Behavior Research Methods (2023) 55:4419–4436
https://doi.org/10.3758/s13428-022-02025-w

FAB: A "Dummy's" program for self-paced forward and backward reading

Tianwei Gong[1] · Xuefei Gao[2,3,4] · Ting Jiang[1]

Accepted: 11 November 2022 / Published online: 22 March 2023

Abstract

The self-paced reading paradigm has been popular and widely used in psycholinguistic research for several decades. The tool described in this paper, FAB (Forward and Backward reading), is a tool created to hopefully and maximally reduce the coding demands and simplify the operation costs for experimental researchers and clinical researchers who are doing experimental work, in their designing, coding, implementing, and analyzing self-paced reading tasks. Its basis in web languages (HTML, JavaScript) also promotes experimental implementation and material sharing in our era of open science. In addition, FAB has a unique forward-and-backward mode that can track regressive-like behaviors that are usually only recordable using eye-tracking or mouse-tracking equipment. In this paper, the specific application and usage of FAB is demonstrated in one laboratory and two online validation experiments. We hope this free and open-sourced tool can benefit research in a diverse range of contexts where self-paced reading is desirable.

Keywords Self-paced reading · Eye tracking · Eye regression · Language processing · Online experiment · Open science

The self-paced reading paradigm (also known as the moving-window paradigm) has seen widespread use in psycholinguistic studies (Aaronson & Ferres, 1984; Hasher & Zacks, 1988; Just et al., 1982). In those studies, the reading materials are segmented into smaller parts (windows) in advance such that participants would read one window at a time followed by another, and pressing a button would reveal the next window and often simultaneously erase the preceding window. Researchers can decide how fine-grained the windows can be (e.g., at the level of the word, phrase, sentence, or any other pre-defined fragments) before collecting and analyzing the *dwelling times* participants spend on each window.

As an informative real-time measurement, self-paced reading has been used to investigate a wide range of issues in language comprehension including but not limited to the processing at the lexical (Acheson & MacDonald, 2011; MacDonald, 1993; Van der Schoot et al., 2009), syntactic (Gibson, 1998; Stowe, 1986; Trueswell et al., 1994), and discourse levels (Daneman & Carpenter, 1983; Graesser et al., 1994; Myers et al., 1987). Compared with the eye-tracking technique which also allows the reading process to be controlled by the participants (vs. the experimenters), self-paced reading is much less expensive and also easier to administer. Indeed, it imposes no apparatus-related requirements besides a computer (or a laptop), while also imposing no extra implementation-related requirements such as frequent calibration and validation. More importantly, self-paced reading has proven to be comparable to eye tracking in its capacity to capture various cognitive processes underlying language understanding (Just et al., 1982; see Keating & Jegerski, 2015; Mitchell, 2004 for review).

In this article, we introduce FAB with which researchers can design, code, conduct, and analyze their own self-paced reading tasks. Compared to previous similar toolkits such as PsyToolkit (https://www.psytoolkit.org/), Linger (http://tedlab.mit.edu/~dr/Linger/), Ibex Farm (Drummond, 2013), or SPaM (Luke & Christianson, 2013), FAB has four

✉ Xuefei Gao
xuefeigao@gmail.com

✉ Ting Jiang
psytingjiang@gmail.com

[1] Faculty of Psychology, Beijing Normal University, Beijing 100875, China

[2] School of Foreign Languages, Fuzhou University of International Studies and Trade, Fuzhou 350202, China

[3] CAS Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China

[4] Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China

unique features setting it apart. Firstly, as it is named, FAB has a special function for participants to go both *forward and backward* in reading. Eye tracking that collects data on regressive eye movements (i.e., regressions; see Staub & Rayner, 2007 for review) is recognized as having high ecological validity because it allows comprehenders to return to previous segments (regressions) or swiftly alternate between segments (saccades) in the manner of ordinary reading (Dehaene, 2009; Rayner et al., 2012). Those regressive measures are also used in recent mouse-tracking experiments (Lin & Lin, 2020; Schoemann et al., 2021) and touchscreen experiments (Hatfield, 2016). Importantly, Paape and Vasishth (2021, 2022) have implemented a *bidirectional self-paced reading paradigm*, which means participants can work through the sentences by pressing one key to move forward and another key to move backward. They showed that the regressive measures, including regression path duration used in our later validation experiments, can successfully capture the syntactic ambiguity effect in literature (Frazier & Rayner, 1982; Just & Carpenter, 1992). Therefore, we enable this bidirectional self-paced reading in FAB and allow it to record "regression-like" reading behaviors. Accordingly, FAB provides algorithms to automatically generate "eye-tracking-like" indices (Rayner, 1998, 2009) for both forwarding and regressing processes, such as first-pass reading time (gaze duration), regression path duration, rereading time, and probability of regression out/in (see below and Appendix A for details).

Secondly, FAB is an open-source tool developed using free web languages (JavaScript, HTML, CSS). There is no need to install any special software to run the FAB for task preparation, experiment conduction or data analysis, with the only tool required being a web browser. After each individual study, researchers on distributed sites can readily compile and upload their customized programs and data for other researchers to check or replicate. FAB's codes are also open for testing, revising, and updating continuously and sustainably. FAB follows the practices of open science (Open Science Collaboration, 2012; Towse et al., 2021), thereby potentially contributing to a more open and transparent research environment.

Thirdly, since FAB is based on web languages, it can be used to conduct online experiments. With the development of online crowdsourcing, researchers have the freedom to recruit participants via the Internet and allow targeted participants to finish the prescribed tasks on their own local devices. Online experiments are not only economical but also enable researchers to obtain a more diverse sample within a shorter time. Recent findings demonstrate that data obtained from online platforms such as Amazon's Mechanical Turk (MTurk), Qualtrics, are mostly in line with laboratory results, even for reaction-time measurements at the level of milliseconds (Crump et al., 2013; Zwaan & Pecher, 2012), including the self-paced reading paradigm itself (Enochson & Culbertson, 2015). Apart from replications, researchers have also conducted new psycholinguistic studies on online platforms (see Fine & Jaeger, 2016; Morgan & Levy, 2016; Villata et al., 2018, for example). By way of a direct contrast between lab vs. online experimentation, the present paper also reports two FAB experiments implemented online.

Fourthly, FAB is user-friendly for researchers who have no experience with programming. It provides user interfaces to customize its parameters, so there is no need for researchers to read or write any code at any point during the process. Using a parameterized interface design, FAB presents rich and detailed options for stimulus type, trial control, text layout, and comprehension question settings.

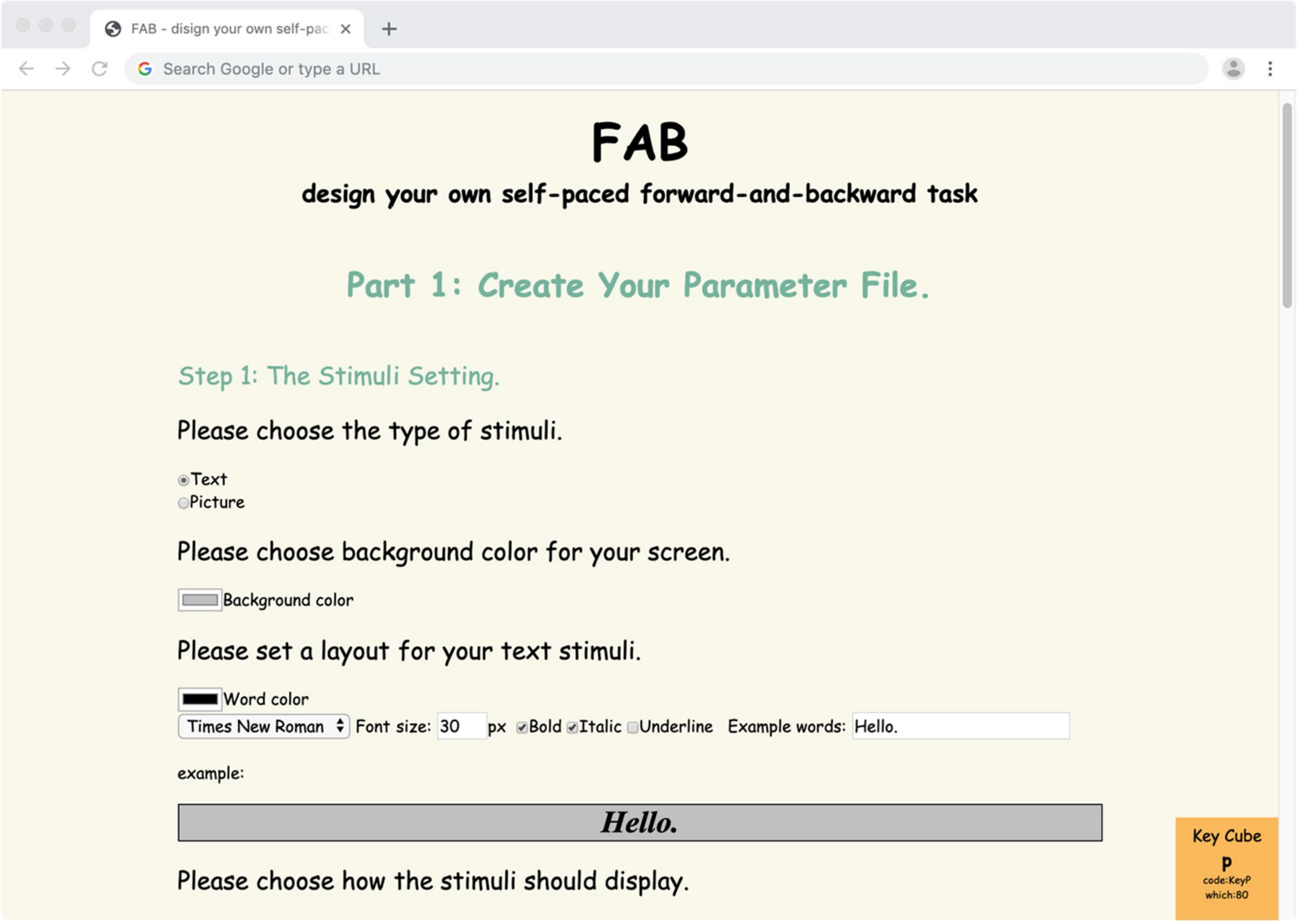
In the following sections, first there is an introduction to the usage of FAB in relation to task preparation, task running and data processing, following which three preregistered validation experiments are reported, with one laboratory task (Experiment 1) and two online tasks run on MTurk (Experiments 2 and 3).

