking, however, requires careful experimental design and extensive data wrangling skills. In this paper, we provide a framework for reproducible, open science visual world paradigm studies using online experiments. We provide step-by-step instructions to building a typical visual world paradigm psycholinguistics study, and walk the reader through a series of data wrangling steps needed to prepare the data for visualization and analysis using the open-source software environment, R. Importantly, we highlight the key decisions researchers need to make and report in order to reproduce an analysis. We demonstrate our approach by carrying out a single change replication of an in-person eye-tracking study, Porretta et al. (2020). We conclude with best practices and recommendations for researchers carrying out bi-/multilingualism web-based visual world paradigm studies.

***Keywords**:* web-based eye-tracking, visual world paradigm, open science, data quality, replication

1. Introduction

Bi-/Multilingual psycholinguistic research is fundamentally constrained by the populations we can test and traditional lab-based research has tested university-aged adults within Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies. Whereas this lab-based approach undoubtably advanced psycholinguistics as a field, there are at least two problems as a result. First, the field struggles to account for individual differences (e.g., Cunnings & Fujita, 2021). This is a natural limitation of largely testing homogenous 18–to-30-year-olds. Yet, researchers continue to probe relationships between speakers, their environment, and their cognition (Kidd et al., 2018, Perpiñán & Montrul, 2023). Second, the field has unintentionally promoted problematic methodological control in many bi-/multilingualism studies (Rothman et al., 2023). Bi-/multilingual studies, for example, tend to compare 'monolinguals' to 'bilinguals' or 'natives' to 'non-natives.' Yet, notions of 'nativeness' or 'bilingualism' naturally vary given the study and setting (Brown et al., 2022, Han et al., 2023).

Fortunately, web-based research has proliferated, thus removing geographical barriers and allowing researchers to collect data from any population of languages users with access to the internet. This in turn allows bi-/multilingualism researchers the potential to recruit more varied populations in search of individual differences and exert more appropriate (theory-driven) experimental control in bi-/multilingualism research. Here we discuss web-based visual world eye-tracking, which has become more accessible and reliable than ever (e.g., Semmelmann & Weigelt, 2017; Vos et al., 2022). Access to this method, however, comes at the cost of multipart data wrangling to properly handle between-participant differences in camera/browser specifications (Prystauka et al., 2023; Vos et al., 2022).

As web-based eye-tracking grows in accessibility and popularity, it is essential to recognize that data wrangling is data analysis; it is data clean-up, transformation in and between data sets, visualization, and statistical analysis (Wickham & Grolemund, 2017). The choices made during web-based eye-tracking data wrangling can and should be standardized and reported, where possible, which in turn can help improve replicability and reliability in the field (e.g., Bolibaugh et al., 2021; Coretta et al., 2023). Here, we provide a framework for handling multilingual web-based visual world paradigm eye-tracking data using R (R Core Team, 2022).

1.1 The Visual World Paradigm

The visual world paradigm (VWP) involves displaying visual stimuli including a target, and competitor(s), and/or distractor(s) with a variety of possible layouts and formats, from pictures to words (e.g., Allopenna et al., 1998; Cooper, 1974; Tanenhaus et al., 1995). While the images are shown, eye-movements are recorded and an audio stimulus (e.g., “beaker”) is played aloud. The participant either needs to select the correct answer based on the perceived audio or simply listen and look as the sound stimulus plays (e.g., passive listening). VWP experiments vary widely in what linguistic process is being investigated e.g., referent prediction, sentence processing, word recognition, phonetic cue integration. However, all VWP experiments carefully control three core constructs—time, audio stimuli, visual stimuli—in order to bring meaning to a fourth core construct: eye-fixations. For the remainder of this paper, these "core four" constructs will be used to guide the reader’s understanding of how variation in eye-movement behavior can be captured, organized, and analyzed.

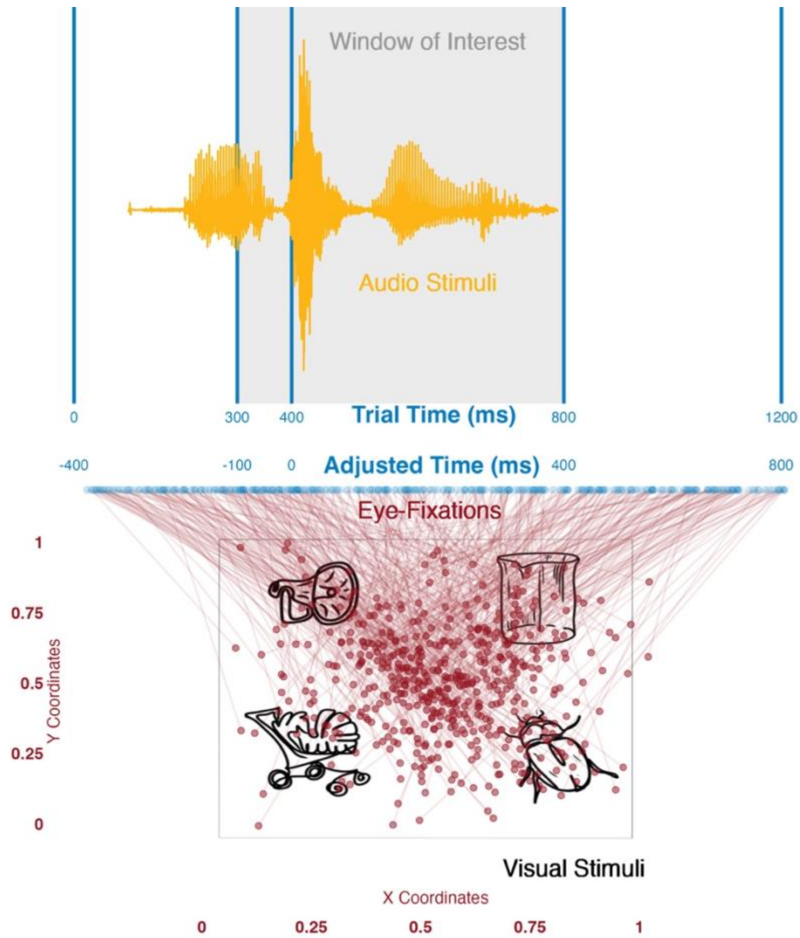

Figure 1. *Illustration of the core four constructs within the VWP. Eye fixations, represented by red dots, and respective times (blue dots).*

1.2 The Core Four Constructs of a VWP Experiment

Time. Eye-tracking is especially valuable because it provides insight into the time-course of cognitive processing. Time can be measured from the beginning of the trial to the end of the trial ('Trial Time' in Figure 1). There are two adjustments, however, that are typically made ('Adjusted Time' in Figure 1). First, it typically takes a listener about 200ms to plan an eye-movement (Matin et al., 1993). Eye-movements within the first 200ms are therefore discarded and researchers typically adjust their analysis accordingly. Second, within each trial there exists a window of interest (grey area in Figure 1), which contains the crucial information necessary to identify the target. For example, time in which any carrier phrase is presented is typically ignored and time after the start of the target word is examined.

Audio Stimuli. The stimulus can be a word, a sentence, or even a non-speech noise. The audio informs the participant about the visual stimuli, often indicating which on-screen visual stimulus is the target or topic of the sentence. The audio stimuli must be carefully locked to time. For example, the end of the gold audio stimuli in Figure 1 is time-locked to end at 800ms (trial time).

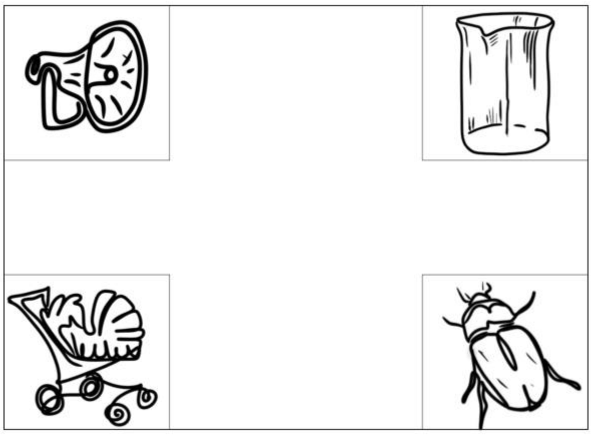

Figure 2. *Example visual stimuli inspired by Allopenna et al. (1998): target 'beaker', onset competitor 'beetle', rhyme competitor 'speaker', and distractor 'stroller'.*

Visual Stimuli. Visual stimuli (Figure 2) can be presented with a preview time or simultaneously with the audio stimuli (Apfelbaum et al., 2021). Ultimately, the specific timing used in a study depends on the research question. Most commonly, visual stimuli are made up of two types: targets and competitors. In the case of four visual stimuli, an additional two visual stimuli can include a second competitor, a single distractor, two distractors, or even target absent designs (Huettig & McQueen, 2007). Visual stimuli are always counterbalanced across the four quadrants so as to reduce the chances of bias in eye-movements in a particular direction. Quadrants are absolute positions on the computer screen (e.g., upper right, bottom left).

Eye-Fixations. Eye-fixations are time-stamped x- and y- screen coordinates that are recorded throughout a trial i.e., where a participant is looking at a particular time. In Figure 1, red dots are specific x- and y- coordinates and red lines tie those fixations to specific times (blue dots). The rate of recording is a function of the measurements recorded per second (e.g., measuring 1000 times in one second = 1000Hz). Eye-fixations get categorized into absolute positions on the screen (quadrants) and then mapped to visual stimuli. Where a participant is looking over time is informed by the audio stimuli.

2. Building a Web-based Visual World Paradigm Experiment

Web-based eye-tracking experiments can be built with a variety of tools including simple web-based GUIs, such as Gorilla/Pavlovia, as well as manual coding on Gorilla or PCIbex Farm, or directly hosting a JavaScript-based experiment online. Readers are invited to follow along on OSF with our detailed Gorilla tutorial (and cloneable experiments). Figure 3 shows an example of a single eye-tracking experiment trial.

