Behav Res (2016) 48:1718–1724
DOI 10.3758/s13428-015-0678-9

Reaction time effects in lab- versus Web-based research: Experimental evidence

Benjamin E. Hilbig[1,2]

Published online: 5 November 2015

Abstract Although Web-based research is now commonplace, it continues to spur skepticism from reviewers and editors, especially whenever reaction times are of primary interest. Such persistent preconceptions are based on arguments referring to increased variation, the limits of certain software and technologies, and a noteworthy lack of comparisons (between Web and lab) in fully randomized experiments. To provide a critical test, participants were randomly assigned to complete a lexical decision task either (a) in the lab using standard experimental software (E-Prime), (b) in the lab using a browser-based version (written in HTML and JavaScript), or (c) via the Web using the same browser-based version. The classical word frequency effect was typical in size and corresponded to a very large effect in all three conditions. There was no indication that the Web- or browser-based data collection was in any way inferior. In fact, if anything, a larger effect was obtained in the browser-based conditions than in the condition relying on standard experimental software. No differences between Web and lab (within the browser-based conditions) could be observed, thus disconfirming any substantial influence of increased technical or situational variation. In summary, the present experiment contradicts the still common preconception that reaction time effects of only a few hundred milliseconds cannot be detected in Web experiments.

Keywords Web · Internet · Response latency · Reaction time · Word frequency effect

Over the past two decades, research conducted online via the Internet has become increasingly frequent. Today, Web-based research is common across the whole range of social and behavioral sciences. This trend is not surprising, given the well-documented advantages of Web-based research, especially the possibility to recruit large, heterogeneous (and more representative) samples in less time and with lower costs than in traditional lab- or paper/pencil-based research (for overviews, see Birnbaum, 2004; Gosling, Vazire, Srivastava, & John, 2004; Kraut et al., 2004; Reips & Birnbaum, 2011; Skitka & Sargis, 2006).

Beyond these well-documented advantages, a growing body of literature has confirmed that data obtained and results found in Web-based studies are generally comparable to those generated by traditional lab- or paper/pencil-based research—for example, in research on personality (Chuah, Drasgow, & Roberts, 2006; Cronk & West, 2002; Lang, John, Lütdke, Schupp, & Wagner, 2011), ability (Ihme et al., 2009), or perception and cognition (Corley & Scheepers, 2002; Germine et al., 2012; Linnman, Carlbring, Ahman, Andersson, & Andersson, 2006; Reimers & Stewart, 2007). Nonetheless, "researchers doing Web-based experiments can encounter skepticism from reviewers and editors" (Germine et al., 2012, p. 848), especially "skepticism about the accuracy of response time measures recorded in a Web browser online" (Reimers & Stewart, 2015, p. 310).

Specifically, although there is no conclusive evidence to show that Web-based measurement of response latencies is

Electronic supplementary material The online version of this article (doi:10.3758/s13428-015-0678-9) contains supplementary material, which is available to authorized users.

✉ Benjamin E. Hilbig
hilbig@uni-landau.de

[1] Cognitive Psychology Laboratory, Department of Psychology, University of Koblenz-Landau, Fortstraße 7, 76829 Landau, Germany

[2] Max Planck Institute for Research on Collective Goods, Bonn, Germany

inherently problematic, Birnbaum (2004) noted that "brief time intervals . . . are *thought* to be less precisely . . . measured via the Web" (p. 824, emphasis added). That is, as comments from editors and/or reviewers regularly reveal, "the measurement of response time in a Web experiment is *perceived* to be problematic" (Brand & Bradley, 2012, p. 350, emphasis added). Essentially, there is a persistent preconception that "the Internet may not be optimal for research that is dependent on detecting . . . small differences in response time" (Skitka & Sargis, 2006, p. 547). Thus, although several well-established reaction time effects have been replicated in Web-based research (e.g., Crump, McDonnell, & Gureckis, 2013; Keller, Gunasekharan, Mayo, & Corley, 2009; Simcox & Fiez, 2014), skepticism remains widespread—for at least three reasons.

