Behavior Research Methods (2025) 57:68
https://doi.org/10.3758/s13428-024-02572-4

ORIGINAL MANUSCRIPT

# A beginner's guide to eye tracking for psycholinguistic studies of reading

**Elizabeth R. Schotter**[1] · **Brian Dillon**[2]

Accepted: 1 December 2024 / Published online: 22 January 2025


**Abstract**
Eye tracking has been a popular methodology used to study the visual, cognitive, and linguistic processes underlying word recognition and sentence parsing during reading for several decades. However, the successful use of eye tracking requires researchers to make deliberate choices about how they apply this technique, and there is wide variability across labs and fields with respect to which choices are "standard." We aim to provide an easy-to-reference guideline that can help new researchers with their entrée into eye-tracking-while-reading research. Because the standards do – and should – vary from field to field or study to study as is appropriate for the research question, we do not set a rigid recipe for handling eye tracking data, but rather provide a conceptual framework within which researchers can make informed decisions about how to treat their data so that it is most informative *for their research question*. Therefore, this paper provides a description of eye movements in reading and an overview of psycholinguistic research on the topic, an overview of experiment design considerations, a description of the data processing pipeline and important choice points and implications, an overview of common dependent measures and their calculation, and a summary of resources for data analysis.

**Keywords** Eye tracking · Data processing · Reading · Word recognition · Sentence processing

## Goals of this paper

The use of eye tracking has been a popular method to study core word recognition and sentence parsing processing in reading (see Clifton et al., 2016; Rayner, 1998, 2009),[1] particularly as eye tracking technologies have become cheaper and easier to use (Holmqvist et al., 2011). However, as with all advanced scientific techniques, the ability to make informed choices about how to apply this technique rests on a firm understanding of basic facts of eye movements during reading and familiarity with a wide range of assumptions, conventions, and practices. Many of these are often implicit, passed down from researcher to researcher within a research laboratory. Our goal in this article is to capture some of this 'lab lore' in an easy-to-reference format in the hopes that it can help new researchers bootstrap their way into eye-tracking-while-reading research. Along the way, we will try to make explicit many issues that we see as potential pitfalls on the way to generating new eye tracking research. Importantly, this paper is not a "cookbook" – we will not describe a list of inviolable rules for conducting these types of studies. There are many potentially valuable ways to design experiments and process and analyze data, and some approaches may be most appropriate in one scenario but not another. To set a standard practice that all future researchers are urged to follow under all circumstances could risk doing more harm to the field than good. Rather, we aim to provide a framework for decision-making that will empower new researchers to think about how to apply best practices *in the context of their own study*. Of course, these suggestions are embedded in a rich research literature, which we cannot exhaustively summarize here, so we urge our readers

[1] There are, of course, other important topics related to reading (e.g., discourse processing, and engagement with and appreciation of the text, etc.) and dependent measures (e.g., blink rate, body posture, etc.) that are beyond the scope of this paper. Many of the recommendations here will apply to those types of studies but some will not. As we emphasize in the main text, researchers must use past research and deliberate decision making to determine the best practices for their particular study.

This work was partially supported by grants BCS-2120507 to ERS and BCS-2020914 to BD. The two authors worked collaboratively on this paper.

✉ Elizabeth R. Schotter
eschotter@usf.edu

[1] Department of Psychology, University of South Florida, 4202 E. Fowler Ave. PCD 4118G, Tampa, FL 33620, USA

[2] Department of Linguistics, University of Massachusetts, 650 North Pleasant Street, Amherst, MA 01003, USA

to make informed decisions based on past research related to their research domain – whether following those studies or not, there should be a good reason for the researcher's choice.

In this paper, we provide (1) a description of eye movements in reading, (2) an overview of the psycholinguistic literature on eye-tracking-while-reading, including common lexical and sentential variables known to affect reading behavior and leading theories generated from the literature, (3) a discussion of experiment design and preparation, including choices that researchers need to make in the process, (4) a description of the data processing pipeline, including important choices that researchers must make and their potential consequences, (5) an overview of common dependent measures used to investigate reading and how to calculate them from raw eye tracking data, and (6) a brief summary of resources for data analysis approaches. We also provide, in the Appendix, a checklist for authors to go through as they are preparing a manuscript to ensure that they report all the information necessary to evaluate an eye-tracking-while-reading paper.

First, we point the reader to some other valuable tutorials on aspects of eye tracking research that we will not focus on here. For general guidance on best practices for experimental science, see Frank et al. (2024). For a tutorial on the best practices in eye tracking research in general, see Carter and Luke (2020). For a tutorial on eye tracking for syntax research, see Kush and Dillon (2020). Finally, for a retrospective on the career, accomplishments, and scientific approach of Keith Rayner, who essentially founded the modern-day field of eye tracking for reading, see Clifton et al. (2016).

## The state of the field

Eye movement data have been used to study reading since Javal (1878) first observed that the eyes do not glide smoothly across the text, but rather "jerk" from location to location (i.e., make *saccades*) and, in between, stay relatively stable (i.e., during *fixations*). At present, the eye-tracking-while-reading technique is used by researchers in a number of distinct research communities, yielding diverse insights into the basic psychology of reading, as well as questions about the cognitive mechanisms and linguistic processes that support language comprehension. A strength of such research – and what sets eye tracking apart from other common reading methodologies like self-paced reading – is that it uses an ecologically valid task and a highly practiced skill (i.e., reading) with a methodology that provides a non-invasive, temporally precise picture of how those comprehension processes unfold in time. More specifically, by measuring where the eyes fixate (i.e., gaze tracking), researchers can measure one aspect of overt attention that is commonly taken to reflect the attention allocated to achieve text comprehension, referred to as the 'eye–mind link' (Just & Carpenter, 1980).

## The work of the eyes in reading

The eyes move in discrete events as they progress through the text (see Schotter & Rayner, 2015 for a review) and it is important to understand these raw data before undertaking an eye-tracking-while reading study. The reason people even make saccades during reading is that there is a limit to visual *acuity* (i.e., perceptual resolution) imposed by the visual system as a function of *eccentricity* (i.e., distance from the point of gaze). Acuity is highest in the *fovea* (from the point of gaze out to 1° of visual angle away from it in all directions) than in the *parafovea* (1–5° away from the point of gaze) or the *periphery* (areas more than 5° away from the point of gaze; Rayner et al., 2016). This means that, in order to process a given word efficiently, readers make saccades in order to bring perceptual input from that word into their fovea by fixating directly on it. This does not mean that readers do not obtain any information from non-foveal areas of the text; in fact, questions about the nature and extent of word recognition in parafoveal perception is an active area of research (see Section "Linguistic influences on reading behavior"). However, the majority of language processing occurs for words that are in the central focus of attention (i.e., the fovea and nearby upcoming parafovea) and it is for this reason that readers move their eyes across the text and that researchers can take advantage of measurements of the reader's gaze location to make inferences about how they are processing the text.

In a typical eye-tracking-while-reading experiment, researchers measure the position and duration of a sequence of fixations across a sentence or a multi-line paragraph. These two aspects of the eye movement record have been referred to as "where" and "when" decisions (Rayner, 1998, 2009; Schotter & Rayner, 2015; see Schotter et al., 2024). During analysis, these data – this running record of fixations and saccades – are transformed into a number of derived dependent measures, and it is typically these derived dependent measures that are reported in studies using eye-tracking-while-reading.

### Where decisions

Saccades are ballistic movements, which means that, once they are initiated, their execution cannot be altered

(Gilchrist, 2011). There are a number of qualitatively different saccade types that make up reading behavior (Fig. 1).

Saccades generally move the eyes forward about 7 to 9 letter spaces and last 20–35 ms for readers of English and the extent varies from language to language, depending on linguistic properties such as average word length (see Kuperman, 2022; Liversedge et al., 2016; Rayner, 2009). Most saccades move the eyes from one word to the next (i.e., produce *single fixations* on the word), but 5–20% move to another location in the same word and are termed *refixations*. Some saccades move completely past a word without ever landing on it, leading to 30% of the words in the text being *skipped*. However, word skipping is not uniformly distributed across all the words in the language; short, common, predictable words are skipped more often than long, rare, unpredictable words (Kliegl et al., 2004). For example, the word *the* is skipped about 50% of the time, whereas some long rare content words are rarely ever skipped (see Angele & Rayner, 2013). Word length has the strongest influence on word skipping (Brysbaert & Vitu, 1998; Drieghe et al., 2004, 2008; Gautier et al., 2000; Heilbron et al., 2023; Slattery & Yates, 2018) and this may be due to oculomotor constraints on the reading process (i.e., *saccadic range error*: the tendency for longer-than-average saccades to undershoot their intended targets and for shorter-than-average saccades to overshoot their target; Kuperman, 2022; McConkie et al., 1988).

Another 10–15% of saccades move the eyes backward to a different word (i.e., right-to-left in languages that are written left-to-right like English) and are termed *regressive saccades* (or *regressions*; Rayner, 1998, 2009). Most regressions are quite short, going back only a word or two, to correct for oculomotor error (e.g., over-shooting the intended saccade target; McConkie et al., 1988), but regressions can also be quite long and there are presumably many underlying reasons for regressive movements (see Section "Linking eye movements to cognitive activity"). When reading multi-line texts, readers must make a *return sweep* to continue going forward in the text when the eyes move from the end of one line to the beginning of the next (Hofmeister et al., 1999; see Section "Special considerations for multi-line text studies"). Return sweeps often do not make it all the way to the saccade target and are followed by an additional short right-to-left saccade.

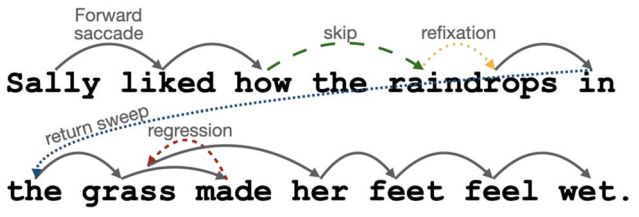


**Fig. 1** Schematic diagram of different types of eye movements during reading

In addition to measuring *which* word the reader fixates, high-precision eye trackers can indicate *where within* a word the reader fixates. In general, readers aim their saccades toward the middle of a word because word recognition is fastest at the *Optimal Viewing Position* (*OVP*); in single-word recognition studies, for every letter that the eyes deviate from the OVP there is an increase in refixation rate, and an increase in processing time of about 20 ms (O'Regan & Jacobs, 1992; O'Regan et al., 1984). However, in natural reading (i.e., within a sentence context), readers tend not to actually land on the OVP, but rather slightly before the middle of the word, at what has been referred to as the *Preferred Viewing Location* (*PVL*; Rayner, 1979). The reason the PVL is shifted slightly leftward of the center of a word (in languages that are read from left to right, like English) is because their saccades tend to fall short when targeting the center due to saccadic range error (McConkie et al., 1988; Rayner, 1979). If the saccadic range error is large, leading the reader to fixate a nonoptimal position, they are more likely to refixate it (O'Regan, 1990; Radach & McConkie, 1998; Rayner et al., 1996; Vitu et al., 1990). Landing position also varies as a function of the launch site from the prior word. It is shifted leftward for far launch sites (e.g., 8–10 letter spaces) and shifted to the right for near launch sites (2–3 letter spaces; McConkie et al., 1988; Rayner et al., 1996).

### When decisions

Little to no visual information is extracted from the text during saccades due to saccadic suppression (Matin, 1974), and therefore the uptake of useful visual information occurs during fixations and is pieced together by the reading system over a series of "snapshots." During reading, fixations usually last 150 to 500 ms (with the average being 200–250 ms; Rayner, 1998, 2009). Fixation durations vary greatly across people and even within an individual. The mean, median, and mode of the distribution is generally around 200–250 ms (Reingold et al., 2012; Staub & Benetar, 2013), but it is not impossible for fixation durations to be as short as 50 ms and as long as 600 ms (Reingold et al., 2012; Schotter & Leinenger, 2016). It remains contested how consistent this average duration is across languages and orthographic systems (see Gagl et al., 2018; Liversedge et al., 2016; Siegelman et al., 2022). Nevertheless, the durations of these fixations are reliably influenced by a range of linguistic variables (see Section "Linguistic influences on reading behavior"), supporting the idea that ongoing language processing is the "engine that drives the eyes" through the text (Reichle et al., 2009a, 2009b, p. 4).

It has been suggested that word skipping is a 'hybrid' measure of when and where decisions because it is both a

decision to move past the word in space and to spend no time directly fixating it (see Schotter et al., 2024). It is important to note that, even if a word is skipped, it is still processed to some extent on the prior fixation (Fisher & Shebilske, 1985), which is often inflated compared to fixations prior to a word that is fixated (Kliegl & Engbert, 2005; Pollatsek et al., 1986; fixations after skipping also tend to be inflated, Reichle et al., 2003). However, it is still not clear how deeply a word is processed on the prior fixation when it is skipped. Although predictability effects on word skipping rates (see Staub, 2015) imply that the word was assessed for meaning, the fact that anomalous words are sometimes skipped (see Angele & Rayner, 2013) and the fact that word length is the strongest predictor of word skipping (see Brysbaert & Vitu, 1998; Heilbron et al., 2023) suggests that there may be some oculomotor component to saccade planning that is separate from word identification.

## Linguistic influences on reading behavior

Because both where and when decisions are related to a wide array of linguistic variables, the use of eye tracking has become popular among psycholinguists, who see it as an ecologically valid tool for probing a range of questions concerning basic language comprehension. One of the strongest indications that eye movements in reading reflect underlying word recognition processes are findings that the durations of fixations on words are influenced by lexical variables. These lexical variables include, but are not limited to, word frequency/familiarity (Chaffin et al., 2001; Inhoff & Rayner, 1986; Kliegl et al., 2004; Rayner & Duffy, 1986; Schilling et al., 1998), ambiguity of meaning (Duffy et al., 1988; Folk & Morris, 2003; Leinenger & Rayner, 2013; Rayner et al., 2006; Sereno et al., 2006; see Duffy et al., 2001), age of acquisition (Juhasz & Rayner, 2003, 2006), phonological properties of words (Ashby & Clifton, 2005; Folk, 1999; Jared et al., 1999; Rayner et al., 1998a, 1998b; Sereno & Rayner, 2000; see Leinenger, 2014), and orthographic neighborhood size (Perea & Pollatsek, 1998). Word length, frequency, and predictability exert some of the strongest influences, and are most commonly studied or controlled for in studies of new variables, leading to them being affectionately called “the big three” (Kliegl et al., 2006; Rayner & Liversedge, 2011). Thus, at a minimum, researchers may want to control these variables if they are not the manipulation of interest.

Readers not only extract information from the word they are currently fixating but there are also influences of the properties of the surrounding (i.e., parafoveal) words. For example, fixation durations are not only longer on low-frequency words than on high-frequency words, but these effects may extend to the fixation durations on the subsequent word (i.e., the effects *spillover* into the next region; Kliegl et al., 2004; Rayner & Duffy, 1986). Some researchers also suggest that effects of word difficulty such as frequency may be observed on the word preceding the manipulated word (i.e., *parafoveal-on-foveal effects*; Kliegl et al., 2004), but the fact that these effects are only observed in corpus analyses, not experimental manipulations, and are also observed in corpus analyses even when the parafoveal word is masked calls this conclusion into question (Angele et al., 2015). Furthermore, it has been claimed, but also challenged, that the difficulty of processing the current word affects the ability for readers to obtain lexical information from the upcoming parafoveal word (i.e., *foveal load effects*; Henderson & Ferreira, 1990, see Veldre & Andrews, 2018b). The extent/magnitude of these effects, as well as what they imply about the distribution of attention during reading is debated (i.e., whether attention is allocated serially or in parallel across multiple words (see Section "Linking eye movements to cognitive activity").

Although many people assume that the decision to skip must be made because the word was fully identified during parafoveal preview (Gordon et al., 2013), there is good reason to believe that it is actually based only on partial recognition. For example, the decision about the to-be-skipped or to-be-fixated word is mostly based on the low-fidelity information obtained from the word in the parafovea (Rayner, 2009), which is degraded by acuity and attentional limitations that reduce the speed of processing (Rayner & Morrison, 1981). Therefore, words can sometimes be inappropriately skipped (Reichle & Drieghe, 2013), even when they make no sense in the context but are very short and high frequency (Abbott et al., 2015; Angele & Rayner, 2013; Angele et al., 2014) and will likely be followed by a regression back to the inappropriately skipped word.

