Behav Res (2017) 49:1241–1260
DOI 10.3758/s13428-016-0783-4

Online psychophysics: reaction time effects in cognitive experiments

Kilian Semmelmann[1] · **Sarah Weigelt**[1]

Published online: 5 August 2016

Abstract Using the Internet to acquire behavioral data is currently on the rise. However, very basic questions regarding the feasibility of online psychophysics are still open. Here, we aimed to replicate five well-known paradigms in experimental psychology (Stroop, Flanker, visual search, masked priming, attentional blink) in three settings (classical "lab", "web-in-lab", "web") to account for possible changes in technology and environment. Lab and web-in-lab data were both acquired in an in-lab setting with lab using "Gold Standard" methods, while web-in-lab used web technology. This allowed for a direct comparison of potential differences in acquisition software. To account for additional environmental differences, the web technology experiments were published online to participate from home (setting web), thereby keeping the software and experimental design identical and only changing the environmental setting. Our main results are: First, we found an expected fixed additive timing offset when using web technology (M = 37 ms, SD = 8.14) and recording online (M = 87 ms, SD = 16.04) in comparison to lab data. Second, all task-specific effects were reproduced except for the priming paradigm, which couldn't be replicated in any setting. Third, there were no differences in error rates, which are independent of the timing offset. This finding further supports the assumption of data equality over all settings. Fourth, we found that browser type might be influencing absolute reaction times. Together, these results contribute to the slowly but steadily growing literature that online psychophysics is a suitable complement – or even substitute – to lab data acquisition.

Keywords Psychophysics · Web technology · Online study · Replication · Cognitive psychology · Reaction time

The Internet has a huge impact on every aspect of today's life. In experimental psychology, the introduction and establishment of online surveys has allowed researchers to reach more people from a broader background more easily, efficiently, and faster than ever before (Reips, 2000). Thanks to the higher heterogeneity of the participants (age, gender, origin, social status; see Birnbaum, 2004) online studies allow us to tackle one of the biggest open questions in psychology: To what extent can the results of a single study be mapped to the whole population? Furthermore, through the worldwide availability of online experiments, scientists have easier access to hard-to-reach populations, benefit from a double-blind situation, and are able to conduct parallel experiments independent of equipment or available experimenters (Gosling, Vazire, Srivastava, & John, 2000; Reips, 2000). For example Joinson (2001) found through the results of three studies that participants in an anonymous, computer-mediated research environment disclose significantly more information than in classical research settings. A work by Reimers (2007) collected data from over 255,000 participants in a study about sex differences in Britain – one of the largest up to date. Collecting such large samples will help scientists to battle the issue of insufficient power, which often renders studies irreplicable. A much more selective, but still numerous, sample was collected by Cohen, Collins, Darkes, and Gwartney (2007), who used online message boards to get nearly 2,000 responses from a very specific population of non-medical anabolic steroid users in the USA. These and more advantages have been covered in excessive

✉ Kilian Semmelmann
kilian.semmelmann@rub.de

[1] Developmental Neuropsychology, Department of Psychology, Ruhr-University Bochum, Universitätsstr. 150, 44801 Bochum, Germany

exploration and validation studies (Birnbaum, 2000; Gosling et al., 2000; Skitka & Sargis, 2006; amongst others). However, experimental psychological research is more than survey data.

Especially in cognitive psychology, the study of mental processes in human beings, one of the most common approaches is to measure reactions to stimuli that are usually presented on a computer screen, thereby gaining an external measurement of internal processes. The two main variables of such psychophysical measurements are error rates (ERs) and reaction times (RTs). In a typical cognitive experiment, a participant is presented with a display of different conditions and reacts to this display via a keypress. Usually, this is done by inviting participants into the lab, seating them in a testing chamber, and instructing them to press a mouse or keyboard button when appropriate. Through this approach researchers can control for external factors like hardware (computer system, keyboard, and mouse), software (operating system, programming language, versions), and the environment (e.g., noise, lightning, distractions). On the other hand, conducting cognitive experiments in a classical in-lab setting requires a high amount of resources, is time-consuming, and often relies on recruiting participants from the local student population, thereby potentially limiting the explanatory power of any findings.

So it is not surprising that the use of web technology and the Internet is becoming more and more prevalent in experimental psychology in the same way social sciences and questionnaire-based psychology used it to transform their way of acquiring data. Recently, researchers additionally began to pick up these new technologies as potential psychophysical research methods for RT experiments. Investigations that rely on high temporal accuracy not only need reliable software as a basis, but are also often influenced by the surrounding in which data conduction takes place (e.g., Ramsey & Morrissey, 1978). There are three main ways to test the potential of web technologies against classical methods currently used in experimental psychological research. Hardware-focused approaches try to identify differences in the physical systems used for recording, software-based investigations do the same for the experimental programs, while research about conduction environment focuses onto differences in data acquisition surroundings. Within these three categories, researchers differentiate between the research approaches (see de Leeuw & Motz, 2015): Most investigations of hardware and software differences are realized through a purely technical setup, such as measuring presentation times of stimuli on the screen through a photo diode. This isolated and well-controlled setup is succeeded by empirical research that employs studies with human participants and tries to identify differences in the data produced. These studies can either focus on the question whether effects produced in the lab can be replicated online, or the question which differences occur in replication attempts, and whether scientists are able to negate those potential deviations. Thus, employing empirical research builds upon the technical investigations and complements it by going one step further into the phase of data acquisition.

Potential differences in hardware

Changing from an in-lab to an online environment obviously changes the computer system that is used to record data. These changes influence hardware (computer system, monitor, and periphery) and software (operating system, data acquisition software). With regard to hardware, studies found that monitors (Elze & Tanner, 2012), type of keyboards and mice (Plant & Turner, 2009), or even keyboards of the same type (Neath, Earle, Hallett, & Surprenant, 2011) influence presentation timing measurements.

However, Reimers and Stewart (2015) argue that the differences in hardware may not influence the results from a psychological study to a large extent. The authors tested 15 different computer systems on presentation timing accuracy and RT accuracy through the use of a diode on the screen and an automatized reaction system. While in both cases the use of web technology overestimated the timing and took longer than intended, they found that the variability over systems could easily be corrected by recording about seven more participants. Furthermore, Reimers and Stewart point out that by using a within-subject design, the influence of different systems becomes irrelevant. In summary, potential hardware influences have either small or no effects on the results. Still, if there are influences, this does not solely apply to online studies, but to the same extent to findings within or between classical lab settings.