First, one of the core arguments fueling skepticism about Web-based response time measurement lies in the inherent and indubitable increase in technical and situational variance as compared to the lab (Reips, 2002). Unlike in the lab, Web-based data necessarily stem from many different computers, displays, input devices, operating systems, and Web browsers. In light of existing evidence that different input devices (mice/keyboards) or variations in the number of parallel processes (e.g., other applications running) will indeed affect reaction time measurement (Plant, Hammond, & Whitehouse, 2003; Plant & Turner, 2009), technical variation may increase unexplained error variance. In addition, Web-based research comes with less control over aspects of the situation (e.g., the lighting, participant's viewing position, time of day, or distractions), which may further increase error variance. On the other hand, simulations have shown that the effects of increased error variance are unlikely to offset the advantages in terms of statistical power and more precise effect estimates due to larger sample sizes (Brand & Bradley, 2012).

A second type of concern is with the software and technologies used. That is, some technologies that are suitable for reaction time measurement—such as JAVA applets (Hecht, Oesker, Kaiser, Civelek, & Stecker, 1999) or Adobe Flash (Linnman et al., 2006; Reimers & Stewart, 2007, 2015)—require special software or plugins that may not be available to all potential participants, and more problematically yet, their availability may vary systematically with the characteristics of the users, thus creating potential confounds (Reips & Krantz, 2010). Also, some technologies have been shown to provide inaccurate timing if no countermeasures are taken (Eichstaedt, 2001). The most widely applicable technology (in terms of availability on client machines) offering millisecond resolution is JavaScript (de Leeuw, 2015; Reips & Krantz, 2010). So far, investigations using nonhuman response systems have demonstrated that JavaScript provides adequately accurate timing under most conditions (Reimers & Stewart, 2015), and a recent experiment with human response data has confirmed this conclusion (de Leeuw & Motz, 2015).

Third, and most importantly, it must be acknowledged that most empirical comparisons of Web- versus lab-based reaction time effects (and indeed other effects) suffer an unfortunate methodological drawback: Typically, the results obtained in different, independent samples are compared. For example, Corley and Scheepers (2002) compared their priming results obtained in a Web-based sample to the lab-based data from a previous, independent study. On the basis of high consistency across the two studies, they concluded that Web-based research is "valid." Similarly, McGraw, Tew, and Williams (2000) compared the results of several Web-based paradigms to well-established effects previously found in the lab and concluded that Web-based data can be trusted, given that they reliably mirror said established effects. More recent investigations have similarly based their conclusions on cross-sample comparisons (e.g., Germine et al., 2012; Linnman et al., 2006). In what has arguably been the largest set of studies to date, Crump et al. (2013) replicated an impressive series of established effects—including Stroop, flanker, Posner cueing, attentional blink, and subliminal priming—on the Web using Amazon Mechanical Turk (see also Simcox & Fiez, 2014). Although insightful and indeed encouraging, the trouble with all of these comparisons is that, strictly speaking, there is no control over possible confounds. Since participants were not randomly assigned to lab- versus Web-based data collection, the comparisons remain inconclusive and cannot be tested statistically. Stated simply, "[f]ailure to find a difference tells us nothing unless we are sure that the samples compared really do not differ on the constructs of interest . . . ," implying that one must "[r]andomly assign participants to Web versus Lab condition when performing such comparisons" (Reips, Buchanan, Krantz, & McGrawn, in press, MS p. 8).

The most notable recent exception has been the experiment by de Leeuw and Motz (2015), who manipulated within subjects whether a visual search task was performed using JavaScript versus MATLAB's Psychophysics Toolbox (for a similar experiment comparing Adobe Flash—in the lab and on the Web—to a program written in C, see Reimers & Stewart, 2007). Thus, by assessing real human responses and systematically manipulating the underlying technology, their comparison allows for conclusions about the equivalence of the technologies in practice—that is, the extent to which actual empirical effects will be found with comparable reliability and precision. Indeed, they found no substantial differences between the software packages and concluded that JavaScript thus "offers suitable sensitivity for the measurement of response time differences between conditions in common psychophysical research."