Beyond the lexical properties of the word itself, eye movements reflect the attempt to make sense of the word in the sentence or discourse context in which it is encountered. For example, word fixation probability and duration is strongly related to the word’s predictability in context (Balota et al., 1985; Carroll & Slowiaczek, 1986; Ehrlich & Rayner, 1981; Morris, 1994; Rayner & Well, 1996; Zola, 1984; see Staub, 2015 for a review) and whether it is anomalous (e.g., nonsensical; Rayner et al., 2004; Staub et al., 2007; Warren & McConnell, 2007). However, when a word is merely implausible, but not truly anomalous, the effect in the eye movement record is typically delayed, showing up in later processing measures (Joseph et al., 2009; Rayner et al., 2004; cf. early effects of plausibility during parafoveal preview, Schotter & Jia, 2016; Veldre & Andrews, 2016, 2017, 2018a, 2018b). These effects are sensitive to higher-level discourse processes; when a word is anomalous in a real-world context (e.g., “Jane used a pump to inflate the large *carrot*”) the reading time is inflated, but when that sentence followed from a cartoon or fantasy-like discourse context

(e.g., Bugs Bunny balloons at the Macy's Thanksgiving Day Parade), the effects are reduced or eliminated (Filik, 2008; Warren et al., 2008).

Since regressive eye movements are associated with post-lexical integration difficulty (Reichle et al., 2009a, 2009b), this measure is often used to index the integration difficulty associated with a particular word or words. One prominent early example where this was observed was in Frazier and Rayner's (1982) study of garden path sentences, or sentences where the reader's expectations about the structure of the sentence are dramatically changed by an unexpected syntactic continuation late in the sentence (e.g., "While Mary bathed the baby spat up on the bed"). When the reader processes words that disambiguate the sentence structure to a less expected alternative (e.g., *spat on* in the previous examples), this causes increased fixation times on the disambiguating word (Frazier & Rayner, 1982; Rayner & Frazier, 1987) as well as a higher probability of regression (Frazier & Rayner, 1982; Meseguer et al., 2002).

Even when the linguistic input is not anomalous or potentially misleading, readers tend to have longer fixations at the end of clauses and sentences (Hirotani et al., 2006; Just & Carpenter, 1980; Rayner et al., 1989, 2000). These "wrap-up" effects may have multiple distinct causes, including (1) lagged syntactic integration processes catching up to the location of the eyes, (2) implicit prosody effects (i.e., the inner voice pausing, as one might do when speaking), (3) consolidation of memory for the previously read material (Kaakinen & Hyönä, 2007), or (4) visual effects due to punctuation and capitalization at sentence boundaries (see Stowe et al. (2018) for a discussion of various perspectives on sentence wrap-up effects).

## Linking eye movements to cognitive activity

Before moving on, it is important to note that there are numerous analytical challenges in using eye movements to study linguistic processing. Although the brain presumably processes language continuously during reading, the behavior that readers exhibit during that process is actually made up of a sequence of discrete events (i.e., the fixations and saccades mentioned above[2]; Rayner et al., 1989). It is important to remember that there is generally assumed to be a 'many to many' relationship between underlying psychological constructs (e.g., lexical access difficulty) and the various dependent measures that researchers analyze. For example, in the E-Z Reader model (see more details below), lexical access difficulty jointly impacts fixation duration and the planning of the next saccade. Because of this assumption, we may expect a single underlying cognitive process to be reflected in several different dependent measures in eye-tracking. One possible response to this is to adopt statistical techniques that directly model the relationship between continuous underlying processes and discrete eye movement behavior (Shain & Shuler, 2021). Another response is to construct standardized dependent measures out of a sequence of fixations and saccades, and seek to understand how variables of interest relate to these second-order dependent variables.

It is this second route that is generally pursued in eye-tracking-while-reading research. In order to analyze the eye movement data, eye tracking researchers have defined a number of different dependent measures that are derived from the durations and sequence of fixations across a text. The number of dependent measures that are now commonly analyzed continues to grow as the field evolves, but this is a double-edged sword. On the one hand, the typical eye tracking experiment yields a rich dataset that can be investigated from multiple perspectives. On the other hand, this multiplicity of dependent measures comes with some peril. Most importantly, it increases researcher degrees of freedom (i.e., the choices of which measure to test), and since researchers typically analyze more than one dependent measure in a given study, standard issues concerning multiple statistical comparisons are of particular concern for reading research (von der Malsburg & Angele, 2017; see Section 6 on Data Analysis).

Given these pitfalls, the appropriate application of theory and background knowledge can be a safeguard for the researcher. Importantly, different dependent measures reflect different stages of the reading process (although not with a one-to-one correspondence; see Section "Typical dependent variables"), and therefore we would not expect that every measure shows an effect of a manipulated linguistic variable (von der Malsburg & Angele, 2017). This potential specificity of a presumed effect, and the narrow scope of its time course of influence, may restrict the number of dependent measures a researcher should analyze, thus decreasing the family-wise error rate, and may even allow them to make predictions about where effects of some variable *should not appear*. However, we again must emphasize that the decisions researchers make with respect to which measures to analyze or make predictions about with respect to the effect of a particular variable must be grounded in the prior literature – not only studies that have investigated the manipulated variable, but also studies that have focused on explaining what about the reading process a given dependent variable reflects.

Charting out the variables that influence the movement of the eyes over text is a key part of studying basic reading processes. However, researchers also use

[2] In addition, there is a *blink* event, when the eyelids rapidly close and open. We cover blinks and how to deal with them during data processing in Section "Artifact rejection procedures".

eye-tracking-while-reading to study the cognitive mechanisms that underpin these processes. For example, eye tracking research has featured prominently in debates over the cognitive mechanisms that provide sensitivity to contextual predictability (e.g., Staub, 2015), the interplay between syntactic and discourse-level influences on linguistic processing (e.g., Clifton et al., 2007), the nature of the cognitive mechanisms that handle ambiguity resolution during incremental reading (e.g., Clifton & Staub, 2008), and how these processes relate to other areas of cognitive processing (e.g., Gagl et al., 2018).

Linking eye movements to underlying cognitive processes requires linking hypotheses: specifications of how cognitive processes of interest are reflected in measures of eye movements. These models are critically important for two reasons. First, they provide the theoretical basis for understanding basic reading processes, which is of course an important theoretical goal in its own right. Second, they serve to help guide researchers using reading measures to study higher-order linguistic processing. As we detail below, the space of possible analytical decisions that face a researcher using eye-tracking-while-reading is massive. These theoretical frameworks are a critical tool in helping researchers decide key analytical questions such as: Which word (or words) should I expect to see an effect of my experimental manipulation? Which measures do I expect to see this effect show up in?

There are a number of integrative models of eye movements during reading that provide a basis for such linking hypotheses. However, these models constitute theories of the architecture of reading that remain under active investigation, and there is not yet consensus on all aspects of this model. For example, the widely used E-Z Reader model proposes that reading is mediated by a serial attention process that can recognize one word a time, eye movements are initiated by the *partial* completion of the word recognition process, and attention is guided by a number of physical and linguistic features of the text (Reichle et al., 1998; Rayner et al., 2004; Reichle et al., 2009a, 2009b; see also EMMA: Salvucci, 2001). In contrast, the SWIFT model (Engbert et al., 2005; see also OB-1 Reader; Snell et al., 2018) proposes that attention is best understood as a gradient resource, such that multiple words can be concurrently processed in a single fixation; instead of eye movements being initiated by lexical processing as in E-Z Reader, they are initiated by a random timer with inhibition from foveal processing difficulty. Even though serial models like E-Z Reader (Reichle et al., 1998) posit that a given word cannot be lexically identified before the word preceding it, that does not mean that readers cannot attend to words prior to fixating them. Rather, because all eye movements, including skips, are initiated part-way through word identification, if the fixation on the current word is not yet terminated by the start of the subsequent saccade by the time that lexical identification completes, the reader can shift attention to the upcoming word and begin processing it prior to fixating it. Therefore, although many researchers and published papers characterize these models as making clear and differentiated predictions about reading behavior, it is often difficult to find points where they make clearly divergent predictions despite the differences in their underlying architectures.

Models like E-Z Reader, EMMA, SWIFT, and OB-1 focus on the nature of the attentional mechanisms that drive reading in text. However, it is equally important to integrate these models with cognitive architectures that specify the underlying linguistic processes that drive attention, and this remains an active area of research (e.g., Bicknell & Levy, 2010a, b, 2012; Engelmann et al., 2013; Rabe et al., 2024; Vasishth & Engelmann, 2021). For example, although regression behavior is thought of as an index of how much time is necessary to resolve processing difficulty (Clifton et al., 2007), there are presumably many underlying reasons for regressive movements. They may be triggered by confusion about the text and used to reinspect a misidentified word or reinterpret a misparsed structure (see Bicknell & Levy, 2011; Booth & Weger, 2013; Frazier & Rayner, 1987; Pollatsek et al., 2006; Schotter et al., 2014b). According to one view, regressive eye movements are engaged in the service of obtaining more information that can aid with integration. For example, the eyes may be directed to reinspect portions of the input that need to be reanalyzed (i.e., *selective reanalysis;* Frazier & Rayner, 1982), or the readers may wish to resample previous input to increase confidence in their analysis of that input (Bicknell & Levy, 2011; Levy et al., 2009). Alternatively, regressions may simply delay the forward progression of the eyes in the text in order to buy readers time to resolve processing difficulty through other processes (e.g., internal reanalysis that does not depend on the re-intake of perceptual information; Mitchell et al., 2008). In situations where there is not a syntactic difficulty manipulation, but rather the researchers are comparing reading of longer passages, regressions may indicate that the reader is consolidating a memory representation of the text (Ariasi et al., 2017; Hyönä & Nurminen, 2006; Hyönä et al., 2002; Kaakinen et al., 2003), among other possible functions. These alternatives are not mutually exclusive, and on either view, this measure is particularly useful in studies that investigate linguistic integration because regressive saccades increase when integration had failed (Bicknell & Levy, 2011; Rayner et al., 2004). Depending on the time course of full word recognition and integration, regression behavior for the post-target area of interest may be more informative than measures on the target word itself. Such spillover effects could occur if post-lexical integration processes lag behind the lexical processes necessary to move the eyes forward (Reichle et al., 2009a, 2009b).

Now that we have covered the core findings and theoretical models from the literature, we turn to a more nuts-and-bolts discussion of the practical aspects of studying reading using eye tracking. The following sections are presented in the "order of operations" of a typical research endeavor (e.g., designing an experiment, processing data, defining dependent variables, performing statistical analysis), but it is important to emphasize that researchers should think deeply and deliberately about all steps of the process before starting their research study. In other words, it is a bad idea to program an experiment and start collecting data without an idea of how the data will be processed and how the dependent variables will be calculated; otherwise, the researcher may risk spending months collecting data only to realize that they are completely useless. For this reason, it is valuable for researchers to pre-register their experiment prior to collecting (or analyzing) the data. Not only does pre-registration help safeguard against the publication of false positive results (see Frank et al., 2024), but it can also help the researcher think through all the choice points discussed in this article before starting, and therefore anticipate – and resolve – potential issues in the design of the study, and processing and analysis of the data before they arise. From this perspective, pre-registration is not a handcuff that restricts the options available to the researcher but rather a framework that limits those options and makes the post hoc choices to deviate from that plan potentially more justifiable (so long as the researcher is transparent about their decision-making along the way).

## Experimental design

Experimental design is a key part of collecting high-quality eye-tracking-while-reading data. The typical eye-tracking-while-reading experiment makes use of many familiar concepts in experimental psychology, such as randomized stimulus presentation, blocked designs, large numbers of fillers to mask critical experimental manipulations, and so on. An exhaustive introduction to experimental design is beyond the scope of this paper, for more general considerations about experimental design we point the reader to Frank et al. (2024). In this review, we will limit our discussion primarily to those aspects of stimulus and experimental design where the specific features of eye-tracking data inform this stimulus and design process.

## Statistical power, number of participants, and items

Standard practices with respect to the typical number of items and participants varies across different sub-fields of reading research. In general, the more observations per condition per subject, the better. However, it is also important to keep experiments to a reasonable length for several reasons. The chief reason is participant fatigue: Any benefit of collecting additional data may be undermined by the drop in data quality associated with fatigue for longer experiments. Especially with challenging or unique sentence structures, an experiment that lasts too long will cause participants to eventually 'zone out', which can lead to a drop in data quality. A secondary challenge for longer experiments is maintaining sufficiently precise tracker calibration (see Section "Preparing the eye tracker to collect data"). A participant's calibration can degrade over the course of the experiment, leading to greater noise in experimental measurements. Unlike fatigue, this challenge can be overcome by regular calibration checks and re-calibrating the eye tracker to the participant when necessary.

Studies of sentence-level effects require a lot of homogeneity in the sentence structures that make up the experimental conditions, and those sentence structures can often be quite marked or challenging to process. Therefore, researchers in the field of sentence parsing avoid the experimental stimuli comprising a large proportion of trials and instead have a large number of fillers to obscure the key experimental manipulations. This means that in practice there may not be a large number of observations per condition in this sub-field. Even so, it is highly unusual to collect fewer than six observations per condition per subject. Furthermore, any fewer than this number is likely to be insufficient to allow the researcher to draw reasonable inferences about the size of the effects under investigation.

In contrast, studies of lexical processing, word identification, or parafoveal preview may have small effect sizes (e.g., on the order of 5–25 ms) and the manipulation may be more subtle than sentence-level studies. Therefore, in this sub-field there is an increasing push among reviewers and practice among researchers to have a large number of items per condition; it is not unusual to see on the order of 30–60 observations per condition per participant. Based on a review of power in mixed effects designs (i.e., designs in which there are repeated measures over both participants and items), a general recommendation is that a properly powered reaction time experiment (and note that fixation durations are technically reaction times) should have at least 1600 observations (40 participants, 40 stimuli) per condition (Brysbaert & Stevens, 2018).

With these considerations in mind, it is important to consider the value of an a priori power analysis, especially considering that most journals require authors to report how statistical power was determined. With a certain amount of information (e.g., how many stimuli *could be created*, how many participants *could be recruited* for this study, and how large of an effect size is reasonable to assume), a priori power analyses can give the researcher a pretty decent idea of what kind of experimental design they will need to adequately test their research question. Westfall

et al. (2014) provide guidance on planning experiments for mixed effects designs, and have created a web interface (https://jakewestfall.shinyapps.io/pangea/) for performing a simple power analysis by entering only four details: (1) the design to be used, (2) the anticipated effect size or the mean condition difference, (3) the anticipated numbers of participants and stimuli, and (4) estimates of the relevant variance components (which they note is the most difficult to define, but they provide reasonable defaults). One of the nice features of this approach is that it can be used to perform a "sensitivity analysis" (Rosenbaum, 2020) – essentially by knowing how many participants and stimuli are reasonable within the researcher's constraints they can fix those values and then make adjustments to the effect size value until power is equal to 0.8 (a generally acceptable level of power; Cohen, 1988). The resulting effect size would then represent the minimal effect size that could be detected by the given experimental design with acceptable power. An alternative approach, described by Kumle et al. (2021), uses simulations to estimate power for mixed effects models in different use cases by simulating a dataset multiple times, performing the statistical test on the simulated data, and then calculating the proportion of statistically significant results out of all results. As the authors note, the accuracy and strength of this approach depends on the values chosen to create the simulated datasets, which will obviously be better with more pre-existing data that is analogous to the design of the intended experiment. However, they do provide methods to still perform simulation-based power analysis even when published data are and are not available (see also Green and MacLeod (2016) for a tutorial on an R-based package for performing these simulations).

## Stimulus design

For psycholinguistic studies of single-line sentences, the unit of analysis is typically a single word or phrase (see Section "Typical dependent variables"). Because the relevant pieces of text could technically be located anywhere on the screen, the researcher must define an *area of interest* (*AOI*) that marks the location of that word or phrase on the screen. To make the concept of an AOI more concrete, consider the sentences in 1a, 1b, 2a and 2b, taken from Ashby et al., (2005), who were interested in whether a constraining sentence context could make up for the difficulty in recognizing low-frequency words (among other variables, such as reader skill).

(1) a. The sailor stopped at the deserted| island| for a week.
b. The sailor stopped at the deserted| casino| for a week.
(2) a. The gambler visited the| casino| as part of his vacation.
b. The gambler visited the| island| as part of his vacation.

In this example, each sentence is marked by an AOI that varies by the same word pair (i.e., island vs. casino, which vary in the frequency of occurrence: island is more common). In one version of the sentence, the higher-frequency word is more expected whereas in the other version of the sentence the lower-frequency word is more expected. Because the contexts differ, the use of the AOI allows the researchers to analyze the same words, regardless of their location in the sentence or position on the computer screen (see Section "Local measures"). Furthermore, because eye-tracking-while-reading experiments use repeated measures (i.e., each participant reads multiple stimuli in each experimental condition, the use of an AOI allows researchers to identify the location of the critical word even when it appears in different locations across different items or across different conditions for the same item.