Running FAB

FAB is based on JavaScript, HTML, and CSS and runs on one's web browser (e.g., Chrome, Firefox, Safari, Edge). It can be obtained from GitHub (https://github.com/tianweigong/fabreading/) or the Open Science Framework (https://osf.io/k3wjv/). A full user manual presenting instructions in a stepwise manner, and a tutorial that illustrates the details of how our three validation tasks were built are also available for reference. Users can further report errors, raise questions, or suggest improvements via email or the forum (i.e., the issue board on GitHub). Here, a basic and brief introduction is provided to task preparation, task running, and data processing.

Task preparation: FABdesign.html

To customize a self-paced reading experiment, researchers need to convey certain *parameters* to control their trials. FAB makes this as straightforward as filling out a form (Fig. 1). The first questions in the "form" relate to inherent characteristics of the self-paced reading, such as whether participants are allowed to go back to previous windows; designating the corresponding forward and backward keys; whether the stimuli are to be segmented by spaces (e.g., English), by characters (e.g., Chinese), or by larger or otherwise idiosyncratic chunks (e.g., at the level of the phrase); and whether the window should be fixed in the center of the screen or "moving" from left to right (see Fig. 2). FAB also contains the parameterized options for the layouts, such as the background color of the screen and the font, size, and



FAB

design your own self-paced forward-and-backward task

Part 1: Create Your Parameter File.

Step 1: The Stimuli Setting.

Please choose the type of stimuli.

● Text
○ Picture

Please choose background color for your screen.

▨ Background color

Please set a layout for your text stimuli.

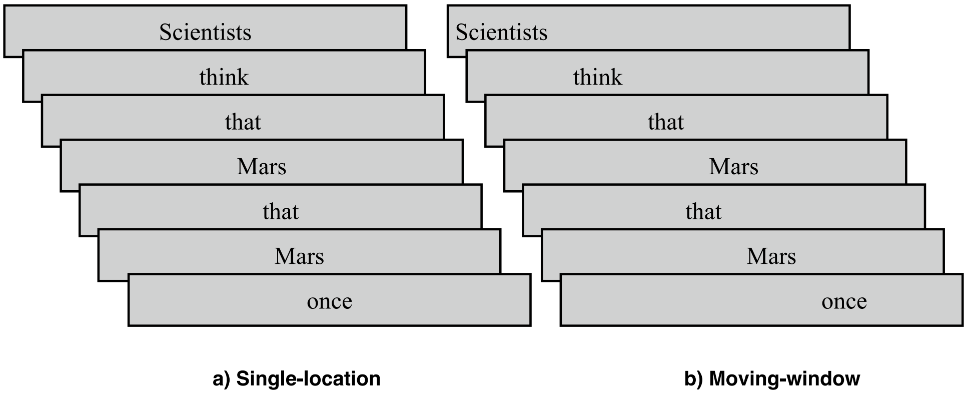
■ Word color
Times New Roman Font size: 30 px ☑ Bold ☑ Italic □ Underline Example words: Hello.

example:

Hello.

Please choose how the stimuli should display.

Fig. 1 A screenshot of FAB’s user interface



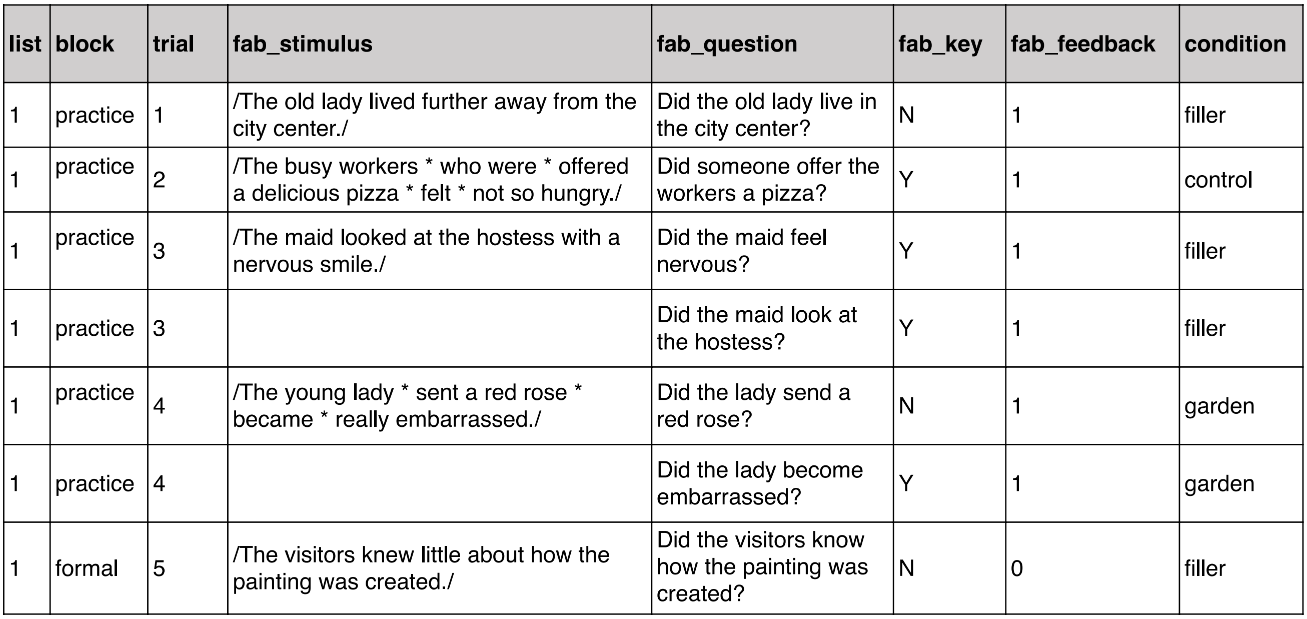
a) Single-location

b) Moving-window

Fig. 2 Two ways of stimulus displays in FAB

color of the stimuli. Researchers are also free to customize the instructions, questions, and feedback.

After confirming the parameters above, researchers can now deliver the *stimulus table*. Since psycholinguistic researchers generally prefer to prepare and counterbalance the reading materials on spreadsheets (e.g., Excel, Numbers), they are encouraged to make direct use of them. Users only need to rename the variables in spreadsheets using names (i.e., “fab_...”) that are system-reserved so that FAB can recognize the presented stimuli (“fab_stimulus”, which



list	block	trial	fab_stimulus	fab_question	fab_key	fab_feedback	condition
1	practice	1	/The old lady lived further away from the city center./	Did the old lady live in the city center?	N	1	filler
1	practice	2	/The busy workers * who were * offered a delicious pizza * felt * not so hungry./	Did someone offer the workers a pizza?	Y	1	control
1	practice	3	/The maid looked at the hostess with a nervous smile./	Did the maid feel nervous?	Y	1	filler
1	practice	3		Did the maid look at the hostess?	Y	1	filler
1	practice	4	/The young lady * sent a red rose * became * really embarrassed./	Did the lady send a red rose?	N	1	garden
1	practice	4		Did the lady become embarrassed?	Y	1	garden
1	formal	5	/The visitors knew little about how the painting was created./	Did the visitors know how the painting was created?	N	0	filler

Fig. 3 A demo of the stimulus list. The "fab_stimulus" column is mandatory. Other columns whose name began with "fab_..." would be referred to in the running program. Other customized columns (without the "fab_" prefix) might help the data analysis process. The interval between the stimulus and the question could be varied. Multiple comprehension questions for one stimulus is enabled by adding rows below

is necessary), questions ("fab_question"), correct answers ("fab_key"), etc. (Fig. 3). For each presentation stimulus, users can also set up interest areas by adding asterisks to delineate these, thereby guiding the later data processing algorithms. Some customizations for particular trials are allowed here through the addition of functional columns. For example, feedback could be specified to appear only for certain trials ("fab_feedback"). The interval between the stimulus and the question ("fab_qwait") could also be varied for different trials/conditions/levels, which may be convenient for typical within-participant designs (Fig. 3). Finally, users are free to add other columns that might be useful in later data analysis (e.g., the condition, the type of question, the origin of stimuli).

At this moment, some functions remain undeveloped in FAB. However, there are more specific means by which to realize these functions. For example, FAB cannot randomize (or pseudorandomize) the order of trials, nor can it deliver different stimuli to participants in a manner informed by the experimental conditions. However, it is easy for researchers to make different stimulus tables on their own and pack these up as different programs for assignment and implementation. At the same time, extra columns can be added to default columns to tag the condition or list ID for further data combination and integration. FAB is also intentionally built for smooth feature expansion and integration. For instance, if users seek to add multiple comprehension questions for one stimulus, they can append the extra questions in successive rows while leaving other cells empty. On the other hand, if users seek only to insert probe questions for certain trials, they can merely skip the cells of other trials without including probe questions. When it is ready, the spreadsheet can be converted to a JavaScript file by FAB and added to the main program folder together with the generated parameter file before running the task.

Task running: FABrunning.html

At the beginning of the task, researchers can input a subject ID, following which this ID will appear in the file name as well as in the dataset. During the task, participants can continuously press forward to reveal each window for reading, and as each new window appears, each preceding window disappears (Fig. 2). However, if researchers enable the "go-back" action, then, unless they have entered the comprehension question or the prompt (often a cross "+") for the next trial, participants can press the backward key to return to any given previous window(s) of that trial, one at a time, at any time during reading.