However, despite these promising results, the experiment by de Leeuw and Motz (2015) is limited to comparisons of software/technology within the lab. That is, their setup did not include a fully Web-based condition, and thus cannot address the concern above regarding technical and situational variance. Aiming to extend their work, the experiment reported in what follows was designed to further tease apart the potential effects of different sources of variation or error. Most importantly, the goal was to test whether Web-based reaction time measurement is offset by the mere technical and situational variance that is usually absent in the lab (see the first point above). At the same time, it is vital to separate such a potential effect from the error that may be inherent in the technologies and software used (see the second point above). Although the latter concern per se is alleviated by the findings of de Leeuw and Motz, it seemed prudent to provide another test, using a different experimental design, other software for comparison, and a different type of task.

Experiment

Design, procedure, and participants

For the present purpose, the well-known word frequency effect in lexical decisions—that frequent words are detected faster as words (over nonwords or pseudowords) than less frequent words (Gordon, 1983; Rubenstein, Garfield, & Millikan, 1970)—was chosen. This effect is robust and reliable, but nonetheless is typically only 150–200 ms in size. At the same time, it is a genuine within-subjects effect that is particularly useful here, since it allows for substantial statistical power: Testing whether the word frequency effect is equivalent across the between-subjects conditions of interest (see below) corresponds to an *F* test of a within–between interaction that in turn requires only a moderate sample size, even for relatively small interaction effects (Faul, Erdfelder, Lang, & Buchner, 2007).

To perform the comparisons of interest outlined above, the present experiment comprised three (between-subjects) conditions: First, the lexical decision task was implemented in the lab, using standard software for psychological experimenting, namely E-Prime (Schneider, Eschman, & Zuccolotto, 2002). This condition (termed "lab/E-Prime" in what follows) can be considered the benchmark or baseline. The second and third conditions implemented the same lexical decision task for the Web browser using a "low-tech" solution (Reips & Krantz, 2010). Specifically, the task was written in HTML (with PHP controlling the task flow and handling HTML forms), and reaction time measurement was implemented via a simple JavaScript using an event-handler function for the "keydown" event. The essence of the code used to achieve the reaction time measurements can be found in the Appendix. Importantly, the second condition was run in the exact same lab as the first, and is therefore referred to as "lab/browser." The only difference to the "lab/E-Prime" condition was thus the technology used for reaction time measurement (E-Prime vs. Web browser with HTML/JavaScript), whereas all other aspects (same lab, computers, etc.) were equivalent. The third condition, by contrast, was a genuine Web-based condition in which the HTML/JavaScript version of the task was completed by participants on whatever computer (in whichever place) they desired. This "Web/browser" condition is thus fully equivalent to the second, except for the place (lab vs. Web) and the differences in technical and situation variation it comes with. In summary, the design allows for an in-depth analysis, not only of whether the lab and Web differ but—if so—also dissecting two aspects: Differences due to software and technology can be tested by comparing "lab/E-Prime" with "lab/browser," whereas differences due to variation (i.e., technical and situational heterogeneity) can be tested by comparing "lab/browser" with "Web/browser." Note that this includes differences due to the presence versus absence of an experimenter (Ollesch, Heineken, & Schulte, 2006): The two lab conditions were equivalent in terms of experimenter presence (the same experimenters ran all lab-based sessions, to which they were randomly assigned), whereas the Web condition did not involve an experimenter (but possibly other unknown individuals).

