Can we use masked priming in an online setting?
The effect of prime exposure duration in online masked priming lexical decision

Bernhard Angele[1,2], Ana Baciero[2,3], Pablo Gomez[4], & Manuel Perea Lara[2,5]

[1] Bournemouth University, Bournemouth, UK

[2] Universidad Antonio de Nebrija, Madrid, Spain

[3] DePaul University, Chicago, USA

[4] California State University San Bernardino, Palm Desert Campus, USA

[5] Universitat de Valencia, Valencia, Spain

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line. Enter author note here.

Correspondence concerning this article should be addressed to Bernhard Angele, Department of Psychology, Faculty of Science and Technology, Talbot Campus, Fern Barrow, Poole BH12 5BB, UK. E-mail: bangele@bournemouth.ac.uk

Abstract

Masked priming is one of the most important paradigms in the study of visual word recognition, but it is usually thought to require a laboratory setup with a known monitor and keyboard. To investigate if this technique can be used in an online setting, we conducted two online masked priming lexical decision task experiments using PsychoPy/PsychoJS (Peirce et al., 2019). In particular, we wanted to compare our online results to the data collected by Gomez, Perea, and Ratcliff (2013), who compared masked and unmasked priming. Furthermore, we also tested the role of prime exposure duration effectively in an online experiment (33 vs. 50 ms in Experiment 1 and 16 vs. 33 ms in Experiment 2). We found that our online data are indeed very similar to the masked priming data reported by Gomez, Perea, and Ratcliff (2013). Additionally, we found a clear effect of prime duration, with the priming effect (measured in terms of response time and accuracy) being stronger at 50 ms than 33 ms and no priming effect at 16 ms prime duration. From these results, we can conclude that modern online browser-based experimental psychophysics packages (e.g., Psychopy) can present stimuli and collect responses on standard consumer devices with enough precision. In sum, these findings provide us with confidence that masked priming can be used online, thus allowing us to reach populations that are hard to test in a laboratory.

Keywords: Masked priming, Lexical decision task, Online experiments, PsychoPy, Prime duration

Word count: XXX

Can we use masked priming in an online setting?

The effect of prime exposure duration in online masked priming lexical decision

Masked priming (K. I. Forster & Davis, 1984) is one of the most important techniques to study the effects of orthography, phonology, morphology, and meaning in visual word recognition(see K. Forster, 1998; Grainger, 2008, for reviews). Priming refers to the influence of a prime stimulus (e.g., *nurse*, *horse*) on a subsequently presented stimulus that the participant has to respond (e.g., "is *DOCTOR* a word?"). It is measured as the difference in a dependent variable (e.g., response time [RT]) between two conditions (e.g., unrelated: *horse-DOCTOR*; related: *nurse-DOCTOR*). In masked priming, the prime stimulus is presented very briefly (for less than 60 ms) and is itself preceded by a pattern mask (e.g., #####) for a much longer duration (typically 500 ms). The rationale of the procedure is to make participants unaware of the identity of the masked prime (K. Forster, 1998; K. I. Forster & Davis, 1984), thus minimizing the role of participants' strategies. Indeed, masked priming experiments do not show the strategical effects that occur with visible, unmasked primes (e.g., Grossi, 2006; Perea & Rosa, 2002).

The masked priming paradigm has been used in a large number of studies over the last decades. For instance, a search of the expression "masked priming" in Google Scholar in March 2021 produced nearly 10,000 hits. Virtually all masked priming experiments have been run in a laboratory setting, often using the DMDX software developed by K. Forster and Forster (2003). The issue we examine in the present paper is whether masked priming experiments can safely be conducted in an online setting. Even before the current exceptional situation due to the COVID pandemic, in which many labs around the world have been closed (or with minimal activity) for many months, online data collection has shown its many advantages: 1) easy access to a much more diverse population than that accessible at the typical university research laboratory; 2) independence from laboratory space constraints, and, often, lower costs as participants only need to be compensated for

their time on the experiment; 3) no time spent commuting, waiting for the experiment to start, etc. Indeed, researchers in decision making and economics have been using online paradigms for several decades now (e.g., Birnbaum & Birnbaum, 2000; Paolacci, Chandler, & Ipeirotis, 2010).

However, cognitive psychologists have been much slower in taking up online paradigms (Cai et al., 2017, for some exceptions; Eerland, Engelen, & Zwaan, 2013; Rodd et al., 2016; see Vandeberg, Eerland, & Zwaan, 2012), often due to concerns about the validity of the results. Such concerns are not limited to cognitive studies (see Aust, Diedenhofen, Ullrich, & Musch, 2013), but they are exacerbated by the reliance on precise presentation times in cognitive psychology. Indeed, in the masked priming technique, it is critical for the onset of the mask, the prime, and the target to occur at the nominal times. In particular, presenting the masked prime for longer than intended would counteract the effect of the mask, making the prime consciously visible to the participant and possibly altering the processes of interest.

There have been attempts to address these concerns. For example, a Web version of DMDX (webDMDX) was developed and showed promising results in a trial experiment (Witzel, Cornelius, Witzel, Forster, & Forster, 2013). However, webDMDX is a self-contained Windows executable file that participants have to download and run rather than a "true" online programming script that could be run inside of a browser. A downside of this format is that participants often are (and should be) understandably skeptical about downloading and running executable files from the Internet. Additionally, many participants may not have access to a Windows PC, or may be discouraged from participating by the extra work it takes to deploy the experiment on their computer. As a consequence, the use of webDMDX has been rather limited so far (CITE somebody who has employed it)

Fortunately, in recent years, there have been significant improvements in how content

can be presented on the World Wide Web. Most notably, the HTML5 standard now makes it possible to use Javascript in order to draw stimuli interactively and monitor participant responses with remarkable flexibility inside the browser. Participants do not have to install any software, and the HTML5 standard is supported by a wide variety of devices, including mobile phones and tablets (Reimers & Stewart, 2015). There is now a variety of software packages taking advantage of the new capabilities to present experimental stimuli and collect data, both commercial, e.g., Gorilla (A. L. Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020) or Testable (Rezlescu, Danaila, Miron, & Amariei, 2020) and open-source, e.g., jsPsych (de Leeuw, 2015) or PsychoJS (the Javascript version of Psychopy 3, Peirce et al., 2019). In addition, online setups allow researchers to target any individual with an internet connection as a participant, from very different countries and backgrounds. Indeed, various on-line platforms (e.g., Prolific, Amazon Mechanical Turk) offer the possibility to recruit participants for on-line experiments based on various specific characteristics set up by the experimenter not necessarily in the country's lab (e.g., native French speakers, not older than 30 years old, not currently in college).

