Bilingualism: Language and Cognition

cambridge.org/bil

Conducting psycholinguistic research online: Comparable evidence of second language lexical and sentence-level processing in web-based versus lab-based studies

Freya Gastmann[1,2], Sarah Schimke[1] and Gregory J. Poarch[2]

[1]Institute of German Philology, Ludwig-Maximilians-Universität München, Munich, Germany and [2]Department of English Language & Culture, Center for Language & Cognition, University of Groningen, Groningen, The Netherlands

Research Article

Cite this article: Gastmann, F., Schimke, S. and Poarch, G.J. (2026). Conducting psycholinguistic research online: Comparable evidence of second language lexical and sentence-level processing in web-based versus lab-based studies. *Bilingualism: Language and Cognition* **29**, 615–628. https://doi.org/10.1017/S136672892510028X

Received: 19 September 2024
Revised: 10 May 2025
Accepted: 20 May 2025
First published online: 18 July 2025

Keywords:
psycholinguistics; web-based data collection; language comprehension; second language processing

Corresponding author:
Freya Gastmann;
Email: f.gastmann@lmu.de

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

Abstract

Although web-based data collection has become increasingly popular in (linguistic) research over the past years, many researchers are still cautious about collecting data via the internet. Thus, this study aims at comparing web-based and lab-based testing of linguistic manipulations that have resulted in robust findings in previous lab-based research on bilingual language processing. A total of 134 L1 German students of L2 English participated in two experiments in a web-based (n = 78) or lab-based setting (n = 56). The study examined potential language co-activation through cognates in an English Lexical Decision Task (Experiment 1) and the use of L2 lexical and syntactic information in English relative clause processing in a Self-paced Reading Task (Experiment 2). We found comparable evidence of lexical and syntactic processing in both groups in both experiments. Critically, this paper provides important methodological implications for web-based data collections with second language learners.

Highlights

- Adult L2 learners were tested either in the lab or in a web-based setting.
- Both groups performed a lexical decision task and a self-paced reading task.
- Effects of word order and plausibility were found in L2 sentence comprehension.
- Cognate effects emerged neither in isolation nor in L2 sentence context.
- Results were comparable across both the lab-based and web-based groups.

1. Introduction

Since the COVID-19 pandemic, web-based data collection has experienced an upsurge in linguistic research. Due to social distancing restrictions and other safety precautions during the pandemic, conducting (psycho)linguistic experiments in person had become almost impossible for a period of time. Therefore, more and more researchers had resorted to not only recruiting but also testing participants via the internet, and since then, the body of research on web-based data collection has grown substantially (see e.g., Gagné & Franzen, 2023; Rodd, 2024; Sauter et al., 2020, for introductions to online behavioral data collection). Not only has the pandemic made it indispensable for many researchers to resort to online methods, but the possibility of collecting data via the web has also made accessing participants more convenient and enabled testing more diverse samples. Thus, remote testing facilitated reaching participants from wider geographical contexts or rather understudied populations (Garcia et al., 2022), and it allowed for data collection in less time compared to in-person laboratory studies. However, fundamental questions have emerged about the reliability and validity of web-based exploration of linguistic processes in behavioral experiments. To this day, many researchers still treat online data collection with caution and view the lack of control during data collection with skepticism (Sauter et al., 2022). Thus, the quality of web-based data collection has been frequently questioned in terms of technical disparities and human factors possibly affecting experimental outcomes. Technical influencing factors may include differences in hardware, such as variation in response time measurements depending on the keyboard (Neath et al., 2011) or the use of different triggering devices, such as touchscreen versus keyboard (Pronk et al., 2020). In addition to differences in hardware, software can likewise have an impact on the course and outcome of behavioral experiments. In a mega-study by Bridges et al. (2020), the comparability of response time measurements with ten different experimental software packages in both online and offline settings was assessed. The authors found that web-based technologies provided a slightly higher variability and thus less precise measures of response times than lab-based systems. Nonetheless, the authors argue that online data collections can still be suitable, particularly when comparing

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

participants' response times across conditions. Similarly, Reimers and Stewart (2015) concluded that different computer systems and browser plug-ins can operate similarly regarding the detection of response time differences across conditions, and potential small effects on measured response times vanish with using within-subject designs. Assessing five widely common cognitive experimental paradigms, such as the Stroop and Flanker task, Semmelmann and Weigelt (2017) compared three environmental settings: traditional lab-based testing, using web technology in the lab and regular web-based data collection. The authors included the "web in lab" setting to account for potential differences between settings induced by the use of (web) technology rather than the environmental setting. Overall, the authors found no difference in error rates regardless of the setting. However, web-based technology caused a timing offset of about 37 ms, which may be ascribed to factors inherent to JavaScript. Nevertheless, despite the timing offset, general task-specific effects were replicated in all three settings – except for the priming paradigm, which was not replicable in any of the settings (for a discussion on the absence of a priming effect, see Semmelmann & Weigelt, 2017). To sum up, potential confounding factors induced by differences in hardware and/or software might be of only little importance depending on the research objective and can be controlled for by focusing on differences in participants' response times between conditions (for similar conclusions, see Anwyl-Irvine et al., 2021; Pronk et al., 2020).

Besides technical influences, human factors can similarly have an impact on web-based data collection. One main aspect that distinguishes online data collection from in-person testing in the lab is the reduced level of control in the web-based format. With web-based testing, it is more difficult to minimize or eliminate potential confounding factors such as environmental noise, to monitor the participants' concentration, and – depending on the nature of the task – to prevent participants from using external resources during task completion. To investigate such potential issues, Germine et al. (2012) examined the comparability of data quality of web versus lab samples with a wide range of cognitive and perceptual tests. Their results contradict the previously widespread assumption of increased performance variability and measurement noise in web samples compared to lab samples that was assumed to be induced by a potential lack of focus and motivation in unsupervised settings: Overall, no differences in mean test performance, performance variance and internal measurement reliability between web and lab sample were observed for most measures. Although differences between the groups were evident in mean performances for two tests tapping into aspects of general intelligence, the authors argue that these differences were not systematic (i.e., mean performance differences were observed in both directions). Hence, the authors suggest these differences are derived from sample-inherent characteristics rather than general data quality. Consequently, their findings suggest that testing unsupervised participants online does not necessarily need to impede data quality but that cognitive processes can be mapped similarly well instead (for similar results, see De Leeuw & Motz, 2016, for visual search; Kochari, 2019, for numerical cognition; Miller et al., 2018, for cognitive paradigms; Weydmann et al., 2023, for reinforcement learning).

In line with the general trend toward an increase in use of online data collection methods for behavioral research, a growing number of psycholinguistic researchers have addressed the topic of web-based testing compared to in-person, lab-based data collection as well. Similarly, the broad consensus in this field of research is that online testing can serve as a suitable means for psycholinguistic data collection – provided that its use is always scrutinized against the backdrop of the respective research question and objective. Recent studies investigating well-established effects on first language (L1) word recognition have found that response time effects obtained in lab-based experiments could be replicated in web-based settings, including effects of word frequency (Hilbig, 2016) and emotionality (Kim et al., 2023). Hilbig (2016) examined the word frequency effect in a lexical decision experiment in three different contexts (similar to Semmelmann & Weigelt, 2017): a classical lab setting with standard experimental software, a lab setting with browser-based software, and a web setting with browser-based software. The author found a large effect of word frequency in reaction times for all three contexts, with no main effect of context or significant interaction of the two parameters. This demonstrates that robust response time effects found in previous lab-based research can be replicated using web-based technology. More recently, Kim et al. (2023) investigated the effect of emotion words on lexical decision times in a web-based and a lab-based setting. The authors found faster response times for positive and negative compared to neutral words in both settings. Similar to Hilbig (2016), there was no evidence of a difference between settings. Thus, the findings of both studies add to the previous literature, providing evidence of the comparability of web- and lab-based reaction time measurements.

The use of web-based technologies for behavioral research has not only increased in L1 research but also in research on second language (L2) processing (e.g., Berger et al., 2019, for L2 lexical processing; Klassen et al., 2022, for L2 syntactic processing; Tiffin-Richards, 2024, for L1 and L2 lexical and syntactic processing). However, fewer L2 studies have directly compared outcomes of web-based data collection relative to in-person lab-based testing. While there may be little reason to assume that comparing these two settings in L2 research will produce fundamentally different results compared to L1 research, findings obtained with native speakers should not automatically be generalized to L2 learners. Importantly, certain learner-intrinsic factors that impact the acquisition and processing of languages, such as target language proficiency and affective factors like motivation or anxiety, are generally assumed to be constant across L1 speakers but may differ widely across L2 learners (Ellis, 2004). Due to this greater variability, L2 learners may be more easily affected by external factors such as the general setting – a factor that is more difficult to control for in web-based settings. Additionally, regarding participant recruitment, there may be greater variability, particularly in L2 proficiency, in online L2 learner recruitment via platforms like MTurk (Amazon Mechanical Turk; Buhrmester et al., 2011) compared to online testing of a preselected group of participants, such as university students, possibly due to a larger diversity in first languages or a wider range of L2 proficiencies on participant platforms. Based on this assumption, an L2 self-paced reading study by Patterson and Nicklin (2023) compared syntactic processing of proper versus common nouns in sentence context across three samples: a crowdsourced population tested online, a student population tested online, and in-person data collected from students in a previous study conducted in the lab. Their findings provide evidence for the overall replicability of lab-based outcomes in both online settings. However, the authors found higher L2 proficiency in their crowdsourced sample and drew attention to the overall poorer controllability of proficiency on crowdsourcing platforms. Consequently, they point out that the method of participant recruitment should always be considered with regard to the research objectives.

https://doi.org/10.1017/S136672892510028X Published online by Cambridge University Press

1.1. The present study

Against the backdrop of the rather sparse literature on comparability of behavioral data in L2 research, the present study aims to fill this gap by focusing on the direct comparison of web-based and lab-based testing[1] in two linguistic domains, namely L2 lexical and sentence processing. For this purpose, several well-established linguistic effects on both the lexical level (i.e., the word/nonword effect and the cognate facilitation effect) and sentence level (i.e., effects of word order and plausibility as well as cognate effects in sentence context) will be further examined.