The lexical decision task requested participants to judge—as speedily and accurately as possible—whether or not six-letter strings represented words, by pressing one of two keys. As materials, a total of 200 German six-letter nouns (half of which were high vs. low in word frequency, respectively) with 200 matched pseudowords (created by replacing one letter from the words) were used, taken from a previous psycholinguistic experiment (Albrecht & Vorberg, 2010, Exp. 2).[1] For each participant, 140 items were randomly selected and shown one at a time in random order (with a 1,000-ms intertrial interval); on exactly half of the trials (70 in total) a word was displayed, whereas a pseudoword was displayed on the remaining half of trials. The entire experiment (including informed consent and demographics, instructions, the lexical decision task, and debriefing) lasted about 10 min, on average.

A total of 67 participants (35 male, 32 female, between 18 and 32 years of age; *M* = 21 years, *SD* = 2.3 years) were recruited from a local participant pool. All were invited via e-mail and registered for the experiment online. The online registration system randomly assigned participants to the three between-subjects conditions outlined above, with the constraint that participants were assigned to the Web/browser condition with a higher probability, so as to counteract the potentially higher drop-out rate (although, ultimately, no drop-outs occurred in any of the conditions). Consequently, there were *n* = 28 participants in the Web/browser condition, *n* = 20 in the lab/browser condition, and *n* = 19 in the lab/E-Prime condition. Participants in the Web/browser condition completed the experiment online at a place and time of their choosing, within one week of having registered. The remaining participants signed up for a lab session within the same week. All participants were paid a flat fee of €2.00 (approximately USD 2.75 at the time).

[1] I thank Thorsten Albrecht for providing his materials.

Results

Reaction times from the lexical decision task served as the dependent variable (the complete raw data are available as supplementary material). To reduce the influence of outliers, the first five trials of each participant were disregarded, as well as all trials in which the reaction time was more than 2.5 standard deviations above or below the individual mean reaction time (2.7 % of trials).[2] Descriptives characterizing the reaction time distributions in each of the three experimental conditions are summarized in Table 1. As can be seen, there was a trend toward shorter reaction times in the lab/E-Prime condition, which is in line with previous findings that JavaScript produces slightly longer times both in an automated response system (Neath, Earle, Hallett, & Surprenant, 2011) and in human data (de Leeuw & Motz, 2015). At the same time, the smallest degree of variability was observed in the lab/browser condition, and the largest in the Web/browser condition, implying that variance is not primarily due to software or technology, but rather is caused by situational and technical variation (which is greater on the Web than in the lab).

All statistical comparisons were based on individual median reaction times (and double checked with individual mean log-transformed reaction times, which yielded equivalent results). Participants' overall accuracy was high (M = 94 %, SE = 0.5 %), and the mean of their median reaction time across all trials (M = 958 ms, SE = 30 ms) was in the range typical for this type of task (cf. Rubenstein et al., 1970). Across all (between-subjects) conditions, responses were made more speedily to words (M = 770 ms, SE = 15 ms) than to pseudowords (M = 968 ms, SE = 33 ms), $t(66) = 8.2$, $p < .001$, Cohen's d = 0.99. More importantly, high-frequency words were more speedily accepted as words (M = 697 ms, SE = 12 ms) than were low-frequency words (M = 878 ms, SE = 23 ms), thus mirroring the primary effect of interest, $t(66) = 12.4$, $p < .001$, Cohen's d = 1.52.

To test the main question of interest, the word frequency effect was considered depending on the between-subjects condition (lab/E-Prime vs. lab/browser vs. Web/browser). The effects (the mean difference between participants' median reaction times for high- vs. low-frequency words) per condition are reported in Table 2. As can be seen, the effect was substantial in all three conditions, albeit somewhat larger in the two browser-based conditions. To test the full pattern, a mixed analysis of variance was conducted on participants' median reaction times for low- versus high-frequency words (repeated measures factor), with Condition as a between-subjects factor.