Potential differences in software

While hardware is a factor that is not under the control of the experimenter (people will not buy a new computer to participate in a psychological study), software is to a certain extent. Therefore deciding which method should be used for stimuli presentation and recording of responses needs to be done very carefully.

Over a decade ago, Schmidt (2001) investigated this matter by using five of the most common ways to present stimuli through web technology (GIF images, Java, JavaScript, Flash, and Authorware) and measuring their presentation times. He recorded the actual number of raster scans through a photo detector (which equates to the real-world presentation time these methods were able to achieve) under different conditions (new/old system, MacOS/Windows, Internet Explorer/Netscape). His results showed that only Java and JavaScript

were "accurate" to some extent. Still, JavaScript's performance deteriorated with system speed like the three other methods (GIF, Flash, and Authorware). Thus, he argued that accurate presentation time is not to be achieved with the current methods.

But with advancements in browser and hardware technology, this study has been revised several times (Barnhoorn, Haasnoot, Bocanegra, & van Steenbergen, 2014; de Leeuw & Motz, 2015; Reimers & Stewart, 2015; Schubert, Murteira, Collins, & Lopes, 2013; amongst others). A very recent example is the study by Reimers and Stewart (2015) mentioned above. Over a decade later than Schmidt they found an offset of presentation time of about 20 ms, when averaged over two presentation methods (Flash, JavaScript), three presentation durations (50, 150, 500 ms), and four different computer systems. Interestingly, the most accurate conditions were achieving accuracies below 1 ms. When using automated response timing to investigate RT time accuracy, they found an offset of about 60 ms averaged over conditions.

This finding is in line with earlier work from Barnhoorn et al. (2014), who tested their JavaScript-based experimental suite QRTEngine in a diode experiment. They showed that in 97 % of cases the stimuli are presented within ±1 frame (16.67 ms) of intended stimuli duration. They report system-logged timing to be below 10 ms, even in high-load conditions.

For Flash-based experiments, Schubert et al. (2013) developed a toolbox that they tested through six experiments. The authors compared their software to other software packages (DMDX, E-Prime, Inquisit, and SuperLab) with regard to presentation and reaction timing accuracy. When using an emulated keyboard, the difference in RT measurements was about 65 ms, while by using a real keyboard, the difference was about 21 ms. In one case, their software produced more accurate results than a classical research software by about 5 ms. In presentation accuracy, they found an increase in timing of 24 ms for the web technology-based experiment.

Another piece of evidence was collected very recently by de Leeuw and Motz (2015), who developed a method to measure JavaScript and Matlab Psychtoolbox trials in an interleaved way. The alternation allowed them to keep other factors identical that might influence RT. Their results showed no difference in variability or measurement sensitivity between the web technology and classical recording software for human reaction timing accuracy. What they did find was a 25-ms offset for the web technology software, which is in line with other studies.

In summary, in contrast to Schmidt, recent studies argue for a fixed offset when using web technology-based experiments. As this offset seems to vary little within the method used and can be accounted for (e.g., within-subject design), web technology allows measuring RTs accurately, as long as absolute RTs are not the focus of interest.

Potential environmental differences

Next to in-lab investigations that concentrate on hard- and software differences between web technology and established psychophysical software, another way to approach the question is to create an experiment through web technology, make it available online and then compare the results with existing literature. This approach mainly targets the bigger construct we denote as "environmental difference" when comparing online experimentation to classical in-lab research environments. By changing from a well-controlled in-lab environment with low degrees of freedom to participating from home, the environmental influence changes. This does not only cover surroundings (e.g., noise, lights, and distractions), but also hardware (different computer systems from high-end multimedia machines to sub-par laptops) and software (operating systems, browsers, additionally installed software).

One of the most prominent examples in this area has been published by Crump et al. (2013). The authors tested eight very well-established experiments through Amazon Turk (AMT) as an experimental platform and compared these results to existing literature of in-lab research. They found all experiments well replicated, except those that required very short and precise millisecond presentation times, namely the attentional blink and masked priming paradigms. Crump et al. inferred from their results that testing with online systems might be critical when using presentation times below 50 ms.

This concern was contradicted by Barnhoorn et al. (2014), who recorded the data of about 50 participants in three cognitive experiments each. They used their self-written QRTEngine that used a different timing method with participants from AMT to perform in their Stroop, attentional blink, and masked priming experiments. Each experimental effect was replicated as expected, even in in the masked priming task, which used presentation times of 16 ms. Notably, these are the exact paradigms that were not be replicated by Crump et al. (2013).

More evidence in this area has been provided by Germine et al. (2012), who measured over 30,000 participants in a variety of cognitive tests (e.g., Cambridge Face Memory Task, Reading the Mind in the Eyes). The authors focused on tests that covered complex visual stimuli, limited presentation times, and required sustained participant attention. Based on their results regarding mean performance, variance, and internal reliability, they found their data did not differ systematically from previous studies that were conducted in a classical lab setting. Germine et al. interpret these results as showing that data quality in online conducted samples is not of a lower quality – which still is a common preconception against online studies (Gosling et al., 2000).

Combining those findings, the way to use web technologies and as a result the Internet as a research method starts to get paved. While initially it may have seemed that web

technology is slower and less accurate, recent studies argue that main effects can be replicated, especially when concentrating on presentation times over 50 ms and recording slightly more data to account for differences in hardware. Additionally, when using online experimentation, researchers cannot only plan to conduct studies on a much higher scale than up to now, it also allows them to ask new questions in a new extent. Replication, confirmation, and generalization are just few of the huge advantages we see when planning to do further methodological exploration with web technology.

Research rationale

With these prospects in mind, we developed a design that combines the ways of analyzing web technologies as a potential research method over technology, replication, and environment. Our study covers five well-known experiments in psychology (Stroop, Flanker, visual search, attentional blink, and masked priming) and tests each experiment in three settings. The first setting is a very classical "Gold Standard" in-lab environment with well-established technology as a basis ("lab" setting). The second setting covers the same experiments, but uses web technology for presenting stimuli and collecting the data (setting "web-in-lab"). As we keep all other factors (hardware, operating system, population, external influences, seating, lab setting) the same between the first two settings, this allows for a direct comparison of performance of experimental software between well-established and web technologies.