A critical AOI may span more than a single word, and the analysis strategy will still follow a similar logic. Consider the sentences in 3a and 3b, taken from Frazier and Rayner (1982), who were interested in how readers analyzed the ambiguous noun phrase "a mile and a half," and part of their investigation of this question involved measuring eye movements on the disambiguating AOI. In their experiment, they defined the disambiguating AOI as the two words following the ambiguous noun phrase. Although this created heterogeneous AOIs of analysis across conditions, in both 3a and 3b this critical AOI serves the same functional role (i.e., it is the point at which the syntactic structure of the sentence is disambiguated).

(1)a. Since Jay always jogs| a mile and a half| this seems| like a short distance to him.
b. Since Jay always jogs| a mile and a half| seems like| a short distance to him.

One technical detail that is worth addressing is the question of how to treat the spaces around the target words when defining AOIs. In the examples above, the space preceding the word(s) has been included in the AOI but the space after has not been. This tends to be the approach taken based on the assumption that if a reader fixates on a blank space (which happens approximately 15% of the time; McConkie et al., 1988), it is more likely that their attention is on the upcoming word rather than the preceding word and therefore the gaze location should be allocated to the word that they are (presumably) attending.

Like many fields of experimental psychology, it is common in reading studies to use cross-factorial experimental designs. In general, eye tracking measures can be fairly noisy, and vary substantially individual to individual: See Staub (2021) for a discussion of the reliability of various eye tracking measures and variability across individuals.

Accordingly, it is generally ideal to employ within-participant experimental manipulations where possible, varying experimental factors within a single individual's testing session, if possible.

It is also standard to adopt within-items designs in eye-tracking-while-reading experiments. Within-items designs are analogous to within-participants designs, except that the sentence is the relevant level of experimental grouping, rather than (or in addition to) the participant. For example, a within-items design implementing a frequency manipulation would create two versions of an experimental sentence: one including the low-frequency word as the target and one including the high-frequency word. The two versions of the same sentence are then assigned to two distinct experimental lists, and any individual participant in the experiment will only read one of these lists. When this method of counterbalancing is applied to an entire stimulus set, subject to the constraint that every list has the exact same number of observations per condition, and that each item set only occurs once on each list, the resulting counterbalancing scheme is known as a *Latin Square* (Reese, 1997). If the same number of participants are run in each list of the experiment generated in this way, then the Latin Square counterbalancing scheme ensures that the actual stimulus does not systematically vary across conditions. Note, however, that adopting this type of within-item and within-participant design is something that must be incorporated into the statistical analysis of an experiment (e.g., via (generalized) linear mixed effects models; Baayen et al., 2008; Barr et al., 2013; see Section "Data Analysis").

In general, it is good practice to make sure any AOIs are not at the beginning or end of a sentence or passage of text because there is some disruption to the typical flow of reading at the beginning or end of a trial (Kuperman et al., 2010). Similarly, unless it is the focus of investigation, researchers should avoid putting the AOIs at the start or end of a line in the case of multi-line text experiments to avoid issues associated with return sweeps or the lack of parafoveal preview across line breaks (see Section "Special considerations for multi-line text studies"). It is also important to make sure AOIs are far away from the end of the sentence or major punctuation, as these are often the sites of so-called 'wrap-up' effects. In eye tracking, this can result in an increase in regressions and re-reading at the end of a sentence or clause (Andrews & Veldre, 2021). In practice, similar effects extend to major punctuation boundaries too, since reading times are often slowed leading up to, and speeded up after major punctuation breaks such as commas (Hirotani et al., 2006).

### Stimulus norming

Because the stimuli are used to implement the manipulation in eye-tracking-while-reading experiments but are also a source of random variance (Clark, 1973; Coleman, 1964; see Jannsen, 2012), the researcher should be careful to quantify the degree to which the individual-level and condition-level stimuli represent the manipulation (i.e., are sufficiently manipulated to differ on the investigated variable) and do not differ on other variables that are not of interest (i.e., are sufficiently controlled on other physical, lexical, and contextual variables). This is because many visual and linguistic variables have a large impact on the duration and frequency of fixations (see Section "Linguistic influences on reading behavior"). There are two general approaches to manipulating stimuli: manipulation of the context (while controlling the target words) and manipulation of the target words (while controlling the context). Each of these requires different approaches to stimulus control, which we describe separately below.

**Manipulation of the context** An advantage to manipulating the context is that the manipulated features of the trial are distal to the analyzed AOI, which is identical across conditions. For example, when investigating the effects of plausibility (e.g., the degree to which two different words make sense in a sentence), the researchers may choose to embed the same target word (e.g., journal) in two different contexts (e.g., plausible: "The man noticed the journal was missing from his desk." and implausible: "The man angered the journal by placing it in the drawer."; Abbott & Staub, 2015). Because the manipulation (i.e., plausibility of (the word in) the sentence) is a subjective evaluation, it is important to conduct a "norming study" prior to the eye tracking experiment to make sure that the plausible and implausible versions are judged – by a unique set of participants from the same population – to seem plausible and implausible, respectively. For example, participants might see one version of the sentence or the other (counterbalanced across lists, much like the Latin square design used for the eye tracking study) and then rate on a Likert scale how plausible/sensible/natural/acceptable the sentence is. Another commonly conducted norming study is a 'cloze task' (Taylor, 1953), in which participants see a fragment of a sentence and enter a word (or words) that could follow or complete it. The cloze task is commonly used to derive information about the predictability of a word in a context – the proportion of responses that are a given word out of the total number of responses represents the "cloze probability."

Data from a norming study can then be used to screen out – or rewrite – stimuli that do not meet some criterion. For example, the researcher might require that a stimulus must be rated above the halfway point on the scale to be considered "plausible" and below the halfway point to be considered "implausible," or have a cloze probability of about 0.7 to be considered "predictable" and below 0.1 to

be considered "unpredictable." However, the specific criteria may vary from study to study depending on the research question and goals.

In addition to using the norming data to finalize the stimulus set prior to running the study, data from the norming study should be reported in the manuscript to give the readers a sense of the strength of the manipulation and control of the stimuli. For example, this may include the means and standard deviations, aggregated at the condition level, for all the norming measures collected from the set of stimuli that were ultimately presented in the eye tracking experiment. The researcher may also report the inclusion criteria (i.e., criterion values) used to determine whether a stimulus would be included in the experiment.

**Manipulation of the target word** If the experimental design demands that the target words be different, it is important to control the stimuli to ensure that they do not differ on dimensions that are not central to the research question. For example, if not the variables of interest, at a minimum, the researcher should ensure that the target words do not differ on the "big three" (i.e., word length, frequency, and predictability in the context; Kliegl et al., 2006; Rayner & Liversedge, 2011), but note that these variables do not exhaust the list of relevant variables that could be controlled for. One helpful source that can be used to extract these (and other) characteristics for words is the English Lexicon Project (ELP; elexicon.wustl.edu; Balota et al., 2007). In general, the variables that must be controlled for must be considered on a study-by-study basis. For instance, controlling character bigram frequency across conditions may or may not be desirable for a given study, depending on whether this variable can reasonably be expected to confound the critical tests. Here, as in other places, existing literature looking at similar questions will provide a reliable guidepost to what variables are worthy of consideration.

As with norming data from the context manipulations, the researcher would use these data for both screening the stimuli and determining the final stimulus set that is used in the experiment, and also for reporting descriptive statistics of the final stimulus set in the manuscript.

## Instructions to participants

Relevant to the question of how to structure an experiment is how a participant's task impacts eye movement behavior. The task a participant is asked to do, such as answering comprehension questions, is commonly called a secondary task – presumably this is because, from the point of view of the reading researcher, the act of reading itself is primary. Previous work has suggested that the difficulty of secondary comprehension questions can influence reading behavior, but in complex ways. If a reader is minimally engaged in linguistic processing, their eyes move through the text faster (Duggan & Payne, 2009; Strukelj & Niehorster, 2018) and show smaller magnitude effects of the lexical variables described above; for example, word frequency effects are diminished when readers 'zone out' (Reichle et al., 2010) and when they skim the text for a gist (White et al., 2015). In contrast, when a reader scrutinizes the text more deeply, their eyes move through the text more slowly (Schotter et al., 2014a; Strukelj & Niehorster, 2018) and show larger magnitude effects of lexical variables; for example, word frequency effects are magnified when readers proofread for spelling errors (Kaakinen & Hyönä, 2010; Schotter et al., 2014a) and predictability effects are magnified if the proofreading task requires the readers to use the sentence context to detect an error (Schotter et al., 2014a).[3] Even when the task is to read for comprehension, the nature of the secondary task can change reading behavior. Wotschack and Kliegl (2013) and Weiss et al. (2018) suggest that difficult comprehension questions can significantly alter the amount of regressive re-reading that participants will engage in, but have relatively modest (if any) effects on first-pass reading. Andrews and Veldre (2021) suggest that task-dependent variation in how participants approach 'wrap up' effects may be partially to blame. The effects of other secondary tasks are more mixed: Zhang et al. (2018) showed that listening to music during reading led to significant re-reading, and Mertzen et al. (2023) showed that a secondary memory load task sped up German and English readers even in first-pass processing measures. Overall, it's important to pay close attention to the secondary task used, as this can significantly impact especially later measures. However, there is evidence that secondary tasks have a more modest impact on early measures, if a reasonable amount of engagement with the experimental materials can be assured.

When manipulating the secondary task, it is important to consider "blocking" trials so that the participant does not have to shift their mental framework from one trial to the next. Therefore, all trials with one secondary task would be completed before all trials from another secondary task. With this design, however, order effects are baked into the comparisons of the task, and it is therefore difficult to tease apart task effects from effects of, for example, fatigue across

[3] Not only can the secondary task have an impact on reading behavior within those trials, the task may carry over to future trials if the readers cannot "turn off" that task goal. For this reason, it would be important to block the task manipulation so that the participant engages in all the trials with one set of task instructions before engaging in another task. Furthermore, the researcher should consider which task is more "natural" or otherwise less likely to carry over to future blocks and may want to have the participants perform that task first (or they may consider counterbalancing the order of the blocks if comparisons of carry-over would be theoretically informative).

the experiment. One option is to counterbalance, across participants, which block (i.e., task) comes first and which comes second. However, it may not always make sense to counterbalance the order of the tasks if one of them is less natural or might carry over to the second one. For example, if readers are asked to look for spelling errors (i.e., proofread) before reading for comprehension, they may (implicitly) continue to look for errors in the second block even though that is not the explicit task.

It is also important to point out that the secondary task can, itself, provide dependent measures that can illuminate the end state of language comprehension. For example, while perhaps tempting to assume that readers generally successfully interpret material they read, this assumption isn't justified, as readers systematically fall prey to a range of misinterpretations and interpretive illusions in reading text (e.g., Ferreira & Yang, 2019). Exploring the relationship between reading behavior and comprehension success has yielded important insights into language processing. For example, it has revealed that regressive eye movement behavior is related to improved comprehension accuracy when readers are able to reinspect the text (e.g., Schotter et al., 2014b). In a similar vein, Huang and Staub (2021) survey a range of results on the relationship between reading behavior and participants' success in noticing various types of errors when presented with an explicit error detection task.

## Preparing the eye tracker to collect data

Before collecting any eye tracking data, a critical first step is to conduct a calibration and validation of the eye tracker. The purpose of this is to ensure that the eye tracker has accurate measurements for that particular equipment setup and that particular participant. Every person has a slightly different posture, head size, eye shape, etc. and all of these variables affect the measurements that the eye tracker uses to infer gaze position (i.e., where the participant is looking). Therefore, calibration is necessary at the beginning of an experimental session (i.e., when a new participant arrives) and if the same participant moves significantly during the experiment (e.g., if they get up out of the chair and walk around). Also, sometimes participants can shift or slouch in the chair (or eye tracker chin/headrest) so much that recalibration is necessary even if the participant did not completely leave their seat.

The most popular eye trackers for eye-tracking-while-reading-research use measurements of reflections of light off of the eye to map the reader's gaze to a particular point on the screen (for more details on how eye trackers work, see Duchowski, 2007; Holmqvist et al., 2011). Different hardware may take slightly different approaches, but in essence, they shine (infrared, i.e., invisible) light onto the eye and then capture images of the reflections of that light back to a high-speed video camera. The particular measurements of interest are the location of the *pupil* (i.e., no reflection because it is a hole in the center of the iris where light enters the eye and typically does not reflect back out) and in some trackers also the location of the *corneal reflection* (i.e., an intense reflection of light because the cornea is a clear shiny curved surface and there will be some point where the angle is perfect to return most of the light coming from the emitting source). As the eye rotates to look at different locations, the location of the pupil shifts within the camera view (or the corneal reflection shifts relative to the pupil) and it is these movements that the eye tracker is capturing. However, these movements are in an arbitrary coordinate space and the calibration procedure is necessary to allow the eye tracker to map them to locations on the screen by displaying a target stimulus (usually a dot or bullseye) in a known location and measuring the movements when the participant looks there. The calibration procedure involves placing a number of targets in different locations and then the eye tracker uses an algorithm to infer the mapping of any other reflections to any other points on the screen.

After performing a calibration, is it necessary to perform a validation, which feels identical to the calibration procedure in the participant's experience, but is importantly different on the eye tracker side. During validation, the target stimulus locations are slightly offset from the locations during calibration so that the eye tracker can make a prediction about what the pupil (and corneal reflection) locations should be and then compares that to the measurements that are taken when the participant looks there. The deviation of the prediction and the observed value gives the researcher a precise idea of the measurement error, which is an important component of data quality. During the experiment, it is beneficial to incorporate frequent "drift checks" – moments where a single target stimulus with a known location is displayed on the screen and the eye tracker compares its prediction of the measurements to the observed measurements. This serves as a validation and opportunity for the experimenter to check on calibration accuracy but does not disrupt the flow of the experiment as much as a full calibration/validation procedure.

It is up to the experimenter to determine what is an acceptable amount of calibration error and set an a priori criterion. If the error is higher than that threshold, the experimenter should redo the calibration and validation procedure (making adjustments and coaching the participant) as necessary until the error is below threshold or it is determined that the participant cannot be in the experiment because accurate calibration is not possible. The exact value of acceptable error may vary from study to study, but it is important to consider the precision that is necessary for the research question. Because many studies focus on reading behaviors around single words, it is important to be able to distinguish

whether a fixation is on one word or another that is immediately next to it. In spaced languages (e.g., English) a single character space separates two adjacent words and therefore it is preferable to have a calibration error less than the space that is subtended by a single character.

## Special considerations for gaze-contingent studies

As mentioned at the outset, the reason why eye tracking provides such a good measure of reading processes is that readers must move their eyes to a word in a text in order to process it most efficiently. However, readers can – and do – obtain information from words prior to looking at them (i.e., during *parafoveal preview*; see Schotter et al., 2012) and the nature of this processing has been an area of interest for decades. In the 1970s, Keith Rayner and George McConkie (as well as other collaborators of theirs) developed a number of *gaze-contingent paradigms*, in which the content of the text displayed on the screen is manipulated based on the location of the reader's gaze position (see Clifton et al., 2016; Rayner, 2014; Schotter et al., 2012). These paradigms allowed them to study how readers obtain information about the text from parafoveal vision and how they use that information in the reading process. The two most prominent of these approaches is the *moving window/mask paradigm* (McConkie & Rayner, 1975; Rayner & Bertera, 1979), in which a change to the text is implemented during every saccade on the trial, and the *boundary paradigm* (Rayner, 1975), in which the change to the text is implemented when the reader's gaze moves toward or past a particular word.

The processes of conducting and cleaning data from these types of studies are a bit more complex than the typical eye-tracking-while-reading studies because they require real-time information from the eye tracker about the reader's gaze position and therefore require extremely high temporal and spatial precision. These experiments also require the researcher to program into the experiment information about the "trigger(s)" for the display change(s) and which information should be displayed where and at what time. In addition to the general guidelines for processing data from eye tracking experiments (see Section "Artifact rejection procedures"), an important additional task is to identify problematic display changes in pre-processing and exclude fixations, saccades, or trials when they have occurred. A problematic display change is one in which the update to the content of the screen did not happen within a few milliseconds of the eye crossing the trigger location (generally by the time the next fixation starts or up to 5 ms after; Slattery et al., 2011). Sometimes, this happens because of a "j-hook" saccade – one in which the measured location of the reader's gaze temporarily crosses the boundary (triggering the display change) but then ends before the boundary. In this scenario, the parafoveal preview manipulation is completely invalidated because the fixation on the pre-target word no longer implements the preview manipulation (i.e., because the target word – rather than the preview word – is already being displayed).