Most eye-tracking experiments can be thought of as a forced-choice task (see Experimental ET Tasks for example: simple forced-choice at Gorilla link). From the participant's perspective, they hear an audio stimulus and select one of the visual stimuli[1]. Timing between the onset and/or offset of the core four constructs is essential: the audio and visual stimuli must be time-locked. When building the experiment, it is essential to focus on the timing of the trials, the types of data you want out of the trial[2], and when the webcam should record eye-fixations.

[1] Look and listen paradigm experiments are similar; however, no overt selection occurs.

[2] Feedback is often used in bi-/multilingual studies; an additional *screen* indicating the correct target, such as a circle around the beaker or written corrective feedback could be added.

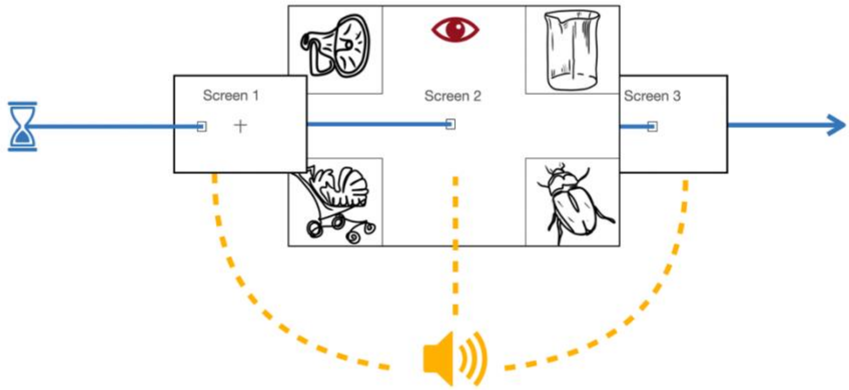

Figure 3. *Sample trial for an eye-tracking study with three screens. Colors match that of Figure 1: blue (time), gold (audio stimuli), black (visual stimuli), red (eye-fixations).*

Figure 3 shows how the exact presentation of your audio stimuli depends on where you want the audio time-locked to the visual stimuli, which is determined by the respective research question. For example, if we were to play the audio in Figure 2 in order to understand spoken word recognition (e.g., Allopenna et al., 1998), we would first show the images and start the beginning of the audio stimulus at a set time after the visuals have been displayed (e.g., 200ms). In this way, participants' eye-fixations for the first 200ms would be evenly distributed over the visual stimuli. Then as the word starts to play, the fixations would gravitate towards the target (i.e., "beaker") and/or competitors (i.e., "speaker" and "beetle") and away from the distractor ("carriage"). As the trial progresses the fixations would tend more and more toward the "beaker."

Most web-based eye tracking studies, including the current study, capture eye-fixations using WebGazer.js (Papoutsaki et al., 2016). WebGazer.js is java script library that uses common webcams to infer the gaze of participants in real time. WebGazer is straightforward to use in both the self-hosted JavaScript based experiments as well as through Gorilla, Psychopy, and PCIbex. Best of all, many of the height and monitor restrictions used in in-person eye-tracking can be ignored because WebGazer uses ridge regression models to infer gaze under a variety of different user set-ups and behaviors.

When creating a WebGazer eye tracking experiment, either a five- or nine-point calibration can be used, with any level set for calibration fail points or repeat calibrations. Nine-point calibration provides a better standard but takes longer and may fail more often. Although it is not necessary because of the manner in which WebGazer.js functions (Chen et al., 2001), we recommended calibration at the beginning of the experiment with reported calibration metrics provided. Importantly, webcams have variable frame-rates (frames per second or FPS) that depend on participant movement, and the participant's device, which can range between 20Hz and 60Hz (Vos et al., 2022). The typical raw eye-fixation samples captured per second is 15, 30, 60, and 120 (standard webcam FPS) but will likely be much lower in the actual data due to the aforementioned reasons.

Additionally, the participant's lighting environment can affect the number of fixations recorded. For example, darker rooms may lower FPS. This means that some trials will capture more/less eye-fixations than other trials (Prystauka et al., 2023). Whereas brighter rooms can result in greater FPS, the directionality of the lighting can also affect calibration. If a light source is behind the participants this can lead to improper exposure. Finally, the timing of eye-fixations can vary within a trial with non-equal measurements between captured eye fixations. This means that the eye-fixations being captured start to drop throughout the trial. This variability in frame-rate can be somewhat attenuated by doing in-person eye-tracking with WebGazer but is nonetheless somewhat unavoidable (e.g., Papoutsaki et al., 2016).

2.1 VWP Raw Data and Tidy Data

Raw web-based eye-tracking data will vary given the platform for data collection (e.g., directly hosting or Gorilla). Raw data from a web-based VWP experiment, generally, has two basic parts: behavioral task data and eye-tracking data (WebGazer data). Behavioral data will include all selections and timings of those selections (e.g., reaction time, condition, trial order). Eye-fixation data will contain trial-by-trial eye-fixation data that is paired with within-trial trial-time.

Task data				
ID	Trial	Audio file	Response	Correct
		Audio Stimuli	Visuali stimuli selected	
1	1	beaker.wav	image_1	1
1	2	stroller.wav	image_3	1
2	1	beatle.wav	image_4	0
2	2	speaker.wav	image_1	1

Eye-tracking data (1 participant x a single trial)				
ID	Trial	Time	Screen location x	Screen location y
		Trial Time	Eye-fixations	
1	1	0.01	0.576	-0.236
1	1	0.029	0.592	-0.222
1	1	0.038	1.067	0.986
1	1	0.082	1.13	0.852

Figure 4. *Behavioral task data (left) and trial-specific eye-tracking data (right).*

The data structure depicted in Figure 4 is relational. That is, for every trial of each participant, there exists a corresponding set of eye-tracking data that is associated with both the trial and the participant. The eye-tracking data provides a detailed account of the gaze locations throughout the duration of the trial. This form of data while maximally informative, is untidy and difficult to understand. We next turn to tidying the data so that each column refers to a single variable (e.g., audio stimuli) and each row is exactly one observation (e.g., "beaker.wav"). In order to better demonstrate this process, we walk the reader through a replication study involving predictive sentence processing of accented and unaccented speech.

3. Replication of Porretta et al. (2020)

3.1 Background and Motivation

We carried out a single change (web-based data collection) replication study of Porretta et al. (2020)'s in-person VWP experiment. The study was chosen for replication for two principled reasons following Marsden et al. (2018): 1) The majority of materials were made available by the researchers, which minimizes heterogeneity. 2) The recency, novelty, and theoretical impact of the initial study warrant replication for the sake of validation and generalizability. Whereas our study changed only the method of collecting data, this single change caused three important differences summarized in Table 1.

Table 1. *Key Differences Between our Web-based Replication Study and Lab-based Porretta et al. (2020).*

	Our web-based replication	Porretta et al. (2020)
Eye-tracker	Variable personal webcams	Eyelink 1000
Participants	60 Prolific participants	60 university students
Data wrangling	Self-wrangled	Pre-processed

Porretta et al. (2020) used a 2-by-2 experimental design to manipulate talker (native/non-native) and verb type (restrictive/non-restrictive, e.g., “the fireman will climb/need the ladder”, *climb* allows for object prediction but *need* does not). These English sentences were spoken by either a native or Chinese-accented talker. There were two research questions: 1) To what extent do restrictive and non-restrictive verbs modulate predictive sentence processing in accented and unaccented speech? 2) To what extent does accent experience modulate prediction in accented speech?

A direct comparison can be made between our study and Porretta et al. (2020) for research question one, which will indicate the usefulness of web-based eye-tracking for capturing prediction in sentence processing. For research question two, our interpretation will be limited given our random sample of Prolific participants (i.e., we are not controlling experience with Chinese-accented English). For this reason, results of the second analysis cannot provide insight into the quality of online eye-tracking data, but our approach may instead provide evidence of the usefulness of web-based eye-tracking for recruiting varied, non-WEIRD populations outside the university setting, which may be particularly useful for advancing bi-/multilingualism psycholinguistics research and exploring individual differences.

3.2 Methods

We used Gorilla Experiment Builder's eye-tracking 2 zone implemented with WebGazer.js (Anwyl-Irvine et al., 2019; Papoutsaki et al., 2016). All research materials, R data analysis, Gorilla experiment and tasks, and data are available on the Open Science Framework (OSF) (Foster & Deardorff, 2017). The study was approved by the authors' Institutional Review Board. All participants were compensated for their participation. Average completion time of the experiment was 16 minutes including a second (pilot) task that is not reported here.

3.2.1 Participants

To ensure direct comparison to Porretta et al. (2020), we tested the same number of participants, 60 (median age = 31). We recruited through Prolific (Palan & Schitter, 2018) using the same criteria: native monolingual English speakers, between the ages of 18 to 40. Not included in the 60 participants that completed the study were 37 rejected participants (eight failed headphone check, 23 failed eye-calibration, 5 timed-out after 90 minutes, one failure to consent). As we demonstrate below, an additional 11 participants were removed during the data tidying, resulting in 49 total participants analyzed. We return to this internet data quality issue and reduced statistical power in the discussion.

3.2.2 Materials

All recordings were taken from Porretta et al. (2020). The experiment contained 250 images, 50 of which were center images and 200 that made up targets and distractors. 99 of the images were identical to the original experiment (all 50 center images and 49 of the visual stimuli for objects across practice, filler, and experimental items). The remaining 151 images were obtained following the same specifications of the initial study (open-source line-drawn images). Four of the images were created in-house due to not being available online. Four presentation lists were made which counterbalanced talker and verb type.