The third setting publishes the web technology-based experiment to an online environment for testing students at home (setting "web"). Compared to the web-in-lab data, we kept the technology completely identical, recruited from the same population, and therefore only changed the surrounding and hardware of participants. This setting is intended to reveal potential environmental differences (including physical environment, soft- and hardware) between a classical in-lab and an online setting. Throughout all settings, we use within-subject designs to produce more robust task-specific effects, as proposed by Reimers and Stewart (2015).

As we are not able to screen the different hardware setups used in the web setting, we introduced an additional short performance test that runs before the start of the actual experiments. This performance task is an automatic measure of internal timing accuracy of the software, similar to Barnhoorn et al. (2014). On this basis we were able to identify and exclude underperforming systems. As an additional measure to compare the three settings, we implemented a short RT task. This was used as a very basic measure of RT accuracy that could identify malfunctioning hardware or high system load up-front. All participants were recruited from the same population and randomly assigned to one of the three settings and asked to do all subtasks to exclude potential influences of age, gender or origin.

In line with de Leeuw and Motz (2015) and others, we focused on Javascript for the following two reasons: First and foremost, every browser nowadays includes JavaScript as a part of its engine. Every website that incorporates JavaScript (and that is nearly all of them) can therefore be used to conduct experimental data without the need to download and install any plugins or applets. We assume that this will be a huge advantage, especially once online experiments are moved from the preliminary lab-concentrated testing to the "real world," where participants always look for a most direct and comfortable approach to contribute and tend to avoid any extra actions. Second, due to the fact that it is the most commonly used web-language, it is also in the main focus of being improved by the respective browser producers. This will allow later research approaches to always be at the newest technological advancements in time.

A similar methodological approach can be found in three prior studies: Reimers and Stewart (2007) for example compared a native C program with the web-language Flash in an offline setting and found a fixed offset of about 20 ms in RT (Flash being slower than the C program) but no varying standard deviation (SD). Taking the Flash experiment out of the lab into an online environment added another 10-ms delay, while keeping the SDs comparable. Schubert et al. (2013) first investigated the presentation accuracy of their toolbox ScriptingRT, before they followed up with a comparison of human RTs in an in-lab and online environment. While they were able to replicate the main effects of the Stroop task, they found a small lag and increased variance in their web technology implementation. While these studies were considering Flash, very recently a study by Hilbig (2015) incorporated these principles by using JavaScript in a methodological investigation of a lexical categorization task. He also used three settings to investigate potential differences in conducting data through a lab (E-Prime), a web-in-lab, and online environment. He found the lab implementation being 27 ms faster than the web-in-lab, which in turn was 83 ms faster than the online setting. Still, the word frequency effect was replicated in each setting. His interpretation of these results was that web- or online-based data was not inferior to classical in-lab data.

Hypotheses

We had three overarching hypotheses in this study that are investigated in each experiment separately.

First, based on previous research (Barnhoorn et al., 2014; de Leeuw & Motz, 2015; Reimers & Stewart, 2015; Schubert et al., 2013) we expected an additive offset of RT measurements between 20 and 65 ms over settings. This offset might

vary with number of presented items (that add up singular offsets) and depend on the computational complexity of the experiment, as well the hardware used by the participants.

Second, as studies almost exclusively showed replication of the tasks through web technology up to now, we expected to replicate the main effects of each of our experiments. While the expectation of an offset in timing would be an argument for lower accuracy and data quality through the use of online experimentation, the replication would confine this assumption. The potential flaw would be restricted to specific designs, where absolute timing accuracy is required, timing measurements are below a certain threshold, and/or a between-subject design is used.

Third, if data quality is comparable over the settings and only a RT accuracy offset is expected, that would lead to the assumption that other not time-based measurements need to be the same. By analyzing ERs in each of our subtasks we aimed at revealing whether the environmental part of the different settings does have any influence on data accuracy.

Taken together, our study was based on the idea of (a) replicating classical paradigms online (e.g., Crump et al., 2013), (b) using a three-step approach directly comparing data acquisition settings (e.g., Hilbig, 2015), and (c) studying a wide variety of experiments that not only cover a singular data point, but contain an increasing need for sensitivity in data recording. This procedure allowed us to establish a solid basis of experiments that were not only replicated online, but differed with regard to the changes of technology and environment, thereby extending previous studies. To analyze the results in more detail, we added two additional measurements – a simple RT task and a performance task that did not rely on user input – to find indications whether potential differences already emerged at a very basic level of presentation or are due to inaccuracies in recording responses. Additionally, we performed Bayesian factor analysis on all relevant effects and interactions and examined the question whether the type of browser that has been used to participate influences the results. Furthermore, to investigate the preconception that online data is of lower quality than in-lab data, we complemented our RT analyses with ER analyses, which are not reliant on the timing measurement accuracy of software. If we were to find no difference over settings in ERs, this would argues that web participants are answering as accurately as in-lab participants and the differences are due to the change of soft- and hardware. In sum, we used established approaches to produce a broad range of data that consolidated the possibility of psychophysical online experimentation on the one hand, but also pushed the scientific discussion towards more detailed analysis about the potential origin of offsets, their influence on results, and how to treat them on the other.

General methods

Participants

In total, 147 participants were recruited at the Ruhr-University Bochum and participated for course credit. Each participant was randomly assigned to one of the three settings, yielding 50 participants in setting “lab”, 49 in “web-in-lab”, and 48 in “web”. In-lab participants (lab and web-in-lab) were seated approximately 60 cm in front of an 18-in. CRT monitor, set to 85 Hz refresh rate and 1,280 × 1,024 px resolution. After a short written overview over the tasks, detailed instructions were shown on-screen before each experiment. Online participants were asked to come to the lab and were instructed in person to mediate a basic level of control on the experimental process before receiving the URL and participation code. These participants received the same detailed on-screen instruction for each task.