Another important point to make about these studies is the need to use a mono-spaced (i.e., fixed-width) font – a font in which the horizontal space subtended by each letter is the same. This is because replacing one letter with another could potentially change the position of subsequent letters and words in the text if, for example, a narrow letter like i is replaced with a wide letter like m. In a monospaced font, because an i and an m take up the same amount of space, replacing one with the other would not affect where on the screen any of the other letters would appear.

### Moving window/mask paradigm studies

In the moving window paradigm (McConkie & Rayner, 1975; see Rayner, 1998, 2014; Schotter et al., 2024), only the text immediately around the reader's fixation location is available and this visible window moves instantaneously with the reader's fixation location. Outside of the visible window, the text is replaced by a mask so that the reader can only extract useful information from the text within their foveal (central) vision and whatever parafoveal regions fall within the visible window. On different trials, the size of the window is varied to allow for different amounts of parafoveal information to be visible in order to determine how wide of an area of the text the reader is using (i.e., the size of their perceptual/attentional/word identification/reading span – depending on the manipulation, see Schotter et al., 2024). The reader's eye movement behavior across the trials (see Section "Global reading measures") is compared between window sizes, and the span size is indicated by the smallest window in which reading does not significantly differ from normal reading (i.e., without a window), or the largest window that shows a significant improvement from the next smallest window. The moving mask paradigm (Rayner & Bertera, 1979) is essentially the inverse of the moving window in that the text at and immediately surrounding fixation is masked and the more eccentric areas of the text are visible. In this way, readers must rely exclusively on parafoveal vision to read because the foveal information is masked.

For these studies, because the manipulation is so noticeable and – particularly for small window or large mask conditions – can be disruptive to the reading process, researchers may want to consider blocking the window/mask conditions so that the reader can acclimate to reading in that scenario. Of course, that is guided by the research question and some studies have intermixed rather than blocked the trials (although not all studies report the method in a way that makes clear whether the conditions were blocked or mixed). Another consideration is the type of window manipulation

to implement, as this varies considerably across studies. For example, in some studies, the window is symmetrical (i.e., includes the same number of visible characters to the left and right of fixation) and in others it is asymmetrical (i.e., manipulates the number of visible characters to the right while keeping the number of visible characters to the left of fixation constant across conditions). The motivation to use an asymmetrical window comes from studies showing that the span only extends about 3–4 characters to the left of fixation, so there might be no need to provide more visible characters. However, the size of the leftward span can be larger for some readers (see Schotter et al., 2024; Stringer et al., 2024), so this choice should be made in the context of the research question and the population under study. Additionally, researchers should consider whether the mask covers the letters, the spaces, or both (see Schotter et al., 2024), and what type of mask is used (i.e., x's, random letters, blurred text, etc.).

#### Boundary paradigm studies

In the boundary paradigm (Rayner, 1975; see Schotter et al., 2012), rather than implementing a manipulation across the entire sentence, the research question focuses on processing of a specific target word. An invisible boundary is specified (usually located at the end of the word before the target word) and the gaze location of the reader is compared with that boundary location to determine whether the target word is displayed (i.e., once the reader crossed the boundary) or whether a preview stimulus is displayed (i.e., before they crossed the boundary). When the reader's eyes cross the boundary location, the preview word instantaneously changes to the target word. The change is generally not noticed by the reader because it occurs during a saccade when vision is effectively suppressed (Matin, 1974). The preview is manipulated to test whether information that was only available during parafoveal preview has any effect on reading behavior on the target word (see Section "Local measures").

In boundary paradigm studies, the choice of preview conditions and comparisons is not theory-neutral. For a long time, comparisons between conditions focused on questions of the nature of *preview benefit effects* – shorter fixation times on a target when the preview was related compared with unrelated to it (see Schotter et al., 2012), or on the flip side whether there are *preview cost effects* for unrelated previews compared to related previews (Kliegl et al., 2013; Marx et al., 2015). These studies hinged on creating manipulations of the preview stimulus that varied in the degree of overlap between the preview and target with respect to different types of information (e.g., visual, orthographic, phonological, morphological, semantic, etc.). However, more recently, some studies have revealed that there are effects of the nature of the parafoveal preview on reading times on the target word that are completely separate from the degree of relationship between the preview and target and rather are related to the degree to which the preview is an interpretable stimulus (e.g., whether it is a word/nonword, or a high or low-frequency word; e.g., Schotter & Leinenger, 2016), or whether it makes sense in relation to the preceding sentence context (Schotter & Jia, 2016; Veldre & Andrews, 2016; see Andrews & Veldre, 2019; Schotter, 2018). Therefore, interpreting data in these studies may not be as simple as assuming that preview benefit/cost effects are the result of *trans-saccadic integration* (see Cutter et al., 2015; Pollatsek et al., 1992; Rayner, 1975), but these latter findings suggest that there may also be *direct preview* effects on eye movement behavior on the target word (see Schotter, 2018).

With respect to the treatment of the data for these studies, because a primary motivation is to investigate parafoveal processing of a word before it is fixated (or skipped) it is important to filter the data during pre-processing so that only trials on which the pretarget word was fixated are included in the analysis. Additionally, researchers may want to measure the duration of the pre-target fixation or its proximity to the target word, as these variables may impact the amount of parafoveal preview of the target word the reader obtains during that fixation (Kliegl et al., 2013; Slattery et al., 2011).

### Special considerations for multi-line text studies

As we noted above, a primary focus of this paper is on best practices for studies that present single sentences on a single line of text. However, this drastically limits the inferences researchers can make about language comprehension and reading, and there are questions as to how much these phenomena generalize to more naturalistic scenarios in which the words and sentences are embedded in larger discourse contexts. To some extent, single-line studies have dominated the field for practical reasons. First, vertical eye tracking is less precise and therefore when multiple lines of text are present it can be difficult to determine which line a given fixation should be allocated to. Some fixations during multi-line text reading may not resemble "typical" fixations investigated in eye-tracking-while-reading studies, particularly those that occur at the end or beginning of a line of text and that do not coincide with the beginning or end of a phrase, sentence, or paragraph. We address each of these below.

#### Correcting vertical drift

One common practice to alleviate the issue of vertical tracking imprecision is to insert a large amount of white space

between one line of text and the lines above and below it to avoid fixations falling ambiguously close to two lines. However, even with this additional drift in the calibration over the trial, it can lead to a pattern of fixations across space that does not accurately reflect the text at which the participant is looking. Therefore, "correction approaches" have been an area of interest for researchers for the past decade or so, and there are now a number of algorithmic approaches to correction that work rather successfully. Carr et al. (2022) compared ten algorithmic methods for dealing with vertical drift using both simulated and natural eye tracking data and concluded that some of the algorithms worked quite well, but the success was dependent on the characteristics of an individual trial. Therefore, there is still a role for the researcher in determining which algorithm is most appropriate for a given dataset and it is important to test how well an algorithm works before using it to manipulate the raw data prior to moving on to subsequent data processing steps.

#### Considering fixations around return sweeps

Fixation behavior surrounding return-sweep saccades, which only occur during multi-line text reading, may be qualitatively different from intra-line fixations. For this reason, unless return sweep saccades are the focus of the investigation, researchers should ensure that their target AOI is not at the end or beginning of a line of text. Fixations that precede return-sweeps (i.e., *line-final fixations*; typically located 5–7 characters from the end of a line) are shorter than intra-line fixations (see Hofmeister et al., 1999; Parker et al., 2023), possibly indicating processing of line breaks rather than (as deep) lexical processing (Kuperman et al., 2010). There is increased error in saccade targeting during return sweep saccades, and the fixation following the return sweep may indicate reduced processing because it is shorter than a typical fixation, approximately 120–160 ms, and is followed by a corrective saccade approximately 40–60% of the time (see Vasilev et al., 2021). *Accurate return-sweeps* (i.e., those that land close to their intended target) are followed by forward saccades whereby readers continue reading through the new line, whereas *under-sweeps* (i.e., those that land short of their target) are followed by a corrective backward saccade (i.e., regression) prior to progressing forward. The durations of accurate return-sweep fixations are longer than intra-line fixations, presumably because of a lack of parafoveal preview, whereas the durations of under-sweep fixations are shorter than intra-line fixations (Parker et al., 2023).

## Typical processing pipeline

Once the raw eye tracking data are collected, there are many steps that researchers must make when *processing* the data (i.e., converting the raw eye movement data into one or more dependent measures that index the processes of interest; see Section "Typical dependent variables") to produce an analysis-ready dataset, and the number of decisions that face the researcher can be daunting. In practice, eye tracking researchers navigate this 'multiverse' of possible datasets through a mixture of convention and careful reasoning about the dependent variables they are constructing. Here, we describe the key steps in this processing pipeline with the aim of making it easier to reason deliberately about how different processing choices may or may not impact the final results.

### Raw data

The most raw output of an eye tracking system consists of a stream of "samples" corresponding to a timestamp of when the gaze location was measured and the corresponding *x*- and *y*- coordinates of the gaze location at that sample.[4] However, these data can be summarized into the fixation events mentioned above, which is the most common use of raw data from an eye tracking experiment of reading: a record of all the fixations made on a given passage of text, their durations, locations, and the order in which they occurred. With this information, the saccades made between those fixations can be inferred, but it is important to note that even this record represents only a subset of the "events" that may occur during an experimental session: there are also blinks, track loss, experimentally triggered display events, participant response events, and so on.

Raw data output may vary depending on the particular hardware used, but a commonly used eye tracking system, the EyeLink (1000plus, duo; SR Research), produces a binary file that is created in real time during the experimental session (i.e., an eye tracking data file;.EDF) that can be read in by proprietary Data Viewer software, or converted into a readable ascii file with a simple EDF2ASC conversion software. When programmed properly, the.ASC file should contain information about the experimental session (e.g., start time, calibration sequence and error, etc.), each individual trial (e.g., trial identifier, display contents, etc.), and the sequence of events (e.g., each fixation start and end timestamp and x and y location) and sometimes the samples that occurred during each trial (e.g., the *x* and *y* location

[4] Some eye trackers (e.g., the EyeLink systems) provide information about pupil dilation (i.e., diameter within the camera image), which is another type of eye tracking data researchers may be interested in, particularly as an index of listening effort for auditory language processing. The theoretical considerations and practicalities of processing and analyzing these data are sufficiently different from what we focus on here that we will not discuss this issue further. For a tutorial on the best practices for using pupillometry to study listening effort see Winn et al. (2018).

and timestamp of every sample of gaze position), depending on the settings selected on the tracker. As mentioned above, the primary data of interest for reading experiments are the sequence, location, and duration of fixation events. Thus, the full record of the data may be reduced to a format that represents each trial of the experiment in a row, which contains identifying information about that trial (e.g., item identifier, condition identifier, etc.) and then a series of cells of data that represent each dependent measure of interest, either at the trial level (i.e., global measures), or at the word or word-group level (i.e., local measures) with the aid of a file that specifies the locations of the AOIs.

Of course, there are other eye tracking hardware, and researchers should use the equipment available to them. However, when choosing eye tracking hardware, the researcher should ensure that it has a very high spatial and temporal resolution, low recording latency, small measurement error, and that it provides clear documentation for the hardware and software, including information about the settings and parameters, including those used for the saccade detection algorithm. This is particularly important if the effects of interest are small in magnitude as the need for high precision becomes even more important.

## Artifact rejection procedures

In the process of creating a fixation event file, data processing pipelines generally clean the data, rejecting fixations or trials that are problematic due to artifacts such as excessively long or short fixations, excessively long saccades, or data loss due to the eye tracker losing measurement of gaze position or due to the participant blinking. Artifact rejection is one part of the process where experimenter degrees of freedom abound and vary across disciplines and labs. In a recent review, Eskenazi (2024) surveyed 185 articles and reported that 89% reported using at least one data-cleaning method. Although the vast majority of researchers do apply some type of data-cleaning method, there is very little consensus on the correct approach. Eskenazi (2024) reports that the vast majority of articles in his sample (81%) use some form of temporal exclusion criterion (e.g., rejecting fixations whose durations fall above or below preset thresholds), but articles in his sample are fairly evenly split on whether they apply this cutoff on its own (34% of articles) or in combination with some other form of data cleaning (48%). Given the significant amount of variation in how eye-tracking practitioners apply data cleaning procedures – and the sparsity of data on exactly how these choices impact data analysis – there are few hard and fast criteria for artifact rejection, and certainly no 'correct' choices. Instead, the sole single guiding principle we would offer beginning researchers is that transparency is key. Researchers should aim to carefully document and motivate the choices they make, and make both cleaned and uncleaned datasets available for other researchers to explore (see also Eskenazi, 2024). To that end, in this section, we summarize the types of choices researchers must make and considerations that go into such choices, to equip researchers to make informed decisions about data cleaning.

### General approach

The first decision researchers must make is whether exclusion of data at the event level (e.g., fixation or saccade) will only affect that event, will also affect later events in that region, or will affect the entire trial. For example, consider a situation in which a reader blinks while looking at an AOI the first time they land on it. Because a blink leads to track loss (i.e., the eye is not visible during a blink and therefore the point of gaze cannot be mapped to the screen), the start and end of the fixation are not calculable and therefore any first-pass duration measures for that AOI will be skewed because that fixation would be excluded. However, because cognitive processing still occurs even during a blink or track loss, the duration that is measured aside from the blink will be skewed (i.e., underestimate the time the reader processes the information in that AOI). If the reader returns to that AOI later on and does not blink, the return fixation would not be excluded and the total reading time measure would only include the return fixation, not any time spent during first-pass. Because of inferential difficulties regarding data for which there is track loss in one measure that percolates to later more inclusive measures, one approach is to exclude all measures within an AOI, or entire trials, for which there was a blink during first-pass reading.

The scenario may be different, however, if there is no problematic fixation during first pass but there is during a return fixation in the AOI. Presumably, the measures in first-pass reading are still valid and should not be excluded, but the measures that include the problematic fixation should be. Note that this would lead to differential data exclusions across dependent measures, which is not necessarily a problem, but researchers should consider whether this aligns with their research questions and should be explicit about their choice when reporting their findings.

### Excluding fixations

There are at least three general methods for cleaning up fixations prior to data analysis. Fixations may be (1) excluded by temporal criteria (i.e., falling above or below certain thresholds), (2) excluded by outlier criteria (i.e., by being too extreme on the distribution of fixation values), or (3) merged (i.e., the duration of an anomalous fixation is summed with some other nearby fixation). All three of these methods can be used in isolation or in combination. In fact, almost half

of the articles in Eskenazi's (2024) survey (49%) used more than one of these methods.

In his sample of eye-tracking articles, Eskenazi reports that most (81%) of articles apply some type of temporal exclusion criteria, but with significant variation in exactly which cutoffs are applied. In his sample, minimum fixation durations range between 0 ms (i.e., no minimum fixation duration length) to 140 ms with the most common value for a minimum fixation duration being 80 ms (used by 52% of articles). To our knowledge, there is no published rationale for these values (Eskenazi, 2024). This exclusion criterion is typically justified by noting that these values are statistical outliers, and that statistical outliers are unlikely to reflect the processes of interest (perhaps reflecting tracking errors or that the reader somehow temporarily disengaged from active processing of the text).

Very short fixations can sometimes reflect the processes under interest. If this is true, then there is an argument to be made that such short fixations should not be excluded. For example, following a suggestion by Morrison (1984), Schotter and Leinenger (2016) demonstrated that readers may make a single fixation on a word before moving forward that may be as short as 50 ms. Importantly, the likelihood of making those short fixations is influenced by properties of the word when it is viewed parafoveally (e.g., the frequency of the parafoveally previewed word in a boundary change paradigm study), and therefore these fixations are not due to oculomotor error, but rather due to linguistic preprocessing the word. In other words, these short fixations (termed *forced fixations*) represent cases in which the reader intended to skip the word, but the saccade plan to that word was already committed to being executed and they instead pre-initiated the subsequent saccade away from the word, leading to a short intermediate fixation. Since forced fixations are informative about (parafoveal) linguistic processing, they aren't clearly outliers. However, we note the prevalence of forced fixations is low enough that the results are unlikely to be skewed based on the choice to exclude or include them. More generally, the choice of whether to use a more liberal lower cutoff value should be deliberately made based on the research question. For example, Schotter and Leinenger (2016) chose not to exclude very short fixations based on the theoretically motivated hypothesis that, under certain circumstances in their experiment, there may be very short linguistically mediated fixations.