With regard to L2 lexical processing, the present study aims to replicate two robust psycholinguistic findings concerning the lexicon. On the one hand, the aim is to demonstrate the well-established word/nonword effect (expressed through a delay in processing of nonwords; e.g., Stanners et al., 1975) in both experimental settings using a Lexical Decision Task (LDT; Experiment 1) in the participants' L2. Furthermore, Experiment 1 aims at replicating language co-activation through cognates, which has been extensively demonstrated in lab-based research on lexical processing in second language learners. Cognate words are translation equivalents that share meaning and similar/identical form across languages (e.g., English *banana*, German *Banane*). There is ample evidence that cognates are processed faster and more accurately by bilinguals than noncognates, which are translation equivalents without such form overlap (e.g., English *pumpkin*, German *Kürbis*). Such a processing advantage of cognate over noncognate words (the so-called cognate facilitation effect) is considered evidence for co-activation of languages and language nonselective access in bilingual speakers (Dijkstra et al., 2010).

Furthermore, the present study examines L2 sentence processing in both settings. Experiment 2 aims at replicating three known effects in L2 sentence comprehension through self-paced reading (SPR): (1) lexical co-activation in sentence context, (2) the effect of word order and (3) the effect of world knowledge (plausibility). Building on Experiment 1, Experiment 2 investigates cognate processing in L2 sentence context to assess lexical co-activation during sentence processing (e.g., Hopp, 2017; Miller, 2014; Tiffin-Richards, 2024). Previous studies have shown that word processing can differ depending on the context in which words are presented (e.g., isolation versus sentential context; Dirix et al., 2019; Lauro & Schwartz, 2017). The additional language context in sentences, bearing semantic and/or syntactic constraints, may affect or even mitigate possible lexical effects during reading (Tiffin-Richards, 2024). Thus, it is warranted to more closely inspect the processing of lexical items not only out of context but also within context.

Furthermore, the SPR task examines the processing of noncanonical word orders by testing the comprehension of English subject relative clauses (SRC; (1)) compared to object relative clauses (ORC; (2)).

(1) *There is a dog that chases a bird.*
(2) *There is a bird that a dog chases.*

Previous research has found ample evidence that canonical SRC structures are processed faster and more accurately than noncanonical ORCs in both native and non-native speakers of English (Lau & Tanaka, 2021; Lim & Christianson, 2013). A reason for processing disadvantages for English ORCs might be their greater structural complexity due to a greater distance between filler and gap position. This complexity may cause higher working memory demands compared to less complex SRCs (Gibson, 1998). For L1 German learners of L2 English, there is an additional difficulty: The German parse for English ORCs is ambiguous. Whereas in English, SRCs (1) and ORCs (2) are disambiguated by word order, in German, relative clause (RC) structures do not differ in verb position but are instead disambiguated via case marking for masculine nouns (SRC: 3, ORC: 5) or remain ambiguous for feminine and neuter nouns (7). Thus, these learners will have to use their L2 syntactic knowledge to parse successfully.

(3) *Da ist ein Hund, der$_{NOM}$ einen$_{ACC}$ Vogel jagt.*
There is a dog that$_{NOM}$ a$_{ACC}$ bird chases
There is a dog that chases a bird.

(4) **Da ist ein Hund, den$_{ACC}$ ein$_{NOM}$ Vogel jagt.*
*There is a dog that$_{ACC}$ a$_{NOM}$ bird chases
**There is a dog that a bird chases.*

(5) *Da ist ein Vogel, den$_{ACC}$ ein$_{NOM}$ Hund jagt.*
There is a bird that$_{ACC}$ a$_{NOM}$ dog chases
There is a bird that a dog chases.

(6) **Da ist ein Vogel, der$_{NOM}$ einen$_{ACC}$ Hund jagt.*
*There is a bird that$_{NOM}$ a$_{ACC}$ dog chases
**There is a bird that chases a dog.*

(7) Da ist eine Katze, die eine Maus jagt.
There is a cat that$_{NOM/ACC}$ a$_{NOM/ACC}$ mouse chases
There is a cat that chases a mouse. or *There is a cat that a mouse chases.*

Additionally, Experiment 2 examines the impact of plausibility on L2 sentence processing. Previous research has shown that sentences that are not consistent with our world knowledge (see examples 4 and 6; asterisks indicate implausibility) are harder to process and more likely to be misinterpreted (Ferreira, 2003; Lim & Christianson, 2013).

The present study will investigate a) the replicability of the aforementioned lexical and sentence-level effects and b) the comparability of web-based and lab-based findings. We ask the following research questions:

RQ1: Can the effects of word status (word/nonword) and cognate status (cognate/non cognate) during L2 lexical processing in isolation be replicated in web-based and lab-based settings? Does processing in the two settings differ?

RQ2: Can the cognate effect during lexical processing in L2 sentence context be replicated in web-based and lab-based settings? Does processing in the two settings differ?

RQ3: Can the effects of word order and plausibility during L2 sentence processing be replicated in web-based and lab-based settings? Does processing in the two settings differ?

These research questions will be addressed by administering a Lexical Decision Task (Experiment 1) and a Self-paced Reading

[1]In the following, the terms *web-based experiment*/*online testing* and *lab-based experiment*/*in-person testing* are used synonymously. Web-based testing refers to remote testing via the internet/a web browser, which explicitly does not take place in the laboratory. Lab-based testing refers to the more traditional testing in the laboratory without the use of internet/browser technologies.

https://doi.org/10.1017/S136672892510028X Published online by Cambridge University Press

Table 1. Participant characteristics (n = 106)

	Online participants (n = 54)			Lab participants (n = 52)		
	M	*SD*	Range	*M*	*SD*	Range
Age (years)	22.6	2.1	19.4–28.3	22.8	2.9	18.9–33.6
Age of English acquisition (years)	8.2	1.8	3.0–12.8	8.0	1.9	3.0–12.4
Length of immersion in an English study program at university (years)	2.6	2.2	0.0–9.1	2.5	1.9	0.6–7.7
English use[a]	5.9	1.2	3.2–8.6	5.8	1.3	2.6–8.6
GE LexTALE	88.1	5.9	72.83–97.5	87.3	6.2	68.8–98.8
EN LexTALE	83.7	9.0	62.5–98.8	81.9	9.6	60.0–97.5
Flanker effect (in ms)[b]	76	6.6	63–88	71	6.6	52–82

[a]English use was aggregated across five situations (with family, with friends, at university, using media such as movies and reading, on social media) on a 10-point rank scale (ranging from 1 = *I do not use English in these situations at all.* to 10 = *I exclusively use English in these situations.*)
[b]Flanker effect as a measure of inhibitory control, calculated by subtracting congruent from incongruent reaction times in the Flanker task.

Task (Experiment 2) in the participants' L2. Both experiments assess decision accuracies as a measure of offline comprehension and response times as a measure of processing speed. We predict the replication of the previously mentioned well-established effects in both settings. Regarding the comparability of findings in the online versus laboratory setting, the present study's exploratory aim is to test whether the absence of differences between settings observed in previous studies on L1 processing can be replicated in a population of bilingual speakers.

2. Experiment 1 – Lexical decision task

2.1. Participants

A total of 134 L1 German L2 English speakers participated online (n = 78) or in the lab (n = 56). All subjects were recruited from a population of students of English linguistics at TU Dortmund University. In the web-based experiment, 13 participants were excluded from further analyses due to missing data and/or technical issues. Additionally, participants were excluded because they were L1 English (2), not L1 German (1), did not study English (7) or had an eye disease (1) that impeded their participation. In the lab-based experiment, participants were excluded due to missing data (1) or previous participation in the web-based study (3). Thus, 54 online participants (43 female, 10 male, 1 nonbinary) and 52 lab participants (41 female, 8 male, nonbinary) remained for further analyses. All participants filled in a background questionnaire assessing their first language(s) and further language learning history, as well as their social background (based on the Language and Social Background Questionnaire, LSBQ; Anderson et al., 2018). Thirteen participants from each group reported having another L1 in addition to German. Furthermore, participants' language proficiencies were assessed through the German and English versions of the LexTALE (Lemhöfer & Broersma, 2012), a vocabulary test in which participants have to distinguish words from pseudowords.[2] Participants were advanced learners of L2 English, indicated by a mean English LexTALE score above 80, which roughly corresponds to C1-level advanced learners according to the Common European Framework (CEF; see Lemhöfer & Broersma, 2012, for a correlation between LexTALE scores and CEF proficiency levels). Participant characteristics are summarized in Table 1. Independent sample t-tests showed no significant differences in background measures between the groups (all ps > .33). Participation was voluntary, and participants joined the study either as part of a seminar or received a small monetary compensation. Informed consent was secured from all participants. Ethical approval was granted by the Ethics Committee of TU Dortmund University (ethics vote no. GEKTUDO_2022_38 & GEKTUDO_2022_39), and the study followed the principles of the 1964 Declaration of Helsinki.

2.2. Materials

The experimental stimuli consisted of 160 letter strings, including 80 words and 80 nonwords.

2.2.1. Words

The words contained 40 cognates and 40 noncognates between German and English and were matched across conditions on length, number of syllables, frequency (SUBTLEX-US log10; Brysbaert & New, 2009) and orthographic and phonological neighborhood size (English Lexicon Project; Balota et al., 2007). Independent sample t-tests yielded no significant difference between conditions (all ps > .15). For cognates and noncognates, normalized orthographic Levenshtein distance (Levenshtein, 1966; Schepens et al., 2012) was calculated as a proxy for overlap between languages. Cognates and noncognates differed significantly (p < .001), with cognates exhibiting more overlap than noncognates. Stimuli characteristics for target words are displayed in Table S1 (Supplementary Materials).