When running the task, FAB takes advantage of embedded features in browsers. If researchers seek to minimize disruption, they can activate the full-screen mode in the browser settings. If researchers plan to customize or enrich particular windows, they can add certain simple HTML or CSS codes into the stimulus cell (e.g., "<br>" for line break, "<b></b>" for bold). Similarly, special symbols can be added to the stimulus by writing corresponding HTML codes (e.g., "&#128077" for an emoji *thumbs-up* icon)[1]. The browser will convert the codes and display corresponding layouts or symbols automatically rather than bare nonsense codes to average readers.

At the end of the task, all subjects' datasets will be saved to the browser's default download folder. If either the researcher or the participant wishes to stop the task halfway through, then the Shift+Q keys can be pressed to quit and save the incomplete data.

Data processing: FABanalysis.html

The subject data would be saved as a CSV file, with each row representing a single window[2] and each column a dependent variable. All dependent variables are further classified into five clusters:

1. Gaze information: The gaze duration; the Unix time; the gaze's position in the gaze sequence of this trial.
2. Window information: The window's content; the window's position in the trial; the window's position in the area of interest.
3. Interest area information (if predefined): The area's position in the trial; the total number of windows in this area.
4. Trial information: The trial's position in the stimulus list; the total number of windows in the trial; all variables displayed in the stimulus table; the answer, accuracy, and response time with respect to the comprehension question.
5. Subject information: The subject's ID.

FAB maintains the output dataset such that it is as exhaustive and detailed as possible to help with further data extraction and analyses. Furthermore, it also provides algorithms to automatically generate "eye-tracking-like" indices at both the window level (predefined) and the areas of interest level (researcher specified) in the case that they are different: for instance, first-pass reading time (i.e., gaze duration), rereading time, total reading time, regression path duration, selected regression path duration, number of regression out/in, and probability of regression out/in. The definition and calculation of each index are presented in Appendix A. In FABanalysis.html, researchers can import the subject data files generated by FAB, choose the proposed indices, and then export the result file for further analyses.

Similar to data cleaning in eye tracking, researchers can define the valid duration (window dwelling time or gaze duration) range for a gaze such that FAB removes the outliers before calculating any further meaningful indices. The swift shift of eye fixations between words/areas is known as "saccades" in eye movement, and typically lasts between 20 and 40 ms (Staub & Rayner, 2007). As a result, previous eye-tracking studies have usually been conducted such that a lower boundary of 80 ms is set to clean the fixations (Rayner, 1998). To validate this result, Fig. 4 presents the gaze duration distributions in word-by-word moving window tasks from two current experiments, both of which follow multimodal distributions. In Experiment 1 (in-lab), the first peak had a very short time window of around 20–40 ms. This was mainly caused by people simply keeping a key pressed. In Experiment 2 (online), the first peak was not as obvious as in Experiment 1 but still visible as having around the same 20–40 ms time window. The behavior of keeping a key pressed may be due to motoric planning failures (e.g., "overshooting" or "undershooting", in the terminology of De la Peña et al., 2008) imposed by the paradigm or the desire to skip the current window and to move onto the next or previous one. They might also reflect shallow reading at the unconscious level (Myers & O'Brien, 1998). Therefore, to avoid this "sticky-key" issue and be consistent with previous eye-tracking studies, the lower boundary was set at 80 ms for data cleaning in our experiments. It should be noted that researchers are encouraged to test on their machines how long the duration would be if they simply hold a key down before conducting experiments and/or to plot the gaze distribution after data collection to ascertain the lower boundary for gaze duration as this can vary greatly from device to device.

On the other hand, the upper boundaries should be decided by researchers in a manner reflective of their experiences. Note that the gaze duration in FAB can be regarded as the lump sum of a collection of fixation durations in one area of interest before leaving that area in eye tracking. As a result, the upper boundary is supposed to be longer than that of a single fixation duration in eye movement (i.e., 800–1200 ms, Staub & Rayner, 2007).

Validation experiments

We conducted three validation experiments using different psycholinguistic materials, with one experiment conducted in the laboratory environment (Experiment 1) and two online

[1] Given that commas are the delimiters for CSV files by default, we currently require the user to instantly replace commas in their language materials with the HTML code for commas (&#44) before exporting the file. See the tutorial on the website for more information.

[2] We keep data format and terminology consistent with eye-tracking conventions as each row representing a single window, which is similar to a row representing an area of interest in the eye-tracking output file.

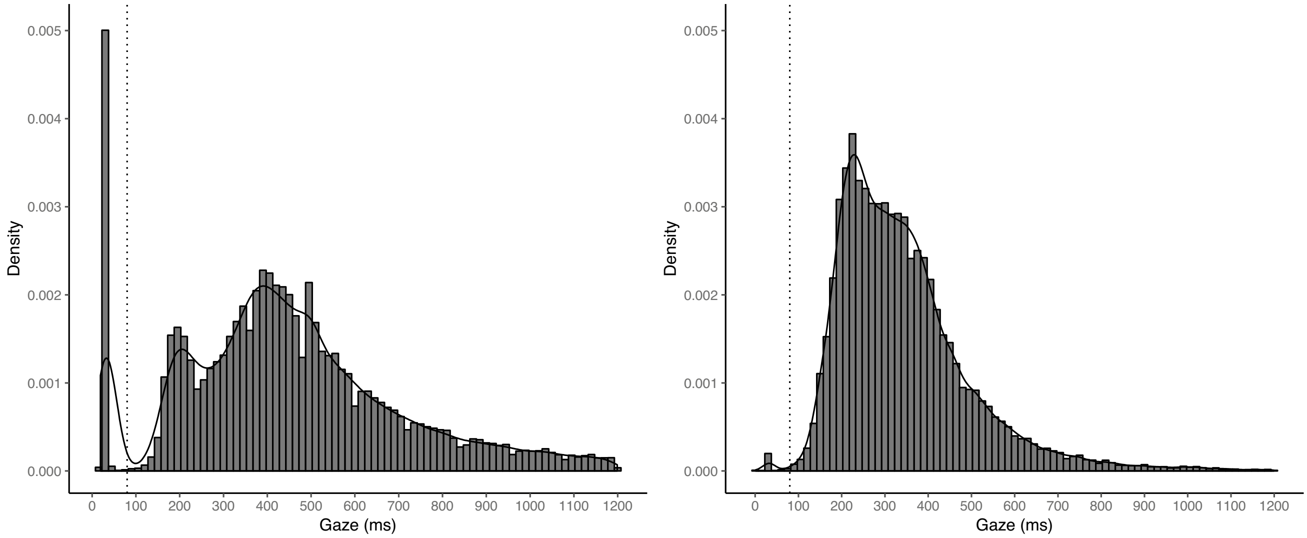


Fig. 4 The gaze distribution in Experiment 1 (*left*) and Experiment 2. The bin width was 15 ms. The *vertical dotted lines* were at 80 ms, a usual lower boundary for saccade cleaning in eye-tracking studies

(Experiments 2 and 3). Two experiments required word-by-word reading (Experiments 1 and 2) and the other one required sentence-by-sentence reading (Experiment 3).

The first experiment investigated sentences with temporary structural ambiguities, known as "garden-path sentences" (e.g., "The older kids showed all the dances felt amazed at the party.") When reading garden-path sentences, comprehenders' initial interpretation of certain components in sentences can later be shown to be false with the subsequent presence of the remaining (con)text. This "garden-path" effect is often measured as eye regressions when reading ambiguous sentences with eye tracking (Christianson et al., 2017; Frazier & Rayner, 1982; Kemper et al., 2004; Trueswell et al., 1994). Participants have been demonstrated to re-process ambiguous regions in garden-path sentences more often than unambiguous regions in controls, as well as giving fewer correct answers to comprehension questions (see Christianson et al., 2006; Ferreira & Henderson, 1991, for self-paced reading results). The processing difficulty of English garden-path sentences affects both English native speakers and second language learners (Chen & Xu, 2010; Juffs, 2004; Juffs & Harrington, 1996).

The second experiment was designed to investigate how the salience of syntactic boundaries could affect passage (re) processing and thus encourage regressive eye movements. Previous research has utilized both word-by-word self-paced reading and eye-tracking tasks to show a "*pay now or pay later*" effect in language processing: A salient marker (e.g., a period) in the middle of the passage can trigger an early wrap-up (Just et al., 1982) to further facilitate downstream sentence processing, while unmarked material triggers a delayed wrap-up at the end of the passage (Stine-Morrow et al., 2010).

The third experiment was designed to explore the relationship between causal relatedness and language processing. In previous sentence-by-sentence self-paced reading studies, reading time has been reported as correlating negatively with the degree of the causal connection between events in a text, with highly causally related events being recalled with greater ease than events that are causally related to a more tenuous extent (Keenan et al., 1984; Myers et al., 1987; see Solstad & Bott, 2017 for review).

Pay-now-or-pay-later and causal-relatedness were chosen because they represent well-established experimental phenomena that require reprocessing of textual materials at the discourse level. Meanwhile, the lengths of the two kinds of materials are similar enough so that they can serve as "fillers" for each other in different experiments. That is, participants in Experiments 2 and 3 would go through both materials via either word-by-word or sentence-by-sentence reading, while we will mainly analyze the pay-now-or-pay-later material in Experiment 2 and the causal-relatedness in Experiment 3, and put the results of filler materials in Appendix C.

In all three experiments, one offline measure (comprehension accuracy), one early measure (first-pass reading time) and five later measures of online processing (regression path duration, rereading time, total reading time, probability of regression out, and probability of regression in) were reported. Notably, we reported as many indices as possible only with the purpose of displaying the functions in FAB. Researchers usually do not need to report all these indices because there may be strong correlations between measures for the same region (e.g., first-pass reading time would be equal to regression path duration when there is no regression behavior), as well as for different regions (e.g., if two regions are adjacent, probability of regression in for one region would be identical to probability of regression out for the other). Accordingly, we only visualize the results of two indices (first-pass reading time and regression path duration).