Three participants were removed because of technical issues (one in web-in-lab, two in web). Additionally, one participant was excluded from the web data because he used a tablet device. On top of that, two participants were removed from the web data due to high timing offset identified through the performance task. Thus, all the following analyses were performed on 50 participants in the lab setting, 48 in web-in-lab, and 43 in web. During the analysis of each experiment, additional exclusion criteria applied but were limited to the sub-analysis and not removed from other analysis by default. For details, see the Results section of each experiment.

Procedure

The study was split into seven experiments (see Fig. 1). First, we conducted an automatic performance measurement of timing accuracy of the participant’s system configuration that did not require any input. Second, the participant performed a short and simple RT task yielding a baseline measurement of RT and performance of the system. Experiments three to seven were our five main tasks, presented in random order. In total, we set each of the five main experiments to a length between 6 and 10 min resulting in an overall experimental duration of about 1 h. Limiting the overall experimental duration is important with regard to exporting the whole experiment online and having volunteers participate from all over the world at a later stage.

Materials

The lab setting was programmed in Matlab R2009b with Psychtoolbox (Brainard, 1997) installed. Having multiple possibilities to present a stimulus through HTML/Javascript, we tried to emulate the method Psychtoolbox uses for the web technology settings as closely as possible. Therefore we

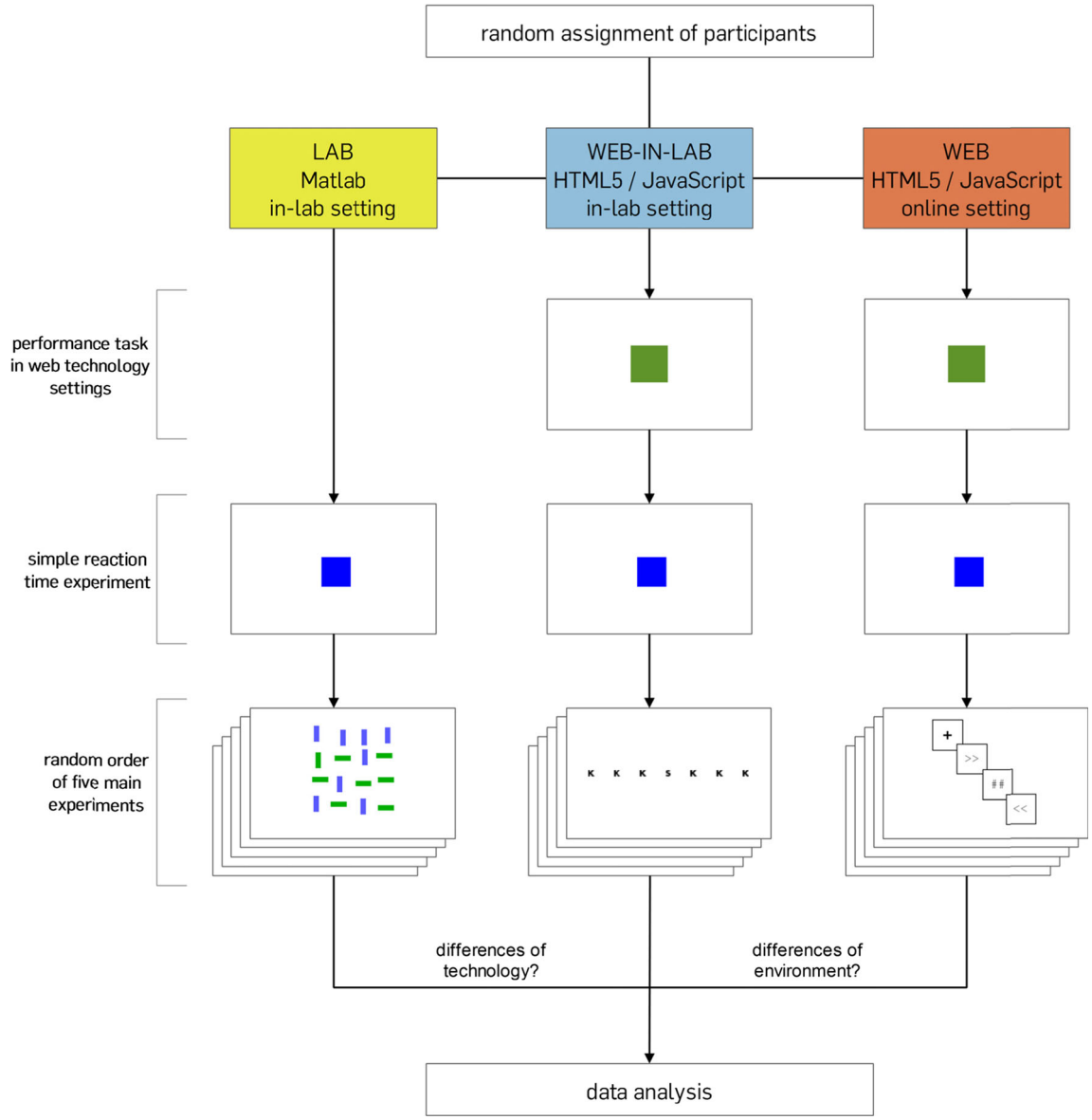

Fig. 1 Experimental design. The *graph* shows the complete experimental design for our study. Each participant was randomly appointed to one setting: lab, web-in-lab, or web. Based on this attribution, the participant performed the subtasks either in-lab (lab, web-in-lab) or at home to his own scheduling (web). Comparing lab and web-in-lab data, we are able to investigate potential differences in technology, as all other factors are kept identical. Comparing web-in-lab data to web data will show potential differences that occur due to the change of environment. Please note that the colors *yellow* (lab), *blue* (web-in-lab), and *orange* (web) will identify the settings in all following graphs

utilized the layer system of modern browsers through modifying the z-index of a stimulus and an overlaying mask to show and hide the stimuli. This allowed us to pre-render the images and include them into the Document Object Model (DOM) before simply "flipping" them to focus. With regard to optimizing the timing, we relied on the *setTimeout()* method, that is supposed to be accurate between 2 and 10 ms according to its official W3C specification.

To avoid any loss of timing accuracy from third party plugins, most importantly jQuery, we tried to use pure Javascript in any timing-related operation. Additionally, we preloaded all stimuli and the experimental design to avoid any latency through providing the participants with the necessary stimuli and counteract possible influences of a slow internet connection. Thus, the web technology experiments used a combination of HTML, Javascript, and php, supported by the plugin jQuery Version 2.0.3. Web-in-lab participants used the Browser Chrome Version 46.0.2490.86, with XAMPP 1.8.5 running as a local webserver for file handling. Online the experiment ran on a server with Apache 2.6.18 and php 5.3.3 installed.