2.2.2. Nonwords

The nonwords consisted of 80 pseudowords created by selecting 80 English nouns (that did not serve as target words) and changing only one letter in each word. It was ensured that the newly created nonwords did not exist in German and followed the rules of English orthography and phonotactics (for a similar procedure, see Dijkstra et al., 2015). Nonwords and words were exactly matched on length and number of syllables (p = 1).

2.3. Procedure

Participants performed a visual English Lexical Decision task in which they had to decide as quickly and accurately as possible

[2]To ensure that participants did not look up any words, German and English versions of the LexTALE were developed for OpenSesame (Mathôt et al., 2012). This made it possible to record the overall task duration to be able to exclude participants who spent an unusually and comparatively large amount of time. No participant had to be excluded for this reason.

https://doi.org/10.1017/S136672892510028X Published online by Cambridge University Press

whether a word presented on screen was an existing English word or not. Decisions were made by pressing one of two designated keys on the keyboard in the web-based experiment or one of two designated buttons on a MilliKey MH-5 button box in the lab-based experiment. For the online experiment, the keys "f" (for NO-presses) and "j" (for YES-presses) were chosen. These keys are comparably positioned across keyboards as they serve as the position keys for typewriting (and hence, usually have small bumps on them). Both groups were presented with on-screen instructions. The stimuli were presented in white ink (#FFFFFF), 40 px Arial font, on a black background (#000000) at the center of the screen. All stimuli were displayed in capitals to avoid language cues, as in German (unlike English), all nouns are always capitalized (see Lemhöfer et al., 2018). Each trial started with a 500 ms fixation cross at screen center. The target stimulus followed and remained on screen for a maximum of 3000 ms or until a key/button was pressed. No feedback regarding participants' responses was provided. The experiment was presented in five blocks: a ten-trial practice block followed by four experimental blocks of 40 items each. The first three experimental blocks were followed by a short break that allowed the participants to rest until they pressed a key/button to continue. Stimuli were presented in a different randomized order per participant. The experiment took approximately 5 minutes and was programmed in OpenSesame (version 3.3.11; Mathôt et al., 2012).

In addition to the Lexical Decision task and the self-paced reading experiment (see Experiment 2), the participants completed several background tasks, the results of which are not covered in this paper. The order of tasks was as follows: (1) Lexical Decision task, (2) Flanker task, (3) Self-paced Reading task, (4) German LexTALE, (5) English LexTALE, (6) Reading Span task, (7) Questionnaire. In total, the testing session took approximately 70 minutes.

2.3.1. Web-based experiment

For the web-based version of the task, OSWeb was used to ensure that the experiment could be run in a browser (Mathôt & March, 2022). A key advantage of OSWeb is that it supports not only different operating systems, such as Windows or macOS, but also multiple browser types. The experiment was hosted on JATOS (Lange et al., 2015), and participants took part with their own computer/laptop. Participation via tablets was not permitted. To make participation more accessible for a larger sample, the operating system was not restricted. In terms of browsers, specific versions for certain browsers were specified in advance to ensure the proper functioning of the online experiments. Prior to the experimental session, participants were briefed by the experimenter in a meeting via an online video conferencing tool. These meetings either took place individually or in small groups to allow for simultaneous testing. The experimenter instructed the participants to perform the study alone, in a quiet environment and to limit potential confounding factors as best as possible by, for instance, closing all nonrelevant computer applications. Although data collection was carried out online, it still took place in a semi-supervised setting: If questions arose during the session, participants could revisit the online meeting and consult the experimenter at any time, preferably not during the experimental blocks. Data collection took place during the week, either in the morning or afternoon.

2.3.2. Lab-based experiment

Participants were tested individually and under supervision of the experimenter in the lab during times equivalent to the web-based data collection.

Table 2. Mean accuracies (proportions) and reaction times (in milliseconds) by group and condition

	Online participants		Lab participants	
	Accuracy	RTs	Accuracy	RTs
Nonwords	0.94 (0.07)	905 (157)	0.93 (0.07)	930 (196)
Words	0.98 (0.02)	705 (81)	0.99 (0.02)	693 (101)
Cognates	0.99 (0.02)	706 (89)	0.99 (0.01)	699 (107)
Noncognates	0.98 (0.03)	704 (76)	0.98 (0.03)	687 (100)

Note: Standard deviations are in parentheses.

2.4. Results

Accuracy rates and reaction times (RTs) for words/nonwords and cognates/noncognates were analyzed in R (version 4.4.0; R Core Team, 2024). One noncognate item (EN: *pupil*) was removed because it is a homonym with both a noncognate translation (*Schüler:in*) and a cognate one (*Pupille*) in German. For the accuracy analysis, RTs below 200 ms were coded as false alarms (web: $n = 1$; lab: $n = 0$). For the RT analysis, incorrect button presses and extreme RTs above 2000 ms (web: $n = 115$ [1.39%]; lab: $n = 104$ [1.31%]) were excluded. Table 2 lists the accuracy rates and RTs for the respective conditions.

Accuracy and RT data were analyzed using mixed-effects regressions with the aid of *lme4* (Bates et al., 2015) and *lmertest* (Kuznetsova et al., 2017). Separate models were run for the analyses of word status and cognate status. For the analyses of words and nonwords, word status, setting and their interaction were entered as fixed effects. For the analyses of cognates and noncognates, nonwords were removed from the data set, and cognate status, setting, as well as their interaction, were entered as fixed effects. All two-level fixed effects were sum-coded (setting: web = 1, lab = −1; word status: nonword = 1, word = −1; cognate status: cognate = 1, noncognate = −1). To identify the maximal converging random effect structure, the "order" function in the *buildmer* package was used for all mixed-effect models (Voeten, 2021).[3] Effect sizes were calculated with the *effectsize* package (Ben-Shachar et al., 2020). Overall, accuracies were close to the ceiling, which speaks to the participants' high L2 proficiency. For words and nonwords, the model returned a significant main effect of word status for accuracy (ß = −0.61, $SE = 0.12$, $z = -4.94$, $p < .001$, $d = -0.96$) and RT data (ß = 111.01, $SE = 9.04$, $t = 12.28$, $p < .001$, $d = 2.4$), with words being processed more accurately and faster than nonwords. Furthermore, accuracy analyses of cognates and noncognates yielded a moderate main effect of cognate status (ß = 0.32, $SE = 0.12$, $z = 2.60$, $p = .009$, $d = 0.51$), with cognates being processed more accurately than noncognates. RT analyses yielded neither main effects nor an interaction. Note that none of the models yielded a main effect of or an interaction with setting. Detailed model output is displayed in Table 3.

2.5. Discussion

Experiment 1 investigated the replicability of the word/nonword effect as well as the effect of cognate status during isolated L2 lexical processing. Besides the replicability of these well-established effects,

[3]By utilizing the "order" command, buildmer commences model selection with an empty random effect structure to which terms are successively added until convergence fails.

https://doi.org/10.1017/S136672892510028X Published online by Cambridge University Press

Table 3. LDT model outputs for word status (word versus nonword) and cognate status (cognate versus noncognate) after "buildmer" model optimization

	Accuracy			Reaction times		
	Estimate	*SE*	*z*	Estimate	*SE*	*t*
(Intercept)	4.12	0.13	32.89***	810.83	13.31	60.91***
Word_stat	**−0.61**	**0.12**	**−4.94*****	**111.01**	**9.04**	**12.28*****
Setting	0.001	0.08	0.01	−3.30	11.96	−0.28
Word_stat x Setting	0.06	0.08	0.71	−8.98	6.89	−1.30
	Formula: Accuracy ~1 + Word_stat + Setting + Word_stat:Setting + (1 | Stimulus) + (1 + Word_stat | Participant)			Formula: RT ~ 1 + Word_stat + Setting + Word_stat:Setting + (1 | Stimulus) + (1 + Word_stat | Participant)		
(Intercept)	4.61	0.17	26.91***	699.79	11.13	62.87***
Cog_stat	**0.32**	**0.12**	**2.60****	3.44	7.04	0.49
Setting	−0.10	0.11	−0.89	5.74	8.86	0.65
Cog_stat x Setting	−0.16	0.09	−1.73	−2.26	2.05	−1.10
	Formula: Accuracy ~1 + Cog_stat + Setting + Cog_stat:Setting + (1 | Stimulus) + (1 | Participant)			Formula: RT ~ 1 + Setting + Cog_stat + Setting:Cog_stat + (1 | Stimulus) + (1 | Participant)		

Note: Significant effects in bold.

the comparability of L2 word recognition across the two settings was examined (RQ1). For word versus nonword processing, differences in accuracies and RTs were observed between the two conditions in both settings, replicating previous lab-based results. Although participants' decision accuracies were above 90% in both conditions, analyses still yielded a significant difference between conditions, with comparatively lower accuracies for nonwords compared to words. Similar to previous findings, RT analyses revealed a processing disadvantage for nonwords over words, with pseudowords exhibiting a delay in processing. Additionally, overall mean RTs differed by only 6 ms between settings. This confirms previous observations by Hilbig (2016) and Kim et al. (2023), who likewise demonstrated no significant difference in RT effects in web-based versus lab-based lexical decision.