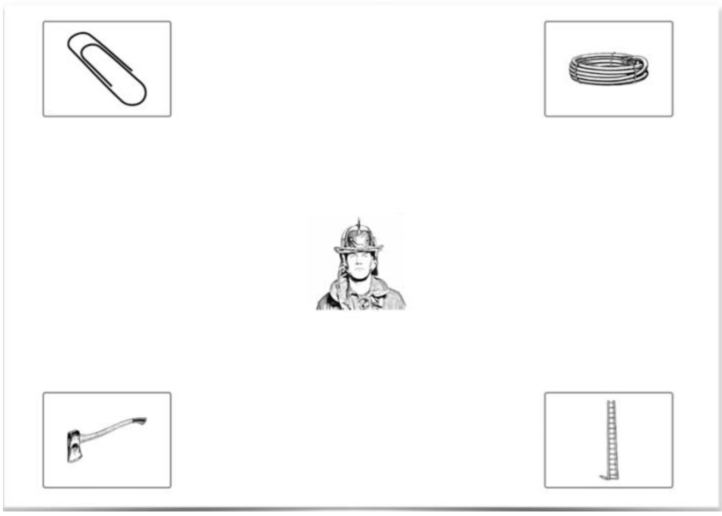

Figure 5. *Example Porretta et al. (2020) visual stimuli and center image. Restrictive sentences (e.g., the fireman climbed the ladder) or nonrestrictive (the fireman needs the ladder) sentences are counter balanced across participants.*

3.2.3 Procedure

After consenting, each participant did two headphone checks: a basic listening task for volume and a dichotic pitch task (Milne et al., 2021). Next, participants did a 5-point eye-calibration set to reject participants below four successful points with a limit of three calibration attempts before rejection. On each trial (24 target, 24 filler), participants were presented with a 500-ms fixation cross followed by a 2x2 visual stimulus with an additional center image that represented the subject of the sentence (Figure 5). Each stimulus was previewed for 200ms. Next, participants heard either a restrictive (e.g., the fireman climbed the ladder) or nonrestrictive (the fireman needs the ladder) sentence spoken with either a native accent or non-native accent. Note competitors and distractors are conflated in this study i.e., everything that is not the target (e.g., the ladder) could be considered a competitor or distractor. Participants then answered a simple comprehension question to ensure attention. After the experimental task, participants filled out a brief questionnaire (identical to Porretta et al.'s) including age, language experience, and estimated Chinese accent experience (captured on a scale of 0-100 with a slider that starts at zero). In order to make a comparison to Porretta et al.'s reported mean of 1.78 (SD = 0.82), accent experience was scaled to 0-30 and then log transformed with a constant of 1. Our population's mean of 0.99 (SD = 0.92), therefore, is lower than that of Porretta et al.'s.

3.3 Data Analysis

In what follows, "L: + line number" (e.g., L:156-157) refer to line numbers in `AOW_r_work_flow.rmd` found on OSF. In L:33, we read in three data frames: The `task_data`, `eye_tracking_data`, and `OSF_data`. To follow along, download the data folder from OSF and select task_data.csv when prompted by R after running L:33. You can load the other data frames by running the following lines. Following Figure 5, the `task_data` is made up of the behavioral data and information obtained during testing; the `eye_tracking_data` is made up of eye-fixations. `task_data` is a messy 97,827 rows by 111 columns, and `eye_tracking_data` is an overwhelming 400,305 rows by 36 columns. As noted earlier, the data are relational. In the next 200 lines of code, we wrangle these structures into data that we can fully use, adapt, and share (see supplementary `combining_data.Rmd` for three methods on combining separate experimental files into a single data frame).

```
31 ## ----Data Reading---
32 #select task_data
33 task_data_select<-file.choose()
34 task_data<-read.csv(task_data_select,header=TRUE, row.names=1)
35 #change for ET data
36 et_data_select<-sub("task_data", "et_data", task_data_select)
37 eyetracking_data<-read.csv(et_data_select,header=TRUE, row.names=1)
38 #change for OSF data
39 OSF_data_select<-sub("task_data", "OSF_data", task_data_select)
40 OSF_data<-read.csv(OSF_data_select,header=TRUE, row.names=1)
```

3.3.1 Questionnaire wrangling

After loading all relevant packages and data, data wrangling always starts with data removal. In a VWP experiment, removal occurs at four levels: questionnaire-based, item-based, behavior-based, fixation-quality-based. Which level you start with is unimportant; we start with questionnaire-based removal and ask which participants should be excluded based on post-experiment questionnaire exclusion criteria, which may be most relevant for bi-/multilingualism studies (e.g., not an L1 English speaker and not between the ages of 18 and 40). In L:43, we start with a clone of our behavioral data frame `task_data` and assess needed variables (`Screen.Name`, `Responses`, `Participant.Private.ID`, `Reaction.Time`(RT)). RT is kept because it allows for removing items that were unnecessarily generated from the experiment structure (i.e., getting rid of rows with 0 RT).

```
42 ## ----Questionnaire: Clean---
43 cleaned_quest_data<-task_data%>%
44 filter(display=="questionairre",na.omit=TRUE)%>%
45 select(Participant.Private.ID,Screen.Name,Response,Reaction.Time)%>%
46 filter(Response != "",Reaction.Time!=0)%>% select(!Reaction.Time)
```

Now that we have a data frame with three columns (`Participant.Private.ID`, `Screen.Name`, `Response`), we can create tidy data with one observation per row and one variable per column. `pivot_wider()` and `pivot_longer()` offer a simple solution to this common data structure problem. Figure 6 demonstrates how experimental data (e.g., Gorilla-tasks, Psychopy, E-Prime) often require widening, whereas questionnaire data (e.g., Gorilla-questionnaires, Google forms, Qualtrics) require pivoting longer. In L:49, we pivot wider to create a single row for each participant with each question having its own column. It is much easier to come up with standards for removal in the `speaks_L2`, `age`, or `hear_impaired` columns than for the Response column, which would require conditional standards based on `Screen.Name`.

```
49 ## ----Questionnaire: Tidy---
50 tidy_quest_data<-cleaned_quest_data%>%
51 group_by(Participant.Private.ID,Screen.Name)%>%
52 summarise_all(toString)%>%
53   pivot_wider(names_from=Screen.Name,values_from=Response)%>%
54   mutate(speaks_L2 =if_else(str_detect(other_languages_spoken,"German")&
55                        !is.na(other_languages_spoken),1,0),
56         across(c(chinese_study_duration,age,experience_chinese_accent),
57                as.numeric),
58         Participant.Private.ID = as.factor(Participant.Private.ID))%>%
59   select(!other_languages_spoken)
```

ID	screen.Name	response
1	age	24
1	speaks_L2	1
1	hear_impaired	0
2	age	36
2	speaks_L2	0
2	hear_impaired	0

Longer

Wider

ID	age	speaks_L2	hear_impaired
1	24	1	0
2	36	0	0

Figure 6. *Examples of long data (left) and wide data (right).*

In L:69, we find that two participants should be removed for language expertise outside English and one for exceeding the age cutoff (both predetermined values based on Porretta et al.). We can now use this data frame to filter out unqualified participants in the `Participant.Private.ID` column of the next removal stage (See L:61-68 in `AOW_r_work_flow.rmd` for an example of helpful visualization).

```
69 ## ----Questionnaire: Filtered---
70 filtered_quest_data<-tidy_quest_data%>% filter(age<=40
71 & age>=18, #1 removed for age range
72 chinese_study_duration==0, #none removed
73 speaks_L2==0,#2 removed that speak other languages
74 language_disorder == "No") #none removed
```

3.3.2 Behavioral-task wrangling

The next cycle of data wrangling begins with the question: Which participants and items should be removed based on the behavioral results? Cleaning is similar to the questionnaire cycle, but we start from scratch with a clone of `task_data` called `experimental_cleaned` because the new question has new goals, which requires different variables. We start this cycle's implementation by filtering the participants in the behavioral-task clone with the questionnaire data from above in order to only keep those participants that qualified in the questionnaire wrangling cycle (L:77). We then remove all rows except ones related to behavioral data questions (L:78-79) and experimental items (L:80), followed by removing columns with all NAs. Lastly, to achieve tidy data, we split the visual image selection and comprehension question into two columns so that each participant has a single observation for each trial (e.g., pivot into a wider structure, L:84). Removal of columns in L:86-88 makes pivoting possible. Pivoting requires that rows do not have uniquely identifiable information outside the data columns being "widened" (This could also be achieved with the column argument of `pivot_wider`).

```
76 ## ----Experimental Data: Clean and Tidy---
77 experimental_cleaned <- task_data%>%
78     filter(Participant.Private.ID %in%
79            filtered_quest_data$Participant.Private.ID)%>%
80     filter(Zone.Type == "response_button_image"|
81            Zone.Type == "response_button_text")%>%
82     filter(verb_type == "Restricting" |verb_type == "NonRestricting")%>%
83     select_if(~sum(!is.na(.)) > 0)
84 
85 experimental_tidy<-experimental_cleaned%>%
86    select(!c(Event.Index:Local.Date,
87              Screen.Number:Zone.Name,
88              Reaction.Time:Response.Type))%>%
89    pivot_wider(names_from = Zone.Type,values_from = Response)%>%
90    mutate(subject_img_file=center_image)#for renamed match in next step
```

Additionally, we must load in a second data frame `OSF_data` (L:94) from the original experiment. We do this because our experiment only has the quadrants or the visual stimuli without the target, competitor, and distractor information, and later we need `SUBTLWF_obj`, which is the log frequency of the object words used in the statistical models.

```
93 ## ----OSF Data: Clean and Tidy----
94 OSF_filt<-OSF_data%>%
95    select(talker,verb_type,subject_img_file,img_1_file, img_2_file,
96           img_3_file, img_4_file,log_SUBTLWF_Obj)
```