Of course, despite the technological advances, many cognitive psychologists still have concerns about timing and measurement precision. While Javascript-based experiments run on the participants' devices and thereby avoid any lag due to connection issues (e.g., to avoid delays, all stimuli are usually downloaded before the start of the experiment), experimenters have little control over which devices the experiment is run on beyond the option of explicitly preventing the experiment to run on specific device types such as mobile devices. Moreover, experimenters have no control at all over what other applications are running on the device, screen size and resolution, viewing distance, properties of the keyboard/touchscreen, etc., as all of these are determined by the device or the participants' preferences. As Reimers and Stewart (2015) point out, there are two ways of testing whether timing and response issues are problematic: (1) comparing a Web-based experiment directly with an established lab-based version by measuring presentation

timings (using a photodiode) and response timings on various device configurations and (2) attempting to replicate existing lab-based findings using a Web-based paradigm. If the results of the Web-based study are comparable to previous lab-based results, this suggests that whatever the deviations in stimulus and response timing are, they are not severe enough to affect the overall findings in the paradigm in question.

The first approach has the advantage that differences in presentation timings can be objectively recorded and evaluated. A very thorough recent example of this approach is the "timing mega-study" by Bridges, Pitiot, MacAskill, and Peirce (2020), who compared the timing in experiments run in lab-based setups with the timing in online packages run in different browsers. The very similar study by A. Anwyl-Irvine, Dalmaijer, Hodges, and Evershed (2020) compares only online packages and browsers with regard to timing. Overall, Bridges, Pitiot, MacAskill, and Peirce (2020) found that online packages were capable of presenting visual stimuli with "reasonable" precision, although the lab-based packages were slightly better in this regard. This first approach is important in order to establish that a certain level of precision and accuracy can be achieved at all. If this is not possible, there is no point in moving forward to the second approach and replicating specific paradigms. However, it is of course impossible to test every possible device and configuration that participants might use. On the other hand, some of the differences in precision and accuracy between setups that can be observed using a photodiode may be too small to have an influence on actual participant performance.

Therefore, we consider replication of previous key lab-based effects a more important test of online paradigms than photodiode measurements. Based on the results by Bridges, Pitiot, MacAskill, and Peirce (2020) and A. Anwyl-Irvine, Dalmaijer, Hodges, and Evershed (2020), modern Javascript-based stimulus presentation systems are capable of sufficiently fast and precise stimulus presentation. To establish whether masked priming studies can be successfully run online, the next step is now to follow the second approach and implement the masked priming paradigm in one of the online experiment packages tested by Bridges,

Pitiot, MacAskill, and Peirce (2020) and A. Anwyl-Irvine, Dalmaijer, Hodges, and Evershed (2020). After evaluating their results, we decided on Psychopy/PsychoJS (Peirce et al., 2019) as it combines relative ease of use with high precision and accuracy across the great majority of platforms. COMMENT: Was there something better (more precise)? It needs better rationale (we want to say it is the best). Possible reviewers' comments:

- Have you checked in different platforms?
- What does it mean "ease of use"?
- How precise is it?

In the present study, we were interested in whether we could replicate, and extend, a key phenomenon in laboratory masked priming lexical decision using an online setup: masked identity priming reflects a savings effect. As first suggested by K. Forster (1998), for a masked identity prime, "the lexical entry is already in the process of being opened, and hence the evaluation of this entry begins sooner," whereas for an unrelated prime, "the entry for the target word would be closed down (since it fails to match the prime), and no savings would occur" (p. 213). Thus, according to the savings account, a target word like *DOCTOR* would enjoy an encoding advantage when preceded by an identity prime such as *doctor* than when preceded by an unrelated prime such as *pencil* (i.e., a head-start). One implication of such benefit is that the RT distributions of the unrelated and identity pairs should reflect a shift rather than a change in shape. Furthermore, this shift should be approximately similar in magnitude to the prime-target stimulus-onset asynchrony (SOA). Empirical evidence supporting this view has been obtained in several papers not only with skilled adults but also with developing readers (Gomez & Perea, 2020; e.g., Gomez, Perea, & Ratcliff, 2013; Taikh & Lupker, 2020; Yang, Jared, Perea, & Lupker, 2021). Moreover, this pattern of results fits very well with the diffusion model (Ratcliff, Gomez, & McKoon, 2004). This model proposes that, when making a two-choice decision, the resultant RT can be explained as the sum of non-decision parameters, which are the encoding time and

response execution (T_{er}) and decision parameters, which refer to the process of accumulation of information until a decision criteria is reached. Importantly, in the decision process, the information gathered from the stimulus can vary in noise, depending on its quality, which modifies the rate at which information is accumulated (i.e., the *drift rate*). With regards to RTs from masked priming tasks, Gomez, Perea, and Ratcliff (2013) found that the difference between identity and unrelated conditions could be accounted by a change in the T_{er} parameter, while there were no differences across conditions in the parameter that corresponds to the quality of evidence gathered (i.e., drift rate)–note that changes in drift rate would necessarily produce a skewer RT distribution in the slower, unrelated condition.

Critically, the above pattern is *specific* to masked priming. When primes are visible (i.e., unmasked priming), identity priming effects are stronger in the upper quantiles of the RT distribution than in the lower quantiles (i.e., a change in shape rather than a shift in RT distributions; see Gomez, Perea, & Ratcliff, 2013). Fits from the diffusion model show that this result corresponds to changes in both the T_{er} parameter and the drift rate (see Gomez, Perea, & Ratcliff, 2013). Hence, when the prime is visible, it does influence the quality of the information accumulated of the target word. Clearly, this dissociation between masked and unmasked priming reflects qualitative differences in the way primes affect the processing of the target: purely encoding in masked priming (with an expected effect close to the prime duration) vs. both encoding + information quality in unmasked priming.