For cognate versus noncognate processing, analyses of participants' decision accuracies revealed a main effect of cognate status. With overall accuracies of 98% and 99% per condition, participants' accuracy was high across the board, thus corroborating their overall high L2 English proficiency as evidenced by their English LexTALE results. The descriptively rather small difference of 1% yielded statistical significance, providing evidence for a cognate facilitation effect. However, contrary to initial predictions, the cognate facilitation effect could not be replicated in participants' response times. Nonetheless, this pertains to both settings and thus suggests that the null results are not setting-induced (see also Semmelmann & Weigelt, 2017, who reported null effects with priming across three settings). The absence of a cognate facilitation effect in RTs may be a ceiling effect caused by participants' overall high L2 proficiency (Bultena et al., 2014) or the stimuli's lexical frequency. The cognate and noncognate nouns used in the LDT were predominantly high-frequency nouns (mean SUBTLEX-US log10 = 3.25), which may have disguised potential cognate effects (Peeters et al., 2013). Consequently, future studies could further explore potentially modulating factors in cognate processing and limitations to the cognate facilitation effect. Although the cognate effect could not be replicated in the present study's RT analysis, the overall findings of Experiment 1 suggest comparability of the results obtained in both the online and laboratory settings. The lexical decision manipulation itself was successful, as evidenced by the word/nonword RT effect observed in both settings.

In summary, these findings show that with overall ceiling performance in accuracies across settings, online testing did not prove to adversely affect participants' concentration on ultimate decision-making. Additionally, RT results did not differ across groups, neither for relative RTs between conditions nor for absolute overall mean RTs. Thus, the results expand the findings of previous research by providing comparable evidence of L2 lexical processing across both web-based and lab-based experimental settings.

3. Experiment 2 – Self-paced reading task

3.1. Participants

The same participants who took part in Experiment 1 also participated in Experiment 2.

3.2. Materials

Experimental sentences were constructed based on a 2 (Sentence type: SRC versus ORC) × 2 (Plausibility: plausible versus implausible) × 2 (Cognate status: cognate versus noncognate) design. This resulted in 40 sentence quadruplets in the cognate condition and 40 sentence quadruplets in the noncognate condition (see 8a–h for examples; slashes indicate the phrases in which the sentences were presented). Additionally, 120 plausible and implausible filler sentences were created (see Appendix S1 in Supplementary Materials for details).

(8) a. There is a/man/that drinks a tea/in a café. (SRC, pl. cog.)
b. There is a/tea/that a man drinks/in a café. (ORC, pl. cog.)
c. There is a/tea/that drinks a man/in a café. (SRC, impl., cog.)
d. There is a/man/that a tea drinks/in a café. (ORC, impl., cog.)

https://doi.org/10.1017/S136672892510028X Published online by Cambridge University Press

e. There is a/woman/that buys a dress/ in the store. (SRC, pl., noncog.)
f. There is a/dress/that a woman buys/ in the store. (ORC, pl., noncog.)
g. There is a/dress/that buys a woman/ in the store. (SRC, impl., noncog.)
h. There is a/woman/that a dress buys/ in the store. (ORC, impl., noncog.)

3.2.1. Sentence type

All target sentences were English present-tense embedded relative clause constructions (SRC and ORC), starting with "There is …" and containing one animate and one inanimate noun, a transitive verb and a locative prepositional phrase at the end. Each noun and verb was repeated twice throughout the experiment but always in different combinations (i.e., they never appeared twice with the same verb, noun or prepositional phrase). Each participant was presented with 80 target sentences, distributed across four lists following a Latin-square design. Hence, participants saw ten items per condition and, including filler sentences, responded to 200 sentences in total.

3.2.2. Cognate status

Cognate status was manipulated for the verb in the relative clause and for both nouns in the second and third phrase for each item. This means that in cognate sentences, all of these words were cognates, while in noncognate sentences, all of these words were noncognates. The manipulated nouns were identical to the target words in Experiment 1 (see Tiffin-Richards, 2024, for a similar procedure).

3.2.3. Plausibility

Sentences were either semantically plausible or implausible. Plausibility of the target sentences was manipulated by reversing the roles of the animate agent and the inanimate patient. In plausible sentences, the agent was animate, and the patient was inanimate, while in implausible sentences, the agent was inanimate, and the patient was animate. The semantic plausibility of the sentences was assessed by L2 English speakers in a separate plausibility norming study prior to the actual experiment (see Ferreira, 2003, and Lim & Christianson, 2013, for similar procedures), which yielded a significant difference between both conditions, with plausible sentences being rated as far more plausible than implausible ones (see Appendix S2 and Table S2 in Supplementary Materials for further details).

3.3. Procedure

Subsequent to the Lexical Decision task and the Flanker task, participants performed a noncumulative self-paced reading task in which they read English sentences phrase-by-phrase and rated the plausibility of each sentence immediately after having read it. Implausibility was defined as the event described in the sentence being very unlikely or even impossible to occur. Instructions were presented on screen. Phrases were presented in white ink (#FFFFFF), 40 px Arial font, on a black background (#000000) at the center of the screen, using the stationary window method. Trials were initiated with a 500 ms fixation cross at screen center. Subsequently, the sentences were displayed in phrases of different lengths (varying from one to four words) with a maximum of six phrases. Experimental sentences always consisted of four phrases. Each phrase remained on screen until a designated key was pressed (the space bar online; a central button on the button box in the lab). The first phrase always began with a capital letter. The last phrase always ended with a full stop to indicate the end of the sentence. It was followed by a visual display of a question mark that prompted participants to judge the plausibility of the previously read sentence by pressing one of two designated keys. Online, participants pressed the "f" (for implausible) and "j" (for plausible) keys on their keyboard (see Experiment 1). In the laboratory, a MilliKey MH-5 button box was used. No feedback regarding responses was provided. The experiment was presented in six blocks: A six-trial practice block preceded five experimental blocks of 40 sentences each. The total number of 200 sentences was randomly distributed across these five blocks. The first four experimental blocks were followed by short breaks that allowed participants to rest until they pressed a key to continue the experiment. The experiment lasted approximately 15–20 minutes and was programmed in OpenSesame (version 3.3.11; Mathôt et al., 2012).

3.3.1. Web-based and lab-based experiment versions

The general set-up for both the web-based and the lab-based implementation of the experiment was identical to Experiment 1.

3.4. Results

Plausibility judgment accuracies and reading times were analyzed in R (version 4.4.0; R Core Team, 2024). Two noncognate items that contained the noun *pupil* were excluded from further analyses (see Experiment 1). Additionally, one item with a low plausibility judgment accuracy of 54% was removed. The remaining items had an average accuracy of 93% (range: 77–99%). For the analyses of reading times, the focus was on two regions of interest: (i) the critical phrase containing the RC and (ii) the post-critical phrase immediately following the RC. For reading time analyses, all items with incorrect plausibility judgment were excluded. Additionally, phrases with reading times below 200 ms and above 5000 ms were removed from further analyses (see Klassen et al., 2022, for a similar procedure). For the critical phrase, this resulted in the removal of 44 trials (web: n = 34 [0.89%]; lab: n = 10 [0.27%]). For the post-critical phrase, 77 trials were removed (web: n = 44 [1.15%]; lab: n = 33 [0.88%]). Table 4 lists the mean plausibility judgment accuracies and reading times for the critical and post-critical phrases per condition for both groups.

Accuracy and reading time data were analyzed using mixed-effects regressions (*lme4*, Bates et al., 2015; *lmertest*, Kuznetsova et al., 2017). Setting, sentence type, plausibility and cognate status, as well as their interactions, were entered as fixed effects into the models. Sum-coding was applied to the two-level fixed effects (setting: web = 1, lab = −1; sentence type: SRC = 1, ORC = −1; plausibility: plausible = 1, implausible = −1; cognate status: cognate = 1, noncognate = −1). By means of the "order" function in the *buildmer* R-package (Voeten, 2021), the maximal converging random effect structure was identified. Effect sizes were calculated with the *effectsize* package (Ben-Shachar et al., 2020). Plausibility judgment accuracy was overall high across the board, suggesting that in both groups, participants were able to successfully parse L2 non-canonical sentence structures. Still, generalized linear mixed effects analyses revealed a main effect of plausibility, with implausible sentences being processed more accurately than plausible ones (ß = −0.45, SE = 0.15, z = −3.09, p = .002, d = −0.6). For reading times in the critical phrase, analyses yielded significant main effects for plausibility (ß = −74.11, SE = 8.23, t = −9.01, p < .001, d = −1.76) and sentence type (ß = −48.27, SE = 7.50, t = −6.44, p < .001, d = −1.26), indicating that in the critical RC region,

https://doi.org/10.1017/S136672892510028X Published online by Cambridge University Press

Table 4. Mean plausibility judgment accuracies (proportions) and reading times (in milliseconds) for critical and post-critical phrases per condition by group

	Online participants			Lab participants		
	Accuracy	RT crit.	RT post-crit.	Accuracy	RT crit.	RT post-crit.
SRC, pl., cog.	0.93 (0.12)	1049 (368)	965 (331)	0.91 (0.10)	1114 (341)	1018 (332)
ORC, pl., cog.	0.87 (0.16)	1225 (538)	1111 (431)	0.91 (0.14)	1187 (328)	1114 (408)
SRC, impl., cog.	0.94 (0.09)	1176 (365)	874 (301)	0.94 (0.13)	1297 (463)	972 (338)
ORC, impl., cog.	0.93 (0.15)	1257 (394)	889 (333)	0.95 (0.11)	1368 (414)	1063 (499)
SRC, pl., noncog.	0.93 (0.09)	1000 (345)	895 (283)	0.94 (0.08)	1073 (301)	985 (265)
ORC, pl., noncog.	0.89 (0.10)	1105 (360)	1037 (368)	0.92 (0.10)	1172 (331)	1157 (424)
SRC, impl., noncog.	0.94 (0.11)	1168 (348)	877 (324)	0.94 (0.11)	1243 (319)	1005 (493)
ORC, impl., noncog.	0.95 (0.13)	1254 (410)	912 (374)	0.95 (0.12)	1345 (479)	1005 (397)

Note: Standard deviations are in parentheses.

plausible sentences were overall processed faster than implausible ones and SRCs faster than ORCs. Analyses of reading times in the post-critical phrase revealed main effects of plausibility (ß = 47.11, SE = 16.21, t = 2.91, p = .004, d = 0.57) and sentence type (ß = −42.50, SE = 7.34, t = −5.79, p < .001, d = −1.13) and an interaction of both fixed effects (ß = −30.00, SE = 7.04, t = −4.26, p < .001, d = −0.83). In the post-critical region, implausible sentences were processed faster than plausible sentences – opposite to processing patterns in the critical phrase. Similar to the critical phrase, SRCs were processed faster than ORCs in the region following the RC. While there was a clear advantage for SRCs over ORCs for plausible sentences, the difference between these two conditions was much smaller for implausible sentences. Neither the main effects of cognate status or setting nor interactions with these two factors were found in any of the three dependent measures. Participants' mean reading times for the critical and post-critical phrases are plotted in Figures 1 and 2. Note that these times were collapsed across cognate status to allow for better visualization, as this factor proved not to have a significant influence. Detailed model output is displayed in Table 5.