In L:99, we filter the `OSF_data` for experimental items and use a `left_join()` based on talker condition `verb_type`, and the center visual image `subject_img_file`, which simultaneously pulls in the variables that we need and filters out nonce items (this step could be avoided by putting these variables in the original experimental spreadsheets). Figure 7 demonstrates filtering through different types of joining.

```
98  ## ----Behavioral Data: Join OSF and Experimental Data---
99  behavioral_data<-experimental_tidy%>%
100   left_join(OSF_filt, by=c( "talker","verb_type","subject_img_file"))
```

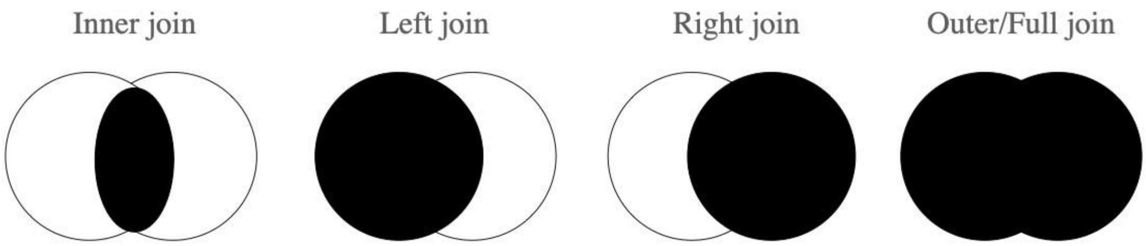

Figure 7. *Solid portions refer to what is kept. Full join retains all rows from both data frames. Left join is more restrictive and includes all the rows from the left (first) data frame and matching values from the right data frame (second). Right join is the inverse of left join. Inner join is the most restrictive, it only retains rows with matching values from both data frames.*

Now that we have the variables we need in `behavioral_data`, we can create variables for the answers being correct/incorrect for our removal process. We will do this for both the item selection (L:105) and comprehension question (L:106).

```
102 ## ----Behavioral Data: Clean and Tidy---
103 behavioral_data <-behavioral_data %>%
104   mutate(participant = as.factor(Participant.Private.ID),
105          image_incorrect= if_else(img_1_file==response_button_image,0,1),
106          text_incorrect = if_else(response_button_text=="Yes",0,1))
```

Importantly, researchers should establish a criterion for removal prior to data collection. Because Porretta et al. (2020) did not report the criteria they used, we based our removal on three standard deviations from the mean inaccuracy of participants/items separately, which results in three participants being removed.

```
108 ## ----Behavioral Data: Removal Standards----
109 #Standard deviations is used to retain maximum amount of quality data
110 #We set all of these to be 3 SDs, code here is only for your future use
111 image_participant_threshold = 3
112 image_item_threshold = 3
113 text_participant_threshold = 3
114 text_item_threshold = 3
```

We aggregated participant inaccuracies by adding together incorrect items by participant and item for both item selection (L:118-129) and comprehension question (L:131-142), respectively. We end here by removing the incorrect trials to prepare for the eye-tracking data wrangling (L:144-145).

```
## ----Behavioral Data: Participant and Item Removal----
#participant removal
participant_agg<-behavioral_data%>%
   group_by(Participant.Private.ID)%>%
   summarize(num_incorrect_image=sum(image_incorrect),
             num_incorrect_text=sum(text_incorrect))%>%
   mutate(mean_image_score = mean(num_incorrect_image),
          sd_image_score = sd(num_incorrect_image),
          mean_text_score = mean(num_incorrect_text),
          sd_text_score = sd(num_incorrect_text))%>%
   filter(num_incorrect_image <= mean_image_score+
           (sd_image_score*image_participant_threshold) &
          num_incorrect_text <= mean_text_score+
           (sd_text_score*text_participant_threshold))
#item removal
item_agg<-behavioral_data%>%
   group_by(center_image)%>%
   summarize(num_incorrect_image=sum(image_incorrect),
             num_incorrect_text=sum(text_incorrect))%>%
   mutate(mean_image_score = mean(num_incorrect_image),
          sd_image_score = sd(num_incorrect_image),
          mean_text_score = mean(num_incorrect_text),
          sd_text_score = sd(num_incorrect_text))%>%
   filter(num_incorrect_image <= mean_image_score+
           (sd_image_score*image_item_threshold) &
          num_incorrect_text <= mean_text_score+
           (sd_text_score*text_item_threshold))

behavioral_data <-behavioral_data%>%
   filter(image_incorrect == 0 & text_incorrect == 0)
```

One important note here is that the removal is done in parallel. That is, we removed participants and items simultaneously. If you sequentially remove participant or item first then removal results would be different in the `behavioral_data` (e.g., more or less items or participants would be removed). Said another way, this removal method assumes that a “bad” item or poor performing participant would be below the distributional counts independently.

3.3.3 Eye-tracking wrangling

Removal and adjustment of eye-tracking data is done through an exploratory lens as there is little current reference for expected results for eye-fixations and frame-rate in web-based eye-tracking. However, recent work has begun to fill this gap (see Prystauka et al., 2023; Vos et al., 2022). Here, two questions guide our approach: How should eye-fixations be classified into quadrants in web-based eye-tracking? And, what quality of frame-rate is needed to capture the effects of interest? We start by filtering out participants from the previous data sets. Here, the retained participants (L:118) and items (L:131) from the previous step are used to define what we want to keep in the `behavioral_data` (L:148-150) with the `%in%` operator.

```
147 ## ----Behavioral Data: Removing with IN Operator---
148 behavioral_data<-behavioral_data%>%
149   filter(Participant.Private.ID%in%
              participant_agg$Participant.Private.ID&
150   center_image %in% item_agg$center_image)%>%
151   select(-c(text_incorrect,image_incorrect,response_button_text))
```

Whereas the `et_data` is much larger than the previous data frames, the same methods are used. Selection of data can be reduced to only the time `time_elapsed`, participant

`participant_id`, and eye-fixations `x_pred_normalised y_pred_normalised` (L:154-156), which is filtered by only usable fixation points (L:157), followed by variable renaming for upcoming joining of `et_data` and `behavioral_data` (L:158-159).

```
153 ## ----ET Data: Tidying and Filtering with an Inner Join---
154 et_data<-eyetracking_data%>%
155   select(time_elapsed,participant_id,spreadsheet_row,
156          type,x_pred_normalised,y_pred_normalised)%>%
157   filter(type =="prediction" )%>%
158   rename("Participant.Private.ID"="participant_id",
159          "Spreadsheet.Row"="spreadsheet_row")
```

Now that both `behavioral_data` and `et_data` are cleaned and tidy, `left_join()` (L:173) is used to create `all_data` from our `behavioral_data` and `eye_tracking` data. This data frame now has all of the eye-tracking data and behavioral-task data from the entire experiment (L:173-174). However, the data from the `et_data` only includes unclassified eye-fixations. Specifically, it includes the x and y coordinates without a link to the visual stimuli that are being viewed. A Shiny app was created to dynamically explore how eye-fixations are distributed with variable amounts of removal at four crucial time points: the beginning of the sentence (-400ms), verb onset (0ms), object onset, and selection of visual stimuli. The app also includes dynamically calculated data loss. Figure 8 is a fixed version of the fixation points from the app (See Eye-fixations Shiny App in OSF). In the discussion, implications of removal standards based on eye-fixation alone are considered and discussed as a signal detection problem.

As displayed in Figure 8, fixations are mostly distributed at the center of the screen, indicating no looks to quadrants. Whereas this remains true for competitor items throughout the trial, target items begin to move toward visual stimuli as early as the verb onset and much more in later time frames. Crucially, however, the fixations do not always reach the actual quadrants. In analyzing the data from the Shiny app, removing data between the center point of the screen and the inner-edges of the quadrants results in ˜83.33% data loss, which is more than twice as high as previously reported for two image web-based studies (Vos et al., 2022). If we move to a more relaxed categorization, then only 6.71% of data is lost. In contrast, maximal outer-edge removal results in very little data loss (max ~32%). When removing inner-edge eye-fixations, the choice comes down to removing signal to avoid noise in spatial ambiguity, or embracing noise to maximally retain the signal. As shown in the competitors-time 800 (upper-right) section of Figure 8, the noise is randomly distributed across quadrants just as it is early in the trial before eye-movements tend toward visual stimuli. Here, we aim to strike the balance of the signal-to-noise trade off by removing most of the data outside the screen size and by maximally retaining inner data that shows trends. This leads us to believe that no bias would occur even if classifying data from the x, y fixation center (0.5, 0.5).

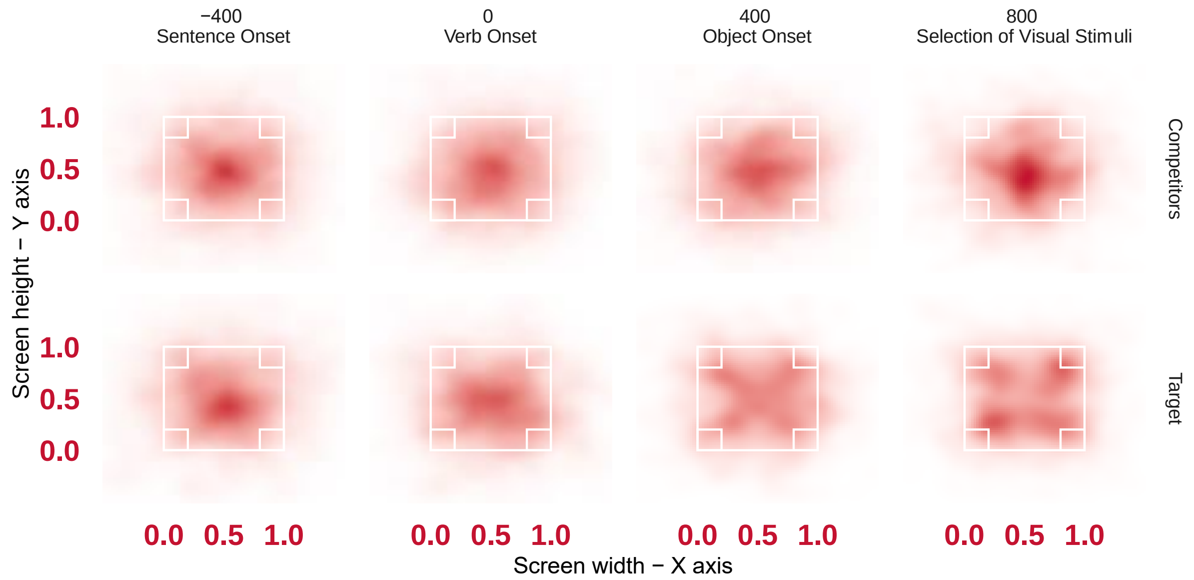

Figure 8. *Quadrant locations and actual screen sizes are denoted with white lines.*