It is common to apply a temporal exclusion criterion for very long fixations as well. The precise value used as an upper cutoff is not as consistent across the literature as is the minimum cutoff. Eskenazi (2024) reports that 27% of articles do not apply any upper cutoff, but of those that do, the cutoff can range between 500 and 3000 ms. Within this range, the three most chosen cutoff values are 800, 1000, and 1200 ms, which together are the upper cutoff values for 61% of the articles in Eskenazi's sample. As with lower cutoffs, these cutoffs are motivated – loosely – by consideration of what constitutes normal eye movement behavior. Most fixations last around 200 ms, but again, there is significant variability. However, it is quite rare for fixations to last much longer than 800 ms, making this a sensible cutoff for individual fixations. There is also significant variability in practice for measures that aggregate across fixations. For example, gaze duration upper cutoffs could be as long as 2000 ms and total times could be as long as 4000 ms. However, these region-level measures vary dramatically based on the region's size and other features specific to particular experiments and stimuli, and it is harder to offer general guidelines as to exclusion criteria here.

In addition to temporal exclusion criteria, it is possible to adopt a cutoff scheme that uses the distribution of the data themselves to determine the cutoffs. Conceptually, this is how the temporal cutoffs were developed, based on the general reading time patterns observed in a large number of studies. The difference here is that the cutoff would be established based on the very same data that would be analyzed for the inferential statistics in the study. This might involve using standard deviations from the data mean to identify outliers, outlier identification based on residuals in a regression model, or percentile distribution cutoffs. Eskenazi (2024) notes that 24% of articles in his sample pursue this method for fixation exclusion.

It is possible to handle outlier fixations not by excluding them outright, but by merging them with adjacent fixations. This approach seeks to identify short fixations that fall within a certain minimum distance from another fixation, and then merges the shorter fixation into the larger, neighboring fixation. The key justification for this is that too-short fixation is thought to be a mis-execution or mis-measurement of a fixation that is properly part of the larger fixation, and so the merging process results in a single fixation measurement that reflects this. However, as with the temporal exclusion cutoffs, the precise values that determine how short a fixation is before it is merged, and how many degrees of visual angle away two fixations can be to be merged, vary across publications with no clear published rationale (Eskenazi, 2024). One potential rationale for this comes from the *double-step paradigm* (Becker & Jürgens, 1979), in which subjects fixated a central point and made saccades to a target presented in the parafovea or periphery, which were sometimes spatially displaced prior to the execution of the eye movement. When the target displacement occurred well prior to the execution of the saccade, the saccade program could be canceled and reprogrammed to target the new location. However, when the displacement occurred immediately prior to the execution of the saccade, the initial saccade (i.e., to the original target) was executed and was followed by a rapid subsequent saccade to the new target location leading to a short intervening fixation. For this reason, very brief fixations are merged with

nearby fixations (e.g., fixations within one character space of the subsequent fixation) because those probably represent cases in which the eyes landed in an inappropriate location and were quickly corrected.

While there are a great deal of choices the analyst faces in data cleaning, the degree to which these choices impact their final conclusions remains under investigation. Eskenazi (2024) found that the choice of exclusion criterion ultimately had a modest impact on the final conclusions. In his study, he did not find that choice of exclusion criteria changed the pattern of statistical significance in his analyses. Adopting more stringent cutoffs did result in smaller simple effects overall, a sensible finding given that more stringent cutoffs mean that longer fixations will be excluded. Interestingly, more stringent cutoffs often (but not always) meant smaller standard errors, which means that despite the effect size mostly remained stable across different cutoff processes.

### Participant exclusions

Data should be excluded if there are data quality problems, like track loss, calibration error (or fixation alignment issues in multi-line studies), or excessive blinking, particularly within the regions of interest. However, there are different ways that data quality issues for one region or measure could be treated with respect to inclusion for other measures or regions. In general, if there is an issue with data quality (e.g., a blink or track loss during a fixation) for an "early" measure like first fixation duration, all subsequent measures that include that measure (e.g., gaze duration, total viewing time, etc.) should also be excluded. If there are multiple instances of track loss during a trial, potentially the entire trial should be excluded. If multiple trials or dependent measures for an individual subject are excluded due to data quality issues, the entire subject should be replaced because it is unlikely that much meaningful information can be derived from the data that remains.

The criterion for what constitutes an acceptable amount of data loss is an area in which practices vary from lab to lab, but in general because data loss is mostly related to blinking or track loss, needing to exclude more than 25% of the data raises questions about data quality. For the most part, it should be possible to collect data from any given participant and keep at least 75% of it. If many participants are being excluded because they do not meet this criterion then the researcher should go back and examine their data collection and data processing procedures and consider whether they need to make improvements there before moving forward.

## Typical dependent variables

Once the data are cleaned, it is next time to derive the dependent measures that will be used in the statistical analysis. Before we move on, it is important for us to emphasize that researchers *should not reify* these dependent measures. Even though we must define these measures for scientific purposes, it is important not to ascribe a particular measure to a particular underlying cognitive-linguistic process. As we noted above, saccades and fixations are merely discrete behavioral events that interface with continually unfolding underlying neural processing. While they may be sensitive to factors that affect those underlying neural processes (e.g., lexical and contextual variables affect fixation probabilities and durations described above) measuring eye movements does not provide a direct measure of the neural processes.

There are two general classes of eye movement measures for reading studies (*local* and *global*), and these measures reflect different aspects of the reading process (Rayner, 1998). Local measures focus on a word or phrase within a sentence and depend on the definition of an AOI (see Section "Stimulus design"). Usually, researchers will carve up the entire sentence into multiple AOIs, although in many instances not all AOIs are formally analyzed. Given a defined AOI, the researcher may then compute a local dependent measure for that AOI on a given trial. Global measures are defined and computed across an entire trial, or presentation of a single experimental stimulus. Depending on the experiment, this may be a single line of text, or multiple lines.

## Local measures

Assuming that a researcher is interested in local reading measures, they will first identify AOIs, and then compute dependent measures that summarizes the reading behavior associated with that AOI from the fixation and saccade events. Because eye tracking data in their rawest form only provide information about the location of the point of gaze on the computer screen[5] (i.e., pixel location), and the word(s) of interest could be located anywhere on the screen (see Section "Stimulus design"), the creation of an AOI (or IA) file is critical. The IA file allows the researcher to mark the location(s) of the AOIs in the same coordinate space as the eye tracking data, which means that AOIs for the text are translated into *x*- and *y*- pixel coordinates. Once translated

[5] We focus on screen-based experiments, but researchers may also be interested in using head mounted trackers for more naturalistic studies of people reading text on paper or hand-held tablets. There is substantially more variability and less experimental control in these studies, which is why the details of implementing these studies are beyond the scope of this paper. However, many of the principles we describe are relevant, but instead of pixel-based coordinates of the display screen, the researcher must code AOIs in terms of coordinates relative to the eye tracker camera location and field of view.

into the same mapping space, a given fixation can be marked as being within a specific AOI, and with information about each fixation's ordinal index, associated AOI, and duration, a large number of potential local reading measures can be defined, each associated with different aspects of the reading process (although see our point about not reifying these measures, above).

### Word skipping

The first behavior with respect to the reading process on a word or AOI is whether it is indeed fixated. Word *skipping* occurs when the AOI is *not* directly fixated. This measure is binary, in the sense that it only codes whether the AOI was skipped (i.e., value = 1) or fixated (i.e., value = 0). Alternatively, researchers can define *fixation probability*, which is merely the inverse of skipping (i.e., the AOI was fixated = 1, or skipped = 0). Usually, a binary measure is sufficient for most purposes, although for particularly long AOIs that are almost guaranteed to be fixated, sometimes a measure of number of fixations may be more informative.

It is important to make a distinction between an AOI that was skipped during *first-pass reading* (i.e., may have been fixated, but only after a later region was fixated) and an AOI that was completely skipped (i.e., was never fixated during the trial). In general, first-pass skipping and total skipping will be highly correlated measures, but first pass skipping rates will be higher than total skipping rates because it is a more inclusive measure. Both may be informative, but if the research question regards initial processing or parafoveal preview, then first-pass skipping may be more relevant.

**Skipped during first-pass reading** This measure is defined at the time when the AOI is first encountered, and a first-pass skip occurs when the saccade moves from an AOI before it to an AOI after it, without stopping on the target AOI itself. More technically, this measure is usually defined algorithmically based on (1) if a later AOI was fixated before the target AOI or (2) there was never a fixation on the target AOI in the entire trial; otherwise, the target AOI was not skipped because the first fixation on it occurred before any later AOI was fixated.

**Completely skipped** This measure is simpler to define than first-pass skipping, because one only needs to check whether the AOI was entered or not on the trial; if not, then the AOI was skipped.

### Landing position

Landing position refers to the location (in characters) in the word where the first fixation lands. Obviously, if the reader skips over the AOI, this measure cannot be defined and would be represented by a missing value. Because readers tend to target the center of words, but there is both random and systematic error in saccade targeting (McConkie et al., 1988), landing position is strongly related to the lengths of the words that the saccade moves from and to (i.e., target word length and the launch distance of the previous fixation, O'Regan, 1980; Rayner, 1979).

### Initial reading time measures

Initial reading time measures regard the time spent fixating the AOI (provided it was not skipped) before moving forward in the text. This consists of a class of various dependent measures that are believed to reflect different types of behavior. As we emphasized above, these dependent measures do not often have a direct theoretical interpretation, but instead represent convenient ways of quantifying a fundamentally continuous behavior (i.e., reading). One way to calculate these measures is to exclude any skipping cases (i.e., to represent a skipping case with a *missing value* as opposed to 0 ms); alternatively, researchers may choose to include 0 ms values for skipping cases in order to account for skipping (e.g., Just & Carpenter, 1980; Rayner et al., 2011; see Rayner, 1998). It is not merely a statistical choice, but also a theoretical choice to consider skipping and first-pass fixation durations as mutually exclusive. When choosing whether to include 0 values for skipped AOIs, the researcher must determine whether they want to consider the skipping decision as part of the same distribution as a duration decision, or whether there are qualitative (i.e., categorical) differences between these behaviors. However, there are also practical statistical implications to consider when making this decision: on the one hand, representing skipped words with missing values will lead to data loss and may do so differently across conditions if skipping rates differ, while on the other hand including 0 values causes a non-unimodal distribution in the measure (i.e., because most fixations are at least ~ 100 ms) and this violates the assumption of many statistical tests that the data will have a normal (or at least unimodal) distribution.

**First fixation duration** First fixation duration is the duration of the first fixation on the AOI, regardless of the number of first-pass fixations. This is the most inclusive measure, because it does not distinguish between single fixation and multiple fixation trials (see below). Because this measure includes a mixture of behaviors it tends to be noisy (Reingold et al., 2010); for example, first fixations of multiple-fixation trials tend to be shorter than single fixations (Reingold et al., 2010).

**Single fixation duration** Single fixation duration is the duration of the fixation on the AOI when only one fixation was

made during the first-pass. A single fixation generally indicates that the word seemed fairly easy to recognize and was of moderate length (i.e., short words are likely to be skipped and long words are likely to be refixated; Kliegl & Engbert, 2005; Schotter & Leinenger, 2016; Vitu & McConkie, 2000). Relatedly, *single fixation probability* represents the proportion of first-pass reading trials in which there was a single fixation (Rayner et al., 1996). Words that are more difficult to identify, for example, lower-frequency words decrease the likelihood of making a single fixation (Reingold et al., 2010).

**Gaze/first-pass duration** Gaze duration, or first-pass duration, is the sum of the duration of all fixations on an AOI before it is exited. This measure may be called gaze duration when the AOI comprises a single word, and first-pass duration when it comprises multiple words. It includes both single fixation durations and multiple fixation cases and is often assumed to reflect the average time required for word identification (Rayner, 1998, 2009). For any given trial, gaze duration will only differ from first fixation duration if there is a refixation on the word before leaving it (and therefore, single fixation duration will not exist). Therefore, analyzing/reporting both single fixation duration and gaze duration is redundant unless there is a substantial refixation rate, in which case the researcher might want to consider why the refixations are occurring and whether the rate differs between conditions. There are two primary causes of refixations: they are more likely for longer words (Vergilino & Beauvillain, 2000) whereas, for shorter words, they are more likely when the first fixation lands far from the OVP (McConkie et al., 1989; Reilly & O'Regan, 1998).

**Go-past time** Go-past time is the sum of all fixations made once an AOI has been entered, up to the point when the eyes *go past* it, so to speak (i.e., exit it, progressing forward in the text; Brysbaert & Mitchell, 1996; Konieczny, 1996; Speer & Clifton, 1998). On trials where no regressive saccade was made, the go-past time is simply equal to the first pass time. However, a first-pass regression, and the subsequent re-reading of an earlier portion of the sentence, will lead to longer go-past times. This may be considered a class of measures, as there are different ways to calculate this, depending on whether the researcher decides to include fixations that occur in material that precedes the AOI during rereading that results from a regressive saccade out of the AOI. Therefore, it may be useful to make a distinction between these measures with different terms: regression path duration and right bounded reading time. Regression path duration (which is most often what researchers mean when they use the term go-past time) indexes both the initial reading of an AOI, along with any re-reading that accompanies regressions launched from the AOI (Konieczny, 1996; Konieczny et al., 1995). Right bounded reading time (Liversedge, 1994; also known as *selective regression path duration*; Payne & Stein-Morrow, 2012) is the sum of all fixations on a given AOI before exiting it to the right, but unlike regression path duration, does not include fixations on previous regions – it indexes both initial reading of an AOI, and any re-reading of that AOI done before any regions to the right of the AOI are fixated. Because this measure ignores fixations that occur during the re-reading process it is not entirely clear what it reflect about higher-level language processes.

**Extended viewing period** The measures reviewed above all regard the initial reading on the AOI, ending when the eyes leave after initially landing on it. This can be contrasted with measures of the total amount of time spent on the AOI, regardless of when the fixations occurred and including any rereading.

**Total time** Total time is the sum of the durations of all fixations on the AOI, including gaze duration and any time spent re-reading it. Some take this measure as an indication of the time required to completely and successfully identify the word or words contained in the AOI (e.g., if the reader has to go back for more information; Bicknell & Levy, 2011). Thus, in general, longer total times would indicate that reinterpretation was more difficult. However, this measure can be highly variable, and may not directly map on to one cognitive process within reading behavior because it includes first-pass and re-reading, and it is unclear or unspecified what different types of processes or behaviors had occurred in between. Therefore, total time is a good example of a dependent measure that is easy to understand and simple to define, but which might only imperfectly measure processing difficulty. For this reason, it may be more informative to interpret effects in total time in the context of the other measures; for example, if a certain pattern is observed in total time that was not observed in an "earlier" measure like gaze duration, then it suggests that the effect arises because of a difference in re-reading between the two conditions.

**Total number of fixations** In addition to duration, one can measure the number of fixations that were allocated to the AOI. However, this measure can be highly biased by the size of the AOI (e.g., there will be more fixations on larger AOIs).

**Rereading time** Because total time on the AOI is in some way confounded with initial reading time because it is a cumulative measure, sometimes researchers may prefer a completely non-overlapping variable in order to compare late processing effects to the patterns observed in first pass (Radach & Kennedy, 2004). For this, it is useful to define rereading time, which is any amount of time spent in the AOI after the first pass.

### Regressions

The measures discussed so far are all duration-based measures, with the exception of word skipping, landing position, and total number of fixations. At the level of individual trials, skipping is a binary dependent variable: a given AOI was either skipped, or it was not. In a similar fashion, we can determine whether a reader exited an AOI to the right or to the left. From this, it is possible to calculate the probability that a reader will not go forward, but rather backward in the text from a given AOI. Backward eye movements, called regressions (see Section "Where decisions"), are what is responsible for rereading. Software packages vary in how they handle trials on which a critical region was skipped in first pass (i.e., either a value of *NA* or a 0); and the way this is handled may differ for different types of regression measures (i.e., whether the saccade enters the region from words further in the sentence – a *regression in* – or leaves the region to go to words earlier in the sentence – a *regression out*). Therefore, we discuss these details further in the sections below, and merely highlight here that researchers should be careful to check exactly how their software package (or self-written code) calculates these measures so they can report their analysis clearly.

**Regressions-in** By the most strict definition, regressions-in are only counted if the previous fixation is further into the sentence (in terms of AOI) than the current fixation. That means that rereading fixations that occur after the reader has gone past the AOI, but then reread earlier parts of the sentence before returning to the AOI will be excluded from the regressions-in measure. To make this more clear, imagine a reader fixates an AOI and then moves past it but becomes confused about its identity and returns to it. If they go directly to the AOI from, for example, the end of the sentence, that fixation will be counted in the measure of regressions-in to the target AOI. If, however, the reader instead starts reading the entire sentence again, from the beginning, the rereading that occurs in that AOI will not be counted in regressions-in because the immediately previous fixation was to the left of the target AOI. This means that this measure may not entirely capture all rereading that occurs; some readers go directly back to source of confusion, some readers backtrack, others return to the beginning and reread the entire sentence again, and others may have more variable scan paths (Frazier & Rayner, 1982; von der Malsburg & Vasishth, 2011).