Validation Experiment 1: Garden-path processing

Participants

Thirty-two Chinese college students (20 female, 12 male, mean age = 21.39 years old, SD = 1.99) were recruited from Beijing Normal University (BNU) and Beijing University of Posts and Telecommunications (BUPT) to complete the prescribed task in the behavioral lab. They were all Chinese native speakers and had learned English as a second language formally for at least 6 years. The experiment lasted around 35–45 minutes and participants were paid 35RMB for their participation. Informed consent and demographic information were obtained from all participants. For all three validation experiments, a criterion was used that participants who gained an overall level of accuracy below 60% would be considered having not adequately engaged with the task and their data would therefore be removed pre-analysis. No such data was removed in Experiment 1. The materials, data, and pre-registrations of the three validation experiments conducted are available at https://osf.io/sgk5v/.

Materials and procedure

Twenty sets of target sentences were adopted and adapted from previous research (Chen & Xu, 2010; MacDonald et al., 1992; Mason et al., 2003; Rayner et al., 1983):

a. Control (complete relative clause) sentences:

The older kids / who were / showed all the dances / felt / amazed at the party.
(Region 1 / Region 2 / Region 3 / Region 4 / Region 5)

b. Garden-path (reduced relative clause) sentences:

The older kids / showed all the dances / felt / amazed at the party.
(Region 1/ Region 3 / Region 4 / Region 5)[3]

The control sentences were divided into five regions and the garden-path sentences were divided into four regions (without Region 2) for subsequent analysis. Target sentences were counterbalanced across two experimental conditions (control versus garden-path) in two unique lists. Each participant saw only one version of each sentence item but saw all items in each condition. Twenty target items were presented along with 40 filler sentences of similar lengths and structures to the target sentences. The order of presentation was pseudorandomized within the list such that no more than two target items from the same condition were presented consecutively and no more than three items had the same comprehension answers (Yes/No) consecutively.

The task was run in Google Chrome and displayed using a 22-inch LCD monitor. Participants were seated approximately 50 cm from the monitor. The background color was pure black and all stimuli were displayed in pure white in Time New Romans 30 pt. Each trial began with a "+" fixation signal. Participants pressed ".>" to reveal the subsequent word and were also free to press ",<" to go back to the previous word. After the final word of each sentence, participants pressed one of two keys to respond to a yes/no comprehension question displayed in pure red (".>" was labeled as "yes" and ",<" was labeled as "no"). Although there were equal numbers of correct Yes and No responses to the comprehension questions, the majority of the target items (70%) had No responses (e.g., "Did the older kids show all the dances?" in the previous example) in order to detect whether participants had truly understood the target sentences by refuting the initial false interpretations (Christianson et al., 2017; Teubner-Rhodes et al., 2016). Accordingly, no feedback was given in order to reduce response strategies.

Before the formal reading task, participants finished six practice items, including one target sentence in each condition and four filler sentences. Feedback was given for the practice trials. Moreover, to ensure that the participants understood the garden-path structure, they were asked to interpret three or four sentences during the practice, including two target sentences and one or two filler sentences. After the reading task, participants completed Grammar Test 1 from the Oxford Placement Test (OPT, Allen, 1992, Chen & Xu, 2010), a self-reported English ability questionnaire (reading, listening, speaking, and writing, on 1–7 point Likert scales) and provided the age at which they began to learn English.

Results

The mean accuracy for the garden-path condition (mean = 59%, SD = 22%) was lower than that for the control condition (mean = 73%, SD = 18%), t (31) = 3.68, $p < .001$, $d = 0.70$, in mixed-effect logistic regression with subjects and items as random factors: $z = 4.04$, $p < .001$, successfully replicating the previous finding (Christianson et al.,

[3] Region 2 was intentionally omitted for direct comparisons between two conditions in other regions.

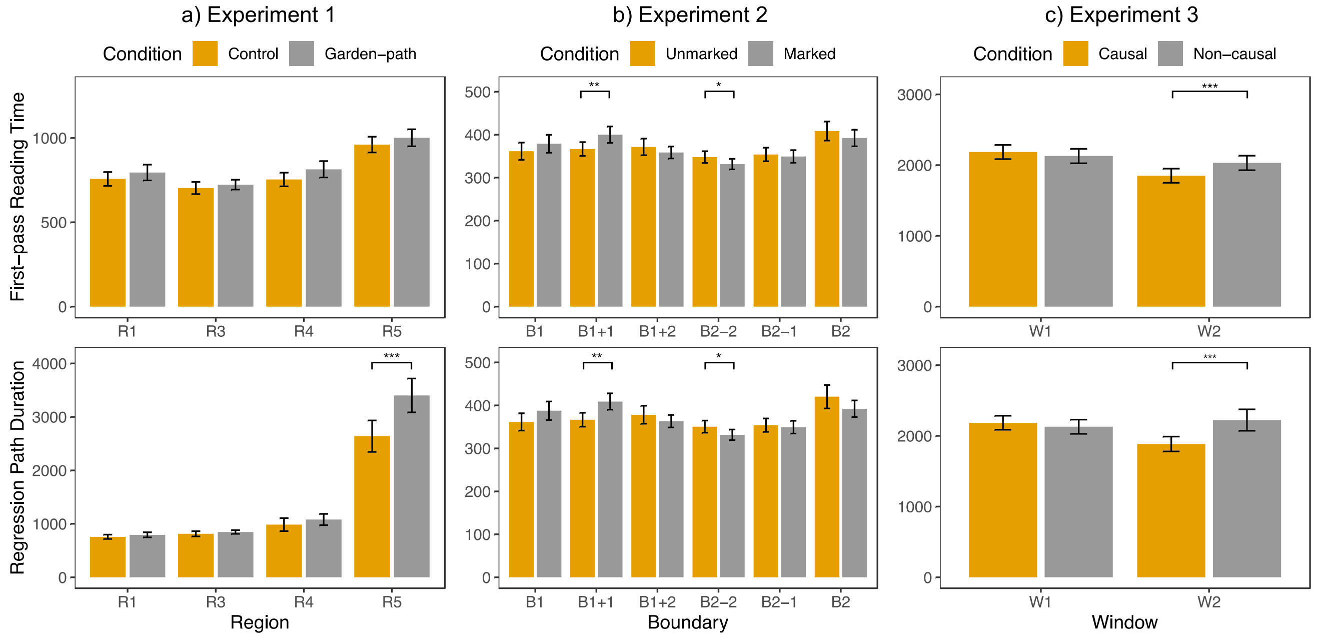


Fig. 5 Results of first-pass reading time and regression path duration in three experiments. *Error bars* indicate standard errors. Significances are based on linear-mixed effect analysis with subjects as well as items as random factors

2017). The online indices were calculated at the region (area of interest) level. Gazes shorter than 80 ms or longer than 5000 ms were ignored in the index calculation. As shown in the previous example, the garden-path sentences were segmented into four regions while the control sentences were segmented into five. Region 2 of the control condition was not compared with the garden-path condition. Again, we used a linear-mixed effect analysis with subjects (as well as items as random factors) to analyze the online indices across all three validation experiments. As shown in Fig. 5a, no difference appeared in first-pass reading time between the two conditions. However, there were differences between the two conditions in all later measures: the regression-path duration was longer for the garden path condition than for the control condition in the final region. For other subsequent indices, the processing deficiency of garden-path sentences was shown in all regions, as summarized in Table 1.

The results related to participants' English ability were shown in Appendix B. In general, participants with higher objective English ability scores were also observed to be subject to stronger garden-path effects and to have performed better in comprehension questions.

Validation Experiment 2: Pay Now or Pay Later

Experiment 2 and 3 were performed online. Online experiments can outperform lab experiments in relation to the speed at which participants can be recruited, sample diversity and representativeness, payment costs, etc. Although FAB has not yet been hosted on its own web server or online database [4], there are many ways for researchers to deploy FAB in the context of online experiments. One of the most straightforward ways is to use a form tool (e.g., Qualtrics, Google Forms) and an online storage tool (e.g., Google Drive, One Drive; researchers should select the tools based on their privacy concerns): Use the form to administer consents and collect demographic questions, include a link in the form for participants to download the experimental procedure, and finally include a file-upload question for participants to send their data in response. Researchers can also combine FAB's codes with other tools designed for online reaction-time experiments such as PsiTurk (Gureckis et al., 2016), albeit this may require programming skills utilized for web development.

Participants

Sixty participants (25 female, 34 male, one reported non-binary, mean age = 42.55 years old, SD = 13.24) were recruited via Amazon Mechanical Turk using Psiturk (Gureckis et al., 2016) and finished the task online. Only participants who lived in the US, had a Human Intelligence Task (HIT) approval rate of at least 95% and a minimum

[4] The main reason is that we are unable to develop a server that could appropriately store data for a range of users while also accommodating users' various privacy and ethical concerns.

number of 120 approved HITs were allowed to participate. They also confirmed that English was their native language. The experiment lasted from 8 to 12 min and participants were paid $1.20 for their participation. Two participants who were tested but were removed from the dataset: one for the overall accuracy was lower than 60% and the other for the data demonstrated a level of inattention during the later experiment (i.e., 66% of the gazes were shorter than 80 ms during the second half of trials with no reread process).

Materials and procedure

Twenty sets of marked or unmarked passages were adopted from Stine-Morrow et al. (2010), for example:

iii. marked:

A jack can be very useful to have in your trunk. / It can raise your car if you have a flat tire.

iv. unmarked:

A jack can be very useful to have in your trunk / by raising your car if you have a flat tire.

The word before the slash (i.e., “trunk”) was the first boundary marked by the punctuation “period” (as in sentence c) or simply unmarked (as in sentence d), and the final word was the second boundary (i.e., “tire”) marked by a “period” for both passage types. Stine-Morrow et al. (2010) found that participants in the marked condition engaged in early wrap-up (at Boundary 1) while participants in the unmarked condition engaged in late wrap-up (at Boundary 2).