From L:180-190, we create a classification system based on no inner-edge removal of the eye-fixations and partial removal of outer-edge eye-fixations (the code was created with inner removal in mind so that future researchers can simply adapt the distance variable L:177, if desired). We use two types of control flow to first classify eye-fixations into quadrants and then create binary variables to link the quadrant to the visual stimuli. `case_when()` is used (L:180-190) because of the multiple conditions and because `case_when()` is Boolean, meaning it provides a specific output in the case of something being true. For example, if we only want to classify images that are within a particular space and leave others blank, then non-binary classification like `case_when()` is optimal. In contrast, if the outcomes of a classification are binary, then `ifelse()` is an effective solution. For example, L:192-200 makes a binary decision on whether an image being viewed is the same or different from the target (L:193), competitors (L:194-195), and distractor (L:196), separately (Note that competitors and distractors are the same in our experiment, so we included this for ease of future use). While complexity of implementation may vary, logically either can be used to achieve the same result in all cases with the use of operators and/or nesting.

```
## ----ET Data: Localizing Visual Stimuli----
#logically equivalent to doing full join and removing non-experimental trials.
all_data <-behavioral_data %>%
   left_join(et_data, by=c("Participant.Private.ID", "Spreadsheet.Row"))

center=.5#center of screen
distance=0#distance to visual stimuli
beyond_screen=1 #distance to beyond_screen

all_data<-all_data%>%
   mutate(image_viewing=
   case_when(x_pred_normalised <= center-distance &
              y_pred_normalised >= center+distance ~ image_1,
             x_pred_normalised >= center+distance &
              y_pred_normalised >= center+distance ~ image_2,
             x_pred_normalised <= center-distance &
              y_pred_normalised <= center-distance ~ image_3,
             x_pred_normalised >= center+distance &
              y_pred_normalised <= center-distance ~ image_4))%>%
filter(!is.na(image_viewing))

all_data<-all_data %>%
   mutate(target = if_else(image_viewing == img_1_file, 1, 0),
          comp_1 = if_else(image_viewing == img_2_file, 1, 0),
          comp_2 = if_else(image_viewing == img_3_file, 1, 0),
          dist  = if_else(image_viewing == img_4_file, 1, 0))%>%
filter(x_pred_normalised>center-beyond_screen &
x_pred_normalised<center+beyond_screen&
y_pred_normalised>center-beyond_screen &
y_pred_normalised<center+beyond_screen)
```

In addition to more variable eye-fixations, web-based eye-tracking also has variable frame-rates. Figure 9 shows a categorization of participants by median frame-rate across trials.

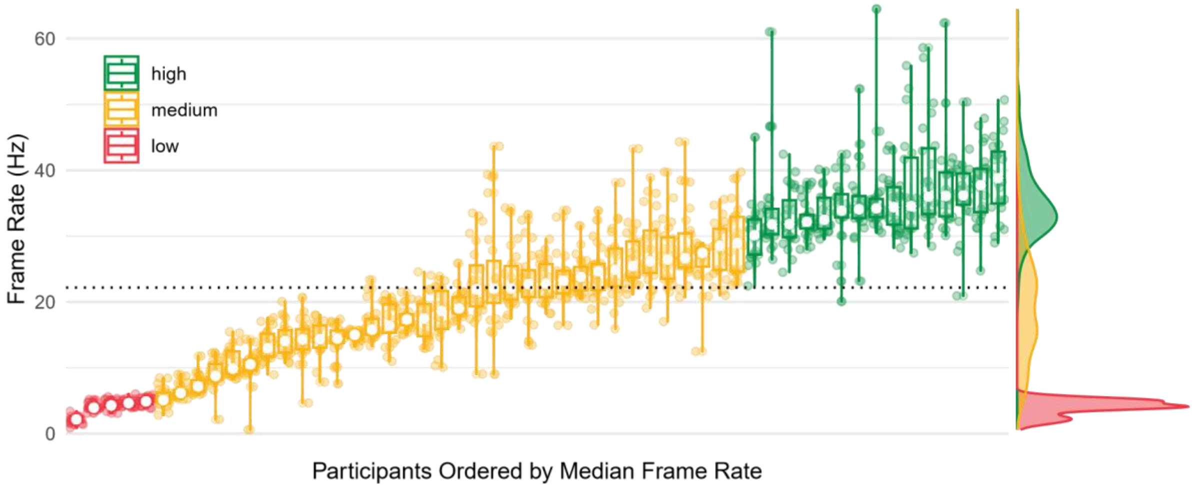

Figure 9. *Participant frame-rate. Mean is marked with dotted horizontal line. High, medium and low categories are defined by the median frame-rate of each participant, making cutoffs by peaks of the distribution. Frame is shown in hertz (Hz) and participants are individually represented by each boxplot.*

Like other recent web-based eye-tracking studies, our mean frame-rate was 20Hz (M = 22.17Hz, SD = 11.61). Here, we remove the five participants with less than 5Hz median frame-rates and create time bins by first creating a standard for removal in L:378 and a binning size (L:379). Median is used because means are more sensitive outliers; e.g., if a participant has one exceptionally low frame-rate this will not be just cause for removal if we use medians (Leys et al., 2013). We then aggregate by participant `Participant.Private.ID`, item `subject_img_file`, and condition `verb_type talker` (L:381) in order to remove all participants that are below our standard predetermined median; i.e., 5Hz (L:381-388) (Vos et al., 2022). Next, time bins are created by normalizing the time range for each item (L:389). Additionally, we subtracted 200ms for human eye movements to occur and thus center the time so that zero is always the onset of the verb of interest (this step was not explicit in Porretta et al. (2020), but we recommend future researchers always make this step explicit to ensure that future studies can reproduce your results). After normalizing, bins are created by dividing the time `time_elapsed` by the bin size `time_binning`, rounding, then multiplying by the bin size `time_binning` (L:390), which is simply rounding items to the nearest bin size number thus allowing you to use any size bin for your data rather than an assumed pre-set bin size.

```
377 ## ----All Data: Clean and Tidy---
378 frame_rate_cut_off<-5
379 time_binning<-50
380 all_data_cleaned<-all_data%>%
381    group_by(Participant.Private.ID,subject_img_file,verb_type,talker)%>%
382    mutate(count = n(),
383           max_time = max(time_elapsed),
384           frame_rate = count/max_time*1000)%>%
385    ungroup()%>%
386    group_by(Participant.Private.ID)%>%
387    mutate(median_frame_rate = median(frame_rate))%>%
388    filter(median_frame_rate>=frame_rate_cut_off)%>%
389    mutate(time_elapsed=time_elapsed-object_start-200)%>%
390    mutate(time_elapsed_rounded=time_binning*round
             ((time_elapsed)/ time_binning))
391
392 all_data_tidy <- all_data_cleaned%>%
393    filter(time_elapsed_rounded>=-400 & time_elapsed_rounded<=800)
```

Creating time bins is fundamentally discretizing a continuous scale. In any fixed set of eye-tracking data, the grain size of the time scale has an inverse relationship to the amount of data in each time bin. If you increase the bin size, you will have more data per bin, but less bins across time. Many statistical analyses can bypass the binning procedure altogether by keeping time a continuous variable. Nevertheless, for analyses that do require time bins and for visualization alone, it is worth exploring whether specific bin sizes affect a researcher's ability to capture an effect. To do this, we created a second Shiny app that is depicted in Figure 10 (see Frame-Rate Shiny App in OSF). The Frame-Rate Shiny App allows the reader to explore the interactions between data removal based on participant median frame-rates, changing bin sizes, and seeing output in the form of empirical logits for either linear lines or GAMM smoothed curves. Here, two crucial discoveries are made.

First, almost any arbitrary sized bin captures the effect of `verb_type`, with the caveat of the bin needing to be several sizes smaller than the window of interest. Second, nearly any frame-rate of data can capture the effect outside very low frame-rates of 5Hz and below. If only examining data that is 6-11Hz, the effect of `verb_type` for `talker` starts to become apparent while the accented speaker effect for `verb_type` becomes apparent between 12-17 Hz.