All tasks were performed on a gray (RGB: 179, 179, 179) background that was visible during all parts of each experiment. Colors used in these experiments were black (RGB: 0, 0, 0), white (RGB: 255, 255, 255), gray, blue (RGB: 0, 0, 255), purple (RGB: 95, 95, 191), green (RGB: 255, 0, 0), olive (RGB: 95, 148, 43), carmine (RGB: 150, 31, 56), fractal (RGB: 147, 147, 147), red (RGB: 0, 255, 0), slate blue (RGB: 89, 89, 255), and slate green (RGB: 0, 175, 0). If not indicated otherwise, all text was displayed in black "Courier New" font due to its monospacing and serif characteristics. Font size did vary and will be noted individually. All instruction was presented on-screen before the respective task. The fixation cross consisted of two black 20 × 4 px bars and was always presented centrally on the display.

Experiment 1: performance task

The performance task was a short automatic measure of stimulus presentation timing accuracy in the settings that used web technology (web-in-lab and web) as experimentation software and was always conducted first. As we were obviously not able to use a diode-based display measurement that would represent a real-life display of the stimulus, we relied on the computer-logged difference between planned and real stimulus presentation time. Thereby, we were able to obtain a baseline measurement of software latency, which is the first factor of potential inaccuracy in stimulus presentation. As web-in-lab participants were using our well-configured in-lab computer system, we expected a higher timing accuracy for web-in-lab than web, based on the influences of the individual systems with regard to hard- and software that is not optimized for psychophysical experimentation. For example, participants could have used insufficient hardware or unmaintained systems that become slower with time. Naturally, we were not able to control which other programs were running (in the background) during the experiment. Thus, there might have been additional influences of parallel software usage on the accuracy of the system as the more tasks a computer has to process, the slower it becomes. Overall, this experiment was mainly targeted at identifying greatly sub-performing systems through the simple measurement of internal timing accuracy of displaying a stimulus.

Methods

Before the start of recording data, the participant was instructed to close all other browser windows and wait for the performance task to be completed without opening other programs. In each of the 100 trials, a 250 × 250 px image was randomly picked from a set of nine colors (black, blue, green, red, purple, olive, carmine, fractal, white) and shown in the center of a gray background for a random time between 0 and 1,000 ms. After the presentation time of one stimulus ended, the next followed immediately. No input was necessary from the participant. Only the intended and real stimulus presentation timing in ms was recorded.

Results

The statistical analysis was performed on the offset that was calculated by subtracting the intended from the real stimulus presentation time. Two outliers ($SD = 1.5$) were removed from the web data and subsequently from any further analysis of the whole data set. This yielded 48 participants in the web-in-lab and 43 in the web setting for all following analysis.

Levene's test indicated unequal variances between the settings, $F = 57.73$, $p < .001$, so a Welch's unequal variances t-test was performed. It indicated a significant higher timing offset for web ($M = 6.74$, $SD = 4.39$) than web-in-lab ($M = 0.78$, $SD = 0.09$), $t(42.03) = 8.90$, $p < .001$, $d = 1.98$. From these results we can conclude that the web-in-lab setting did indeed exhibit a higher timing accuracy in the performance task than the web data by about 6 ms.

Experiment 2: reaction time task

The RT task was always conducted after the initial performance task and preceded the other five experiments. The intention of this task was to detect potential differences in RTs between the settings (lab, web-in-lab, web) at a very crude level and without high cognitive load. This allows us to differentiate between the three settings based on a task that solely relies on perception and reaction compared to the necessary information processing in the five main experiments. Based on the literature (de Leeuw & Motz, 2015; Hilbig, 2015; Reimers & Stewart, 2014) we expected longer RTs for the experiments based on web technology (web-in-lab and web) in comparison to the standard technology (lab). An additional increase in RT was expected (Hilbig, 2015) for the online conducted data (web).

Methods

Before the start of the experiment, the participants were instructed to react as quickly as possible to the appearance of a blue square through pressing the space bar. Each trial started with a fixation cross between 500 and 1,000 ms. The fixation was followed by a blue 100 × 100 px square. Each of these targets had a random position offset (up to 50 px in either direction). The target was shown until response (spacebar), followed by an inter-trial-interval (ITI) of 250 ms. In total, 50 trials were conducted per participant. RT time data were recorded in ms.

Results

Before analysis, extreme outlier trials with a RT of less than 200 ms and more than 2,000 ms were removed. All RT analyses were performed on the median RT per participant, due to the high skewness of RT data (e.g., Whelan, 2008). One outlier was removed from the web data due to long RTs (*SD* = 1.5). Analysis of variance showed a main effect for setting, $F(2, 137) = 36.87$ $p < .001$. Bonferroni-corrected post-hoc t-tests indicated significantly faster RTs for lab ($M = 253$ ms, $SD = 24.52$) than web-in-lab ($M = 280$ ms, $SD = 33.30$), $p < .001$, $d = -0.93$, and web ($M = 318$ ms, $SD = 49.17$), $p < .001$, $d = -1.73$, and significantly faster RTs for web-in-lab than web, $p < .001$, $d = -0.92$. In summary, as expected, our results show 27-ms increased RT data for web-in-lab data compared to lab data. Web data were an additional 38 ms (65 ms total) slower than web-in-lab (for details, see Fig. 2).

Experiment 3: digit stroop task

The Stroop effect shows an interference between different features of a stimulus, thereby slowing down RTs (Stroop, 1935). Most prominently, participants are instructed to state the semantic meaning of words of colors that are printed in different hues. If the color of the word contradicts its meaning (e.g. "yellow" printed in the color blue), the response takes longer and might lead to an error more easily. This is called the "incongruent" condition and shows the interference effect. A congruent condition would be defined as both (color and semantics) exhibiting the same feature (e.g. "blue" printed in the color blue). Between those conditions, a neutral condition (e.g. "blue" printed in a neutral color like black) might be used as a baseline measurement.