Table 1 Descriptives of reaction time distributions in the raw data (excluding the first five trials and outliers as described in the main text)

	Lab/E-Prime	Lab/Browser	Web/Browser
Mean (*SE*)[a]	961 (13)	988 (10)	1,071 (14)
5 %-trimmed mean [a]	873	916	960
SD[a]	666	520	832
Median[a]	784	847	856
IQR[a]	397	353	453
Skewness (*SE*)[a]	7.7 (.05)	4.1 (.05)	7.6 (.04)
Vincentized percentiles[b]			
10th	589	662	637
20th	649	715	705
30th	709	765	767
40th	766	816	829
50th	833	878	901
60th	914	948	995
70th	1,009	1,047	1,106
80th	1,155	1,187	1,287
90th	1,492	1,437	1,653

IQR = interquartile range. [a] Computed across all trials and participants within a condition. [b] Percentiles of individual reaction time distributions averaged across participants within a condition (Ratcliff, 1979)

As expected, the word frequency effect was clearly replicated [$F(1, 64) = 150$, $p < .001$, Cohen's $f = 1.5$]. By contrast, no main effect of condition emerged [$F(2, 64) = 1.1$, $p = .34$, Cohen's $f = 0.19$], showing that the descriptive trend in the raw reaction time distributions was not statistically reliable. Most importantly, there was no interaction between word frequency and condition [$F(2, 64) = 0.49$, $p = .62$, Cohen's $f = 0.12$], confirming that the word frequency effects were essentially comparable in magnitude across all three conditions.

To rule out that the lack of statistical support for the interaction was due to insufficient power, a criterion power analysis was computed (Faul et al., 2007). The analysis revealed a critical F value of 2.2 (and, thus, a Type I error of $\alpha = .12$) to detect the observed effect ($f = 0.12$), with a power of $1 - \beta = .95$ (and thus a Type II error probability of .05), given the present sample size and correlation among the repeated measures (Spearman's $\rho = .85$ across all conditions). Clearly, the observed F value is well below this critical value, implying that the null hypothesis can be accepted within a conventional level of statistical error.[3]

Although the analyses above did not yield any indication of noteworthy differences between the lab- and Web-based reaction time measurements, more specific analyses using Helmert contrasts were conducted to compare the effect of software

[2] All results held when these criteria for inclusion were dropped. The results also held when only trials with accurate responses were considered. The numbers of outliers per participant did not differ between experimental conditions, $F(2, 64) = 0.08$, $p = .92$, Cohen's $f = 0.05$.

[3] The same even held when assuming a small effect ($f = 0.10$) by Cohen's (1988) conventions, rather than the observed effect. In this case, the critical F value was 1.23 (implying a Type I error rate of $\alpha = .30$). As reported, the observed F value is still well below this criterion.

Table 2 Word frequency effect separated by experimental (between-subjects) condition

	Mean Difference* (ms)	95 % Confidence Interval of the Mean Difference	*t* Test	Cohen's *d*
Lab/E-Prime	204	132; 277	$t(18) = 5.9^{***}$	1.35
Lab/browser	176	129; 222	$t(19) = 7.9^{***}$	1.76
Web/browser	170	127; 213	$t(27) = 8.1^{***}$	1.52

* Mean of the difference between each participant's median reaction time for low- versus high-frequency words. *** $p < .001$

and technology (E-Prime vs. browser/JavaScript) with the effect of situational and technical variation (lab vs. Web—within the browser/JavaScript conditions). Regressing the individual difference in median reaction times between high- and low-frequency words on the correspondingly coded contrasts revealed that the effect of software and technology (E-Prime vs. browser/JavaScript) was small and statistically nonsignificant ($\beta = .12$, $p = .35$), despite the descriptive tendency for a larger word frequency effect in the browser-based conditions (see above). Within the browser-based conditions, absolutely no evidence emerged ($\beta = .02$, $p = .87$) for an effect of lab versus Web (i.e., of technical and/or situational variation).