We counterbalanced the experimental passages as well as the filler passages (see Experiment 3), resulting in four unique lists. Each participant saw only one version of each passage item but saw all the items in each condition. Their order of presentation was pseudorandomized within the list such that no more than two target passages from the same condition were presented consecutively and no more than three items had the same comprehension answer (Yes or No) consecutively.

Participants were only allowed to use their computers (i.e., not their mobile devices) to complete the task. The background color was pure black and all stimuli were displayed in pure white using Time New Romans, 30pt. The forward key was set as “P” and the backward key as “O”. Each trial began with a “Ready?” signal as well as a reminder of the two keys. After the final word of each material, participants pressed one of the two keys to respond to a Yes/No comprehension question displayed in pure red (“P” was labeled as Yes, and “O” as No). There were equal numbers of Yes/No responses to comprehension questions

Table 1 Important indices for each region in Experiment 1. The first four indices were averaged by number of words in each region

Region 1			Region 2			Region 3			Region 4			Region 5		
Control	Garden	Sig.	Control	Garden	Sig.	Control	Garden	Sig.	Control	Garden	Sig.	Control	Garden	Sig.
First-pass reading time (by word, in ms with SD in parentheses):														
757 (231)	795 (264)	.24	547 (106)	-	-	703 (202)	723 (166)	.46	754 (228)	814 (275)	.13	960 (265)	1001 (283)	.23
Regression path duration (by word, in ms):														
757 (231)	795 (264)	.24	635 (148)	-	-	814 (279)	847 (201)	.46	985 (691)	1081 (599)	.37	2640 (1667)	3402 (1789)	<.001
Rereading time (by word, in ms):														
334 (333)	493 (380)	<.001	406 (358)	-	-	441 (389)	711 (514)	<.001	483 (461)	857 (619)	<.001	352 (324)	589 (421)	<.001
Total Reading time (by word, in ms):														
1091 (459)	1288 (508)	<.001	954 (384)	-	-	1144 (440)	1434 (493)	<.001	1236 (552)	1671 (729)	<.001	1312 (499)	1590 (558)	<.001
Probability of regression out (%):														
-	-	-	10.2 (8.6)	-	-	11.9 (10.8)	16.4 (10.5)	<.001	11.2 (10.6)	17.1 (11.2)	<.001	24.4 (18.2)	30.9 (17.4)	<.001
Probability of regression in (%):														
23.3 (19.7)	31.6 (19.2)	<.001	12.8 (12.0)	-	-	10.9 (9.8)	17.0 (10.7)	<.001	9.8 (8.5)	15.0 (9.5)	<.001	-	-	-

Significance in linear-mixed effect analysis with subjects as well as items as random factors (e.g., "Can a jack be useful?"). Incorrect answers resulted in feedback being displayed during the practice as well as the formal task to prevent mind-wandering and mindless reading in semi-controlled online experiments. Before the formal experiment, participants finished five practice items, including one passage in each condition and three filler passages.

Results

For comprehension questions, no significant difference appeared with respect to the comprehension accuracy between the unmarked condition (mean = 88%, SD = 11%) and the marked condition (mean = 89%, SD = 11%), t (59) = 0.58, p = .563, mixed-effect logistic regression: z = 0.74, p = .461, similar to previous results (Stine-Morrow et al., 2010).

We reported the online indices for the first boundary (Boundary 1), the second boundary (Boundary 2), two words after the first boundary (Boundary 1+1 and Boundary 1+2, for potential spill-over effects) and two words before the second boundary (Boundary 2-2 and Boundary 2-1, for potential earlier wrap-up effects). Gazes shorter than 80 ms or longer than 5000 ms were ignored when calculating indices. The results of the online measures were shown in Table 2, with the results of first-pass reading time and regression path duration further visualized in Fig. 5b. It is noted that participants rarely returned to previous windows, resulting in a trivial probability of regression out, probability of regression in, and rereading time. Therefore, non-parametric Wilcoxon tests were conducted on the probability of regression out and regression in, and no statistical test was conducted on the rereading time. Participants did not differ in relation to whether they chose to go back in two conditions in all areas of interest (ps > .05).

After controlling for the subject and item as random factors (Table 2), the duration (first-pass reading time, regression path duration, total reading time) of Boundary 1 did not significantly differ between the marked and unmarked conditions. However, the duration of Boundary 1+1 was longer in marked than unmarked passages for first-pass reading time, regression path duration, and total reading time, while the duration of Boundary 2-2 was longer in unmarked than marked passages for all three indices. The inverse data patterns observed for the areas near Boundary 1 and Boundary 2 were predicted by the "pay now or pay later" effect.

We conducted a 2 (Location: Boundary 1, Boundary 2) × 2 (Boundary Salience: marked, unmarked) ANOVA analysis on regression path duration in accordance with Stine-Morrow et al. (2010). The main effect of Location was significant, F (1, 59) = 6.24, p = .015, η^2_p = .096, and the regression path duration was longer at the second boundary than the first one. The main effect of Boundary Salience was not significant, F (1,

Table 2 Important indices for each boundary of "pay now or pay later" material in Experiment 2 (word-by-word reading)

Boundary 1			Boundary 1+1			Boundary 1+2			Boundary 2-2			Boundary 2-1			Boundary 2		
Unmarked	Marked	Sig.	Unmarked	Marked	Sig.	Unmarked	Marked	Sig.	Unmarked	Marked	Sig.	Unmarked	Marked	Sig.	Unmarked	Marked	Sig.
First-pass reading time (in ms):																	
361 (155)	379 (161)	.21	367 (125)	400 (147)	<.01	372 (149)	358 (108)	.28	348 (107)	331 (95)	<.05	354 (122)	349 (114)	.69	408 (171)	392 (149)	.19
Regression path duration (in ms):																	
361 (155)	388 (167)	.08	367 (125)	410 (148)	<.01	378 (163)	363 (112)	.28	351 (109)	331 (95)	<.05	354 (122)	349 (114)	.69	420 (212)	392 (150)	.13
Rereading time (in ms):																	
4 (15)	3 (9)	-	5 (21)	2 (9)	-	4 (14)	3 (10)	-	1 (6)	0 (0)	-	2 (12)	0 (0)	-	1 (8)	0 (0)	-
Total reading time (in ms):																	
365 (155)	381 (160)	.25	372 (126)	402 (147)	<.01	375 (152)	361 (110)	.25	349 (107)	331 (95)	<.05	355 (123)	349 (114)	.61	409 (173)	392 (150)	.17
Probability of regression out (%):																	
0.11 (0.60)	0.37 (1.27)	-	0.37 (3.40)	0.17 (1.79)	-	0.60 (5.03)	0.17 (1.79)	-	0.14 (0.77)	0.00 (0.00)	-	0.06 (0.43)	0.00 (0.00)	-	0.17 (1.29)	0.00 (0.00)	-
Probability of regression in (%):																	
0.36 (1.54)	0.20 (0.93)	-	0.36 (1.47)	0.22 (0.99)	-	0.19 (0.87)	0.14 (0.77)	-	0.06 (0.43)	0.00 (0.00)	-	0.14 (1.08)	0.00 (0.00)	-	-	-	-

Significance in linear-mixed effect analysis with subjects as well as items as random factor

59) = 0.01, *p* = .926. However, there was a significant Location × Boundary Salience interaction, *F* (1, 59) = 5.01, *p* = .029, η^2_p = .078. Simple effect analysis showed that at Boundary 1, the regression path duration was longer for the marked condition than the unmarked condition, *F* (1, 59) = 4.25, *p* = .044. However, no significant difference could be observed as Boundary 2, *F* (1, 59) = 2.33, *p* = .132. For unmarked passages, the regression path duration was longer at Boundary 2 than at Boundary 1, *F* (1, 59) = 7.88, *p* = .007, while the boundary difference was not significant for marked passages.

By way of an exploratory analysis, the ANOVA analysis above was also replicated on the first-pass reading time and total reading time. For the first-pass reading time, the main effect of Location was significant, *F* (1, 59) = 10.46, *p* = .002, η^2_p = .151, with the regression path duration being longer at the second boundary than the first boundary. The main effect of Boundary Salience was not significant, *F* (1, 59) = 0.01, *p* = .926. The interaction effect was also (marginally) insignificant, *F* (1, 59) = 0.06, *p* = .056, η^2_p = .061. Similar to the total reading time, there was only a main effect of Location, *F* (1, 59) = 8.15, *p* = .006, η^2_p = .121, with a (marginally) insignificant effect of interaction, *F* (1, 59) = 3.90, *p* = .053, η^2_p = .062.

Validation Experiment 3: Causal relatedness

Participants

Sixty participants (29 female, 30 male, one reported non-binary, mean age = 36.42 years old, SD = 10.26) were recruited via Amazon Mechanical Turk using PsiTurk and engaged with the task online. The requirements and payment issued were identical to those detailed for Experiment 2. The appropriate command in PsiTurk was used to ensure that the participants did not also participate in Experiment 2. Seven additional participants were tested but removed from the dataset due to their overall accuracy being lower than 60%.

Materials and procedure

Twenty sets of causally (i.e., highest causally) or non-causally (i.e., lowest causally) related passages were adopted from Myers et al. (1987), for example:

e. causally related:

Jerry found typing errors in his manuscript. / The next day he yelled at his secretary.

f. non-causally related:

Jerry finished working on his manuscript. / The next day he yelled at his secretary.