In the present paper, we took advantage of the above marker to examine whether online masked priming studies follow the same pattern as in-lab masked priming studies. Specifically, we manipulated prime exposure duration in identity vs. unrelated primes: 33.3 vs. 50 ms in Experiment 1, and 16.6 vs. 33.3 ms in Experiment 2—note that targets were presented immediately after the primes (i.e., prime exposure duration was equal to the prime-target SOA). The rationale of Experiment 1 is that if the actual exposure duration of the primes is the nominal exposure duration, then we would expect the typical shift

between the identity and unrelated response time distributions, which according to the savings hypothesis (K. Forster, 1998) would be greater for 50 ms than for 33.3 ms exposure duration (i.e., the head-start would be greater for 50 ms identity primes than for 33.3 ms identity primes). This outcome would indicate that the on-line masked priming studies reproduce a characteristic signature of laboratory masked priming studies. Alternatively, if the online presentation conditions lead to a greater actual exposure duration on the participant's device compared to that specified in our experiment, (e.g., prime durations are approximately 20 ms longer on average than programmed) then the 50 ms primes may no longer be masked, but may rather be consciously perceived. If this is the case, the prime could affect not only the encoding, but also core decision processes (i.e., the drift rate), which would be reflected as a stronger priming effect in the higher quantiles of the distribution (i.e., the two RT distributions would have a different shape). In this scenario, one should be very cautious when running online masked priming experiments—at least with the typical software and hardware currently available.

To further constrain the research questions, Experiment 2 was designed to be analogous to Experiment 1, except for the use of a very short prime exposure duration (i.e., 16.6ms). Similar prime exposure durations have shown rather weak masked priming lexical decision experiments in a laboratory setting: smaller than 5 ms for prime durations of 14 ms (Ziegler, Ferrand, Jacobs, Rey, & Grainger, 2000) and smaller than 9 ms for a prime duration of 20 ms (Tzur & Frost, 2007)—in the Tzur and Frost (2007) experiment, this difference increased to 16 ms when using a very high level of contrast in the computer screen. Thus, if the size of the priming effect is roughly similar to the prime exposure duration, we would expect a much larger priming effect at the 33.3 ms prime exposure duration than at the 16.6 ms prime exposure duration. If we do observe a large effect at the 16.6 ms prime exposure duration (e.g., above 20-25 ms), this would suggest, again, that there is a qualitative difference between using the masked priming technique in the laboratory and in online experiments. Keep in mind that we are using a software that has

good control over the exposure duration (CITE).

ADD at some point:

- Even with no control over many context variables, we obtain the effects (peerj:Parker)
- Hypotheses in a figure (as in Gomez & Perea, 2020).
- In addition:

We conducted a simulation study on the impact of suboptimal response time accuracy on statistical power when varying the number of participants, the number of items/condition, and the degree of inaccuracy of the response device.

Experiment 1

In the first experiment, we tested whether we could observe reliable effects of masked priming at prime durations of 33 ms and 50 ms (roughly corresponding to two and three frames at a refresh rate of 60 Hz).

We report here how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the two experiments.

Method

Participants. Participants were recruited through Prolific (www.prolific.co, 2021). The experiment was accessed by 101 participants. Out of these, 177 provided experimental data. Four of these participants did not complete the experiment. A further 20 were excluded because of low accuracy (less than .8). The remaining 77 participants were aged from 18 to 71 (mean age: 31.14). Of the participants, 41 identified as male and 36 identified as female. The experiment was only shown to participants who indicated that English was their first language in the Prolific screening questions. Based on their IP addresses, 47 participants were based in the UK, 23 were based in the US, three participants were based

in Canada, and two participants were based in Ireland. Two participants could not be localized in this way. Participants who completed the experiment were paid £1.25 for their participation (corresponding to £5/hour). All participants were naïve to the purpose of the experiment. Participants could use a desktop or laptop computer or a mobile device. Because of a technical display issue with PsychoJS and the Safari browser, participants who tried to access the experiment using that browser, including all participants on iOS devices, were advised to change browser or device and restart the experiment.

Rationale for sample size and stopping rule.

Brysbaert and Stevens (2018) recommended that masked priming experiments should have at least 1,600 observations per condition. In order to account for potentially smaller effect sizes in an online experiment, we set a target of 3,000 observations per condition (12,000 observations total), which corresponds to a minimum of 50 participants given that there were 240 word stimuli in the experiment (see below). Our stopping rule was to keep collecting data until 3,000 valid observations were found. This goal was met and exceeded in our initial data collection with a budget of £200.

Materials. We selected 240 six-letter English words from the English Lexicon Project (Balota et al., 2007). The mean Zipf frequency based on the HAL corpus (Lund & Burgess, 1996) was 3.8 (range: 1.9–5.5). The mean OLD20 (Yarkoni, Balota, & Yap, 2008) was 2.1 (range: 1.4–3). We also selected 240 matched, orthographically legal six-letter nonwords. For each target, we created an identical prime (e.g. `region` — `REGION` and `fainch` — `FAINCH`) and an unrelated prime consisting of another word from the list (e.g. `launch` — `REGION` and `miluer` — `FAINCH`). Appendix A contains a list of the items.

Procedure. Participants were able to sign up for the experiment on the Prolific website. On signing up, they were redirected to the participant agreement form on the Qualtrics online survey development environment (Qualtrics, 2020). After indicating their agreement to participate, participants were forwarded to Pavlovia (Pavlovia, 2020), where

the actual lexical decision task was implemented in PsychoJS. In the experiment, all stimuli were presented in the center of the screen in black Courier New font on a white background. As we do not know the exact dimensions of each participant's screen, all stimulus sizes and positions were defined in Psychopy's "height" units, with the bottom left of a 16:10 aspect ratio screen being represented as (-.8, -.5) and the top right being (.8, .5). The height for all text stimuli was 0.1 units. Each trial began with a six-character pattern mask (`######`) set to be presented for 500 ms, followed first by the lowercase prime (e.g. `region`) set to be presented for either 33 or 50 ms, and then by the uppercase target (e.g. `REGION`). Participants were instructed to respond to the target stimulus as quickly as possible either by using the keyboard, if their device had one, and pressing the "Z" key if the target was not a valid English word or the "M" key if the target was a valid English word. Participants on a device without a keyboard were instructed to respond by touching one of two rectangular touch areas labeled "Z = Non-word" (presented at -0.4, -0.3) and "M = Word" (presented at 0.4, -0.3). The touch areas each had a width of 0.4 and a height of 0.2 and were presented in white with a black outline. If participants did not respond within two seconds of the target onset, a "Too slow!" feedback message was shown for 500 ms and the trial ended. The experimental trials were preceded by 16 practice trials during which participants also received feedback on the accuracy of their responses. No feedback apart from the trial timeout feedback was given during the experimental trials. Every 120 trials, participants were asked to take a short break before continuing the experiment. After completing the experiment, participants were redirected to a debriefing form on Qualtrics and from there back to Prolific in order to receive their participation payment.