We used this well-established task as it is one of the most robust effects in cognitive psychology and is also used in clinical context (MacLeod, 1991). It allowed us to investigate whether we are able to replicate the task-specific effects of a very basic paradigm on all three settings. To account for the idea that online experimentation will be available to people from multiple regions and the accompanying language effects, we used a number-based Stroop task (Windes, 1968) instead of the more classical version that uses words and colors.

Statistically, a Stroop effect is shown in significantly longer RTs and higher ERs for the incongruent than the congruent condition. Usually, neutral feature combinations are neither facilitated (like congruent) nor impeded (like incongruent), and therefore exhibit RTs and ERs between the other two conditions. Next to this task-specific effect, we expect higher RTs for web technology experiments and comparable ERs over settings, as explained in the Hypotheses section of the Introduction.

Methods

Before the experiment started, participants were instructed to identify the number of items (1, 2, or 3) shown on the screen by pressing the corresponding number on their keyboard as accurately and quickly as possible. This was supported by an example and the announcement of a short training. The training block consisted of 18 trials with feedback ("correct"/"wrong") for 100 ms and preceded a total of four test blocks with 36 trials each. Each trial started with a black fixation cross that was shown for 1,000 ms. The target stimulus followed and was shown until response. A target could consist of neutral (e.g., "XX"), congruent (e.g., "2 2"), or incongruent (e.g., " 3") black characters in 40 px font on a gray background. The number and position of the single characters were randomly selected, while keeping an equal amount of trials

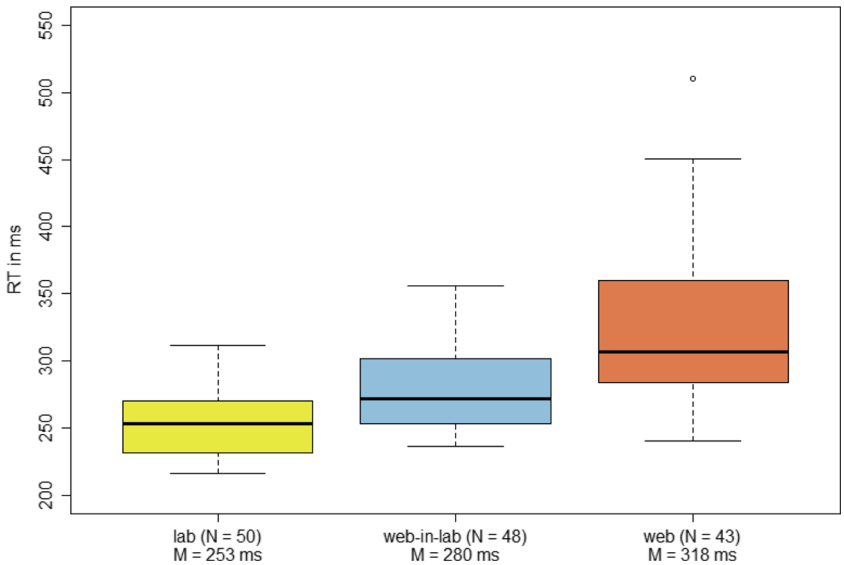

Fig. 2 Results of the reaction time (RT) task. Each *boxplot* represents one setting (lab, web-in-lab, web). RT-based outliers are indicated by a *circle*

per condition. The stimulus was shown until response. Upon response, an ITI followed for a random duration between 250 and 750 ms.

Results

After the removal of extreme outlier trials, four outliers were removed from the analysis due to high ERs ($SD = 3$), namely one in lab, one in web-in-lab, and two in web. Three outliers were removed from the analysis due to high median RTs ($SD = 1.5$), one in lab and two in web-in-lab. Thus, further analysis in this task was on 48 participants in the lab, 45 in web-in-lab, and 41 in web setting.

With regard to RT, Mauchly's test showed a violation of sphericity, $\chi^2(2) = 41.72$, $p < .001$, for the within factor condition (neutral, congruent, incongruent), thus a Greenhouse-Geisser correction ($\varepsilon = 0.79$) was performed. A mixed-design analysis of variance with the between-factor setting (lab, web-in-lab, web) and the within-factor condition revealed main effects for setting, $F(2, 131) = 33.11$, $p < .001$, and condition, $F(1.46, 191.26) = 316.38$, $p < .001$. No significant setting × condition interaction was found, $F(2.92, 191.26) = 0.82$, $p = .52$.

Post-hoc Bonferroni corrected t-tests indicated a significant difference between the conditions congruent and neutral, $p < .001$, $d = -0.54$, congruent and incongruent, $p < .001$, $d = -0.93$, and neutral and incongruent, $p < .001$, $d = -0.43$, and significant differences between the settings lab and web-in-lab, $p = .0036$, $d = -0.87$, lab and web, $p < .001$, $d = -1.67$, and between web-in-lab and web, $p < .001$, $d = -0.89$. Averaged over conditions, web-in-lab RT data ($M = 552$ ms, $SD = 63.37$) was 46 ms slower than lab data ($M = 506$ ms, $SD = 59.37$) and web ($M = 606$ ms, $SD = 76.57$) was 54 ms higher than web-in-lab data (for details see Fig. 3).

Investigating ER data, Mauchly's test revealed a violation of sphericity, $\chi^2(2) = 147.08$, $p < .001$, for the within-subject factor condition, therefore a Greenhouse-Geisser correction ($\varepsilon = 0.60$) was performed. A mixed ANOVA revealed a significant main effect for condition, $F(1.2, 157.20) = 146.20$, $p < .001$, but not for setting, $F(2, 131) = 1.86$, $p = .16$, and no significant setting × condition interaction, $F(2.4, 157.20) = 0.69$, $p = .60$. Significant differences between all conditions, $p < .001$, were confirmed by a post-hoc Bonferroni corrected t-test. Overall, lab ER was 5.24 % ($SD = 6.40$ %), web-in-lab 4.22 % ($SD = 6.22$ %), and web 3.96 % ($SD = 5.90$ %).

Summing up, we found a significant offset of reaction timing accuracy which is in line with our first overarching hypothesis. Second, the task-specific effects of the Stroop task in both RT and ER were replicated. Third, we did not find differences in ER when comparing between settings. This points to equal concentration/attention of participants, independent of the experimental setup.