Finally, the overall mean reading speed was compared between settings. Mean reading speed was calculated across the sum of reading times of all phrases for both experimental and filler sentences. Incorrect trials were excluded from further analyses, and extreme values above 10000 ms were considered outliers and thus removed. The fixed factor setting was sum-coded. Whereas the two groups differed descriptively by about 200 ms, with the online group (M = 3005 ms) being faster than the lab group (M = 3212 ms), linear mixed effects analysis[4] on overall mean reading speed revealed no significant impact of setting on overall reading times (ß = −103.08, SE = 68.51, t = −1.51, p = .132, d = −0.29).

[4]The maximal converging model identified by the "order" command of the *buildmer* package (Voeten, 2021) was the following: RT_all_phrases ~1 + Setting + (1 | Participant) + (1 | item).

3.5. Discussion

Experiment 2 examined the replicability of cognate effects in L2 sentence context (RQ2) and that of word order and plausibility effects during L2 sentence processing (RQ3). Moreover, the comparability of online and laboratory testing was further investigated based on these linguistic phenomena. Whereas lexical co-activation through cognates as observed in prior studies could not be replicated, the present study observed effects of both word order and plausibility on sentence processing. Importantly, the study found similar results for both settings in the accuracy of sentence final judgments as well as two reading time measures, suggesting the comparability of both testing contexts. In the following, we will first discuss the absence of cross-linguistic lexical effects induced by cognates during L2 sentence processing and then review the effects of word order and plausibility on L2 sentence comprehension. Both research questions (RQ2 and RQ3) will be considered against the methodological backdrop of the testing environment.

In view of RQ2 concerning lexical processing in L2 sentence context, the results resemble the null findings for RTs obtained in Experiment 1. Similarly, no effect of cognate status was found in plausibility judgment accuracies or reading times of either the critical or the post-critical phrase in either of the two settings. Thus, in contrast to previous studies (Hopp, 2017; Miller, 2014), a cognate facilitation effect during L2 sentence processing could not be replicated. Instead, descriptively, critical phrases containing cognate words were processed more slowly than those containing noncognate words across all conditions. While this cognate disadvantage was not statistically significant, it is a pattern that can be observed across both groups. The fact that results from both settings showed similar descriptive and inferential results provides further evidence of the comparability of the web-based and lab-based settings. Regarding linguistic implications, it can neither be confirmed nor ruled out whether syntactic or semantic context or cross-linguistic syntactic L1 interference induced the absence of cognate facilitation (Dirix et al., 2019; Lauro & Schwartz, 2017). In fact, the syntactic

https://doi.org/10.1017/S136672892510028X Published online by Cambridge University Press

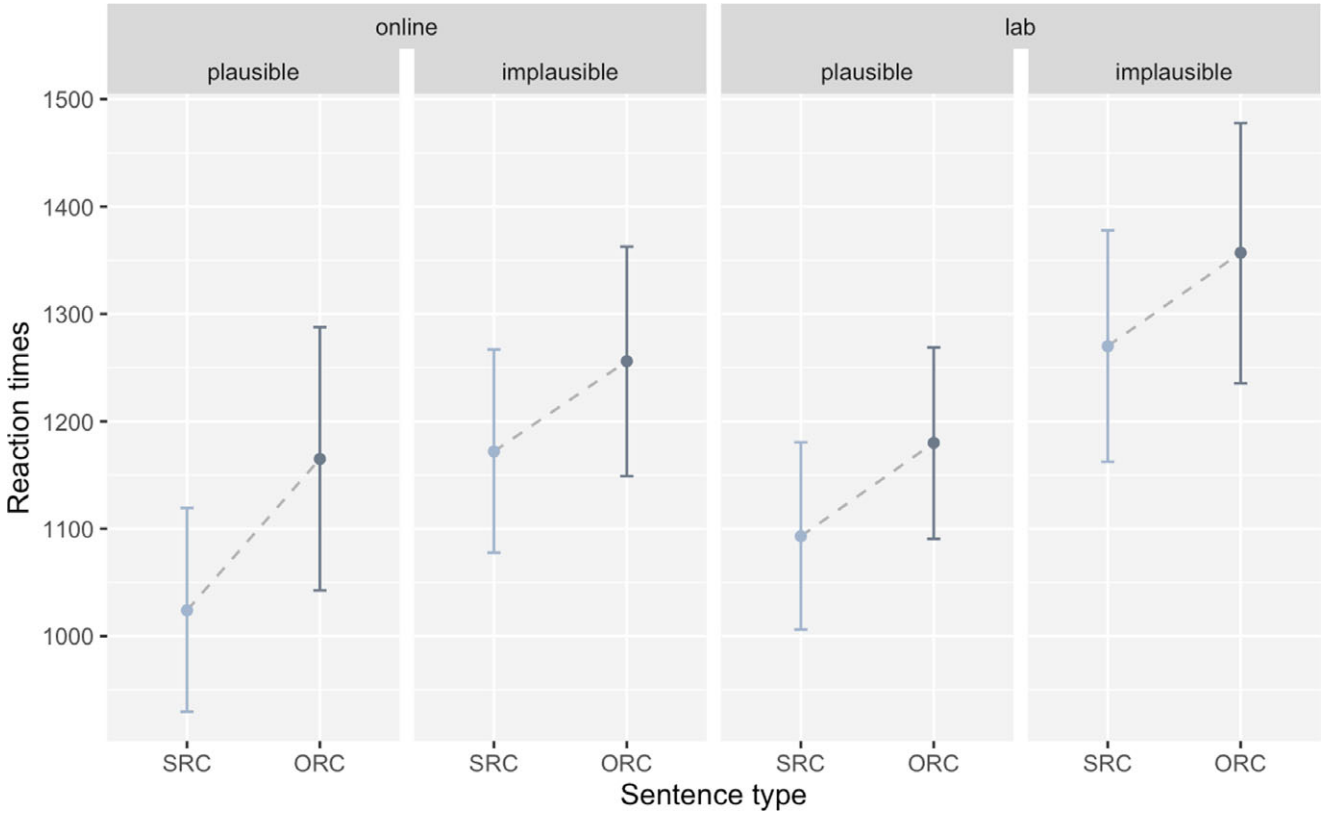


Figure 1. Mean reading times (in milliseconds) for the critical phrase by word order and by plausibility for both settings.
Note: Error bars show the 95% confidence interval.

environment is rather unlikely to have caused the null findings since no cognate facilitation was found for these words out of context either (see Experiment 1). Further research is needed to better understand the interplay of lexical and syntactic L1 activation and its influence on L2 sentence processing.

With respect to RQ3, the present study identified the impacts of word order and plausibility on L2 English relative clause processing. In general, accuracies for plausibility judgment were high across the board in all conditions for both groups, reflecting participants' advanced L2 sentence comprehension. Nevertheless, the plausibility judgment accuracies revealed that implausible sentences were processed more accurately than plausible ones, contradicting the assumption of a processing advantage for semantically plausible sentences. The disadvantage of plausible sentences was particularly evident in the processing of plausible ORCs by the online group, with comparatively lower accuracies in this condition ($M = 88\%$). A possible explanation could lie in thematic roles and positing a somewhat enhanced agent-first preference in these learners. Such a preference is characterized by the first-mentioned noun phrase (NP) of a sentence being more likely to be interpreted as the agent (and not patient) of a transitive event (Bever, 1970; VanPatten, 2015). Applying this to plausible ORCs (first NP is patient and not agent), the initial interpretation tends to be implausible. However, with target-like syntactic parsing, the initial implausible interpretation needs to be overcome, forcing learners to reanalyze the sentence to allow successful interpretation. In this case, it seemed more difficult to judge a sentence that had already been assessed implausible as still plausible than vice versa (i.e., judging a sentence that had been assessed plausible as implausible, e.g., when erroneously applying the agent-first strategy to implausible ORCs that start with the patient). Applying such a processing heuristic thus leads to more revision difficulties when the sentence has already been discarded. However, this descriptive disadvantage for plausible ORC accuracies is only present in the online group. This might indicate a general speed–accuracy trade-off in web-based participants' reading times: The online group may have failed in reanalyzing the initially seemingly implausible sentence as plausible slightly more often than the lab-based group because they were reading descriptively faster, which applies to both overall reading times and to RTs in the critical and post-critical phrase.

With regard to reading times of the critical phrase, analyses revealed main effects of plausibility and sentence type. The present study thus replicated previous in-person research by demonstrating processing advantages for plausible sentences over implausible ones and advantages for SRCs over ORCs in both settings (for similar findings, see Lim & Christianson, 2013). No significant difference in RTs between settings was found.