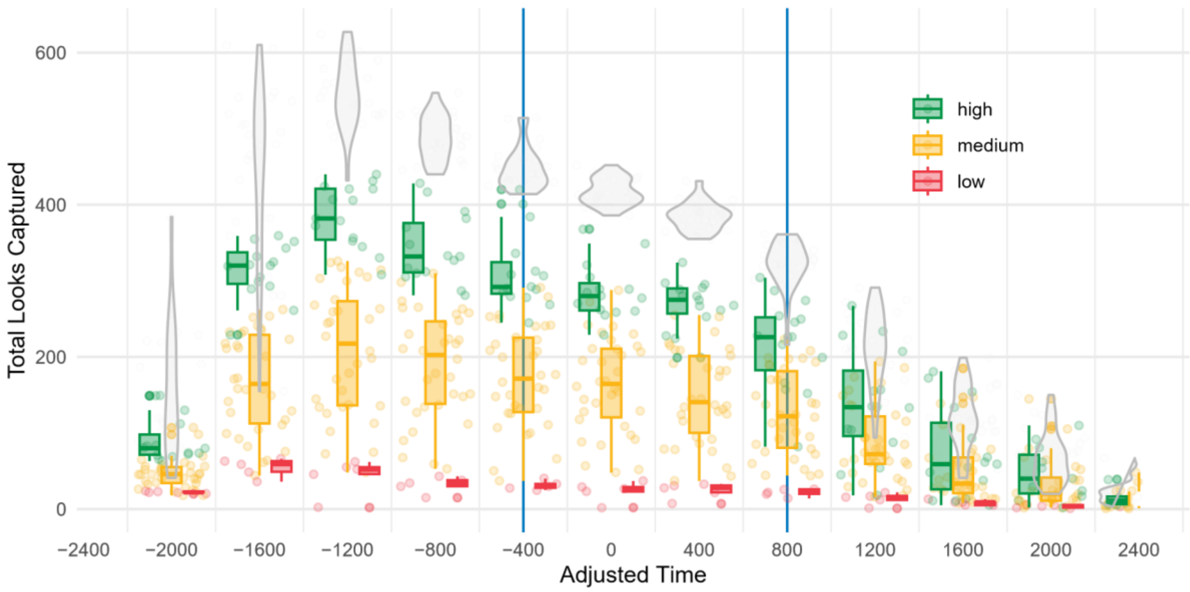

Figure 10. *Total looks per bin size. Like figure 9, high, medium, and low categories are defined by the median frame-rate of each participant. Adjusted time is in milliseconds (ms). Looks captured are raw counts across participants/items.*

The last step before visualization and statistical analysis is a final tidying. Like the first wrangling that we did, we create a tidy data frame through removal. Here, all eye-fixations that are outside the window of interest (-400ms and 800ms) are removed. Now, our new tidy data is structured based on the core four constructs. For each participant, each audio stimuli and visual stimuli set is classified by `talker` and `verb_type`. Finally, we have removed all times outside the window of interest. By tidying in this way, eye-fixations become meaningful in that each row is classified into looks to targets, competitors, and distractors, and each row is a classified eye-fixation based on a specific time, for each participant, and for varying conditions. Between the two data frames `all_data_cleaned` and `all_data_tidy`, we have all of the behavioral data ready for any analysis or exploration that can be done.

4. Modeling ET data

In all previous steps, wrangling can be thought of as a condensing process, where the primary object is to remove, clean, and transform the data into a structure that is usable. However, once the data is put into tidy form, then the data must be transformed for specific visualizations and statistical analyses (hereafter, models). In this section, we think of `all_data_cleaned` and `all_data_tidy` as launching points to gain an understanding of our data[3].

We start by creating two data frames from `all_data_tidy` : `mem_data` in L:453 and `gamm_data` in L:459. In general, maximally retaining informative columns is essential to creating a usable data frame. When building models, however, it is often best to remove

[3] If you wish to start from here then read in the `all_data_tidy` and `all_data_cleaned` from cleaned data on OSF.

variables that you will not be using. This is because some models can have complications interpreting unprocessed data types (e.g., `NAs`). For `mem_data`, we start by selecting all necessary columns for the model (L:454-455). Factor type conversion occurs next (L:456). Finally, to get background information we join `tidy_quest_data`. In addition to the `mem_data`, we create `gamm_data` by simply cloning `mem_data` in L:459 and by adding a single variable needed in the GAMM models.

```
452 ## ----All Data: Preparing for Models---
453 mem_data<-all_data_tidy%>%
454   select(Participant.Private.ID,verb_type,talker,
455          subject_img_file,target,Trial.Number,log_SUBTLWF_Obj,
         target_obj,time_elapsed)%>%
456   mutate(Participant.Private.ID=as.factor(Participant.Private.ID))%>%
457   left_join(filtered_quest_data)
458
459 gamm_data<-mem_data%>%
460   mutate(Condition = paste(talker,verb_type,sep="."))
```

There are a handful of excellent papers that outline the advantages and disadvantages of different methods of eye-fixation analysis and relevant considerations for each method of analysis (Barr, 2008; Ito & Knoeferle, 2022; McMurray, 2023; Mirman et al., 2008). Here, we continue to focus on the data wrangling process and present the data wrangling steps—and decisions—needed to carry out two of the more widely used statistical analyses in the field: generalized linear mixed effect models (GLMMs) and generalized additive mixed effects models (GAMMs), which does not require the assumption of linearity. Both GLMMs and GAMMs require specific contrast coding (e.g., dummy, orthogonal) of the data before running models to get expected results. After contrast coding, all model building starts with maximal models, as justified by the design, working down to simpler models for model comparison (see Barr et al., 2013).

4.1 GLMMs

4.1.1 GLMMs: coding

For GLMMs coding, start with data type conversion (L:464-465), then re-level both `talker` (Native, Non-Native) and `verb_type` (Restrictive or Non-Restrictive) so that `verb_type` Restrictive and `talker` Native are both set as reference levels (L:466-467). We can then rename the contrasts to improve model output readability (L:468-471) and later visualization. In L:473 through L:476, we normalize the `time_elapsed`. Lastly, we create a data frame for the accent model (L:477).

```
463 ## ----GLMM: Leveling the Data---
464 mem_data$verb_type<-as.factor(mem_data$verb_type)
465 mem_data$talker<-as.factor(mem_data$talker)
466 contrasts(mem_data$verb_type)<-c(-.5,.5)
467 contrasts(mem_data$talker)<-c(-.5,.5)
468 colnames(contrasts(mem_data$talker))<- c('Native:')
469 rownames(contrasts(mem_data$talker))<-c("Native","NonNative")
470 colnames(contrasts(mem_data$verb_type))<- c('Restricting:')
471 rownames(contrasts(mem_data$verb_type))<-
                          c("Non-Restricting","Restricting")
472 mem_data$experience_chinese<-mem_data$experience_chinese_accent
473 mem_data <- mem_data %>%
474    mutate(time_normalized =
475           (time_elapsed - min(time_elapsed)) /
476           (max(time_elapsed) - min(time_elapsed)))
477 accent_mem_data<-mem_data%>%filter(talker == "NonNativeMale")
```

4.1.2 GLMMs: models

Two GLMMs were built using the `lme4` package (Bates et al., 2014). Looks to the target (coded as 1, 0) served as the dependent variable. The `Main Model` included three fixed effects: `verb_type` (Restrictive or Non-Restrictive), `talker` (Native, Non-Native) and their interaction (L:509). Random intercepts for `subject_img_file`, `Participant.Private.ID`, and `time_normalized` were included, as were random slopes for `talker` and `verb_type`. The logit link function ("binomial") was specified in the model, equivalent to modeling logit-transformed response probability with identity link function. Model comparison[4] showed preference for the full model with ANOVA comparisons ($p < .001$) and lower AIC and BIC.

```
508 ## ----GLMM: Main Model---
509 glmm1_1<-glmer(target~talker*verb_type+
510                  (talker|subject_img_file)+
511                  (verb_type|Participant.Private.ID)+
512                  (1|time_normalized),
513                family="binomial",data=mem_data)
514 summary(glmm1_1)
```

Similar to the above model, an accent only model was run on `accent_mem_data`. Model specifications are identical to `Main Model` outside of changing fixed effects to `experience_chinese` (L:540). Additionally, `talker` is removed as a random slope because `accent_mem_data` only has one `talker`: accented. Full models were shown to outperform simpler models from ANOVA comparisons ($p < .001$) and lower AIC and BIC, as well as non-convergence of simpler models.

[4] See `AOW_r_work_flow.rmd` for all model comparisons

```
539 ## ----GLMM: Accent Model---
540 glmm2_1<-glmer(target~experience_chinese+
541                (1|subject_img_file)+
542                (1|Participant.Private.ID)+
543                 (1|time_normalized),family="binomial",data=accent_mem_data)
544 summary(glmm2_1)
```

4.2 GAMMs

Like GLMM data, GAMM data must be first coded/prepared (L:546-559). Here, we turn variables into factors and level them at the same time (e.g., L:550-553). However, it is important to note that GAMMs interpret sum-coded variables most effectively, L:550-552. We create `event` as a combination between conditions (L:554-555). Then we only `select()` columns necessary for the analysis (L:557-559). Lastly, we split off the accent data for the accent GAMM (L:560).

```
## ----GAMM: Leveling the Data---
gamm_data <- gamm_data %>%
  mutate(
    Condition = as.factor(Condition),
    subject_img_coded = as.numeric(factor(subject_img_file)) - 1,
    talker_coded = as.numeric(factor(talker)) - 1,
    verb_type_coded = as.numeric(factor(verb_type)) - 1,
    Participant.Private.ID = as.factor(Participant.Private.ID),
    Event = as.factor(paste(
      Participant.Private.ID,Trial.Number,sep= ".")),
    experience_chinese = experience_chinese_accent)%>%
  select(Event,Participant.Private.ID,Trial.Number,verb_type_coded,
         talker_coded,subject_img_coded,Condition,target,time_elapsed,
         log_SUBTLWF_Obj,experience_chinese,Event)
gamm_data_accented<-gamm_data%>%filter(talker_coded == 1)
```

GAMM Models were built using the `mgcv` package (Wood, 2017). Model comparisons suggest that random intercept of `Event` significantly improved the maximal model. Like the GLMM model, the GAMM models treat looks to the target (L:603) as the dependent variable with independent variables including three fixed effects: `talker_coded` (L:603), `verb_type_coded` (L:605) and their interaction (L:607). Random effects included `Event` (L:612). Smooth terms were included for `time_elapsed` by levels of `talker_coded` (L:604), `verb_type_coded` (L:606), and `Condition` (L:608). Smooth terms allow for a non-linear relationship between `time_elapsed` and the response variable `verb_type_coded`, with a different smooth function for each level of variable. An additional smooth term for `log_SUBTLWF_Obj` (L:609) was included. Smooth terms for `time_elapsed` were included for grouping levels: `Participant.Private.ID` and `subject_image_file` (L:610-611). The logit link function ("binomial") was specified in the model, equivalent to modeling logit-transformed response probability with identity link function.

```r
## ----GAMM: Main Model---
mod1 <- bam(target ~ talker_coded +
            s(time_elapsed, by=talker_coded) +
            verb_type_coded +
            s(time_elapsed, by=verb_type_coded) +
            talker_coded:verb_type_coded +
            s(time_elapsed, by=Condition)+
            s(log_SUBTLWF_Obj)+
            s(time_elapsed, Participant.Private.ID, bs="fs", m=1)+
            s(time_elapsed, subject_img_coded, bs="fs", m=1)+
            s(Event, bs="re"),
            family="binomial", data=gamm_data, discrete=TRUE, method="fREML")
summary(mod1)
```