Discussion

Despite a growing body of evidence suggesting that Web-based data will yield results that are comparable to those obtained with more traditional methods (Germine et al., 2012; Gosling et al., 2004; Reips & Birnbaum, 2011), skepticism is still commonplace, especially concerning Web-based measurement of reaction times (Reimers & Stewart, 2015; Simcox & Fiez, 2014). Such reservations are fueled by (i) the indubitably increased technical and situational variance on the Web and (ii) limits in terms of software and technologies. Most importantly, (iii) there has been a lack of direct experimental comparisons between lab and Web—that is, comparisons based on random assignment (Reips et al., in press).

One of the few experimental investigations into the comparability of software packages was recently conducted by de Leeuw and Motz (2015), who demonstrated that JavaScript was largely equivalent, in terms of reaction time measurement, to the Psychophysics Toolbox. However, since their experiment only compared technologies within the lab, it seemed vital to extend their approach to comparisons of Web versus lab and of technologies, thus teasing apart the potential effects of different sources of variation or error. Consequently, the present experiment was designed to critically test whether a classic reaction time effect—the word frequency effect in lexical decisions (Rubenstein et al., 1970)—can be uncovered as reliably on the Web as in the lab, on the basis of full random assignment to the different conditions. Most importantly, I tested three conditions to allow for a more fine-grained analysis of potential differences: The first was lab-based and relied on the widely used E-Prime software ("lab/E-Prime"). The second was also lab-based, but implemented the task in HTML with a simple JavaScript for reaction time measurement ("lab/browser"). The third used the same technological implementation (HTML with JavaScript), but was conducted on the Web ("Web/browser"). Thereby, the effects of software and technology (lab/E-Prime vs. lab/browser) and of situational and technical variation (lab/browser vs. Web/browser) can be teased apart.

The results showed that the effect in question (the word frequency effect in reaction times) was typical in size (170–200 ms), statistically significant, and large (in terms of standardized effect size) in all conditions. Indeed, there was no indication of an interaction between word frequency (within subjects) and condition (between subjects), which confirms that the effects were equivalent across conditions. This finding was statistically confirmed by a criterion power analysis (Faul et al., 2007). Interestingly, if anything, the browser-based conditions produced the larger word frequency effect, although this was a mere descriptive trend, without strong statistical support. Nonetheless, it does imply that reaction time measurement using a browser and HTML/JavaScript is certainly no less appropriate than commonly used software such as E-Prime. This can be considered a conceptual replication of the results of de Leeuw and Motz (2015), using a different experimental design, software for comparison, and paradigm (and effect of interest). In addition, the comparison within the browser-based conditions further revealed that the increase in technical and situational variance inherent in the Web had practically no effect at all. This finding is well-aligned with previous work concluding that technical variation is little cause for worry (Brand & Bradley, 2012), but the first to demonstrate this using human response data and based on experimental manipulation (i.e., all else—including the underlying sample—being equal).

Note that, exactly because the design chosen herein compared different settings for the same population, it cannot provide an estimate of how much noisier Web studies will be due to sample differences in general. This, however, has been addressed by the many studies that have replicated lab-based effects with typical Web samples (e.g., Crump et al., 2013; Germine et al., 2012; Linnman et al., 2006). Thus, the present approach is complementary to the latter and to investigations of whether Web technologies can be considered adequately precise using automated response systems (Reimers & Stewart, 2015): It estimates the effects of technical and situational variation in human response data (holding any sample differences equal). Arguably, the best possible assessment of whether and when Web studies are adequate alternatives to classical lab experiments will come from considering the results of all of these approaches in combination. Note, also, that conclusions from one single task or paradigm need not generalize. Some confidence should come from the fact that the present investigation replicates the results of de Leeuw and Motz (2015) using in a different paradigm, but nonetheless, more experiments using still other tasks will be needed.