Experiment 4: letter flanker task

The Flanker task assesses the ability to separate the irrelevant from the relevant context and to respond appropriately (Eriksen & Eriksen, 1974). Usually, the stimulus consists of a central target letter surrounded by six distractor letters, three on each side. The participant responds to the stimulus by identifying the middle letter and pressing an appropriate response key. The combination of letters could form a "congruent" (same target letter as distractor letters), a "keycongruent" (same response key as distractor letters), or an "incongruent" (different response as distractor letters) condition.

In comparison to the Stroop task, the Flanker task relies on more fine-grained differences and thus is more difficult for the

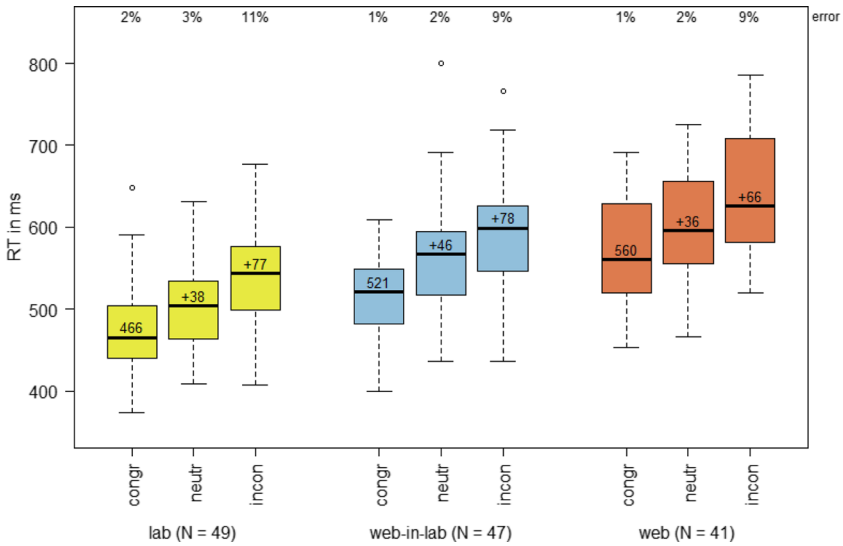

Fig. 3 Digit Stroop task results. Box plotted reaction time (RT) data for the Stroop experiments per setting (lab, web-in-lab, web) and condition (congruent, neutral, incongruent). RT-based outliers are indicated by a circle. Error rates per cell are shown at the top of the graph

participant. This allowed us to investigate if we are able to accurately measure these fine-grained differences in RTs through web technology.

We expected that congruent stimuli elicit the fastest RTs, followed by keycongruent and then incongruent displays (Sanders & Lamers, 2002). This additional RT is due to the need for inhibiting the distractors while assessing the appropriate response to the target letter. Additionally, our overarching hypotheses that there is a RT offset for web technology experiments and no differences in ER rates over settings apply.

Methods

Before the experiment started, the participant was instructed to identify the middle letter of the stimuli and press either the left arrow key (for letters S, C) or the right arrow key (H, K) on the keyboard. At the start of a trial a black fixation cross appeared that lasted between 250 and 750 ms and was followed by the stimulus that was presented until response. Each stimulus consisted of seven black letters in 60 px font centered on a gray background. The distance between each letter was 11 px. These letters could form a congruent (e.g., SSSSSSS), a keycongruent (e.g., HHHKHHH), or an incongruent stimulus (e.g., KKKCKKK). Trials were separated by an ITI of 500 ms. In total, nine blocks with 24 trials each were conducted. Before the actual experiment, a short training of 24 trials with feedback ("correct"/"incorrect") for 100 ms was performed. Each of the conditions was presented equiprobably.

Results

One lab and two web-in-lab participants were excluded from the analysis due to high ERs (SD = 3), after the removal of extreme outlier trials. One additional participant was removed due to high median RTs (SD = 1.5) in the lab setting, five in the web-in-lab and one in the web condition. This yielded a total of 48, 41, and 42 participants for lab, web-in-lab, and web, respectively.

When analyzing the RT data, a violation of sphericity was indicated by Mauchly's test, $\chi^2(2) = 15.88$, $p < .001$, and thus the degrees of freedom were corrected through the Greenhouse-Geisser method ($\varepsilon = 0.90$). Main effects were identified through a mixed-design ANOVA for the between-subject factor setting, $F(2, 128) = 26.28$, $p < .001$, and within-subject factor condition, $F(1.80, 230.40) = 271.82$, $p < .001$. Importantly, the setting × condition interaction was not significant, $F(3.60, 230.40) = 0.94$, $p = .44$.

Pairwise post-hoc Bonferroni corrected t-tests were conducted to follow up on the main effect for setting and indicated a significant difference between the settings lab and web-in-lab, $p = .002$, $d = -0.93$, lab and web, $p < .001$, $d = -1.39$, and web-in-lab and web, $p = .002$, $d = 0.70$. The averaged offsets over conditions showed a 41-ms slower RT for web-in-lab (M = 544 ms, SD = 48.64ms) than lab (M = 503 ms, SD = 54.59 ms), and a 43-ms slower timing for web (M = 587 ms, SD = 75.78 ms) than web-in-lab. As expected for the Flanker task, significant differences were found between the conditions congruent and incongruent, $p < .001$, $d = -0.82$, congruent and keycongruent, $p < .001$, $d = -0.26$, and incongruent and keycongruent, $p < .001$, $d = -0.57$, with congruent (M = 519 ms, SD = 65.23 ms) being faster than keycongruent (M = 536 ms, SD = 64.88 ms), which in turn was faster than incongruent (M = 574 ms, SD = 67.49 ms; for details see Fig. 4).

The ER data also showed a violation of sphericity through Mauchly's test, $\chi^2(2) = 32.30$, $p < .001$, for the within factor condition and therefore was Greenhouse-Geisser corrected ($\varepsilon = 0.82$). A mixed ANOVA revealed a significant main effect for condition, $F(1.64, 209.92) = 99.02$, $p < .001$, but not for setting, $F(2, 128) = 0.49$, $p = .61$. No setting × condition interaction was found, $F(3.28, 209.92) = 0.11$, $p = .98$. A Bonferroni corrected t-test revealed a significant difference between all conditions, $p < .001$. Summed over conditions, ERs in lab data were 4.61 % (SD = 4.24 %), in web-in-lab 5.24 % (SD = 5.19 %), and in web 5.14 % (SD = 4.54 %).