Data analysis. We analyzed the data by fitting Bayesian linear and generalized linear mixed models, using the *brms* package (Bürkner, 2017, 2018) in R (R Core Team, 2021)[1]. We only analyzed trials where the target stimulus was a word. For the response

[1] The full list of software we used for our analyses is as follows: R [Version 4.0.5; R Core Team (2021)] and the R-packages *bayestestR* [Version 0.9.0; Makowski, Ben-Shachar, & Lüdecke (2019)], *brms* [Version

time (RT) analysis, we excluded trials with RTs lower than 250 ms and incorrect responses (5.69 % of trials). For the accuracy analysis, we only excluded trials with RTs lower than 250 ms (0.28 % of trials). As the trials automatically ended after 2000 ms, there were no RTs longer than this. For both RTs and accuracy, we fitted a model with priming condition (unrelated vs. identical) and prime duration (33 ms vs. 50 ms) as well as their interaction as the fixed effects. For the discrete predictors, we used contrasts as follows: For priming condition, identical was coded as -1 and unrelated was coded as 1. For priming duration, 33 ms was coded as -1 and 50 ms was coded as 1. We used the maximal random effects structure possible, with random intercepts and slopes for condition, prime duration, and the interaction for participants and items. We used the ex-Gaussian distribution to model response times and the Bernoulli distribution (with a logit link) to model response accuracy. We used the default priors suggested by *brms* except for the coefficients for the fixed effects, for which we applied weakly informative priors of $\beta \sim N(0, 100)$ in order to rule out improbably large effect sizes. Each model was fitted using four chains with 5000 iterations each (1000 warmup iterations). We consider an effect as credible if the 95% credible interval (CrI) estimated from the posterior distribution does not contain zero.

2.15.0; Bürkner (2017); Bürkner (2018)], *dplyr* [Version 1.0.5; Wickham, François, Henry, & Müller (2021)], *forcats* [Version 0.5.1; Wickham (2021a)], *ggplot2* [Version 3.3.3; Wickham (2016)], *papaja* [Version 0.1.0.9942; Aust & Barth (2020)], *purrr* [Version 0.3.4; Henry & Wickham (2020)], *Rcpp* [Version 1.0.6; Eddelbuettel & François (2011); Eddelbuettel & Balamuta (2018)], *readr* [Version 1.4.0; Wickham & Hester (2020)], *readxl* [Version 1.3.1; Wickham & Bryan (2019)], *rworldmap* [Version 1.3.6; South (2011)], *see* [Version 0.6.3; Lüdecke, Ben-Shachar, Waggoner, & Makowski (2020)], *sp* [Version 1.4.5; Pebesma & Bivand (2005)], *stringr* [Version 1.4.0; Wickham (2019)], *tibble* [Version 3.1.1; Müller & Wickham (2021)], *tidyr* [Version 1.1.3; Wickham (2021b)], *tidyverse* [Version 1.3.1; Wickham et al. (2019)], and *xfun* [Version 0.22; Xie (2021)].

Table 1

Mean, median, standard deviation, and range of correct response times (in ms) and response accuracy (proportion of all responses) in Experiment 1 for words and nonwords and by prime duration and priming condition.

Stimulus Type	Prime Duration	Condition	Mean	Median	SD	Min	Max	Accuracy
Nonword	33 ms	Identical	700	664	189	263	1,885	0.91
Nonword	33 ms	Unrelated	704	664	194	256	1,949	0.92
Nonword	50 ms	Identical	699	657	191	256	1,938	0.91
Nonword	50 ms	Unrelated	700	656	192	251	1,952	0.93
Word	33 ms	Identical	630	598	168	254	1,942	0.94
Word	33 ms	Unrelated	652	618	170	254	1,887	0.93
Word	50 ms	Identical	608	574	161	265	1,866	0.96
Word	50 ms	Unrelated	648	619	158	294	1,801	0.93

Note. Trials with reaction times below 250 ms were excluded from the analysis. For the calculation of response times, incorrect responses were also removed.

Results

Descriptive statistics for response times and accuracy for both words and nonwords in the experimental conditions are reported in Table 1 (although note that we only analyzed the word trials).

Response times. Table 2 shows the mean, standard error, lower and upper bounds of the 95% credible interval (CrI) of the estimate of each fixed effect in the RT model, as well as the $\hat{R}$ for each estimate, which indicate that the model was fitted successfully as they are all close to 1.

Table 2

Posterior mean, standard error (SE), 95% credible interval and $\hat{R}$ statistic for the fixed effects of the model fitted for correct word response times in Experiment 1.

Parameter	mean	SE	lower bound	upper bound	$\hat{R}$
Intercept (μ)	633.620	6.504	620.936	646.430	1.009
Intercept (β)	4.669	0.031	4.607	4.729	1.003
Condition (μ)	31.353	1.878	27.659	35.006	1.001
Prime Duration (μ)	-11.382	1.704	-14.746	-8.018	1.001
Condition:Prime Duration (μ)	19.309	3.699	12.141	26.455	1.000
Condition (β)	0.032	0.020	-0.007	0.070	1.000
Prime Duration (β)	-0.081	0.022	-0.124	-0.038	1.000
Condition:Prime Duration (β)	0.029	0.041	-0.052	0.110	1.000

Note. β is the scale parameter (the inverse of the rate parameter λ) of the ex-Gaussian distribution.