With regard to reading times of the post-critical phrase, analyses yielded main effects of plausibility and sentence type as well as an interaction of the two factors. Overall, implausible sentences were processed faster than plausible ones. This processing advantage for implausible sentences contrasts with the pattern observed in the critical phrase, which is most likely due to the nature of the plausibility judgment task itself. As soon as an implausible sentence has been deemed implausible, for which participants already have sufficient information in the critical phrase, there is no reason to pay further attention to subsequent regions. Thus, in this case, readers are more likely to skim-read or even skip the post-critical phrase. In contrast, a plausible sentence may turn into an implausible one at any moment depending on incoming information. This

https://doi.org/10.1017/S136672892510028X Published online by Cambridge University Press

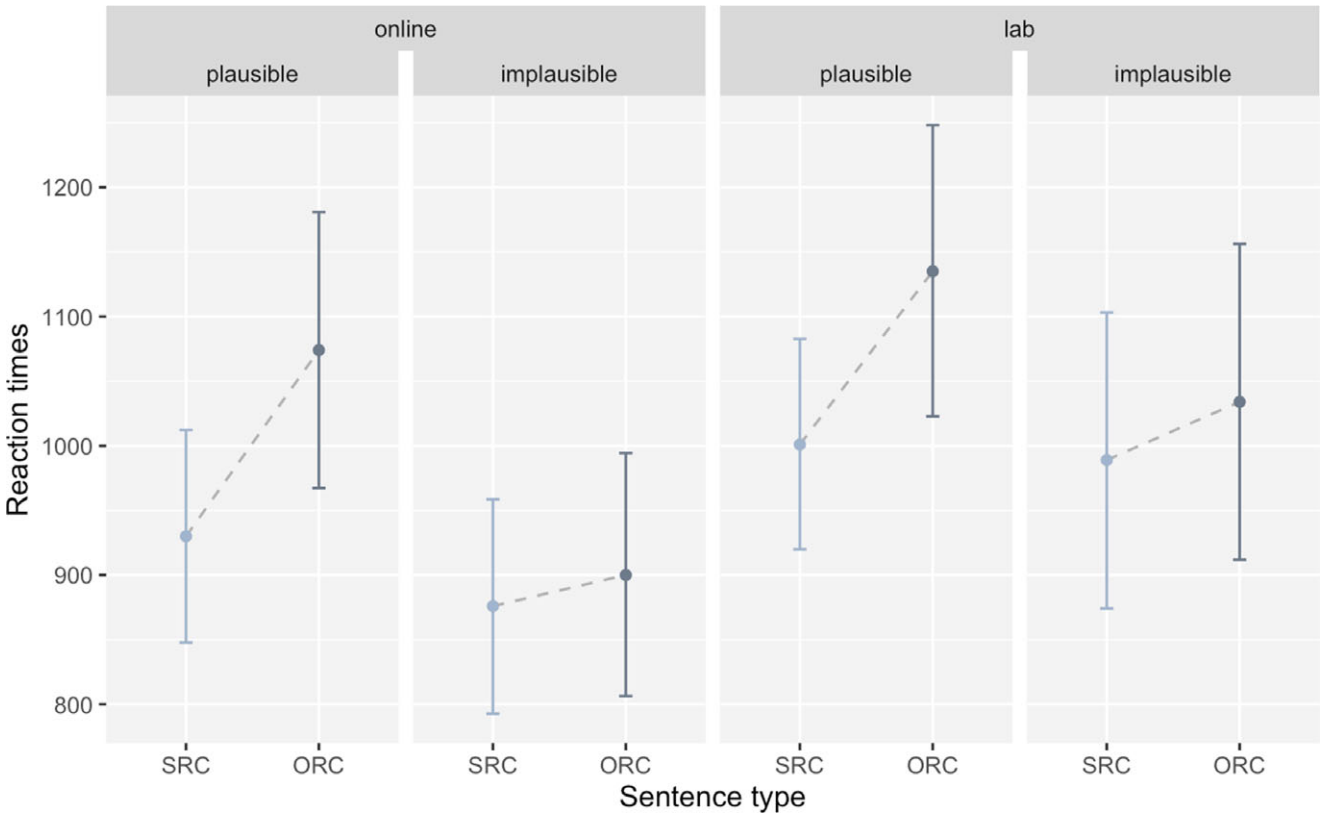


Figure 2. Mean reading times (in milliseconds) for the post-critical phrase by word order and by plausibility for both settings.
Note: Error bars show the 95% confidence interval.

may provide a general advantage for implausible sentences on final sentence segments (for similar task effects, see Williams, 2006). Additionally, similarly to the critical phrase, SRCs were processed faster than ORCs. This is driven mostly by the plausible sentences, expressed by the two-way interaction, and thus (again) reflecting plausible ORCs as the most difficult condition. As suggested above, the particular difficulty of plausible ORCs is likely due to these sentences initially seeming implausible when following a heuristic agent-first strategy, and this initial judgment needs to be revised to correctly judge these sentences as plausible. As for the group, mixed effects analyses once more revealed no differences across settings and no interaction with any of the experimental manipulations.

With respect to the reading times of both the critical and post-critical phrases and the mean reading speed across all phrases, the web-based group demonstrated descriptively faster RTs than the lab-based group. Note, however, that this difference in absolute RTs only appeared descriptively, and neither a main effect of setting nor an interaction with it was found in any of the abovementioned analyses. Since significant effects of word order and plausibility were found for all three dependent measures in both settings, it can be concluded that the self-paced reading data collected online is on par with the data collected in the lab.

In summary, the self-paced reading results complement the findings from lexical decision (Experiment 1) and extend previous research demonstrating comparable findings of behavioral data collections conducted via the web and in the lab (Hilbig, 2016; Kim et al., 2023; Patterson & Nicklin, 2023). The results of Experiment 2 revealed that the linguistic effects of sentence plausibility and sentence structure obtained in the lab can be replicated via online data collection. Methodological consequences resulting from the decision to conduct web-based research and important aspects that need to be considered for the preparation and implementation of online testing will be discussed in more detail below.

4. General discussion and directions for future web-based research

The present study assessed possible differences and similarities in second language processing between traditional in-person laboratory testing and the collection of experimental data via the web. For this purpose, we utilized two well-established linguistic paradigms – lexical decision and self-paced reading – to examine L2 word recognition and sentence comprehension and, by these means, aimed to determine whether the testing environment affects (behavioral) measurements, specifically decision accuracies and response latencies. Participants in both groups were recruited from the same population (i.e., students at the same university) and corresponded in background measures, such as age, age of English acquisition, length of immersion in an English study program and their proficiency in German and English. For L2 word recognition, the present study replicated the expected word/nonword effect in both groups. A cognate effect could be demonstrated in participants' decision accuracies but not in RTs. This pattern was consistent across both settings. Similarly, no cognate effect was found in L2 sentence reading in either group. Regarding the processing of canonical and noncanonical L2 sentence structures, the present study replicated word order and plausibility effects in both settings. Thus, the consistent processing patterns found across groups suggest comparable evidence of web-based and lab-based data collection and support that remote testing is a viable option for behavioral psycholinguistic L2

https://doi.org/10.1017/S136672892510028X Published online by Cambridge University Press

Table 5. SPR model outputs after "buildmer" model optimization

	Plausibility judgment accuracy				Reading times – Critical phrase				Reading times – Post-critical phrase			
	Estimate	*SE*	*z*		Estimate	*SE*	*t*		Estimate	*SE*	*t*	
(Intercept)	3.51	0.14	25.14	***	1187.11	34.19	34.72	***	993.15	28.06	35.40	***
Plausibility	**−0.45**	**0.15**	**−3.09**	**	**−74.11**	**8.23**	**−9.01**	***	**47.11**	**16.21**	**2.91**	**
Sent_type	0.10	0.06	1.70		**−48.27**	**7.50**	**−6.44**	***	**−42.50**	**7.34**	**−5.79**	***
Cog_stat	−0.13	0.08	−1.60		15.35	13.16	1.17		6.25	12.01	0.52	
Setting	−0.04	0.12	−0.35		−38.37	32.06	−1.20		−47.70	26.38	−1.81	
Plausibility x Sent_type	0.09	0.06	1.56		−4.75	5.64	−0.84		**−30.00**	**7.04**	**−4.26**	***
Plausibility x Cog_stat	0.02	0.10	0.19		8.08	6.77	1.19		5.93	10.64	0.56	
Plausibility x Setting	−0.01	0.11	−0.12		11.43	7.32	1.56		13.60	14.11	0.96	
Sent_type x Cog_stat	0.01	0.05	0.22		0.74	6.13	0.12		1.29	7.03	0.18	
Sent_type x Setting	0.10	0.05	1.93		−1.48	7.10	−0.21		2.64	7.34	0.36	
Cog_stat x Setting	0.01	0.06	0.10		2.76	5.63	0.49		5.48	7.29	0.75	
Plausibility x Sent_type x Cog_stat	−0.01	0.05	−0.22		−3.49	5.64	−0.62		5.95	7.03	0.85	
Plausibility x Sent_type x Setting	0.04	0.05	0.82		−8.55	5.64	−1.52		−4.76	7.04	−0.68	
Plausibility x Cog_stat x Setting	0.02	0.05	0.39		5.47	5.64	0.97		10.77	7.03	1.53	
Sent_type x Cog_stat x Setting	0.04	0.05	0.86		−3.59	5.63	−0.64		3.15	7.03	0.45	
Plausibility x Sent_type x Cog_stat x Setting	0.01	0.05	0.20		−2.88	5.64	−0.51		−7.92	7.03	−1.13	
	Formula: Accuracy ~1 + Plausibility + Sent_type + Plausibility: Sent_type + Cog_stat + Setting + Sent_type:Setting + Plausibility:Cog_stat + Plausibility:Setting + Plausibility:Sent_type:Setting + Cog_stat:Setting + Plausibility: Cog_stat:Setting + Sent_type: Cog_stat + Sent_type:Cog_stat: Setting + Plausibility:Sent_type: Cog_stat + Plausibility: Sent_type:Cog_stat:Setting + (1 + Plausibility + Cog_stat | Participant) + (1 + Plausibility + Sent_type + Setting | item)				Formula: RT_phrase3 ~ 1 + Plausibility + Sent_type + Setting + Cog_stat + Plausibility:Cog_stat + Plausibility:Setting + Plausibility: Sent_type + Setting:Cog_stat + Plausibility:Setting:Cog_stat + Sent_type:Cog_stat + Sent_type: Setting + Sent_type:Setting: Cog_stat + Plausibility: Sent_type:Setting + Plausibility: Sent_type:Cog_stat + Plausibility:Sent_type:Setting: Cog_stat + (1 + Plausibility + Sent_type | Participant) + (1 + Plausibility + Sent_type | item)				Formula: RT_phrase4 ~ 1 + Plausibility + Setting + Sent_type + Plausibility: Sent_type + Plausibility:Setting + Cog_stat + Plausibility:Cog_stat + Setting:Sent_type + Setting: Cog_stat + Plausibility:Setting: Cog_stat + Plausibility:Setting: Sent_type + Sent_type:Cog_stat + Plausibility:Sent_type:Cog_stat + Setting:Sent_type:Cog_stat + Plausibility:Setting:Sent_type: Cog_stat + (1 + Plausibility + Sent_type | Participant) + (1 + Plausibility + Setting | item)			

Note: Significant effects in bold.

research despite potentially higher variability in L2 compared to L1 populations.