Overall, the present findings confirm previous research demonstrating the comparability of lab- and Web-based reaction time measurements (Corley & Scheepers, 2002; Crump et al., 2013; Germine et al., 2012; Linnman et al., 2006; McGraw et al., 2000; Reimers & Stewart, 2015; Simcox & Fiez, 2014), in this case using a simple, "low-tech" solution that can be applied without requiring additional software or plugins beyond a browser (Reips & Krantz, 2010). At the same time, due to reliance on random assignment, the present comparison complements the typical cross-study comparisons (Reips et al., in press) and goes beyond prior experiments (de Leeuw & Motz, 2015) by teasing apart different potential sources of error. In conclusion, the still commonplace skepticism whenever data—and even reaction time data requiring sufficient accuracy to uncover an effect less than 200 ms in size—are collected via the Web is no longer appropriate. Importantly, neither the prior investigations nor the present results discredit classical lab-based approaches in any way; rather, they demonstrate that Web/browser-based methods are a viable alternative that should not be treated with general a priori skepticism or suspicion.

Author Note I thank Martin Brandt for countless discussions and helpful suggestions, Theresa Strobel for preparing the lexical decision task in E-Prime, and a substantial number of anonymous reviewers (mostly of other manuscripts) for sometimes forceful reminders that this type of research is still sorely needed.

Appendix. JavaScript and HTML code (simplified extracts) used for reaction time measurement

```
…
  <script type = "text/javascript">
  function response (e){
  …
  var stoptime= (new Date()).getTime();
  var latency= stoptime-starttime;
  document.forms[ "form"] .elements[ "delay"] .value=latency

  document.forms[ "form"] .submit();
  …
  }
  var starttime= (new Date()).getTime();
  </script>
  …
  <body onkeydown = "return response (event)">
  …
  <form name="form" METHOD="POST" … >
  <input type="hidden" name="delay" value="0">
  …
```

References

Albrecht, T., & Vorberg, D. (2010). Long-lasting effects of briefly flashed words and pseudowords in ultrarapid serial visual presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1339–1345. doi:10.1037/a0019999

Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology, 55,* 803–832. doi:10.1146/annurev.psych.55.090902.141601

Brand, A., & Bradley, M. T. (2012). Assessing the effects of technical variance on the statistical outcomes of web experiments measuring response times. *Social Science Computer Review, 30,* 350–357. doi:10.1177/0894439311415604

Chuah, S. C., Drasgow, F., & Roberts, B. W. (2006). Personality assessment: Does the medium matter? No. *Journal of Research in Personality, 40,* 359–376. doi:10.1016/j.jrp.2005.01.006

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Corley, M., & Scheepers, C. (2002). Syntactic priming in English sentence production: Categorical and latency evidence from an Internet-based study. *Psychonomic Bulletin & Review, 9,* 126–131. doi:10.3758/bf03196267

Cronk, B. C., & West, J. L. (2002). Personality research on the Internet: A comparison of Web-based and traditional instruments in take-home and in-class settings. *Behavior Research Methods, Instruments, & Computers, 34,* 177–180.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE, 8*(e57410), 1–18. doi:10.1371/journal.pone.0057410

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

de Leeuw, J. R., & Motz, B. A. (2015). Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods*. doi:10.3758/s13428-015-0567-2

Eichstaedt, J. (2001). An inaccurate-timing filter for reaction time measurement by JAVA applets implementing Internet-based experiments. *Behavior Research Methods, Instruments, & Computers, 33,* 179–186.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191. doi:10.3758/BF03193146

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. doi:10.3758/s13423-012-0296-9

Gordon, B. (1983). Lexical access and lexical decision: Mechanisms of frequency sensitivity. *Journal of Verbal Learning and Verbal Behavior, 22,* 24–44. doi:10.1016/S0022-5371(83)80004-8

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59,* 93–104.