Table 3 Important indices for each window of "causal relatedness" material in Experiment 3 (sentence-by-sentence reading)

Window 1			Window 2		
Causal	Non-causal	Sig.	Causal	Non-causal	Sig.
First-pass reading time (in ms):					
2185 (772)	2129 (787)	.22	1851 (774)	2032 (794)	<.001
Regression path duration (in ms):					
2185 (772)	2129 (787)	.22	1885 (815)	2223 (1174)	<.001
Rereading time (in ms):					
18 (58)	116 (339)	-	16 (51)	74 (198)	-
Total reading time (in ms):					
2203 (789)	2245 (957)	.45	1867 (795)	2106 (928)	<.001
Probability of regression out (%):					
-	-	-	0.75 (2.40)	2.69 (6.36)	-
Probability of regression in (%):					
0.75 (2.40)	2.69 (6.36)	-	-	-	-

Significance in linear-mixed effect analysis with subjects as well as items as random factors

According to Myers et al. (1987), participants' reading time should be identical at the first boundary for both conditions but longer at the second boundary when the sentences are non-causally related.

The procedure utilized was identical to that of Experiment 2, except that now each passage was segmented into only two windows according to the sentences (or boundaries for the filler materials, i.e., the pay-now-or-pay-later materials in Experiment 2). Participants pressed "P" to go forward and were free to press "O" to return to the first section after having progressed to the second.

Results

For comprehension questions, the accuracy observed for the causal condition (mean = 96%, SD = 6%) was higher than that of the non-causal condition (mean = 92%, SD = 9%), *t* (59) = 4.25, *p* < .001, *d* = 0.57, mixed-effect logistic regression: *z* = 3.15, *p* < .01.

We reported the online indices in each window. Gazes shorter than 80 ms or longer than 15000 ms were ignored when calculating indices. The results of online measures are shown in Table 3 with the first-pass reading time and the regression path duration visualized in Fig. 5c. It was rare for participants to return to previous sections. We also followed the method in Experiment 2 to conduct Wilcoxon tests on the probability of regression out and the probability of regression in. No significant difference between the two conditions was found for these two indices (*ps* > .05). For first-pass time, regression path duration, and total reading time, there was no temporal difference in the first window, while the duration in the second window was longer for the non-causal condition than for the causal condition across all three indices (Table 3).

Discussion

Experiment 1 replicated the previous results of garden-path processing difficulty with respect to both online regressive measures and offline comprehension measures (Christianson et al., 2017; Ferreira & Henderson, 1991; Paape & Vasishth, 2021, 2022; Von der Malsburg & Vasishth, 2011). Furthermore, the indices recorded and generated by FAB can reflect participants' individual differences with respect to their abilities in English, indicating that FAB has the ability to detect individual differences in online sentence processing.

Results from Experiment 2 demonstrated that the unmarked boundary buffered the early wrap-up, with participants "paying off" at the end of the passages. The marked boundary triggered early wrap-up and demonstrated significant downstream facilitation near the end of the passage (Boundary 2-2). Similar results showed up when participants were asked to read the materials sentence by sentence (Appendix C). Such results demonstrated the utility of FAB in exploring more complicated passage- or discourse- level language processing where spill-over effects and early wrap-up effects are not uncommon.

In Experiment 3, the negative correlation between causal relatedness and language processing difficulty was successfully replicated in not only the offline measure but also the later online indices. The later durations measured by different indices were longer for the non-causal condition than for the causal condition under sentence-by-sentence reading as well as word-by-word reading (Appendix C).

General discussion

In this paper, we have created and introduced FAB, an open-source browser-based tool for self-paced learning experiments. FAB provides a user-friendly interface for researchers seeking to prepare and conduct self-paced tasks. It also provides algorithms with which researchers can clean their data and generate eye-movement-like output in the form of various indices (Paape & Vasishth, 2021, 2022). FAB can be used on different operating systems since it is written using web languages. Researchers can implement FAB tasks using labs, shared online links, and crowdsourcing platforms.

To evaluate FAB, we conducted three validation experiments adopting three well-defined and well-established psycholinguistic paradigms, namely syntactic and discourse-level language processing, while allowing FAB to interface with laboratory and online environments, respectively. All the main results obtained replicated previous findings measured by eye-tracking equipment or traditional (forward) self-paced reading tasks.

Compared to the garden-path material considered in laboratory Experiment 1, regressive eye movements were less pronounced for online Experiments 2 and 3, and this discrepancy could be caused by inherent processing difficulties of garden-path sentences in Experiment 1, or the difference between online and laboratory environments. The factors undergirding such results are left for exploration in future studies, which might investigate the processing costs incurred for various types of comprehension materials, and perhaps the difference in motivation experienced by participants while engaging with online vs. laboratory-based reading tasks.

FAB was designed with the consideration of self-paced reading tasks in language studies. However, FAB also appears to be sufficiently flexible to be used in many other experimental contexts. For example, it can track the encoding processes of various stimuli that are sequential (e.g., pictures, icons, digits, items to be memorized) with both adults and children. Therefore, it could be applied to investigate scenarios including self-paced memory tasks (Bower & Clark, 1969; Hu et al., 2009; Tullis & Benjamin, 2011), self-paced vocabulary learning (De Jonge et al., 2015), and self-paced visual narrative comprehension (Klomberg & Cohn, 2022).

Being free, easy-to-install, instant code/data share, and user-friendly, FAB has the potential to pose as an alternative to the existing reaction-time-based psychological software, apps, and/or web resources for researchers who are not familiar with programming or who seek to conduct their (first) experiments in a laboratory or via the Internet. However, it should be pointed out that ready-to-use experimental software could also have a downside because the inner mechanism remains opaque to the average user. As such, we still encourage intermediate and advanced users to read the source code of FAB for more fine-tuned decision-making across their research process.

In the era of big data and citizen scientists (Crump et al., 2013; Open Science Collaboration, 2012; Towse et al., 2021), it is hoped that FAB can serve as a useful tool to make linguistic, cognitive, and psychological research more transparent, productive, reliable, and reproducible, and most of all, more accessible and affordable to all, by bringing scientists and average citizens together to co-create society-driven evidence-based research (Rayner et al., 2001).

Appendix 1

In this appendix, we introduce the calculations of all online indices in FAB based on an example of garden-path sentences (Appendix Fig. 6).

Gazes are defined as the single dwelling time in any given window. In this demo, we set the outlier as gazes below 80 ms or beyond 5000 ms. We firstly look at the indices in the window level shown in Appendix Fig. 6(1). *First-pass reading time* refers to the first gaze on the window (e.g., 344 ms for "kids", 1361 ms for "felt"). *Rereading time* refers to the sum of the duration of second and later gazes on the window (e.g., 485+207 = 692 ms for "kids", 1361+863 = 2224 ms for "felt"). *Total reading time* refers to the sum of the duration of all gazes on the window (e.g., 344 + 485 + 207 = 1036 ms for "kids" and 1361+863 = 2224 ms for "felt"). Total reading time is identical to the addition of the first-pass time and the rereading time.

Number of regression-out refers to how many times the comprehender goes back to the previous window from the present window (e.g., 1 for "kids", 1 for "felt"). In the second line of gaze sequences in Appendix Fig. 6, the gazes of "showed", "all", and "the" are outliers. Therefore, the regression path should be regarded as from "dances" to "kids" and from "kids" to "older". Accordingly, the number of regression-out is zero for "show", "all", "the". *Number of regression-in* refers to how many times the comprehender goes back in the present window from the later window (e.g., 1 for "kids", 0 for "felt"). Similarly, the number of regression-in "showed", "all", "the" is zero. *Probability of regression-out* refers to the quotient of the number of regression out and the total gaze numbers of the present window (e.g., 0.33 for "kids", 0.5 for "felt"). *Probability of regression-in* refers to the quotient of the number of regressions in and the total gaze numbers of the present window (e.g., 0.33 for "kids", 0 for "felt"). Because comprehenders cannot regress out of the first window or into the final window, the number/probability of regression out for the first window and the number/probability of regression in for the final window are both always zero.

Regression path duration, also called "go-past" time, refers to the time from first fixating the window to first moving past the window to the right, including time spent on rereading earlier windows. For "felt" in the example sentence, the regression path duration is 1361+499+485+712+207+287+440+663+704+863 = 6221. Since the comprehender in this example moves directly forward after the first gaze for other windows, the regression path duration for other windows is equal to the first-pass reading time. *Selected regression path duration* is a special index derived from the regression path duration. It removes time spent on rereading earlier windows from the regression path duration, and hence is 1361+863= 2224 for "felt" and is equal to the first-pass reading time for all other windows. Since the regression path does not exist for the first window, the regression path duration and the selected regression path duration are always equal to the first-pass reading time for the first window.

The previous measures are adopted from global eye-tracking measures. Besides, FAB also provides two exploratory measures. *Number of gazes* refers to how many valid gazes there are for the present window (e.g., 3 for "kids", 2 for "felt"). *Mean gaze duration* refers to the average duration of all gazes in the present window, equal to the total reading time divided by the number of gazes (e.g., 1036/3=345 ms for "kids", 2224/2= 1112 ms for "felt").

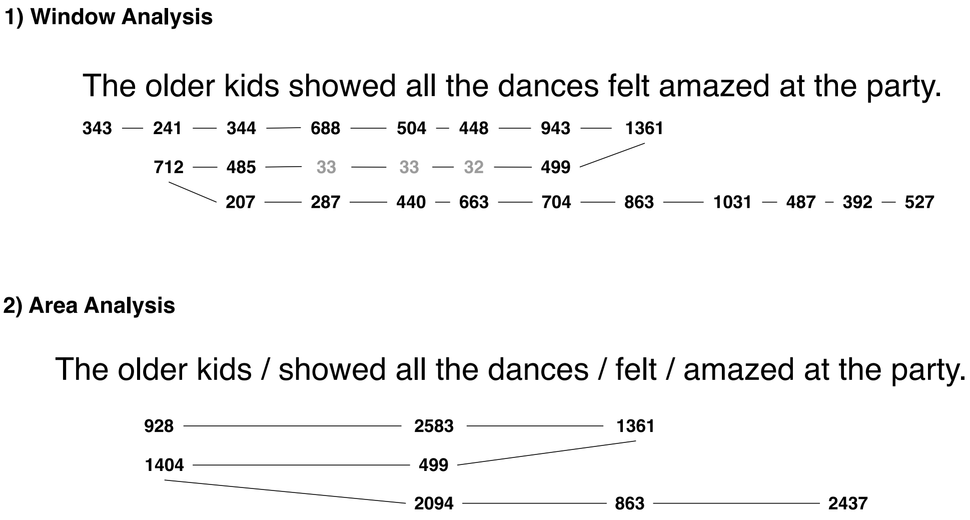


Fig. 6 The sequence of gazes at the window level (1) and the area level (2) for one sentence. Outliers were marked as *grey*

In order to conduct analysis on the area level, FAB would remove the gaze outliers and then merge all gazes by areas as shown in Appendix Fig. 6(2). After that, FAB treats areas as windows and calculates all indices following the same routines mentioned above.