The RT model indicates that response times were higher in the unrelated condition than the identical (b = 31.35, 95% CrI [27.66, 35.01]) and higher in the 33 ms prime duration condition than in the 50 ms prime duration condition (b = -11.38, 95% CrI [-14.75, -8.02]). The interaction term indicates that the priming effect was stronger in the 50 ms condition than the 33 ms condition (b = 19.31, 95% CrI [12.14, 26.46]). Figure 1 shows the posterior distributions of the fixed effects parameters, with various highest density intervals highlighted.

Accuracy. Table 3 shows the mean, standard error, lower and upper bounds of the 95% credible interval (CrI) of the estimate of each fixed effect in the accuracy model, as well as the $\hat{R}$ for each estimate, which indicate that the model was fitted successfully as they are all close to 1.

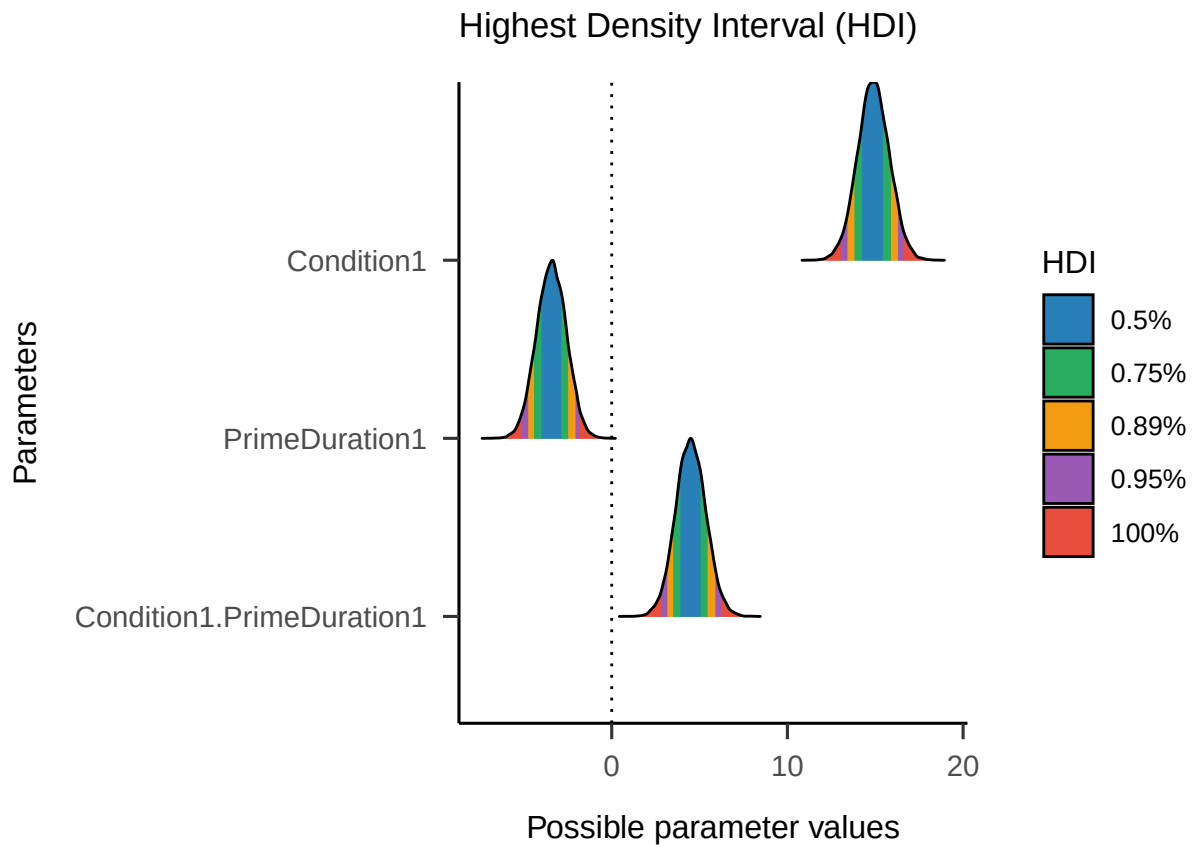

Figure 1. 50%, 75%, 89%, 95%, and 100% highest density intervals (HDIs) of the posterior distributions for each of the parameters of the response time model in Experiment 1.

Table 3

Posterior mean, standard error, 95% credible interval and $\hat{R}$ statistic for the fixed effects of the model fitted for response accuracy on word trials in Experiment 1.

Parameter	mean	SE	lower bound	upper bound	$\hat{R}$
Intercept	3.451	0.122	3.217	3.695	1.000
Condition	-0.496	0.105	-0.714	-0.301	1.000
Prime Duration	0.195	0.109	-0.012	0.417	1.000
Condition:Prime Duration	-0.533	0.204	-0.943	-0.145	1.000

The accuracy model indicates that participants were less likely to produce a correct response in the unrelated condition than the identical condition (b = -0.50, 95% CrI [-0.71, -0.30]). The mean of the posterior distribution for prime duration suggests that accuracy was was slightly lower in the 33 ms prime duration condition than in the 50 ms prime duration condition, but as the CrI included 0, this is not credible (b = 0.20, 95% CrI [-0.01, 0.42]). The interaction term indicates that the effect of priming condition on response accuracy (with the identical condition leading to higher accuracy) was stronger in the 50 ms condition than the 33 ms condition (b = -0.53, 95% CrI [-0.94, -0.14]).

Discussion

The results from Experiment 1 show that we were able to replicate masked priming effects using an online experiment. The size of the effect is similar to that observed in previous studies. We also saw a clear difference between the 33 ms prime duration and the 50 ms prime duration, indicating that the experiment can reliably implement timing differences of up to one frame across a variety of participant devices. We did not observe an effect of either the priming condition or the prime exposure duration on the shape parameter β of the exponential distribution, suggesting, according to the savings hypothesis by K. Forster (1998), that only encoding processes were affected.

Of course, just because there was a difference between the conditions, this does not necessarily mean that the timings in the two prime duration conditions actually corresponded to the display durations set in the experiment script, just that they were different. In order to further explore this, we performed a second experiment in which we set the prime to be displayed for an even shorter duration. As described above, a 16.6 ms prime duration should yield qualitatively different effects from the 33.3 and 50 ms durations used in Experiment 1. If it does not, this casts doubt on the timing accuracy in online experiments.