The accent GAMM had identical structure to the main GAMM with the expectation of having only 1 main effect, `experience_chinese` (L:642), and removing the smoothing term leveled by `talker_coded`. `gamm_data_accented` was the data frame (L:648). Model comparisons suggest that random intercept of `Event` significantly improves in the maximum model.

```r
## ----GAMM: Accent Model---
mod2 <- bam(target ~ experience_chinese +
              s(time_elapsed, by=verb_type_coded) +
              s(log_SUBTLWF_Obj)+
              s(time_elapsed, Participant.Private.ID, bs="fs", m=1)+
              s(time_elapsed, subject_img_coded, bs="fs", m=1)+
              s(Event, bs="re"),family="binomial",
              data=gamm_data_accented, discrete=TRUE, method=" fREML")
summary(mod2)
```

4.3 Results

We observed nearly identical time course of predictive processing (Figure 11) in which restricted sentences resulted in earlier looks to the target object than nonrestrictive sentences. Further, this effect is partially reduced in accented speech in a similar manner to Porretta et al. (2020). For `ggplot()` code and data wrangling for visualizations, see `AOW_r_work_flow.rmd`.

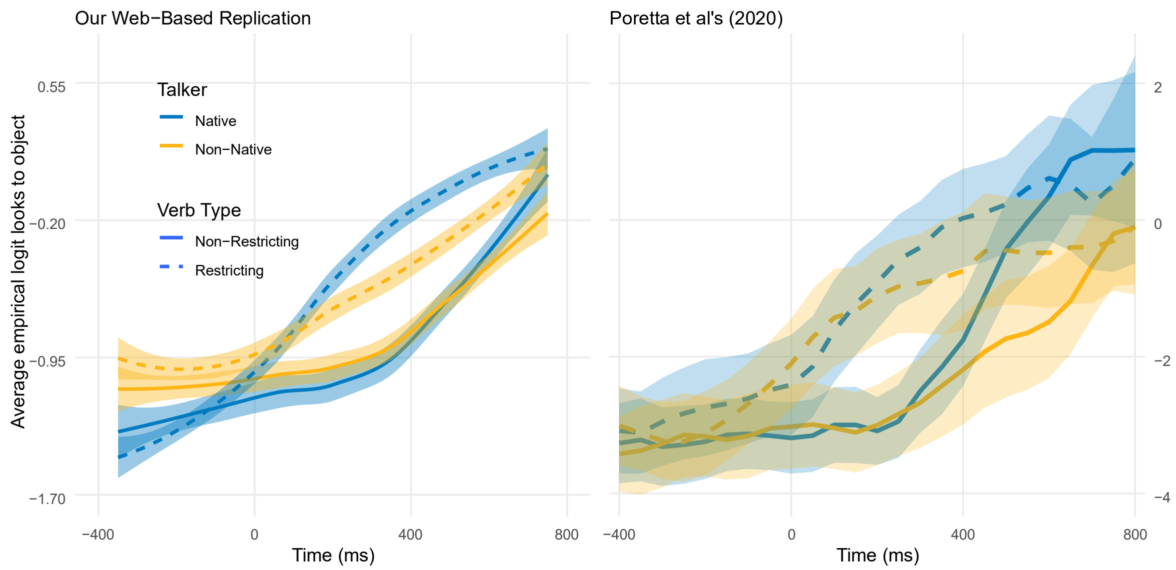

Figure 11. *Looks to native speaker stimuli are shown in blue and non-native are shown in yellow. Dotted lines represent restricting items, while solid lines represent non-restricting*

items. The left y-axis quantifies values from our data, while the right y-axis quantifies data from Porretta et al. (2020).

4.3.1 GLMM Results

Results from the Main GLMM revealed a significant effect of `verb_type` ($\beta = 0.281$, SE = 0.067, z = 4.191, $p < .001$), indicating more looks to targets for restrictive `verb_type` over non-restrictive `verb_type` (Figure 12, left). Additionally, an interaction between speaker and verb type was found ($\beta = -0.136$, SE = 0.053, z = -2.554, $p = 0.011$), indicating less looks when listening to the accented speaker for restricted items. Results from the Accent GLMM failed to reject the null hypothesis at an alpha-level of .05 (Figure 12, right).

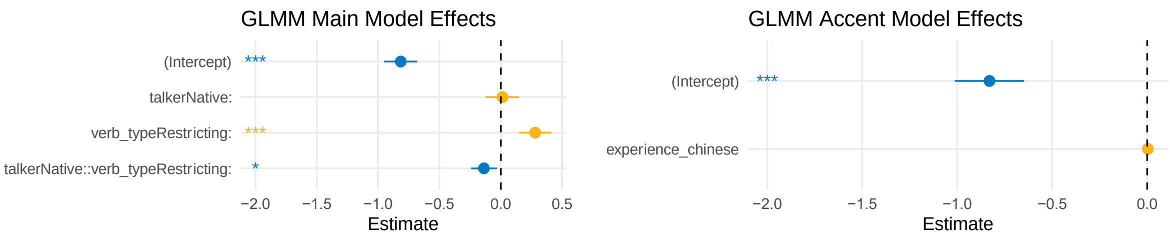

Figure 12. *Model output for parsimonious GLMM models.*

4.3.2 GAMM Results

Like the GLMM modeling results, results from the Main GAMM revealed a significant effect of `verb_type` ($\beta = 0.398$, SE = 0.129, z = 3.078, p = .002), indicating more looks to targets for restrictive `verb_type` over non-restrictive `verb_type` (Figure 13, left). Results from the Accent GAMM failed to reject the null hypothesis at an alpha-level of .05 (Figure 13, right).

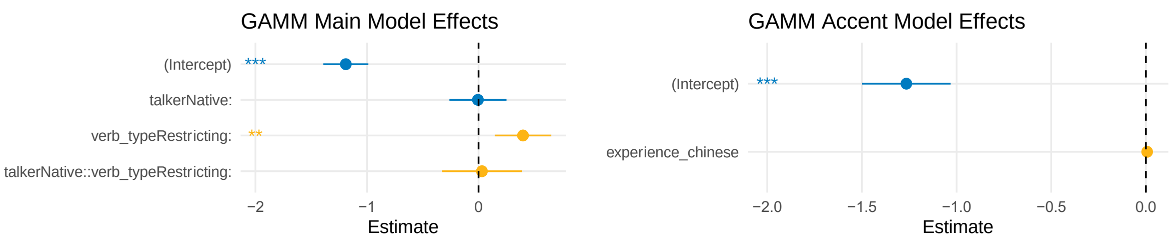

Figure 13. *Model output for parsimonious GAMM models.*

5. Discussion

5.1 Web-based Eye-Tracking May Provide Access to Unique Populations

Our replication results indicate that web-based eye-tracking can capture the same predictive processing as in-person eye-tracking (e.g., Prystauka et al., 2023; Vos et al., 2022). Our main models show that predictive sentence processing is modulated by restrictive and non-restrictive verb type in line with Porretta et al. (2020) and that accented speech impedes predictive processing but does not preclude it. Interestingly, our accent models did not find evidence of accent-experience modulating predictive processing. Why might this be? Our wider (non-university recruited) sample of participants had far less experience with Chinese accents (range = 0-3.43, $M = 0.99$) as compared to the students reported in Porretta et al. (2020) (range = 0–3.43, $M = 1.78$). It is possible that the students tested in Porretta et al. (2020) were exposed to greater Chinese-accented speech more as a result of being on a university campus with international students, while our crowdsourced Prolific participants had far less exposure to Chinese-accented speech in their daily lives. If this difference in experience with Chinese-accented English was behind the lack of evidence for an effect, this may suggest that the population available to test online is different from the population available to test at a traditional WEIRD university setting (Rodd, in press). This speaks to the potential to recruit and test far more varied bi-/multilingual populations, and potentially advance theory and research on individual differences in exciting, new ways.

The null effect may also be due to our low statistical power. With only 49 participants doing 24 trials, we had fewer observations per condition than is recommended (e.g., ~1,600 per condition: Brysbaert & Stevens, 2018). See our ‘main power-analysis simulation’ and ‘accent power-analysis simulation’ R scripts on OSF for post-hoc power analyses to guide replications and extensions. Insights from the 'main power-analysis simulation' indicate that at least 25 to 30 items per condition, with corresponding participant counts of 45 to 50, is necessary to achieve 80% power. For accent models, the participant number must be closer to 90. These simulations underscore the importance of adequate sample sizes for detecting true effects and avoiding Type-II errors.

Finally, measurement error may have contributed to the null effect. The sliding scale used to report Chinese experience was set to start at 0 (Gorilla pre-set setting, which can be controlled in configuration settings). It could be that some of the 13 participants reporting ‘0’ simply selected next to move on quickly. Future studies should clearly state the exact type of method used for capturing such data and make materials fully available to avoid this confusion for metrics that are essential for analyses. Our results are, therefore, inconclusive with respect to the accent models.

5.2 Best Practices for Web-Based Visual World Paradigm Eye-Tracking Research

Alone, eye-fixations are meaningless. Deriving meaning from x- and y-coordinates is achieved through time, visual stimuli, and audio stimuli. These *core four* constructs correspond directly with the variables of our experiment, research questions, and data analyses. However, managing these constructs is complex. Data wrangling through lines of code knits these constructs together, gradually constructing bridges of understanding. In what follows, we summarize best practices that are essential for bi-/multilingual reproducible web-based eye-tracking studies.