Appendix 2

In this appendix we report the individual difference results in Experiment 1. The average accuracy for the Oxford Placement Test (OPT) was 78% (SD = 10%). The average starting age of learning English was 7.13 years old (SD = 2.74). Cronbach's alpha of self-reported English ability (reading, listening, speaking, and writing) was 0.70. One composite score of English ability converged from these four dimensions was added and it was then explored whether these individual differences correlated with participants' performance in the reading task. Alpha was set as 0.01 here to avoid the type-one error in multiple tests. For offline comprehension, the OPT score significantly correlated with garden-path accuracy ($r = .49$, $p = .005$) as well as the accuracy of control sentences ($r = .52$, $p = .002$) but did not correlate with the deviation between the accuracy of the two conditions. No English ability measure correlated significantly with the online measures in the two conditions. However, the OPT scores were able to predict the discrepancies between the control and garden-path conditions (control minus garden-path) in respect of several FAB dependent measures (Appendix Fig. 7): Region 5's rereading time ($r = .52$, $p = .002$), Region 5's probability of regression out ($r = .48$, $p = .005$), and Region 1's probability of regression in ($r = .51$, $p = .003$). The first two indices showed that OPT was positively correlated with the regression behavior. The third index indicated that participants with higher OPT preferred to reread the sentence from the beginning rather than selectively reanalyzing parts of it, which is similar to the regressive pattern found in native speakers (Paape & Vasishth, 2021, 2022). In sum, participants with higher objective English ability scores showed stronger garden-path effects and higher accuracy in comprehension questions.

Appendix 3

In this appendix we report the results of "filler" materials in Experiments 2 and 3. The target materials in Experiment 2 (pay-not-or-pay-later passages) were the fillers in Experiment 3 and the target materials in Experiment 3 (causal-relatedness passages) were the fillers in Experiment 2.

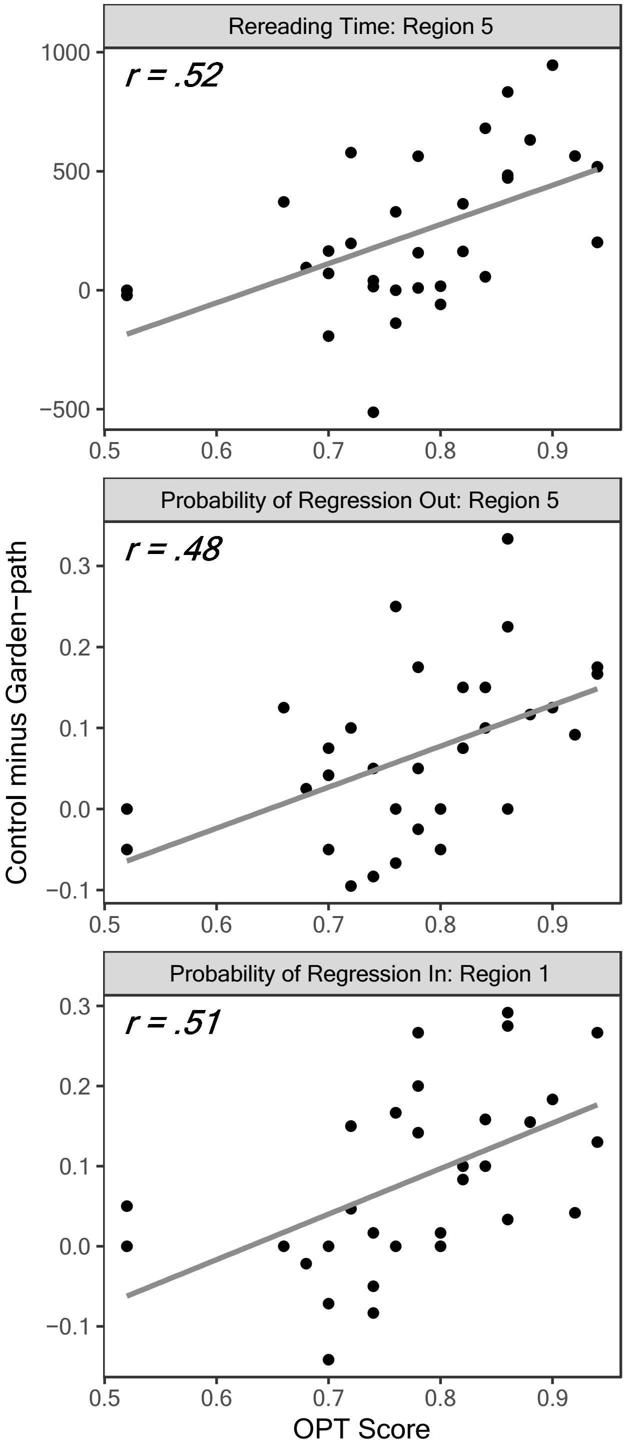


Fig. 7 The correlation between the Oxford Placement Test (OPT) score and self-paced reading indices. We also classified participants into high and low performance groups according to their OPT scores and showed the raw indices here to indicate potential interaction orientations. In the low OPT group, control vs. garden-path sentence: RRT (R5) = 435 ± 410 vs. 502 ± 424 (ms); PRO (R5) = 27 ± 22 vs. 29 ± 20 (%); PRI (R5) = 27 ± 24 vs. 29 ± 23 (%). In the high OPT group, control vs. garden-path sentence: RRT (R5) = 278 ± 210 vs. 666 ± 416 (ms); PRO (R5) = 22 ± 14 vs. 33 ± 15 (%); PRI (R5) = 20 ± 14 vs. 34 ± 16 (%).

Table 4 Important indices for each boundary of causal relatedness material in Experiment 2 (word-by-word reading)

Boundary 1			Boundary 1+1			Boundary 1+2			Boundary 2-2			Boundary 2-1			Boundary 2		
Causal	Non-causal	Sig.	Causal	Non-causal	Sig.	Causal	Non-causal	Sig.	Causal	Non-causal	Sig.	Causal	Non-causal	Sig.	Causal	Non-causal	Sig.
First-pass reading time (in ms):																	
370 (170)	366 (180)	.74	405 (166)	388 (130)	.24	349 (120)	344 (108)	.59	353 (106)	374 (140)	.06	347 (94)	358 (112)	.14	414 (196)	442 (217)	.10
Regression path duration (in ms):																	
370 (170)	369 (186)	.93	406 (166)	388 (130)	.21	349 (120)	374 (261)	.43	362 (128)	384 (164)	.24	350 (97)	366 (119)	.09	434 (296)	526 (422)	<.05
Rereading time (in ms):																	
3 (13)	16 (87)	-	3 (15)	14 (62)	-	6 (24)	14 (37)	-	3 (12)	12 (28)	-	2 (12)	10 (23)	-	1 (6)	4 (12)	-
Total reading time (in ms):																	
373 (169)	382 (216)	.60	408 (169)	402 (152)	.73	355 (127)	357 (119)	.84	355 (107)	386 (146)	<.05	349 (95)	368 (118)	<.05	415 (198)	446 (222)	.07
Probability of regression out (%):																	
0.11 (0.86)	0.05 (2.31)	-	0.22 (2.06)	0.48 (2.89)	-	0.25 (1.98)	0.67 (3.67)	-	0.25 (1.14)	0.68 (1.84)	-	0.19 (1.07)	0.92 (1.77)	-	0.17 (1.29)	0.67 (1.71)	-
Probability of regression in (%):																	
0.19 (1.07)	0.46 (1.77)	-	0.28 (1.63)	0.63 (2.09)	-	0.39 (1.61)	0.78 (1.83)	-	0.17 (0.96)	0.68 (1.65)	-	0.11 (0.86)	0.44 (1.14)	-	-	-	-

Significance in linear-mixed effect analysis with subjects as well as items as random factors

Table 5 Important indices for each window of "pay now or pay later" material in Experiment 3 (sentence-by-sentence reading)

Window 1			Window 2		
Unmarked	marked	Sig.	Unmarked	marked	Sig.
First-pass reading time (in ms):					
2233 (893)	2239 (984)	.49	2035 (783)	1894 (703)	<.001
Regression path duration (in ms):					
2233 (893)	2239 (984)	.49	2122 (909)	1982 (778)	<.01
Rereading time (in ms):					
41 (130)	51 (125)	-	45 (141)	37 (88)	-
Total reading time (in ms):					
2275 (939)	2290 (1014)	.41	2080 (843)	1931 (731)	<.001
Probability of regression out (%):					
-	-	-	1.14 (3.51)	1.56 (3.90)	-
Probability of regression in (%):					
1.14 (3.51)	1.56 (3.90)	-	-	-	-

Significance in linear-mixed effect analysis with subjects as well as items as random factors

Causal relatedness in Experiment 2 The accuracy for the causal condition (mean = 97%, SD = 9%) was higher than that for the non-causal condition (mean = 93%, SD = 10%), t (59) = 3.66, $p < .001$, $d = 0.35$, mixed-effect logistic regression: $z = 2.80$, $p < .01$. Gazes shorter than 80 ms or longer than 5000 ms were ignored when calculating indices. The results of the online measures are given in Appendix Table 4. The method utilized was that detailed in the main text, analyzing the regions near (±1 and ±2) or at the two boundaries. More participants went out (i.e., probability of regression out) of the Boundary 1+2 backwards (Wilcoxon test, $p = .037$), Boundary 2-1 ($p = .013$), and Boundary 2 ($p = .011$) in the non-causal condition than in the causal condition. More participants went back (i.e., probability of regression in) into Boundary 2-2 ($p = .013$) and Boundary 2-1 ($p = .011$) in the non-causal condition than in the causal condition. The duration indices were longer in the non-causal condition than the causal condition for the regression path duration of Boundary 2, and the total reading time of Boundary 2-2 and Boundary 2-1 (Appendix Table 4).