Experiment 2

In the second experiment, we tested whether we could observe reliable effects of masked priming at prime durations of 16 ms and 33 ms (roughly corresponding to one and two frames at a refresh rate of 60 Hz).

Method

Participants. As in Experiment 1, participants were recruited through Prolific (www.prolific.co, 2021). The experiment was accessed by 102 participants. Out of these, 87 provided experimental data. One of these participants did not complete the experiment. A further seven were excluded because of low accuracy (less than .8). The remaining 79 participants were aged from 18 to 69 (mean age: 31.14). Of the participants, 40 identified as male, 39 identified as female, and 0 identified as other. The experiment was only shown to participants who indicated that English was their first language in the Prolific screening questions. Based on their IP addresses, 56 participants were based in the UK, 14 were based in the US, two participants were based in Canada, two participants were based in South Africa, and one participant each was based in Hungary and Ireland. Three participants could not be localized in this way. As in Experiment 1, participants who completed the experiment were paid £1.25 for their participation (corresponding to £5/hour). All participants were naïve to the purpose of the experiment. Participants could use a desktop or laptop computer or a mobile device. Because of a technical display issue with PsychoJS and the Safari browser, participants who tried to access the experiment using that browser, including all participants on iOS devices, were advised to change browser or device and restart the experiment.

Rationale for sample size and stopping rule.

As in Experiment 1, our stopping rule was to keep collecting data until 3,000 valid observations were collected. This goal was met and exceeded in our initial data collection

with a budget of £200.

Materials. The materials were identical to those used in Experiment 1.

Procedure. The procedure was identical to Experiment 1, the only difference being that the primes were set to be displayed for either 16 ms or 33 ms.

Data analysis. We analyzed the data in the same way as in Experiment 1, again only analyzing trials where the target stimulus was a word. For the response time (RT) analysis, we excluded trials with RTs lower than 250 ms and incorrect responses (4.86 % of trials). For the accuracy analysis, we only excluded trials with RTs lower than 250 ms (0.08 % of trials). For priming duration, 16 ms was coded as -1 and 33 ms was coded as 1. Otherwise, the model specifications were identical to those in Experiment 1.

Results

Table 5

Posterior mean, standard error, 95% credible interval and $\hat{R}$ statistic for the fixed effects of the model fitted for correct word response times in Experiment 2.

Parameter	mean	SE	lower bound	upper bound	$\hat{R}$
Intercept (μ)	660.049	7.857	643.917	675.148	1.007
Intercept (β)	4.755	0.010	4.735	4.775	1.000
Condition (μ)	9.822	1.929	6.032	13.625	1.001
Prime Duration (μ)	-5.112	1.995	-9.003	-1.238	1.000
Condition:Prime Duration (μ)	14.800	3.808	7.262	22.247	1.000
Condition (β)	0.038	0.018	0.002	0.073	1.000
Prime Duration (β)	-0.007	0.018	-0.043	0.029	1.000
Condition:Prime Duration (β)	-0.014	0.036	-0.084	0.057	1.000

Note. β is the scale parameter (the inverse of the rate parameter λ) of the ex-Gaussian distribution.

Table 4

Mean, median, standard deviation, and range of correct response times (in ms) and response accuracy (proportion of all responses) in Experiment 2 for words and nonwords and by prime duration and priming condition.

Stimulus Type	Prime Duration	Condition	Mean	Median	SD	Min	Max	Accuracy
Nonword	16 ms	Identical	721	680	191	330	1,929	0.93
Nonword	16 ms	Unrelated	719	677	187	284	1,920	0.94
Nonword	33 ms	Identical	719	668	198	351	1,924	0.94
Nonword	33 ms	Unrelated	718	672	197	281	1,912	0.94
Word	16 ms	Identical	660	625	162	311	1,892	0.94
Word	16 ms	Unrelated	662	623	169	308	1,899	0.95
Word	33 ms	Identical	650	618	158	280	1,873	0.96
Word	33 ms	Unrelated	667	633	164	269	1,920	0.94

Note. Trials with reaction times below 250 ms were excluded from the analysis. For the calculation of response times, incorrect responses were also removed.

The model indicates that response times were higher in the unrelated condition than the identical condition (b = 9.82, 95% CrI [6.03, 13.62]), although across the two prime duration conditions, the effect was much smaller than in Experiment 1. Response times were also higher in the 16 ms than in the 33 ms prime duration condition (b = -5.11, 95% CrI [-9.00, -1.24]) although, as the interaction effect shows, this was mainly due to an absence of the priming effect in the 16 ms prime duration condition, as opposed to the 33 ms condition, which showed a robust priming effect (b = 14.80, 95% CrI [7.26, 22.25]). Figure 2 shows the posterior distributions of the fixed effects parameters, with various highest density intervals highlighted.

```
## Warning: Identical densities found along different segments of the distribution,
## choosing rightmost.
```