With respect to calculating this measure when the AOI was skipped during first-pass, it may not make sense to use NA values. That is because a regression into a word, conceptually, may represent the same process regardless of whether the word was skipped or not. At the very least, the binary values (i.e., 1 for there was a regression in and 0 for there was not) should be maintained and, if the researcher wants to investigate whether first-pass skipping makes a difference, they can split the trials based on skipping or use the skipping measure as a predictor.

**Regressions-out** In general, the issue of variability in regressions-out is less severe than for regression-in because studies using regressions-out as a dependent measure can be designed to have a disambiguating AOI. For example, in garden path sentences, a "disambiguating region" can generally be defined as the word(s) that render one of the two interpretations syntactically impossible and thereby triggers reanalysis (Frazier & Rayner, 1982). However, some manipulations that trigger reanalysis are more difficult to pinpoint to a single word. Moreover, the assumption for these regions is that the information that triggers the regression is processed immediately. However, there is some evidence that the ability to act on regressions interacts with the current (forward) saccade plan, depending on whether that plan can be canceled or is past a point of no return. If the progressive saccade is committed to execution, the regression will not be triggered until the subsequent region (Schotter, von der Malsburg & Leinenger, 2018).

With respect to calculating this measure when the AOI was skipped during first-pass, it may be necessary to further sub-classify this measure into regressions-out that occur during first-pass reading and all regressions out (akin to how there are multiple possible skipping measures).

**First-pass regressions-out** For first-pass regressions-out, a regression would only be counted if it was launched during the first pass through the AOI. Therefore, it would make sense to treat this measure as NA when the AOI was skipped during first-pass because it is technically impossible to calculate a regression-out when there is no fixation from which to launch a regression.

**Total regressions-out** For total regressions-out, any regression out of the region would be counted, regardless of when it occurred. Therefore, it would make sense to treat this measure as 0 even if the AOI was skipped during first-pass because it would represent that the region never initiated rereading. However, the researcher may decide that using NA values in first-pass skipping cases is more appropriate for their research question and study design. In that case, they should make sure that their data are appropriately coded and that they are transparent about the calculation of the measure when reporting their results.

### Scanpath analysis

A scanpath refers to a series of fixations, along with their associated positions in a text. In other words, it refers to the

whole pattern of fixations at a given point in reading. Scanpath analysis involves characterizing these sequences of fixations in time and space to understand how experimental manipulations change the characteristic patterns of reading behavior in different contexts. Contemporary scanpath analysis was developed and popularized by von der Malsburg and colleagues (e.g., von der Malsburg et al., 2011, 2012, 2015), building on long-standing intuitions that the precise sequence of fixations across a text could be informative about the mental processes involved in reading comprehension (e.g., Frazier & Rayner, 1982). Scanpath analysis broadly proceeds in two steps: First, the similarities between sequences of fixations in space in time is computed, and second, different sequences of fixations are clustered to identify stable patterns of scanpath behavior within participants (see von der Malsburg et al., 2015 for in-depth discussion). Given such a clustering, it is possible to ask whether a given experimental manipulation systematically changes the type of reading behavior occasioned by a critical item: For example, von der Malsburg and Vasishth (2011) showed that syntactic reanalysis often triggered re-reading of the critical sentence from the beginning. However, there is variability in scanpath patterns, both when comparing the most prevalent pattern across different participants (von der Malsburg & Vasishth, 2011) and when comparing different items or trials within a participant (von der Malsburg et al., 2015). Therefore, more work is needed to determine the reliability of scanpath analysis as a diagnostic tool, for example in order to determine whether a reader did or did not correctly re-parse a sentence after rereading.

## Global reading measures

Global eye tracking measures can give the researcher a very broad indication of the difficulty of reading under different conditions or for different types of readers. For example, these measures are commonly reported for studies that implement a manipulation of the visual availability of the text for every word or on every fixation, for example the *moving window paradigm* (McConkie & Rayner, 1976; see Rayner, 2014), *moving mask paradigm* (Rayner & Bertera, 1979), *disappearing text paradigm* (Rayner et al., 2003), and experiments that manipulate *text legibility* (Jordan et al., 2017). Generally, these measures increase as the text becomes more difficult to read (e.g., as the size of the moving window decreases, the size of the moving mask increases, the text disappears sooner after the onset of fixation, or the text becomes less clear) or as the reader is younger or less skilled (Rayner, 1986). Exceptions to this principle are saccade length and skipping probability, which decrease with additional difficulty. These measures (and others, see Hyönä et al., 2003) are also useful in studying global text processing. However, caution is warranted as they do not *only* reflect processing difficulty but they may also index reader engagement (Ballenghein et al., 2023).

Global measures may provide a general index of reading difficulty, but because they are defined across an entire trial, they do not allow the researcher to localize any processing differences across conditions to a particular word or phrase. For the calculation of all of the fine-grained reading measures (i.e., those other than reading rate), researchers must consider how they want to treat data exclusions (e.g., fixations interrupted by blinks or excluded for not meeting duration cutoffs; see Section "Artifact rejection procedures") because the exclusion of an individual fixation can complicate the calculation of some of these measures either by either truncating the duration of a fixation or interfering with the proper calculation of the relative positions of two fixations (e.g., saccade length or regressions).

### Reading rate (words per minute: wpm)

Reading rate is calculated as the number of words in the sentence divided by the total sentence reading time in minutes (i.e., the number of milliseconds between when the sentence was first presented until the participant finished reading, which is then divided by 60,000, or the number of milliseconds in a minute).

### Total sentence/passage reading time

As an alternative to reading rate, some researchers report total sentence or passage reading time, which is similar but does not normalize across the amount of content. This may be fine for descriptive purposes, especially when the comparisons of interest are between different visual manipulations of the same text, or if the texts that are being compared have similar properties in terms of text difficulty (e.g., number of words, average word frequency, etc.).

### Number of fixations

Number of fixations is measured as the total number of valid fixations on the trial. It is important to consider that some participants may make additional fixations prior to or after reading the sentence; the researcher should consider ways to identify the moments that they actually start and finish reading and only include fixations between these two events to ensure that only reading-related fixations are included.

### Mean fixation duration

Mean fixation duration is measured as the average duration (in ms) of all the valid fixations included on a trial.

### Mean saccade length

Mean saccade length is measured as the number of characters between one fixation and the immediately preceding fixation. Number of characters is the best unit of analysis here because it provides a linguistically meaningful and generalizable measure (i.e., as opposed to the number of pixels or degrees of visual angle). In general, this measure is restricted to *forward* saccades (i.e., only those for which the preceding fixation was further to the left than the current one) as forward and backward saccades tend to have different distributions with respect to length (Vitu & McConkie, 2000).

### Percent/total number of skips

These are measured as the total number or percentage of AOIs on a given trial that were either never fixated (i.e., total skipping), or were fixated after there had been a fixation on an AOI further to the right at any point earlier in the trial (i.e., first pass skipping). In general, first-pass skipping and total skipping will be highly correlated measures, but first pass skipping rates will be higher than total skipping rates.

### Percent/total number of regressions

These are measured as the total number or percentage of the fixations on a given trial that were located on a word earlier on in the sentence than the fixation immediately preceding it. For reading languages like English that are written left-to-right, "earlier" in the sentence is generally measured as further to the left, except when conducting a multi-line text study, in which case the ordinal position of the word in the text is more useful. For languages like Hebrew or Arabic that are written right-to-left, the heuristic of further to the right is more appropriate, but again only in the case of single line text studies.

## Data analysis

Once a researcher has obtained and processed their data, and calculated dependent measures following the guidelines we have laid out above, the final task is to subject those data to statistical analyses, interpret the results, and report them to other scientists. The analysis and interpretation aspect of the scientific enterprise is not trivial and many of the decisions and potential issues surrounding data analysis have been discussed elsewhere. Therefore, below we raise the issues that eye tracking researchers should consider and give a very brief overview of some resources to which they should refer.

While the statistical choices that face researchers doing reading research are, by and large, shared by researchers in many different areas, we note that there is one particular statistical pitfall that reading research is particularly prone to: the issue of multiple comparisons. As described above, it is standard to analyze multiple dependent measures, and perhaps even multiple AOIs, in a single reading experiment. Because each measure and AOI is typically analyzed independently, researchers generally carry out many statistical comparisons for a given experiment. If researchers adopt a simple strategy of declaring an effect 'significant' if they observe statistical significance *in any one of these* comparisons, then the false positive rate (type I error) for a single reading experiment can be unacceptably high (von der Malsburg & Angele, 2017). This is important to counteract in some way. For example, von der Malsburg and Angele (2017) show that the use of the Bonferroni correction for multiple comparisons, in addition to simple 'rule of thumb' heuristics such as 'significance is required in at least two measures' are both reasonable options for counteracting the type I error issue. However, a better remedy would be to use theoretical considerations to limit and select the dependent variables that are measured. For example, if a research question regards the effect of a parafoveal preview manipulation on initial word recognition, only "early" reading measures such as word skipping or first or single fixation duration are really informative in addressing the question. A researcher might also want to analyze total reading time or regressions for a complete picture of how the manipulations impact the reading time course, but it should be clear at the outset that an effect in those measures should not be used to make inferences about the question of initial word recognition.

As noted by Jannsen (2012), it has long been argued that in all language experiments there are two random factors (participants and items) and therefore many of the "standard" statistical approaches in psychology (e.g., ANOVA) are not quite appropriate (Clark, 1973; Coleman, 1964). For this reason, it is now common practice to analyze data from eye tracking while reading experiments – where multiple participants respond to multiple stimuli within the same study and therefore the researcher has obtained repeated measures for both participants and items – with mixed effects regression models that estimate random effects for both of these sources of variance simultaneously. Brown (2021) provides an approachable theoretical introduction to mixed-effects models and a practical introduction to how to implement them in R. Meteyard and Davies (2020) provide guidelines for reporting mixed-effects models, based on a survey and review of current papers showing wide variability in the existing literature regarding how different researchers build, evaluate, and report models. Mixed effects models have become the "gold standard" in statistical analysis for eye tracking while reading research in the past decade and that is why there is such a focus on that approach. There are, however, disagreements on the best practices in using these models, for example in specifying the random effects structure. Barr et al. (2013) argue that the random effects structure affects

the generalizability of a mixed effects regression analysis, advocate for the maximal random effects structure justified by the design, and discuss approaches to reduce model complexity if the maximum structure is not feasible given the dataset. However, Matuschek et al. (2017) argue against a maximal random effects structure because it can lead to a loss of power, and argue for a (non-maximal) random effect structure supported by the data. A separate issue in mixed effects models regards how to specify the comparisons for the fixed effects. Brehm and Alday (2022) discuss how the contrast codings in mixed-effect models affect the interpretation of model terms using simulations and provide recommendations for best practices in contrast coding. Schad et al. (2020) provides a tutorial on using custom a priori contrasts to test experimental hypotheses in mixed effects models, including the mathematics underlying different contrasts and how they are applied in R.

The research landscape is ever-evolving and new trends in statistical analysis often arise. Therefore, researchers should make sure to keep up with the state of the field and use the analytical technique that is most appropriate for their data and research question. For example, in reaction to null hypothesis statistical testing (of which mixed effects models are an example), some researchers advocate for a Bayesian analysis approach because it has the added value of quantifying the evidence in favor of one hypothesis over another rather than just rejecting – or failing to reject – the null hypothesis. For a summary, as well as an introduction to establishing a top-to-bottom workflow for Bayesian data analysis, see Schad et al. (2022). Because of the large variability in eye tracking measures, it may be more appropriate to use statistical analyses that take account of the distribution of the fixation durations, rather than just the mean, as is the primary focus in mixed effects models. Staub et al. (2010) present an example of how ex-Gaussian analysis can be used to investigate the effect of lexical variables on the distribution of fixation durations (e.g., separating the effects to a shift in the mean and skew of the tail). Reingold and Sheridan (2014) introduce how divergence point analysis can be used to estimate the onset of the influence of an experimental variable on fixation durations by creating survival curves for two conditions and using bootstrapping to compare them.

## Conclusion

In this article, we aimed to give a brief, high-level summary of the need-to-know information that a researcher must have before starting an eye-tracking-while-reading experiment. We provided background on the basics of eye movements in reading, the variables known to affect reading behavior and prominent theories used to explain those phenomena, guidelines for designing experiments and cleaning eye tracking data, commonly used dependent variables and how to calculate them, as well as pointers to resources for data analysis. Obviously, a single paper cannot provide all that is necessary for researchers to know and there is some degree of hands-on experience that really makes these ideas and principles tangibly salient. However, we hope this paper provides the reader with enough information and the confidence to get going in the exciting world of eye-tracking-while-reading research.

## Appendix

*Authors' Checklist.* The checklist is intended to provide a brief overview of important guidelines for what should be reported in a paper, and questions that each section in a write-up should address. Authors may wish to use it prior to submission, to ensure that the manuscript provides key information.

### Open Practices

- ☐ Specify whether the study was pre-registered and provide a link to the publicly available pre-registration document.
- ☐ Specify how and where the data and/or analysis scripts will be shared, including providing a link to the publicly available repository (e.g., OSF).

### Hypotheses

- ☐ Specify which specific dependent measures (and on which specific AOIs) are expected to differ between conditions (and those that are not expected to differ), and the direction of those expected differences.

### Participants

- ☐ Describe the inclusion/exclusion criteria of the participants (e.g., language background/proficiency, visual abilities/impairments, etc.).
- ☐ If any assessment data were collected (e.g., reading comprehension ability, spelling ability, language proficiency, non-verbal IQ, etc.), provide a table of means, standard deviations, and range. This is particularly important when comparing effects between participant groups.

### Recording Characteristics and Instruments

- ☐ What was the eye tracking hardware (e.g., company, model) and setup configuration (e.g., tracker mount settings)?

- ☐ What was the configuration of the eye-tracking software? (e.g., sampling rate, precision of calibration error)
- ☐ What are the details of the monitor displaying the stimuli? (e.g., make, model, screen resolution, refresh rate, size of the displayed area, distance from participant)

**Stimulus Parameters**

- ☐ What were the linguistic characteristics of the overall stimuli? (e.g., sentence length in words and/or characters)
- ☐ What was the nature of the linguistic manipulation and how was this confirmed? Provide a table of descriptive statistics across conditions where possible (e.g., lexical variables like length, frequency, predictability, plausibility, results from norming studies, etc.)
- ☐ What were the characteristics of the critical AOI(s)? (e.g., location in the sentence)
- ☐ What were the display characteristics? (e.g., font size, font type, font color, background color)
- ☐ How large were the stimuli physically? (e.g., how many characters subtended one degree of visual angle, or how large in degrees of visual angle was a single character)
- ☐ Where did the stimuli appear? (e.g., was the sentence displayed on a single line or did it take up multiple lines)

**Procedure**

- ☐ What were the instructions/task given to the participant? If there were multiple tasks, was it a between-participants or within-participants manipulation? ☐ Were the tasks blocked or intermixed? If blocked, was the order counterbalanced?
- ☐ Was there a calibration/drift check and where was the target point located?
- ☐ Was there a gaze-contingent start to the trial such that the participant needed to fixate in a certain location for a certain amount of time?
- ☐ Was there a time limit imposed on reading the stimuli?
- ☐ If there was a visual manipulation (e.g., gaze-contingent boundary, moving window), describe the parameters of the trigger (e.g., location of the boundary, delay of the change)
- ☐ What was the event that ended the trial? (e.g., participant response, experimenter response)
- ☐ If the participants responded to questions about the stimuli, what was the nature of the question and was there feedback provided? On what percent of the trials was the question asked?

**Data Processing Steps**

- ☐ Were practice trials excluded from the analysis?
- ☐ What criteria, if any, were used to merge or exclude individual fixations? (e.g., blinks, response events, extremely long or short durations, etc.)
- ☐ What criteria, if any, were used to exclude entire trials (e.g., too many or too few fixations, blinks, or track loss on a target region, etc.)
- ☐ How many trials (and what percent) were excluded at each step of the process?
- ☐ How is data loss distributed across conditions? Do the conditions significantly differ in the number of retained trials?
- ☐ How many participants were excluded because of excessive data loss? What was the criterion for exclusion?