Set clear exclusion criteria for participants prior to data collection. Removal of participants given language background information or demographics should be made prior to data collection, and should involve a simple filtering step at the beginning of data wrangling. We encourage pre-registration, if possible.

Include and report behavioral/attention task checks. The decisions and standards of participant and item removal should always be done before data analysis begins. We recommend removal by calculating distribution-based removal standards with median absolute deviation (Leys et al., 2013) or standard deviation with a distribution value set prior to beginning wrangling. Crucially, report what criterion you used for removal (e.g., 3 SD).

Report accuracy cutoffs for participant background information. As noted, we removed one participant for reporting a different age outside our preset filter and two for reporting non-monolingual status, again not in line with our preset filter. It is our experience that some Prolific users may have registered their account with inaccurate information in order to qualify for more studies. Ideally, researchers could pre-register cutoffs and exclusion criteria.

Include and report eye-calibration. Prior to obtaining our 60 participants, 23 potential participants failed our five-point eye-calibration. In other words, roughly 20% of possible participants were unable to participate. We echo recent suggestions requiring participants to pass a specific threshold during eye-tracking calibration. Our standard of 4 out of 5 was sufficient for ensuring high quality data.

Require a minimum median frame-rate greater than 5Hz. In our study, below 5Hz is ‘unusable’. Whereas the research question and effect of interest will dictate the required frame-rate—consider a sentence processing study like ours which captured the native-talker predictive effects within 6-10Hz, versus a word recognition study involving subtle voice-onset time differences which may require 20Hz to detect differences—we echo Vos et al.’s (2022) recommendation to remove participants below the 5Hz range. However, we recommend using median frame-rate or over mean to avoid removal based on extreme trial values. Removal should be reported, as well as the ranges of frame-rates. In cases of more extensive removal, analyses should be run with both the removed participants and the full data to justify removing more data.

Additionally, in an exploratory attempt, we observed that device OS and age of the browser potentially explains variability between participants with newer device OS and more updated browsers having better frame-rates. Additionally, Chromebooks generally provide the lowest frame-rates in our data. Cutoffs for types of browsers could be useful in collecting higher quality data and reducing the need to remove large amounts of participants found in other web-based eye-tracking studies (Prystauka et al., 2023).

Identify a quadrant classification method. Previous web-based eye-tracking studies have shown that removal to the boundary of visual stimuli still enables the researcher to capture results even with strict standards for removal of eye-fixations (28% in Vos et al. (2022)). That is, eye-fixations outside the target areas in Figure 8 are excluded regardless of how close they are to the area (i.e., classifying web-based eye-fixation the same way that lab-based eye tracking does). However, ranges of removal at this strict standard suggest removal of up to 93.61% of the data.

Our suggestion is twofold: firstly, embrace the noise. If unmeaningful eye-fixations are random or equally distributed from the center, then including them will not hinder analysis. Secondly, report and explore standards for maximizing signal and minimizing noise retention of eye-fixations. We suggest that future research maximize retained signal, rather than maximizing removed noise.

Report all time adjustments. Report any time adjustments including the 200ms required to program a saccade (Matin et al., 1993) and any adjustment given a carrier phrase.

Use a meaningful eye-fixation bin size given the research question. There is an intrinsic relationship between frame-rate and the amount of data per bin. Consider the scenario where you are using a bin size of 50 with a participant with 20Hz frame-rate (i.e., one eye-fixation per 50ms on average). In this scenario, each bin would only have one eye-fixation per bin for that participant. Along with reporting standards for binning, we recommend that the researcher find a balance between fewer bins with more data and more bins with less data. Vos et al. (2022) and the current study used 50ms time bins. However, larger bin sizes could be useful with audio stimuli with longer duration. The crucial decision comes down to understanding the time-window of interest. Excluding extreme scenarios where the bin size is approaching the size of the time-window of interest, our data suggests that varying bin size has little to no effect on outcomes.

6. Conclusion

Web-based eye-tracking is here to stay, and with that comes a demand for mastering data-wrangling skills. The choices made during web-based eye-tracking data wrangling should be documented and transparent, with key decisions always reported. We hope that the ***Art of Wrangling*** is a first step towards a more uniform approach to web-based eye-tracking in language research.

Data availability statement

All data and scripts are available through OSF. All data is within the data folder of the OSF stored repository. All scripts are linked through Github under files on OSF (you may need to refresh the page). The primary script for data wrangling and analysis is `AOW_r_work_flow.Rmd`:

https://osf.io/a3e5s/?view_only=822c5f28422444768729f5342fd16848

Competing interests declaration

The authors declare none.

Email and Postal Addresses

Adam A. Bramlett: abramlet@andrew.cmu.edu

Seth Wiener: sethw1@andrew.cmu.edu

Department of Languages, Cultures & Applied Linguistics

341 Posner Hall, 5000 Forbes Avenue

Pittsburgh, PA 15213

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*(4), 419–439.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2019). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388–407.

Apfelbaum, K. S., Klein-Packard, J., & McMurray, B. (2021). The pictures who shall not be named: Empirical support for benefits of preview in the visual world paradigm. *Journal of Memory and Language, 121*, 104279.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*(4), 457–474.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear Mixed-Effects models using lme4.

Bolibaugh, C., Vanek, N., & Marsden, E. J. (2021). Towards a credibility revolution in bilingualism research: Open data and materials as stepping stones to more reproducible and replicable research. *Bilingualism: Language and Cognition, 24*(5), 801–806.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1):9.

Brown, B., Tusmagambet, B., Rahming, V., Tu, C.-Y., DeSalvo, M. B., & Wiener, S. (2023). Searching for the "native" speaker: A preregistered conceptual replication and extension of Reid, Trofimovich, and O'Brien (2019). *Applied Psycholinguistics*, *44*(4), 475–494.

Chen, M. C., Anderson, J. R., & Sohn, M. H. (2001). What can a mouse cursor tell us more? *CHI '01 Extended Abstracts on Human Factors in Computing Systems*.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language. *Cognitive Psychology, 6*(1), 84–107.

Coretta, S., Casillas, J. V., Roessig, S., Franke, M., Ahn, B., Al-Hoorie, A. H., Al-Tamimi, J., Alotaibi, N. E., AlShakhori, M. K., Altmiller, R. M., & et al. (2023). Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human-speech analyses. *Advances in Methods and Practices in Psychological Science, 6*(3). https://doi.org/10.1177/25152459231162567

Cunnings, I., & Fujita, H. (2021). Quantifying individual differences in native and nonnative sentence processing. *Applied Psycholinguistics*, *42*(3), 579–599.

Foster, E. D., & Deardorff, A. (2017). Open science framework (osf). *Journal of the Medical Library Association, 105*(2). https://doi.org/10.5195/jmla.2017.88

Han, J., Kim, J., & Tsukada, K. (2023). Foreign accent in L1 (first language). *Linguistic Approaches to Bilingualism.* https://doi.org/10.1075/lab.22028.han

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language*, 57(4), 460–482.

Ito, A., & Knoeferle, P. (2022). Analysing data from the psycholinguistic visual-world paradigm: Comparison of different analysis methods. *Behavior Research Methods.* https://link.springer.com/article/10.3758/s13428-022-01969-3

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in cognitive sciences*, *22*(2), 154-169.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764–766.

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning, 68*(2), 321–391.

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception & Psychophysics, 53*(4), 372–380.

McMurray, B. (2023). I'm not sure that curve means what you think it means: Toward a [more] realistic understanding of the role of eye-movement generation in the visual world paradigm. *Psychonomic Bulletin & Review, 30*(1), 102–146.

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods, 53*(4), 1551–1562.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language, 59*(4), 475–494.

Palan, S., & Schitter, C. (2018). Prolific.ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*, 22–27.

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). Webgazer: Scalable webcam eye tracking using user interactions. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 3839–3845.

Perpiñán, S. and Montrul, S. (2023). Does your regional variety help you acquire an additional language? *Linguistic Approaches to Bilingualism, 13*(5), 663-692. https://doi.org/10.1075/lab.22057.per

Porretta, V., Buchanan, L., & Järvikivi, J. (2020). When processing costs impact predictive processing: The case of foreign-accented speech and accent experience. *Attention, Perception, & Psychophysics, 82*(4), 1558–1565.

Prystauka, Y., Altmann, G. T., & Rothman, J. (2023). Online eye tracking and real-time sentence processing: On opportunities and efficacy for capturing psycholinguistic effects of different magnitudes and diversity. *Behavior Research Methods*.

R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Rodd, J. M. (in press). Moving experimental psychology online: How to obtain high quality data when we can't see our participants. *Journal of Memory and Language, 134*(104472), 104472.

Rothman, J., Bayram, F., DeLuca, V., Di Pisa, G., Duñabeitia, J. A., Gharibi, K., … Wulff, S. (2023). Monolingual comparative normativity in bilingualism research is out of "control": Arguments and alternatives. *Applied Psycholinguistics*, *44*(3), 316–329.

Semmelmann, K., & Weigelt, S. (2017). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods, 50*(2), 451–465.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217), 1632–1634.

Vos, M., Minor, S., & Ramchand, G. C. (2022). Comparing infrared and webcam eye tracking in the visual world paradigm. *Glossa Psycholinguistics, 1*(1). https://doi.org/10.5070/g6011131

Wickham, H. (2014). Tidy data. *Journal of Statistical Software, 59*(10). https://doi.org/10.18637/jss.v059.i10

Wickham, H., & Grolemund, G. (2017, January). *R for data science: Import, tidy, transform, visualize, and model data* (1st ed.). O'Reilly Media. http://r4ds.had.co.nz/

Wood, S. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman; Hall/CRC.