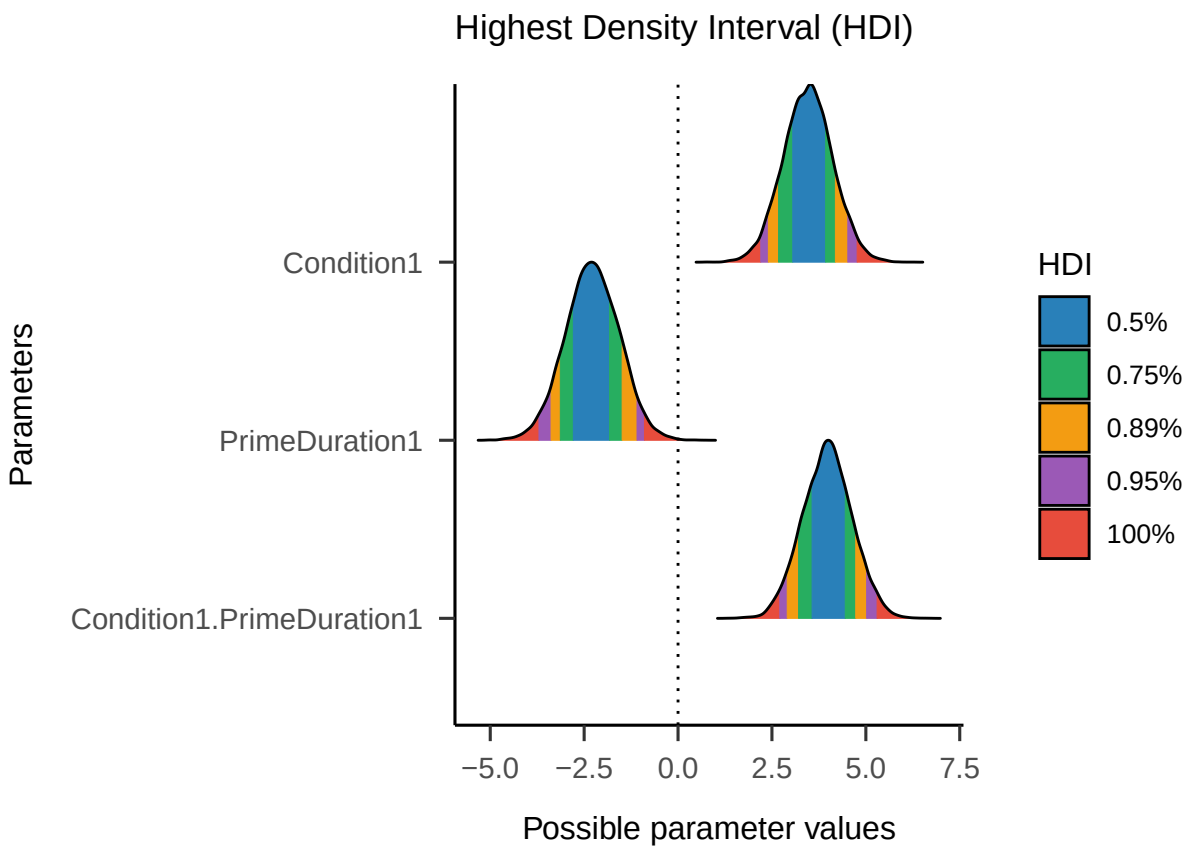

Figure 2. 50%, 75%, 89%, 95%, and 100% highest density intervals (HDIs) of the posterior distributions for each of the parameters of the response time model in Experiment 2.

Table 6

Posterior mean, standard error, 95% credible interval and $\hat{R}$ statistic for the fixed effects of the model fitted for response accuracy on word trials in Experiment 2.

Parameter	mean	SE	lower bound	upper bound	$\hat{R}$
Intercept	3.736	0.135	3.475	4.007	1.000
Condition	-0.218	0.109	-0.434	-0.004	1.000
Prime Duration	-0.054	0.126	-0.313	0.183	1.000
Condition:Prime Duration	-0.708	0.219	-1.165	-0.302	1.000

The accuracy model indicates that participants were less likely to produce a correct response in the unrelated condition than the identical condition (b = -0.22, 95% CrI [-0.43, 0.00]). The mean of the posterior distribution for prime duration suggests that accuracy was was slightly lower in the 16 ms prime duration condition than in the 33 ms prime duration condition, but as the CrI included 0, this is not credible (b = -0.05, 95% CrI [-0.31, 0.18]). The interaction term indicates that the expected effect of priming condition on response accuracy (with the identical condition leading to higher accuracy) only present in the 33 ms condition, and reversed in the 16 ms condition (b = -0.71, 95% CrI [-1.16, -0.30]), although the effect in the 16 ms condition was very weak.

Discussion

In Experiment 2, we observed the pattern we expected from previous research: In the the mean μ of the Gaussian distribution, we saw a robust priming effect in response time and accuracy in the 33 ms condition, but a very weak effect in the 16 ms condition. This suggests that the timing in our online experiments was likely to be quite close to the timing set in the experiment script. Again, we did not observe an effect of either the priming

condition or the prime exposure duration on the shape parameter β of the exponential distribution.

General Discussion

In this study, we set out to test whether using an online, browser-based experiment software, we could replicate the masked priming effect both qualitatively and quantitatively, i.e., in terms of effect sizes. Our results show that this is the case: not only did we observe results comparable to the lab-based studies in the literature, we also observed the effect sizes predicted by Forster's (1998) savings hypothesis: Our data show a shift in the mean of the Gaussian distribution, but no change in the shape parameter β of the exponential distribution, suggesting that our priming manipulations affected — as intended — encoding processes, but not conscious decision making processes.

We also found that the size of the priming effect was directly influenced by the prime exposure duration. This indicates that the experimental software was able to control the display timing of the prime accurately. Accurate display timings are expected in a laboratory-based experiment, where the equipment is known and can be measured, but are much less certain in a situation where the experiment runs on a participant's own device, which could be any of a wide variety of consumer devices including Windows PCs, Macs, tablets, and mobile phones sold in the last decade. The fact that we can observe results that very closely resemble lab results demonstrates the sophistication in modern browsers' Javascript performance, including browsers on mobile devices, as well as the quality of the Javascript implementation of Psychopy.

Our results open the door to a wider use of online experiments in cognitive research, especially in reaction-time sensitive fields like word recognition. Not only does this enable researchers to continue collecting data in times of social distancing, but it also makes it possible to collect data from a larger population than previously possible. For example,

many people in developing countries may not have computers at home, but they do have smartphones. Using a Javascript-based experiment software, anyone with a smartphone can be an experiment participant. Deploying experiments online also means that data can be collected very quickly and efficiently, allowing research to progress more rapidly.

In conclusion, our results give us confidence that high-quality behavioral data can be collected online using Javascript-based experiment platforms. We hope that future research takes advantage of these new methods in order to make research faster, more inclusive, and more efficient.

References

Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2020). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01501-5

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

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

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*(2), 527–535. https://doi.org/10.3758/s13428-012-0265-2

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

Birnbaum, M. H., & Birnbaum, M. O. (2000). *Psychological Experiments on the Internet*. Elsevier. Retrieved from http://books.google.com?id=sw4NHyThFFkC

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., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition*, *1*(1), 9. https://doi.org/10.5334/joc.10

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Cai, Z. G., Gilbert, R. A., Davis, M. H., Gaskell, M. G., Farrar, L., Adler, S., & Rodd, J. M. (2017). Accent modulates access to word meaning: Evidence for a speaker-model account of spoken word recognition. https://doi.org/10.31234/osf.io/5x3tb

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Eddelbuettel, D., & Balamuta, J. J. (2018). Extending extitR with extitC++: A Brief Introduction to extitRcpp. *The American Statistician*, *72*(1), 28–36. https://doi.org/10.1080/00031305.2017.1375990

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. https://doi.org/10.18637/jss.v040.i08

Eerland, A., Engelen, J. A. A., & Zwaan, R. A. (2013). The influence of direct and indirect speech on mental representations. *PLoS ONE*, *8*(6), e65480. https://doi.org/10.1371/journal.pone.0065480

Forster, K. (1998). The pros and cons of masked priming. *Journal of Psycholinguistic Research*, *27*(2), 203–233.