**Presentation of Dependent Measures**

- ☐ Describe how each of the measures was defined when it was calculated from the raw data. Note, it is not sufficient to describe the general definition of this measure – the authors should check that in their data processing pipeline this is indeed how it is calculated.
- ☐ Include appropriate descriptive statistics for each measure analyzed, either in the form of a summary table or in a suitable visualization.

**Statistical Analyses**

- ☐ What analytical approach was used? (e.g., variables for the fixed effects, including contrast coding scheme; structure of the random effects)
- ☐ Were there any controls for multiple comparisons?

**Authors' contributions** The two authors worked collaboratively on this paper.

**Funding** This work was partially supported by National Science Foundation grants BCS-2120507 to ERS and BCS-2020914 to BD.

**Data availability** Not applicable.

**Code availability** Not applicable.

## Declarations

**Ethics approval** Not applicable.

**Consent to participate** Not applicable.

**Consent for publication** Not applicable.

**Conflict of interest** The authors declare no conflicts of interest.

## References


Abbott, M. J., Angele, B., Ahn, Y. D., & Rayner, K. (2015). Skipping syntactically illegal the previews: The role of predictability. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(6), 1703.

Abbott, M. J., & Staub, A. (2015). The effect of plausibility on eye movements in reading: Testing EZ Reader's null predictions. *Journal of Memory and Language, 85*, 76–87.

Andrews, S., & Veldre, A. (2019). What is the most plausible account of the role of parafoveal processing in reading? *Language and Linguistics Compass, 13*(7), e12344.

Andrews, S., & Veldre, A. (2021). Wrapping up sentence comprehension: The role of task demands and individual differences. *Scientific Studies of Reading, 25*(2), 123–140.

Angele, B., & Rayner, K. (2013). Processing the in the parafovea: Are articles skipped automatically? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(2), 649.

Angele, B., Laishley, A. E., Rayner, K., & Liversedge, S. P. (2014). The effect of high-and low-frequency previews and sentential fit on word skipping during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(4), 1181.

Angele, B., Schotter, E. R., Slattery, T. J., Tenenbaum, T. L., Bicknell, K., & Rayner, K. (2015). Do successor effects in reading reflect lexical parafoveal processing? Evidence from corpus-based and experimental eye movement data. *Journal of Memory and Language, 79*, 76–96.

Ariasi, N., Hyönä, J., Kaakinen, J. K., & Mason, L. (2017). An eye-movement analysis of the refutation effect in reading science text. *Journal of Computer Assisted Learning, 33*(3), 202–221.

Ashby, J., & Clifton, C., Jr. (2005). The prosodic property of lexical stress affects eye movements during silent reading. *Cognition, 96*, B89–B100.

Ashby, J., Rayner, K., & Clifton, C., Jr. (2005). Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *The Quarterly Journal of Experimental Psychology Section A, 58*(6), 1065–1086.

Baayen, R. H. (2008). Analyzing linguistic data: A practical introduction to statistics using R. Cambridge University Press.

Ballenghein, U., Kaakinen, J. K., Tissier, G., & Baccino, T. (2023). Fluctuation in cognitive engagement during listening and reading of erotica and horror stories. *Cognition and Emotion, 37*(5), 874–890.

Balota, D. A., Pollatsek, S., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology, 17*, 364–390.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445–459.

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.

Becker, W., & Jürgens, R. (1979). An analysis of the saccadic system by means of double step stimuli. *Vision Research, 19*(9), 967–983.

Bicknell, K., & Levy, R. (2010a). Rational eye movements in reading combining uncertainty about previous words with contextual probability. In *Proceedings of the annual meeting of the cognitive science society* (vol. 32, no. 32).

Bicknell, K., & Levy, R. (2010b). A rational model of eye movement control in reading. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 1168–1178).

Bicknell, K., & Levy, R. (2011). Why readers regress to previous words: A statistical analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, No. 33).

Bicknell, K., & Levy, R. (2012). Why long words take longer to read: the role of uncertainty about word length. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics* (pp. 21–30). Association for Computational Linguistics.

Booth, R. W., & Weger, U. W. (2013). The function of regressions in reading: Backward eye movements allow rereading. *Memory & Cognition, 41*(1), 82–97.

Brehm, L., & Alday, P. M. (2022). Contrast coding choices in a decade of mixed models. *Journal of Memory and Language, 125*, 104334.

Brown, V. A. (2021). An introduction to linear mixed-effects modeling in R. *Advances in Methods and Practices in Psychological Science, 4*(1), 2515245920960351.

Brysbaert, M., & Mitchell, D. C. (1996). Modifier attachment in sentence parsing: Evidence from Dutch. *The Quarterly Journal of Experimental Psychology Section A, 49*(3), 664–695.

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

Brysbaert, M., & Vitu, F. (1998). Word skipping: Implications for theories of eye movement control in reading. In *Eye guidance in Reading and Scene Perception* (pp. 125–147). Elsevier Science Ltd.

Carr, J. W., Pescuma, V. N., Furlan, M., Ktori, M., & Crepaldi, D. (2022). Algorithms for the automated correction of vertical drift in eye-tracking data. *Behavior Research Methods, 54*(1), 287–310.

Carroll, P., & Slowiaczek, M. L. (1986). Constraints on semantic priming in reading: A fixation time analysis. *Memory & Cognition, 14*, 509–522.

Carter, B. T., & Luke, S. G. (2020). Best practices in eye tracking research. *International Journal of Psychophysiology, 155*, 49–62.

Chaffin, R., Morris, R. K., & Seely, R. E. (2001). Learning new word meanings from context: A study of eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 225–235.

Clark, H. H. (1973). The language-as-a-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335–359.

Clifton, C., & Staub, A. (2008). Parallelism and competition in syntactic ambiguity resolution. *Abstract Language and Linguistics Compass, 2*(2), 234–250. https://doi.org/10.1111/j.1749-818X.2008.00055.x

Clifton, C., Jr., Ferreira, F., Henderson, J. M., Inhoff, A. W., Liversedge, S. P., Reichle, E. D., & Schotter, E. R. (2016). Eye movements in reading and information processing: Keith Rayner's 40 year legacy. *Journal of Memory and Language, 86*, 1–19.

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In *Eye Movements* (pp. 341–371).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Coleman, E. B. (1964). Generalizing to a language population. *Psychological Reports, 14*, 219–226.

Cutter, M. G., Drieghe, D., & Liversedge, S. P. (2015). How is information integrated across fixations in reading. *The Oxford handbook of reading*, 245–260.

Drieghe, D., Brysbaert, M., Desmet, T., & De Baecke, C. (2004). Word skipping in reading: On the interplay of linguistic and visual factors. *European Journal of Cognitive Psychology, 16*(1–2), 79–103.

Drieghe, D., Pollatsek, A., Staub, A., & Rayner, K. (2008). The word grouping hypothesis and eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(6), 1552.

Duchowski, A. (2007). Eye tracking techniques. In *Eye Tracking Methodology* (pp. 51–59). Springer, London.

Duffy, S. A., Kambe, G., & Rayner, K. (2001). The effect of prior disambiguating context on the comprehension of ambiguous words: Evidence from eye movements. In D. S. Gorfein (Ed.), *Decade of behavior. On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 27–43). Washington, DC, US: American Psychological Association.

Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language, 27*(4), 429–446.

Duggan, G. B., & Payne, S. J. (2009). Text skimming: The process and effectiveness of foraging through text under time pressure. *Journal of Experimental Psychology: Applied, 15*(3), 228.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior, 20*, 641–655.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review, 112*(4), 777.

Engelmann, F., Vasishth, S., Engbert, R., & Kliegl, R. (2013). A framework for modeling the interaction of syntactic processing and eye movement control. *Abstract Topics in Cognitive Science, 5*(3), 452–474. https://doi.org/10.1111/tops.12026

Eskenazi, M. A. (2024). Best practices for cleaning eye movement data in reading research. *Behavior Research Methods, 56*, 2083–2093.

Ferreira, F., & Yang, Z. (2019). The problem of comprehension in psycholinguistics. *Discourse Processes, 56*(7), 485–495.

Filik, R. (2008). Contextual override of pragmatic anomalies: Evidence from eye movements. *Cognition, 106*(2), 1038–1046.

Fisher, D.F. & Shebilske, W.L. (1985). There is more than meets the eye than the eye-mind assumption. In: *Eye Movements and Human Information Processing*, ed. R. Groner, G. McConkie, & C. Menz. Amsterdam: North Holland.

Folk, J. R. (1999). Phonological codes are used to access the lexicon during silent reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 892–906.

Folk, J. R., & Morris, R. K. (2003). Effects of syntactic category assignment on lexical ambiguity resolution in reading: An eye movement analysis. *Memory & Cognition, 31*, 87–99.

Frank, M. C., Braginsky, M., Cachia, J., Coles, N. A., Hardwicke, T. E., Hawkins, R. D., & Mathur, M. B. (2024). *and Rondeline Williams*. An Open Science Approach to Experimental Psychology Methods. MIT Press.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology, 14*(2), 178–210.

Frazier, L., & Rayner, K. (1987). Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory and Language, 26*(5), 505.

Gagl, B., Golch, J., Hawelka, S., Sassenhagen, J., Poeppel, D., & Fiebach, C. J. (2018). Reading at the speed of speech: the rate of eye movements aligns with auditory language processing. *bioRxiv*, 391896.

Gautier, V., O'Regan, J. K., & LaGargasson, J. F. (2000). The skipping" revisited in French: Programming saccades to skip the article "les. *Vision Research, 40*, 2517–2531.

Gilchrist, I. (2011). Saccades. In *The Oxford handbook of eye movements*.

Gordon, P. C., Plummer, P., & Choi, W. (2013). See before you jump: Full recognition of parafoveal words precedes skips during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(2), 633.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498.

Heilbron, M., van Haren, J., Hagoort, P., & de Lange, F. P. (2023). Lexical processing strongly affects reading times but not skipping during natural reading. *Open Mind, 7*, 757–783.

Henderson, J. M., & Ferreira, F. (1990). Effects of foveal processing difficulty on the perceptual span in reading: Implications for attention and eye movement control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(3), 417.

Hirotani, M., Frazier, L., & Rayner, K. (2006). Punctuation and intonation effects on clause and sentence wrap-up: Evidence from eye movements. *Journal of Memory and Language, 54*, 425–443.

Hofmeister, J., Heller, D., & Radach, R. (1999). The return sweep in reading. In *Current oculomotor research* (pp. 349–357). Springer, Boston, MA.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures. OUP Oxford.*

Huang, K. J., & Staub, A. (2021). Using eye tracking to investigate failure to notice word transpositions in reading. *Cognition, 216*, 104846.

Hyönä, J., Lorch, R. F., Jr., & Kaakinen, J. K. (2002). Individual differences in reading to summarize expository text: Evidence from eye fixation patterns. *Journal of Educational Psychology, 94*(1), 44.

Hyönä, J., Lorch Jr, R. F., & Rinck, M. (2003). Eye movement measures to study global text processing. In *The mind's eye* (pp. 313–334). North-Holland.

Hyönä, J., & Nurminen, A. M. (2006). Do adult readers know how they read? Evidence from eye movement patterns and verbal reports. *British Journal of Psychology, 97*(1), 31–50.

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics, 40*, 431–439.

Janssen, D. P. (2012). Twice random, once mixed: Applying mixed models to simultaneously analyze random effects of language and participants. *Behavior Research Methods, 44*, 232–247.

Jared, D., Levy, B. A., & Rayner, K. (1999). The role of phonology in the activation of word meanings during reading: Evidence from proofreading and eye movements. *Journal of Experimental Psychology: General, 128*, 219–264.

Javal, L. E. (1878). Essai sur la physiologie de la lecture. *Annales D'oculistique, 80*, 61–73.

Jordan, T. R., McGowan, V. A., Kurtev, S., & Paterson, K. B. (2017). Investigating the effectiveness of spatial frequencies to the left and right of central vision during reading: Evidence from reading times and eye movements. *Frontiers in Psychology, 8*, 807.

Joseph, H. S. S. L., Liversedge, S. P., Blythe, H. I., White, S. J., & Rayner, K. (2009). Word length and landing position effects during reading in children and adults. *Vision Research, 49*, 2078–2086.

Juhasz, B. J., & Rayner, K. (2003). Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1312–1318.

Juhasz, B. J., & Rayner, K. (2006). The role of age of acquisition and word frequency in reading: Evidence from eye fixation durations. *Visual Cognition, 13*, 846–863.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*, 329–354.

Kaakinen, J. K., Hyönä, J., & Keenan, J. M. (2003). How prior knowledge, WMC, and relevance of information affect eye fixations in expository text. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(3), 447.

Kaakinen, J. K., & Hyönä, J. (2007). Perspective effects in repeated reading: An eye movement study. *Memory & Cognition, 35*, 1323–1336.

Kaakinen, J. K., & Hyönä, J. (2010). Task effects on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1561–1566.

Kliegl, R., & Engbert, R. (2005). Fixation durations before word skipping in reading. *Psychonomic Bulletin & Review, 12*(1), 132–138.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology, 16*, 262–284.

Kliegl, R., Hohenstein, S., Yan, M., & McDonald, S. A. (2013). How preview space/time translates into preview cost/benefit for fixation durations during reading. *Quarterly Journal of Experimental Psychology, 66*(3), 581–600.

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General, 135*(1), 12.

Konieczny, L. (1996). Human sentence processing: A semantics-oriented parsing approach. *PhD thesis, University of Freiburg, Freiburg, Germany*.

Konieczny, L., Hemforth, B., Scheepers, C., & Strube, G. (1995). PP-attachment in German: Results from eye movement studies. In Studies in Visual Information Processing (Vol. 6, pp. 405–420). North-Holland. Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition, 1(1).*

Kumle, L., Võ, M. L. H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods, 53*(6), 2528–2543.

Kuperman, V. (2022). A cross-linguistic study of spatial parameters of eye-movement control during reading. *Journal of Experimental Psychology: Human Perception and Performance, 48*(11), 1213.

Kuperman, V., Dambacher, M., Nuthmann, A., & Kliegl, R. (2010). The effect of word position on eye-movements in sentence and paragraph reading. *Quarterly Journal of Experimental Psychology, 63*(9), 1838–1857.

Kush, D & Dillon, B. (2020). Eye tracking and experimental syntax. In J. Sprouse (ed.) *Experimental Syntax.*

Leinenger, M. (2014). Phonological coding during reading. *Psychological Bulletin, 140*(6), 1534–1555.

Leinenger, M., & Rayner, K. (2013). Eye movements while reading biased homographs: Effects of prior encounter and biasing context on reducing the subordinate bias effect. *Journal of Cognitive Psychology, 25*(6), 665–681.

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences, 106*(50), 21086–21090.

Liversedge, S. P. (1994). *Referential context, relative clauses and syntactic parsing* (Doctoral dissertation, University of Dundee).

Liversedge, S. P., Drieghe, D., Li, X., Yan, G., Bai, X., & Hyönä, J. (2016). Universality in eye movements and reading: A trilingual investigation. *Cognition, 147*, 1–20.

Marx, C., Hawelka, S., Schuster, S., & Hutzler, F. (2015). An incremental boundary study on parafoveal preprocessing in children reading aloud: Parafoveal masks overestimate the preview benefit. *Journal of Cognitive Psychology, 27*(5), 549–561.

Matin, E. (1974). Saccadic suppression: A review and an analysis. *Psychological Bulletin, 81*(12), 899–917.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315.

McConkie, G. W., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics, 17*, 578–586.

McConkie, G. W., & Rayner, K. (1976). Asymmetry of the perceptual span in reading. *Bulletin of the Psychonomic Society, 8*(5), 365–368.

McConkie, G. W., Kerr, P. W., Reddix, M. D., & Zola, D. (1988). Eye movement control during reading: I. The location of initial eye fixations on words. Vision research, 28(10), 1107–1118.

McConkie, G. W., Kerr, P. W., Reddix, M. D., Zola, D., & Jacobs, A. M. (1989). Eye movement control during reading: II. Frequency of refixating a word. *Perception & Psychophysics*, *46*(3), 245–253.

Mertzen, D., Paape, D., Dillon, B., Engbert, R., & Vasishth, S. (2023). Syntactic and semantic interference in sentence comprehension: Support from English and German eye-tracking data. *Glossa Psycholinguistics, 2*(1).

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language, 112*, 104092.

Mitchell, D. C., Shen, X., Green, M. J., & Hodgson, T. L. (2008). Accounting for regressive eye-movements in models of sentence processing: A reappraisal of the Selective Reanalysis hypothesis. *Journal of Memory and Language, 59*(3), 266–293.

Meseguer, E., Carreiras, M., & Clifton, C. (2002). Overt reanalysis strategies and eye movements during the reading of mild garden path sentences. *Memory & Cognition, 30*(4), 551–561.