Hecht, H., Oesker, M., Kaiser, A., Civelek, H., & Stecker, T. (1999). A perception experiment with time-critical graphics animation on the World-Wide Web. *Behavior Research Methods, Instruments, & Computers, 31,* 439–445. doi:10.3758/bf03200724

Ihme, J. M., Lemke, F., Lieder, K., Martin, F., Muller, J. C., & Schmidt, S. (2009). Comparison of ability tests administered online and in the laboratory. *Behavior Research Methods, 41,* 1183–1189. doi:10.3758/BRM.41.4.1183

Keller, F., Gunasekharan, S., Mayo, N., & Corley, M. (2009). Timing accuracy of Web experiments: A case study using the WebExp software package. *Behavior Research Methods, 41,* 1–12. doi:10.3758/BRM.41.1.12

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological research online: Report of Board of Scientific Affairs' Advisory Group on the Conduct of Research on the Internet. *American Psychologist, 59,* 105–117. doi:10.1037/0003-066x.59.2.105

Lang, F. R., John, D., Lütdke, O., Schupp, J., & Wagner, G. G. (2011). Short assessment of the Big Five: Robust across survey methods except telephone interviewing. *Behavior Research Methods, 43,* 548–567. doi:10.3758/s13428-011-0066-z

Linnman, C., Carlbring, P., Ahman, A., Andersson, H., & Andersson, G. (2006). The Stroop effect on the internet. *Computers in Human Behavior, 22,* 448–455. doi:10.1016/j.chb.2004.09.010

McGraw, K. O., Tew, M. D., & Williams, J. E. (2000). The integrity of Web-delivered experiments: Can you trust the data? *Psychological Science, 11,* 502–506. doi:10.1111/1467-9280.00296

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

Ollesch, H., Heineken, E., & Schulte, F. P. (2006). Physical or virtual presence of the experimenter: Psychological online-experiments in different settings. *International Journal of Internet Science, 1,* 71–81.

Plant, R. R., Hammond, N., & Whitehouse, T. (2003). How choice of mouse may affect response timing in psychological studies. *Behavior Research Methods, Instruments, & Computers, 35,* 276–284. doi:10.3758/bf03202553

Plant, R. R., & Turner, G. (2009). Millisecond precision psychological research in a world of commodity computers: New hardware, new problems? *Behavior Research Methods, 41,* 598–614. doi:10.3758/BRM.41.3.598

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin, 86,* 446–461. doi:10.1037/0033-2909.86.3.446

Reimers, S., & Stewart, N. (2007). Adobe flash as a medium for online experimentation: A test of reaction time measurement capabilities. *Behavior Research Methods, 39,* 365–370. doi:10.3758/bf03193004

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

Reips, U.-D. (2002). Internet-based psychological experimenting: Five dos and five don't@!s. *Social Science Computer Review, 20,* 241–249. doi:10.1177/08939302020003002

Reips, U.-D., & Birnbaum, M. H. (2011). Behavioral research and data collection via the internet. In R. W. Proctor & K.-P. L. Vu (Eds.), *The handbook of human factors in Web design* (2nd ed., pp. 563–585). Mahwah, NJ: Erlbaum.

Reips, U.-D., Buchanan, T., Krantz, J. H., & McGrawn, K. (in press). Methodological challenges in the use of the Internet for scientific research: Ten solutions and recommendations. *Studia Psychologica*. http://www.uni-konstanz.de/iscience/reips/pubs/papers/StudiaPsy_final.pdf

Reips, U.-D., & Krantz, J. H. (2010). Conducting true experiments on the Web. In S. D. Gosling & J. A. Johnson (Eds.), *Advanced methods for conducting online behavioral research* (pp. 193–216). Washington, DC: American Psychological Association.

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior, 9,* 487–494. doi:10.1016/S0022-5371(70)80091-3

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime: User's guide*. Pittsburgh, PA: Psychology Software Tools Inc.

Simcox, T., & Fiez, J. (2014). Collecting response times using Amazon Mechanical Turk and Adobe Flash. *Behavior Research Methods, 46,* 95–111. doi:10.3758/s13428-013-0345-y

Skitka, L. J., & Sargis, E. G. (2006). The Internet as psychological laboratory. *Annual Review of Psychology, 57,* 529–555. doi:10.1146/annurev.psych.57.102904.190048