Notwithstanding the overarching comparability, there is one aspect in which the groups did differ: The online group was characterized by a higher dropout/exclusion rate for participants. This is generally in line with previous research demonstrating increased dropout rates among online participants (Yetano & Royo, 2017). To prevent higher dropout rates and subsequently incomplete data sets in advance, the duration of online experiments should be kept as short as possible. Recent studies recommend overall study durations of a maximum 45 minutes (Gagné & Franzen, 2023) or even less than 30 minutes (Sauter et al., 2020). Increased duration, in contrast, might lead to a drop in participants' motivation and concentration, resulting in tasks being skipped or the entire study being terminated prematurely. The threshold to abort the experiment is much lower in online testing compared to lab-based testing as, during the latter, the experimenter is present, and the risk of distraction can be mitigated. Moreover, it should be noted that the likelihood of technical challenges, such as unstable internet connections, to occur is much higher when testing online. In the present study, almost 17% of the web participants needed to be removed from further analyses due to missing data and/or technical problems. Additionally, around 9% had to be excluded as they did not meet the participation requirements (i.e., studying English). Thus, increasing the sample size for online data collection is recommended.

Besides the higher dropout rate in the web-based setting, the present study has demonstrated that online and laboratory testing can be comparable. Particularly since online participation was not limited to a specific operating system and different browser systems were used, this finding is encouraging for future web-based research. It suggests that behavioral online studies can be implemented without imposing a specific operating system and browser, which facilitates implementation and makes participation considerably more accessible. Note, however, that technologies are subject to constant change and, in most cases, improvement. It is thus crucial for future research to keep pace with technical progress and

https://doi.org/10.1017/S136672892510028X Published online by Cambridge University Press

regularly reassess the data quality of web-based behavioral data collection as technical development moves forward.

Overall, there are further aspects that researchers should bear in mind when collecting data online: to attenuate potential confounding factors, participants should be provided with precise guidelines prior to the experimental session, including detailed information on the technical setup (e.g., allowed hardware, i.e., laptops/computer versus tablets) but also the general framework (e.g., participation in a quiet environment). Additionally, even during online data collection, the experimenter should be available at any time during the experimental session in case questions arise. Therefore, it is advisable to schedule fixed testing appointments, ideally preceded by a brief online meeting prior to the actual experiments, to guide participants through the testing procedure and to which they can return in case questions arise during the testing session. Moreover, to probe the risk of increased distraction or a potential lack of focus in web-based experiments, it is advisable to implement attention checks into online experiments to better monitor participants' concentration (Gagné & Franzen, 2023). Furthermore, it is recommended to pilot experiments on different operating systems and various browsers to preempt potential technical issues during data collection. Whereas the present study's findings from lexical decision and self-paced reading were comparable across both groups, other tasks might be less suitable for online data collection. These include working memory experiments such as Digit Span tasks, which may be more prone to participants using illicit means such as noting down digits for easier later recall. Nevertheless, such issues can at least partially be prevented by including time-outs in tasks. Moreover, tasks eliciting speech production, in turn, imply other challenges (but see He et al., 2021, for recommendations for web-based spoken language production research). With regard to the studied population, web-based data collection comprises two sides of the same coin. On the one hand, online testing is a helpful means to increase access to comparatively understudied populations (Garcia et al., 2022). On the other hand, certain populations might not have access to the means for online data collection, for example, due to their socioeconomic background. For this study's population of university students, online data collection was a suitable means as the participants had access to the necessary technical resources. In contrast, other populations might not have such access or are not as technically inclined and may therefore have more difficulty with performing experiments online (see van der Ploeg et al., 2023, for limitations of web-based research with older adults).

Notwithstanding these potential problems, online research provides benefits compared to in-lab testing. Online testing enables broader access to participants and may thus lead to larger sample sizes. Relatedly, it also provides a smaller threshold to participate than travelling to a lab, and online participants may experience increased anonymity through remoteness as a perk. Additionally, the use of online testing can free up experimenters' resources by not having to rely on lab resources, and concomitantly, the online format can save researchers' time, since participants can be tested simultaneously.

5. Conclusion

In summary, the present study's findings add to the literature advocating for the comparability of web-based and lab-based behavioral data collections. More specifically, the study provides new insights into the comparability of research across these settings with a population of adult second language learners and expands the current literature by directly comparing the groups on L2 word recognition and sentence comprehension. The present study replicated effects of word order (RC processing) and plausibility for L2 sentence reading and the word/nonword effect for L2 lexical comprehension across both settings. Although cognate effects have been extensively demonstrated in previous lab-based research, the current study was not able to replicate these in reaction times, both for words within and without context. Nevertheless, these results were consistent across groups, suggesting that they were not setting-induced but instead more likely to be attributed to other factors such as learners' advanced L2 proficiency or the frequency of the target words. Future studies should thus further investigate potential modulating factors in cognate processing. With respect to methodological implications, future research should keep an eye on technical progress and its consequences for web-based data collection. Nonetheless, this study has demonstrated that even today, online data collection can be a viable and reliable means to collect behavioral data remotely, yet it needs to be prepared with great care. It is crucial to try to best anticipate contingently upcoming issues so that they can be prevented beforehand (such as uncertainties regarding the instructions), and it is vital to always make decisions with the research objective in mind. If these aspects are considered, web-based data collection can serve as a suitable option for psycholinguistic research on not just first but also second language processing.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S136672892510028X.

Data availability statement. Materials, data, and code used in this study can be accessed via OSF (https://osf.io/j86m9/). Further information can be obtained by contacting the corresponding author.

Acknowlegements. This work was supported by the German Research Foundation (DFG; grant no. 436221639) and was conducted while Freya Gastmann and Sarah Schimke were at TU Dortmund University. We thank the student assistants at TU Dortmund University for their help with lab-based data collection and all students who participated in this study.

Competing interests. The authors declare none.

References

Anderson, J. A. E., Mak, L., Keyvani Chahi, A., & Bialystok, E. (2018). The language and social background questionnaire: Assessing degree of bilingualism in a diverse population. *Behavior Research Methods*, **50**, 250–263. https://doi.org/10.3758/s13428-017-0867-9.

Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*, **53**, 1407–1425. https://doi.org/10.3758/s13428-020-01501-5.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., 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. https://doi.org/10.3758/BF03193014.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48. https://doi.org/10.48550/arXiv.1406.5823.

Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). Effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, **5**(56), 2815. https://doi.org/10.21105/joss.02815.

Berger, C., Crossley, S., & Skalicky, S. (2019). Using lexical features to investigate second language lexical decision performance. *Studies in Second Language Acquisition*, **41**(5), 911–935. https://doi.org/10.1017/S0272263119000019.

Bever, T. G. (1970). The cognitive basis for linguistic structures. In R. Hayes (Ed.), *Cognition and language development* (pp. 279–362). Wiley & Sons.

Bridges, D., **Pitiot, A.**, **MacAskill, M. R.**, & **Peirce, J. W.** (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, **8**, e9414. https://doi.org/10.7717/peerj.9414.

Brysbaert, M., & **New, B.** (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, **41**(4), 977–990. https://doi.org/10.3758/BRM.41.4.977.

Buhrmester, M., **Kwang, T.**, & **Gosling, S. D.** (2011). Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, **6**(1), 3–5. https://doi.org/10.1177/1745691610393980.

Bultena, S., **Dijkstra, T.**, & **Van Hell, J. G.** (2014). Cognate effects in sentence context depend on word class, L2 proficiency, and task. *Quarterly Journal of Experimental Psychology*, **67**(6), 1214–1241. https://doi.org/10.1080/17470218.2013.853090.

De Leeuw, J. R., & **Motz, B. A.** (2016). Psychophysics in a web browser? Comparing response times collected with JavaScript and psychophysics toolbox in a visual search task. *Behavior Research Methods*, **48**, 1–12. https://doi.org/10.3758/s13428-015-0567-2.

Dijkstra, T., **Miwa, K.**, **Brummelhuis, B.**, **Sappelli, M.**, & **Baayen, H.** (2010). How cross-language similarity and task demands affect cognate recognition. *Journal of Memory and Language*, **62**(3), 284–301. https://doi.org/10.1016/j.jml.2009.12.003.

Dijkstra, T., **Van Hell, J. G.**, & **Brenders, P.** (2015). Sentence context effects in bilingual word recognition: Cognate status, sentence language, and semantic constraint. *Bilingualism: Language and Cognition*, **18**(4), 597–613. https://doi.org/10.1017/S1366728914000388.