Morris, R. K. (1994). Lexical and message-level sentence context effects on fixation times in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(1), 92.

Morrison, R. E. (1984). Manipulation of stimulus onset delay in reading: Evidence for parallel programming of saccades. *Journal of Experimental Psychology: Human Perception and Performance, 10*(5), 667.

O'Regan, J. K. (1980). The control of saccade size and fixation duration in reading: The limits of linguistic control. *Perception & Psychophysics, 28*(2), 112–117.

O'Regan, J. K. (1990). Eye movements and reading. *Reviews of Oculomotor Research, 4*, 395–453.

O'Regan, J. K., & Jacobs, A. M. (1992). Optimal viewing position effect in word recognition: A challenge to current theory. *Journal of Experimental Psychology: Human Perception and Performance, 18*(1), 185.

O'Regan, J. K., Lévy-Schoen, A., Pynte, J., & Brugaillère, B. É. (1984). Convenient fixation location within isolated words of different length and structure. *Journal of Experimental Psychology: Human Perception and Performance, 10*(2), 250.

Parker, A. J., Räsänen, M., & Slattery, T. J. (2023). What is the optimal position of low-frequency words across line boundaries? An eye movement investigation. *Applied Cognitive Psychology, 37*(1), 161–173.

Payne, B. R., & Stine-Morrow, E. A. L. (2012). Aging, parafoveal preview, and semantic integration in sentence processing: Testing the cognitive workload of wrap-up. *Psychology and Aging, 27*(3), 638–649.

Perea, M., & Pollatsek, A. (1998). The effects of neighborhood frequency in reading and lexical decision. *Journal of Experimental Psychology: Human Perception and Performance, 24*(3), 767.

Pollatsek, A., Rayner, K., & Balota, D. A. (1986). Inferences about eye movement control from the perceptual span in reading. *Perception & Psychophysics, 40*(2), 123–130.

Pollatsek, A., Lesch, M., Morris, R. K., & Rayner, K. (1992). Phonological codes are used in integrating information across saccades in word identification and reading. *Journal of Experimental Psychology: Human perception and performance, 18*(1), 148.

Pollatsek, A., Reichle, E. D., & Rayner, K. (2006). Tests of the EZ Reader model: Exploring the interface between cognition and eye-movement control. *Cognitive Psychology, 52*(1), 1–56.

Rabe, M. M., Paape, D., Mertzen, D., Vasishth, S., & Engbert, R. (2024). SEAM: An integrated activation-coupled model of sentence processing and eye movements in reading. *Journal of Memory and Language, 135*, 104496.

Radach, R., & Kennedy, A. (2004). Theoretical perspectives on eye movements in reading: Past controversies, current issues, and an agenda for future research. *European Journal of Cognitive Psychology, 16*(1–2), 3–26.

Radach, R., & McConkie, G.W. (1998). Determinants of fixation positions in words during reading. in *Eye Guidance in Reading and Scene Perception,* 77–100.

Rayner, K. (1975). The perceptual span and peripheral cues in reading. *Cognitive Psychology, 7*(1), 65–81.

Rayner, K. (1979). Eye guidance in reading: Fixation locations within words. *Perception, 8*(1), 21–30.

Rayner, K. (1986). Eye movements and the perceptual span in beginning and skilled readers. *Journal of Experimental Child Psychology, 41*(2), 211–236.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372–422.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology, 62*(8), 1457–1506.

Rayner, K. (2014). The gaze-contingent moving window in reading: Development and review. *Visual Cognition, 22*(3–4), 242–258.

Rayner, K., & Bertera, J. H. (1979). Reading without a fovea. *Science, 206*(4417), 468–469.

Rayner, K., Cook, A. E., Juhasz, B. J., & Frazier, L. (2006). Immediate disambiguation of lexically ambiguous words during reading: Evidence from eye movements. *British Journal of Psychology, 97*(4), 467–482.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition, 14*(3), 191–201.

Rayner, K., & Fischer, M. H. (1996). Mindless reading revisited: Eye movements during reading and scanning are different. *Perception & Psychophysics, 58*(5), 734–747.

Rayner, K., Fischer, M. H., & Pollatsek, A. (1998a). Unspaced text interferes with both word identification and eye movement control. *Vision Research, 38*(8), 1129–1144.

Rayner, K., & Frazier, L. (1987). Parsing temporarily ambiguous complements. *The Quarterly Journal of Experimental Psychology, 39*(4), 657–673.

Rayner, K., Kambe, G., & Duffy, S. A. (2000). The effect of clause wrap-up on eye movements during reading. *The Quarterly Journal of Experimental Psychology Section A, 53*(4), 1061–1080.

Rayner, K., & Liversedge, S. P. (2011). Linguistic and cognitive influences on eye movements during reading', in Simon P. Liversedge, Iain Gilchrist, and Stefan Everling (eds), *The Oxford Handbook of Eye Movements*, Oxford Library of Psychology (2011)

Rayner, K., Liversedge, S. P., White, S. J., & Vergilino-Perez, D. (2003). Reading disappearing text: Cognitive control of eye movements. *Psychological Science, 14*(4), 385–388.

Rayner, K., & Morrison, R. E. (1981). Eye movements and identifying words in parafoveal vision. *Bulletin of the Psychonomic Society, 17*(3), 135–138.

Rayner, K., Pollatsek, A., & Binder, K. S. (1998b). Phonological codes and eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(2), 476.

Rayner, K., Schotter, E.R., Masson, M., Potter, M.C., & Treiman, R. (2016). So much to read, so little time: How do we read, and can speed reading help? *Psychological Science in the Public Interest, 17*, 4–34.

Rayner, K., Sereno, S. C., Morris, R. K., Schmauder, A. R., & Clifton Jr, C. (1989). Eye movements and on-line language comprehension processes. *Language and Cognitive Processes*, *4*(3–4), SI21–SI49.

Rayner, K., Sereno, S. C., & Raney, G. E. (1996). Eye movement control in reading: A comparison of two types of models. *Journal of Experimental Psychology: Human Perception and Performance, 22*(5), 1188–1200.

Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance, 37*(2), 514.

Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(6), 1290.

Reese, H. W. (1997). Counterbalancing and other uses of repeated-measures Latin-square designs: Analyses and interpretations. *Journal of Experimental Child Psychology, 64*(1), 137–158.

Reichle, E. D., & Drieghe, D. (2013). Using E-Z reader to examine word skipping during reading. *Journal of Experimental Psychology: Learning Memory and Cognition, 39*(4), 1311–1320. https://doi.org/10.1037/a0030910

Reichle, E. D., Liversedge, S. P., Pollatsek, A., & Rayner, K. (2009a). Encoding multiple words simultaneously in reading is implausible. *Trends in Cognitive Sciences, 13*(3), 115–119.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review, 105*(1), 125.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences, 26*(4), 445–476.

Reichle, E. D., Reineberg, A. E., & Schooler, J. W. (2010). Eye movements during mindless reading. *Psychological Science, 21*(9), 1300–1310.

Reichle, E. D., Warren, T., & McConnell, K. (2009b). Using EZ Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review, 16*(1), 1–21.

Reilly, R. G., & O'Regan, J. K. (1998). Eye movement control during reading: A simulation of some word-targeting strategies. *Vision Research, 38*(2), 303–317.

Reingold, E. M., Reichle, E. D., Glaholt, M. G., & Sheridan, H. (2012). Direct lexical control of eye movements in reading: Evidence from a survival analysis of fixation durations. *Cognitive Psychology, 65*(2), 177–206.

Reingold, E. M., & Sheridan, H. (2014). Estimating the divergence point: A novel distributional analysis procedure for determining the onset of the influence of experimental variables. *Frontiers in Psychology, 5*, 1432.

Reingold, E. M., Yang, J., & Rayner, K. (2010). The time course of word frequency and case alternation effects on fixation times in reading: Evidence for lexical control of eye movements. *Journal of Experimental Psychology: Human Perception and Performance, 36*(6), 1677.

Rosenbaum, P. R. (2020). The power of a sensitivity analysis and its limit. *Design of observational studies* (pp. 317–336). Springer International Publishing.

Salvucci, D. D. (2001). An integrated model of eye movements and visual encoding. *Cognitive Systems Research, 1*(4), 201–220. https://doi.org/10.1016/S1389-0417(00)00015-2

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language, 110*, 104038.

Schad, D. J., Betancourt, M., & Vasishth, S. (2022). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods, 26*(1), 103.

Schilling, H. E., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition, 26*(6), 1270–1281.

Schotter, E. R. (2018). Reading ahead by hedging our bets on seeing the future: Eye tracking and electrophysiology evidence for parafoveal lexical processing and saccadic control by partial word recognition. In *Psychology of learning and motivation* (vol. 68, pp. 263–298). Academic Press.

Schotter, E. R., & Rayner, K. (2015). The work of the eyes during reading. In *The Oxford Handbook of Reading* (p. 44). Oxford: Oxford University Press.

Schotter, E. R., & Jia, A. (2016). Semantic and plausibility preview benefit effects in English: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(12), 1839.

Schotter, E. R., & Leinenger, M. (2016). Reversed preview benefit effects: Forced fixations emphasize the importance of parafoveal vision for efficient reading. *Journal of Experimental Psychology: Human Perception and Performance, 42*(12), 2039.

Schotter, E. R., Angele, B., & Rayner, K. (2012). Parafoveal processing in reading. *Attention, Perception, & Psychophysics, 74*, 5–35.

Schotter, E. R., Bicknell, K., Howard, I., Levy, R., & Rayner, K. (2014a). Task effects reveal cognitive flexibility responding to frequency and predictability: Evidence from eye movements in reading and proofreading. *Cognition, 131*(1), 1–27.

Schotter, E. R., Tran, R., & Rayner, K. (2014b). Don't believe what you read (only once) comprehension is supported by regressions during reading. *Psychological Science, 25*(6), 1218–1226.

Schotter, E. R., von der Malsburg, T., Leinenger, M., & Schotter, E. R. (2018). Forced fixations, trans-saccadic integration, and word recognition: Evidence for a hybrid mechanism of saccade triggering in reading. *Journal of Experimental Psychology. Learning, Memory, and Cognition*.

Schotter, E.R., Stringer, C., Saunders, E., Cooley, F.G., Sinclair, G., & Emmorey, K. (2024). The role of perceptual and word identification spans in reading efficiency: Evidence from hearing and deaf readers. *Journal of Experimental Psychology: General,* in press.

Sereno, S. C., & O'donnell, P. J., & Rayner, K. (2006). Eye movements and lexical ambiguity resolution: Investigating the subordinate-bias effect. *Journal of Experimental Psychology: Human Perception and Performance, 32*(2), 335.

Sereno, S. C., & Rayner, K. (2000). Spelling-sound regularity effects on eye fixations in reading. *Perception & Psychophysics, 62*(2), 402–409.

Shain, C., & Schuler, W. (2021). Continuous-time deconvolutional regression for psycholinguistic modeling. *Cognition, 215*, 104735.

Siegelman, N., Schroeder, S., Acartürk, C., Ahn, H. D., Alexeeva, S., Amenta, S., ... & Kuperman, V. (2022). Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods, 54(6),* 2843–2863.

Slattery, T. J., Angele, B., & Rayner, K. (2011). Eye movements and display change detection during reading. *Human Perception and Performance, 37*, 1924–1938.

Slattery, T. J., & Yates, M. (2018). Word skipping: Effects of word length, predictability, spelling and reading skill. *Quarterly Journal of Experimental Psychology, 71*(1), 250–259.

Snell, J., van Leipsig, S., Grainger, J., & Meeter, M. (2018). OB1-reader: A model of word recognition and eye movements in text reading. *Psychological Review, 125*(6), 969.

Speer, S. R., & Clifton, C. (1998). Plausibility and argument structure in sentence comprehension. *Memory & Cognition, 26*(5), 965–978.

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass, 9*(8), 311–327.

Staub, A. (2021). How reliable are individual differences in eye movements in reading? *Journal of Memory and Language, 116*, 104190.

Staub, A., & Benatar, A. (2013). Individual differences in fixation duration distributions in reading. *Psychonomic Bulletin & Review, 20*(6), 1304–1311.

Staub, A., Rayner, K., Pollatsek, A., Hyönä, J., & Majewski, H. (2007). The time course of plausibility effects on eye movements in reading: Evidence from noun–noun compounds. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 1162.

Staub, A., White, S. J., Drieghe, D., Hollway, E. C., & Rayner, K. (2010). Distributional effects of word frequency on eye fixation durations. *Journal of Experimental Psychology: Human Perception and Performance, 36*(5), 1280.

Stringer, C., Cooley, F., Saunders, E., Emmorey, K., & Schotter, E. R. (2024). Deaf readers use leftward information to read more efficiently: Evidence from eye tracking. *Quarterly Journal of Experimental Psychology*, 17470218241232407.

Stowe, L. A., Kaan, E., Sabourin, L., & Taylor, R. C. (2018). The sentence wrap-up dogma. *Cognition, 176*, 232–247.

Strukelj, A., & Niehorster, D. C. (2018). One page of text: Eye movements during regular and thorough reading, skimming, and spell checking. *Journal of Eye Movement Research, 11(1).*

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly, 30*(4), 415–433.

Vasishth, S., & Engelmann, F. (2021). Sentence comprehension as a cognitive process: A computational approach. Cambridge University Press.

Vasilev, M. R., Adedeji, V. I., Laursen, C., Budka, M., & Slattery, T. J. (2021). Do readers use character information when programming return-sweep saccades? *Vision Research, 183*, 30–40.

Veldre, A., & Andrews, S. (2016). Is semantic preview benefit due to relatedness or plausibility? *Journal of Experimental Psychology: Human Perception and Performance, 42*, 939.

Veldre, A., & Andrews, S. (2017). Parafoveal preview benefit in sentence reading: Independent effects of plausibility and orthographic relatedness. *Psychonomic Bulletin & Review, 24*, 519–528.

Veldre, A., & Andrews, S. (2018a). Beyond cloze probability: Parafoveal processing of semantic and syntactic information during reading. *Journal of Memory and Language, 100*, 1–17.

Veldre, A., & Andrews, S. (2018b). How does foveal processing difficulty affect parafoveal processing during reading? *Journal of Memory and Language, 103*, 74–90.

Vergilino, D., & Beauvillain, C. (2000). The planning of refixation saccades in reading. *Vision Research, 40*(25), 3527–3538.

Vitu, F., & McConkie, G. W. (2000). Regressive saccades and word perception in adult reading. In *Reading as a perceptual process* (pp. 301–326).

Vitu, F., O'Regan, J. K., & Mittau, M. (1990). Optimal landing position in reading isolated words and continuous text. *Perception & Psychophysics, 47*(6), 583–600.

von der Malsburg, T., & Vasishth, S. (2011). What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language, 65*(2), 109–127.

von der Malsburg, T., & Angele, B. (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language, 94*, 119–133.

von der Malsburg, T., Vasishth, S., & Kliegl, R. (2012). Scanpaths in reading are informative about sentence processing. In *Proceedings of the first workshop on eye-tracking and natural language processing* (pp. 37–54).

von der Malsburg, T., Kliegl, R., & Vasishth, S. (2015). Determinants of scanpath regularity in reading. *Cognitive Science, 39*(7), 1675–1703.

Warren, T., & McConnell, K. (2007). Investigating effects of selectional restriction violations and plausibility violation severity on eye-movements in reading. *Psychonomic Bulletin & Review, 14*, 770–775.

Warren, T., McConnell, K., & Rayner, K. (2008). Effects of context on eye movements when reading about possible and impossible events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(4), 1001.

Weiss, A. F., Kretzschmar, F., Schlesewsky, M., Bornkessel-Schlesewsky, I., & Staub, A. (2018). Comprehension demands modulate re-reading, but not first-pass reading behavior. *Quarterly Journal of Experimental Psychology, 71*(1), 198–210.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General, 143*(5), 2020–2045.

Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started. *Trends in Hearing, 22*, 2331216518800869.

White, S. J., Warrington, K. L., McGowan, V. A., & Paterson, K. B. (2015). Eye movements during reading and topic scanning: Effects of word frequency. *Journal of Experimental Psychology: Human Perception and Performance, 41*(1), 233.

Wotschack, C., & Kliegl, R. (2013). Reading strategy modulates parafoveal-on-foveal effects in sentence reading. *Quarterly Journal of Experimental Psychology, 66*(3), 548–562.

Zhang, H., Miller, K., Cleveland, R., & Cortina, K. (2018). How listening to music affects reading: Evidence from eye tracking. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(11), 1778–1791.

Zola, D. (1984). Redundancy and word perception during reading. *Perception & Psychophysics, 36*(3), 277–284.