Pay-now-or-pay-later in Experiment 3 No difference was found between the comprehension accuracy in the unmarked condition (mean = 88%, SD = 12%) or the marked condition (mean = 87%, SD = 10%), t (59) = 0.65, $p = .519$, mixed-effect logistic regression: $z = 0.41$, $p = .681$. The results of the online measures were shown in Appendix Table 5. Gazes shorter than 80 ms or longer than 15000 ms were ignored when calculating indices. Similar to the causal relatedness material, no significant difference was found in the probability of regression out/in, nor any temporal difference in the first window. However, the first-pass reading time, regression path duration, and total reading time were all longer for unmarked material than for marked material in the second window.

Acknowledgments The authors would like to thank Baoguo Chen, Yaohan Li, and Guanyu Peng for their contribution to the language material in Experiment 1, and Junruoyu Zheng as an awesome programming tutor.

References

Aaronson, D., & Ferres, S. (1984). The word-by-word reading paradigm: An experimental and theoretical approach. In D. E. Kieras & M. A. Just (Eds.), *New methods in reading comprehension research* (pp. 31–68). Erlbaum.

Acheson, D. J., & MacDonald, M. C. (2011). The rhymes that the reader perused confused the meaning: Phonological effects during on-line sentence comprehension. *Journal of Memory and Language, 65*(2), 193–207.

Allen, D. (1992). *Oxford placement test*. Oxford: Oxford University Press.

Bower, G. H., & Clark, M. C. (1969). Narrative stories as mediators for serial learning. *Psychonomic Science, 14*(4), 181–182.

Chen, B. G., & Xu, H. H. (2010). Influence of working memory capacity on processing English temporary syntactic ambiguity sentences for Chinese–English bilinguals. *Acta Psychologica Sinica, 42*(2), 185–192.

Christianson, K., Williams, C. C., Zacks, R. T., & Ferreira, F. (2006). Younger and older adults'" good-enough" interpretations of garden-path sentences. *Discourse Processes, 42*(2), 205–238.

Christianson, K., Luke, S. G., Hussey, E. K., & Wochna, K. L. (2017). Why reread? Evidence from garden-path and local coherence structures. *The Quarterly Journal of Experimental Psychology, 70*(7), 1380–1405.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One, 8*(3), e57410.

Daneman, M., & Carpenter, P. A. (1983). Individual differences in integrating information between and within sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*(4), 561–584.

De Jonge, M., Tabbers, H. K., Pecher, D., Jang, Y., & Zeelenberg, R. (2015). The efficacy of self-paced study in multitrial learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 851–858.

De la Peña, D., Murray, N. P., & Janelle, C. M. (2008). Implicit overcompensation: The influence of negative self-instructions on performance of a self-paced motor task. *Journal of Sports Sciences, 26*(12), 1323–1331.

Dehaene, S. (2009). *Reading in the brain*. New York: Viking.

Drummond, A. (2013). *Ibex Farm*. Available at: http://spellout.net/ibexfarm/. Accessed 22 Aug 2020

Enochson, K., & Culbertson, J. (2015). Collecting psycholinguistic response time data using Amazon Mechanical Turk. *PloS One, 10*(3), e0116946.

Ferreira, F., & Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language, 30*(6), 725–745.

Fine, A. B., & Jaeger, T. F. (2016). The role of verb repetition in cumulative structural priming in comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(9), 1362–1376.

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.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68*(1), 1–76.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*(3), 371–395.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., … Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods, 48*(3), 829–842.

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory* (pp. 193–225). Academic Press.

Hatfield, H. (2016). Self-guided reading: Touch-based measures of syntactic processing. *Journal of Psycholinguistic Research, 45*(1), 121–141.

Hu, Y., Ericsson, K. A., Yang, D., & Lu, C. (2009). Superior self-paced memorization of digits in spite of a normal digit span: The structure of a memoirist's skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1426.

Juffs, A. (2004). Representation, processing and working memory in a second language. *Transactions of the Philological Society, 102*(2), 199–225.

Juffs, A., & Harrington, M. (1996). Garden path sentences and error data in second language sentence processing. *Language Learning, 46*(2), 283–323.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122–149.

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General, 111*(2), 228–238.

Keating, G. D., & Jegerski, J. (2015). Experimental designs in sentence processing research: A methodological review and user's guide. *Studies in Second Language Acquisition, 37*(1), 1–32.

Keenan, J. M., Baillet, S. D., & Brown, P. (1984). The effects of causal cohesion on comprehension and memory. *Journal of Verbal Learning and Verbal Behavior, 23*(2), 115–126.

Kemper, S., Crow, A., & Kemtes, K. (2004). Eye-fixation patterns of high-and low-span young and older adults: Down the garden path and back again. *Psychology and Aging, 19*(1), 157–170.

Klomberg, B., & Cohn, N. (2022). Picture perfect peaks: Comprehension of inferential techniques in visual narratives. *Language and Cognition, 14*(4), 596–621.

Lin, Y. C., & Lin, P. Y. (2020). Reading minds in motion: Mouse tracking reveals transposed-character effects in Chinese compound word recognition. *Applied Psycholinguistics, 41*(4), 727–751.

Luke, S. G., & Christianson, K. (2013). SPaM: A combined self-paced reading and masked-priming paradigm. *Behavior Research Methods, 45*(1), 143–150.

MacDonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language, 32*(5), 692–715.

MacDonald, M. C., Just, M. A., & Carpenter, P. A. (1992). Working memory constraints on the processing of syntactic ambiguity. *Cognitive Psychology, 24*(1), 56–98.

Mason, R. A., Just, M. A., Keller, T. A., & Carpenter, P. A. (2003). Ambiguity in the brain: What brain imaging reveals about the processing of syntactically ambiguous sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(6), 1319–1338.

Mitchell, D. C. (2004). On-line methods in language processing: introduction and historical review. In M. Carreiras & C. E. Clifton (Eds.), *The On-line Study of Sentence Comprehension: Eyetracking, ERP and Beyond* (pp. 15–32). Psychology Press.

Morgan, E., & Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition, 157*, 384–402.

Myers, J. L., & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes, 26*(2–3), 131–157.

Myers, J. L., Shinjo, M., & Duffy, S. A. (1987). Degree of causal relatedness and memory. *Journal of Memory and Language, 26*(4), 453–465.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*(6), 657–660.

Paape, D., & Vasishth, S. (2021). *Conscious rereading is confirmatory: Evidence from bidirectional self-paced reading*. PsyArXiv. https://doi.org/10.5070/G6011182

Paape, D., & Vasishth, S. (2022). Is reanalysis selective when regressions are consciously controlled. *Glossa Psycholinguistics, 1*(1), 1–34.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 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., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of Verbal Learning and Verbal Behavior, 22*(3), 358–374.

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest, 2*(2), 31–74.

Rayner, K., Pollatsek, A., Ashby, J., & Clifton, C., Jr. (2012). *The psychology of reading*. London: Psychology Press.

Schoemann, M., O'Hora, D., Dale, R., & Scherbaum, S. (2021). Using mouse cursor tracking to investigate online cognition: Preserving methodological ingenuity while moving toward reproducible science. *Psychonomic Bulletin & Review, 28*(3), 766–787.

Solstad, T., & Bott, O. (2017). Causality and causal reasoning in natural language. In M. R. Waldmann (Ed.), *The Oxford Handbook of Causal Reasoning* (pp. 619–644). Oxford University Press.

Staub, A., & Rayner, K. (2007). Eye movements and on-line comprehension processes. In G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 327–342). Oxford: Oxford University Press.

Stine-Morrow, E. A., Shake, M. C., Miles, J. R., Lee, K., Gao, X., & McConkie, G. (2010). Pay now or pay later: Aging and the role of boundary salience in self-regulation of conceptual integration in sentence processing. *Psychology and Aging, 25*(1), 168.

Stowe, L. A. (1986). Parsing WH-constructions: Evidence for online gap location. *Language and Cognitive Processes, 1*(3), 227–245.

Teubner-Rhodes, S. E., Mishler, A., Corbett, R., Andreu, L., Sanz-Torrent, M., Trueswell, J. C., & Novick, J. M. (2016). The effects of bilingualism on conflict monitoring, cognitive control, and garden-path recovery. *Cognition, 150*, 213–231.

Towse, J. N., Ellis, D. A., & Towse, A. S. (2021). Opening Pandora's Box: Peeking inside Psychology's data sharing practices, and seven recommendations for change. *Behavior Research Methods, 53*(4), 1455–1468.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language, 33*(3), 285–318.

Tullis, J. G., & Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *Journal of Memory and Language, 64*(2), 109–118.

Van der Schoot, M., Vasbinder, A. L., Horsley, T. M., Reijntjes, A., & van Lieshout, E. C. (2009). Lexical ambiguity resolution in good and poor comprehenders: An eye fixation and self-paced reading study in primary school children. *Journal of Educational Psychology, 101*(1), 21–36.

Villata, S., Tabor, W., & Franck, J. (2018). Encoding and retrieval interference in sentence comprehension: Evidence from agreement. *Frontiers in Psychology, 9*, 2.

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

Zwaan, R. A., & Pecher, D. (2012). Revisiting mental simulation in language comprehension: Six replication attempts. *PloS One, 7*(12), e51382.

Open practices statements The data and materials for all experiments are available at and are preregistered with Open Science Framework (https://osf.io/sgk5v/).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.