Dirix, N., **Brysbaert, M.**, & **Duyck, W.** (2019). How well do word recognition measures correlate? Effects of language context and repeated presentations. *Behavior Research Methods*, **51**, 2800–2816. https://doi.org/10.3758/s13428-018-1158-9.

Ellis, R. (2004). Individual differences in second language learning. In Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 525–551). Blackwell Publishing. https://doi.org/10.1002/9780470757000.ch21.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, **47**(2), 164–203. https://doi.org/10.1016/S0010-0285(03)00005-7.

Gagné, N., & **Franzen, L.** (2023). How to run behavioural experiments online: Best practice suggestions for cognitive psychology and neuroscience. *Swiss Psychology Open*, **3**(1), 1–21. https://doi.org/10.5334/spo.34.

Garcia, R., **Roeser, J.**, & **Kidd, E.** (2022). Online data collection to address language sampling bias: Lessons from the COVID-19 pandemic. *Linguistics Vanguard*. https://doi.org/10.1515/lingvan-2021-0040

Germine, L., **Nakayama, K.**, **Duchaine, B. C.**, **Chabris, C. F.**, **Chatterjee, G.**, & **Wilmer, J. B.** (2012). Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, **19**, 847–857. https://doi.org/10.3758/s13423-012-0296-9.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, **68**(1), 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1.

He, J., **Meyer, A. S.**, **Creemers, A.**, & **Brehm, L.** (2021). Conducting language production research online: A web-based study of semantic context and name agreement effects in multi-word production. *Collabra: Psychology*, **7**(1), 29935. https://doi.org/10.1525/collabra.29935.

Hilbig, B. E. (2016). Reaction time effects in lab- versus web-based research: Experimental evidence. *Behavior Research Methods*, **48**, 1718–1724. https://doi.org/10.3758/s13428-015-0678-9.

Hopp, H. (2017). Cross-linguistic lexical and syntactic co-activation in L2 sentence processing. *Linguistic Approaches to Bilingualism*, **7**(1), 96–130. https://doi.org/10.1075/lab.14027.hop.

Kim, D., **Lowder, M. W.**, & **Choi, W.** (2023). Emotionality effects in Korean visual word recognition: Evidence from lab-based and web-based lexical decision tasks. *Acta Psychologica*, **237**, 103944. https://doi.org/10.1016/j.actpsy.2023.103944.

Klassen, R., **Kolb, N.**, **Hopp, H.**, & **Westergaard, M.** (2022). Interactions between lexical and syntactic L1-L2 overlap: Effects of gender congruency on L2 sentence processing in L1 Spanish-L2 German speakers. *Applied PsychoLinguistics*, **43**(6), 1221–1256. https://doi.org/10.1017/S0142716422000236.

Kochari, A. R. (2019). Conducting web-based experiments for numerical cognition research. *Journal of Cognition*, **2**(1). https://doi.org/10.5334/joc.85.

Kuznetsova, A., **Brockhoff, P. B.**, & **Christensen, R. H. B.** (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, **82**, 1–26. https://doi.org/10.18637/jss.v082.i13.

Lange, K., **Kühn, S.**, & **Filevich, E.** (2015). Just another tool for online studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLoS One*, **10**(6), e0130834. https://doi.org/10.1371/journal.pone.0130834.

Lau, E., & **Tanaka, N.** (2021). The subject advantage in relative clauses: A review. *Glossa*, **6**(1), 1–34. https://doi.org/10.5334/gjgl.1343.

Lauro, J., & **Schwartz, A. I.** (2017). Bilingual non-selective lexical access in sentence contexts: A meta-analytic review. *Journal of Memory and Language*, **92**, 217–233. https://doi.org/10.1016/j.jml.2016.06.010.

Lemhöfer, K., & **Broersma, M.** (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, **44**, 325–343. https://doi.org/10.3758/s13428-011-0146-0.

Lemhöfer, K., **Huestegge, L.**, & **Mulder, K.** (2018). Another cup of TEE? The processing of second language near-cognates in first language reading. *Language, Cognition and Neuroscience*, **33**(8), 968–991. https://doi.org/10.1080/23273798.2018.1433863.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, **10**, 707–710.

Lim, J. H., & **Christianson, K.** (2013). Second language sentence processing in reading for comprehension and translation. *Bilingualism: Language and Cognition*, **16**(3), 518–537. https://doi.org/10.1017/S1366728912000351.

Mathôt, S., & **March, J.** (2022). Conducting linguistic experiments online with OpenSesame and OSWeb. *Language Learning*, **72**(4), 1017–1048. https://doi.org/10.1111/lang.12509.

Mathôt, S., **Schreij, D.**, & **Theeuwes, J.** (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, **44**, 314–324. https://doi.org/10.3758/s13428-011-0168-7.

Miller, A. K. (2014). Accessing and maintaining referents in L2 processing of WH-dependencies. *Linguistic Approaches to Bilingualism*, **4**(2), 167–191. https://doi.org/10.1075/lab.4.2.02mil.

Miller, R., **Schmidt, K.**, **Kirschbaum, C.**, & **Enge, S.** (2018). Comparability, stability, and reliability of internet-based mental chronometry in domestic and laboratory settings. *Behavior Research Methods*, **50**, 1345–1358. https://doi.org/10.3758/s13428-018-1036-5.

Neath, I., **Earle, A.**, **Hallett, D.**, & **Surprenant, A. M.** (2011). Response time accuracy in Apple Macintosh computers. *Behavior Research Methods*, **43**, 353–362. https://doi.org/10.3758/s13428-011-0069-9.

Patterson, A. S., & **Nicklin, C.** (2023). L2 self-paced reading data collection across three contexts: In-person, online, and crowdsourcing. *Research Methods in Applied Linguistics*, **2**(1), 100045. https://doi.org/10.1016/j.rmal.2023.100045.

Peeters, D., **Dijkstra, T.**, & **Grainger, J.** (2013). The representation and processing of identical cognates by late bilinguals: RT and ERP effects. *Journal of Memory and Language*, **68**(4), 315–332. https://doi.org/10.1016/j.jml.2012.12.003.

Pronk, T., **Wiers, R. W.**, **Molenkamp, B.**, & **Murre, J.** (2020). Mental chronometry in the pocket? Timing accuracy of web applications on touchscreen and keyboard devices. *Behavior Research Methods*, **52**, 1371–1382. https://doi.org/10.3758/s13428-019-01321-2.

Reimers, S., & **Stewart, N.** (2015). Presentation and response timing accuracy in adobe flash and HTML5/JavaScript web experiments. *Behavior Research Methods*, **47**, 309–327. https://doi.org/10.3758/s13428-014-0471-1.

Rodd, J. M. (2024). Moving experimental psychology online: How to obtain high quality data when we can't see our participants. *Journal of Memory and Language*, **134**, 104472. https://doi.org/10.1016/j.jml.2023.104472.

Sauter, M., **Draschkow, D.**, & **Mack, W.** (2020). Building, hosting and recruiting: A brief introduction to running behavioral experiments online. *Brain Sciences*, **10**(4), 251. https://doi.org/10.3390/brainsci10040251.

Sauter, M., **Stefani, M.**, & **Mack, W.** (2022). Equal quality for online and lab data: A direct comparison from two dual-task paradigms. *Open Psychology*, **4**(1), 47–59. https://doi.org/10.1515/psych-2022-0003.

Schepens, J., **Dijkstra, T.**, & **Grootjen, F.** (2012). Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition*, **15**(1), 157–166. https://doi.org/10.1017/S1366728910000623.

https://doi.org/10.1017/S136672892510028X Published online by Cambridge University Press

Semmelmann, K., & **Weigelt, S.** (2017). Online psychophysics: Reaction time effects in cognitive experiments. *Behavior Research Methods*, **49**, 1241–1260. https://doi.org/10.3758/s13428-016-0783-4.

Stanners, R. F., **Jastrzembski, J. E.**, & **Westbrook, A.** (1975). Frequency and visual quality in a word-nonword classification task. *Journal of Verbal Learning and Verbal Behavior*, **14**(3), 259–264. https://doi.org/10.1016/S0022-5371(75)80069-7.

Team, R. C. (2024). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Tiffin-Richards, S. P. (2024). Cognate facilitation in bilingual reading: The influence of orthographic and phonological similarity on lexical decisions and eye-movements. *Bilingualism: Language and Cognition*, 1–18. https://doi.org/10.1017/S1366728923000949

van der Ploeg, M., **Lowie, W.**, & **Keijzer, M.** (2023). The effects of language teaching pedagogy on cognitive functioning in older adults. *Behavioral Sciences*, **13**(3), 199. https://doi.org/10.3390/bs13030199.

VanPatten, B. (2015). Input processing in adult SLA. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 113–134). Routledge.

Voeten, C. C. (2021). *Buildmer: Stepwise elimination and term reordering for mixed-effects regression* [Computer software]. R Foundation for Statistical Computing.

Weydmann, G., **Palmieri, I.**, **Simões, R. A.**, **Centurion Cabral, J. C.**, **Eckhardt, J.**, **Tavares, P.**, **Moro, C.**, **Alves, P.**, **Buchmann, S.**, **Schmidt, E.**, **Friedman, R.**, & **Bizarro, L.** (2023). Switching to online: Testing the validity of supervised remote testing for online reinforcement learning experiments. *Behavior Research Methods*, **55**(7), 3645–3657. https://doi.org/10.3758/s13428-022-01982-6.

Williams, J. N. (2006). Incremental interpretation in second language sentence processing. *Bilingualism: Language and Cognition*, **9**(1), 71–88. https://doi.org/10.1017/S1366728905002385.

Yetano, A., & **Royo, S.** (2017). Keeping citizens engaged: A comparison between online and offline participants. *Administration & Society*, **49**(3), 394–422. https://doi.org/10.1177/0095399715581625.

https://doi.org/10.1017/S136672892510028X Published online by Cambridge University Press