Forster, K., & Forster, J. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*,

35(1), 116–124.

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 680–698. https://doi.org/10.1037/0278-7393.10.4.680

Gomez, P., & Perea, M. (2020). Masked identity priming reflects an encoding advantage in developing readers. *Journal of Experimental Child Psychology*, *199*, 104911. https://doi.org/10.1016/j.jecp.2020.104911

Gomez, P., Perea, M., & Ratcliff, R. (2013). A diffusion model account of masked versus unmasked priming: Are they qualitatively different? *Journal of Experimental Psychology: Human Perception and Performance*, *39*(6), 1731–1740. https://doi.org/10.1037/a0032333

Grainger, J. (2008). Cracking the orthographic code: An introduction. *Language and Cognitive Processes*, *23*(1), 1–35. https://doi.org/10.1080/01690960701578013

Grossi, G. (2006). Relatedness proportion effects on masked associative priming: An ERP study. *Psychophysiology*, *43*(1), 21–30. https://doi.org/10.1111/j.1469-8986.2006.00383.x

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

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208. https://doi.org/10.3758/BF03204766

Lüdecke, D., Ben-Shachar, M. S., Waggoner, P., & Makowski, D. (2020). See: Visualisation toolbox for 'easystats' and extra geoms, themes and color palettes for 'ggplot2'. *CRAN*. https://doi.org/10.5281/zenodo.3952153

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, *4*(40), 1541. https://doi.org/10.21105/joss.01541

Müller, K., & Wickham, H. (2021). *Tibble: Simple data frames*. Retrieved from https://CRAN.R-project.org/package=tibble

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). *Running Experiments on Amazon Mechanical Turk* (SSRN Scholarly Paper No. ID 1626226). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=1626226

Pavlovia. (2020). Pavlovia. Retrieved from https://pavlovia.org/

Pebesma, E. J., & Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*, *5*(2), 9–13. Retrieved from https://CRAN.R-project.org/doc/Rnews/

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Perea, M., & Rosa, E. (2002). Does the proportion of associatively related pairs modulate the associative priming effect at very brief stimulus-onset asynchronies? *Acta Psychologica*, *110*(1), 103–124. https://doi.org/10.1016/s0001-6918(01)00074-9

Prolific. (2021). Prolific | online participant recruitment for surveys and market research. Retrieved from https://www.prolific.co/

Qualtrics. (2020). Qualtrics XM // the leading experience management software. Retrieved from https://www.qualtrics.com/uk/

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

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*(1), 159–182. https://doi.org/10.1037/0033-295X.111.1.159

Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript Web experiments. *Behavior Research Methods*, *47*(2), 309–327. https://doi.org/10.3758/s13428-014-0471-1

Rezlescu, C., Danaila, I., Miron, A., & Amariei, C. (2020). Chapter 13 - More time for science: Using Testable to create and share behavioral experiments faster, recruit better participants, and engage students in hands-on research. In B. L. Parkin (Ed.) (Vol. 253, pp. 243–262). Elsevier. https://doi.org/10.1016/bs.pbr.2020.06.005

Rodd, J. M., Cai, Z. G., Betts, H. N., Hanby, B., Hutchinson, C., & Adler, A. (2016). The impact of recent and long-term experience on access to word meanings: Evidence from large-scale internet-based experiments. *Journal of Memory and Language*, *87*, 16–37. https://doi.org/10.1016/j.jml.2015.10.006

South, A. (2011). Rworldmap: A new r package for mapping global data. *The R Journal*, *3*(1), 35–43. Retrieved from http://journal.r-project.org/archive/2011-1/RJournal_2011-1_South.pdf

Taikh, A., & Lupker, S. J. (2020). Do visible semantic primes preactivate lexical representations? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(8), 1533–1569. https://doi.org/10.1037/xlm0000825

Tzur, B., & Frost, R. (2007). SOA does not reveal the absolute time course of cognitive processing in fast priming experiments73. *Journal of Memory and*

Language, *56*(3), 321–335. https://doi.org/10.1016/j.jml.2006.11.007

Vandeberg, L., Eerland, A., & Zwaan, R. A. (2012). Out of mind, out of sight: Language affects perceptual vividness in memory. *PLoS ONE*, *7*(4), e36154. https://doi.org/10.1371/journal.pone.0036154

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H. (2021a). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from https://CRAN.R-project.org/package=forcats

Wickham, H. (2021b). *Tidyr: Tidy messy data*. Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Bryan, J. (2019). *Readxl: Read excel files*. Retrieved from https://CRAN.R-project.org/package=readxl

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Hester, J. (2020). *Readr: Read rectangular text data*. Retrieved from https://CRAN.R-project.org/package=readr

Witzel, J., Cornelius, S., Witzel, N., Forster, K. I., & Forster, J. C. (2013). Testing the viability of webDMDX for masked priming experiments. *The Mental Lexicon*, *8*(3), 421–449. https://doi.org/10.1075/ml.8.3.07wit

Xie, Y. (2021). *Xfun: Miscellaneous functions to support packages maintained by 'yihui xie'*. Retrieved from https://CRAN.R-project.org/package=xfun

Yang, H., Jared, D., Perea, M., & Lupker, S. J. (2021). Is letter position coding when reading in L2 affected by the nature of position coding used when bilinguals read in their L1? *Memory & Cognition*. https://doi.org/10.3758/s13421-020-01126-1

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971–979. https://doi.org/10.3758/PBR.15.5.971

Ziegler, J. C., Ferrand, L., Jacobs, A. M., Rey, A., & Grainger, J. (2000). Visual and Phonological Codes in Letter and Word Recognition: Evidence from Incremental Priming. *The Quarterly Journal of Experimental Psychology*, *53A*(3), 671–692.