In other words, there was a significant difference in RTs between settings (RT offset), but the main effects of the Flanker task were replicated in each setting, in both RT and ER. Additionally, there were no differences in ERs between settings, thereby confirming all of our three overarching hypotheses.

Experiment 5: visual search task

Visual search refers to the task in which a participant has to identify a target among distractors (Wolfe, 1998). If the features of the target are distinct from those of the distractors, this is called a "popout" search. Popout search-tasks are explained by allowing for parallel search investigation over all available positions, thus having a nearly constant search time over different amounts of stimuli shown. A prominent example would be to search for a vertical bar amongst horizontal bars. In the "conjunction" search on the other hand, target and distractors share features. This requires a serial search for each presented object, therefore increasing overall search time with size of display (i.e., number of items presented). To further the example above, the target could be defined as a green horizontal bar amongst blue horizontal and green vertical bars. As participants need to integrate the information of both features (color and orientation) for each visible item, attention needs to be directed to each target before the decision can be made about whether it is indeed a target or not. The effects of this paradigm hence become very visible when comparing different display sizes. If there is a small display size (e.g., four items), both conditions are usually fast as there are few necessary

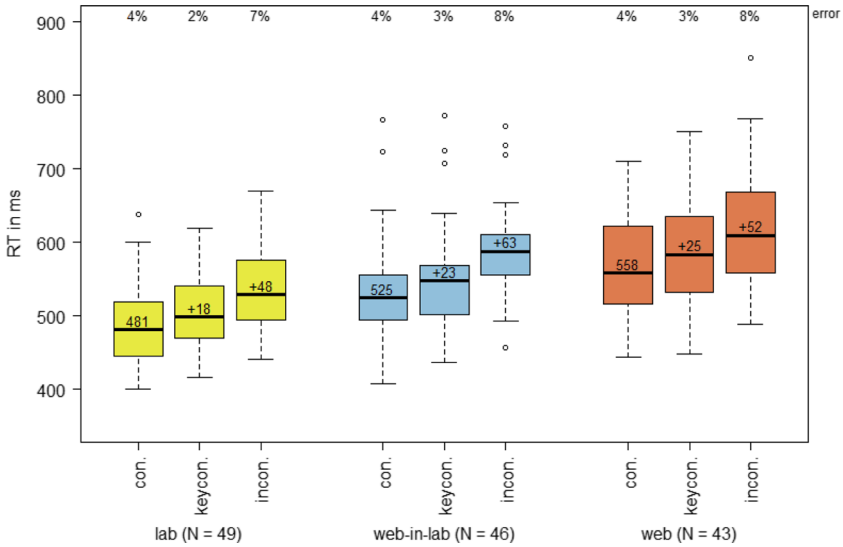

Fig. 4 Flanker task results. Reaction time (RT) data from the Flanker task presented as box plots per setting (lab, web-in-lab, web) and condition (congruent, keycongruent, incongruent). Outliers are marked as circles. Error rates per cell are shown on the top of the graph

operations before giving an answer. With increasing display size (e.g., 16 items) on the other hand, popout search barely gets slowed down due to the high visibility of the necessary feature, whereas in serial search participants take a much longer time to compare their target representation against all items present.

There are multiple factors of this paradigm that make it interesting to test through our three settings. First, it is a very robust and well-established paradigm that has been replicated and modified dozens of times. Second, only a few trials are necessary to find the effects described above enabling a short experimentation time, which is needed for our design. Third, the search effect is defined through the systematic variation of the type of search (popout or conjunction) and display size (number of items) that yield an interaction in RT.

Beyond the usual RT offset for web technology experiments, we expect a higher RT and ER for conjunction trials, especially at a higher display size. This yields an interaction between condition (popout, conjunction) and number of items (four, 16). On the other hand, despite the RT difference over settings, there should be no differences in ER, following our main assumptions that timing accuracies are due to inherent factors of the software.

Methods

Before the start of the experiment, participants were instructed to search for a green vertical bar. They were told to react as fast and accurately as possible by either pressing the left arrow key (target present) or the right arrow key (target absent). Each trial started with a 1,000-ms fixation cross. Then the target display, either from the popout or conjunction condition, was presented. In the popout condition, the target was accompanied by green horizontal bars, while in the conjunction condition it was presented with green horizontal and blue vertical bars. Each item was arranged in a grid-like structure with a small jitter. The display sizes were either four or 16 items and were randomly distributed. Following an answer from the participant, an ITI between 250 and 750 ms was used, before the next trial started. The experiment started with 20 random training trials with feedback before going on to the test phase, in which we recorded 200 trials divided in four blocks per subject. The blocked design was counterbalanced by either being ABBA or BAAB, whereas A was conjunction and B was popout search.

Results

Three participants were excluded beforehand due to technical issues, one in each setting. Extreme outlier removal was performed. Based on ER (SD = 3), three participants were excluded in lab, four in web-in-lab, and one in web. Three outliers (SD = 1.5) due to high median RTs were identified in lab, four in web-in-lab, and one in web. After exclusion, we had 43, 39, and 41 participants for further analysis in lab, web-in-lab, and web, respectively.

A mixed-design ANOVA with RT as dependent variable and setting as between-subject factor and condition and number of items as within-subject factors revealed a significant main effect for setting, $F(2, 120) = 22.51$, $p < .001$, condition, $F(1, 120) = 1212.77$, $p < .001$, and number of items, $F(1, 120) = 892.71$, $p < .001$, as well as significant interactions between setting and condition, $F(2, 120) = 10.85$, $p < .001$, $\eta^2 = 0.018$, and condition and number of items, $F(1, 120) = 581.81$, $p < .001$, $\eta^2 = 0.211$. No significant interaction was found for setting × items, $F(2, 120) = 0.89$, $p = .41$), and setting × items × condition, $F(2, 120) = 2.56$, $p = .82$.

Bonferroni corrected t-tests indicated that the setting × condition interaction is due to no difference between lab conjunction and web-in-lab conjunction data, $t(78.32) = -0.15$, $p = .88$, $d = -0.03$, with lab data exhibiting a higher RT in the conjunction condition than expected, as can be seen in Fig. 5. All other

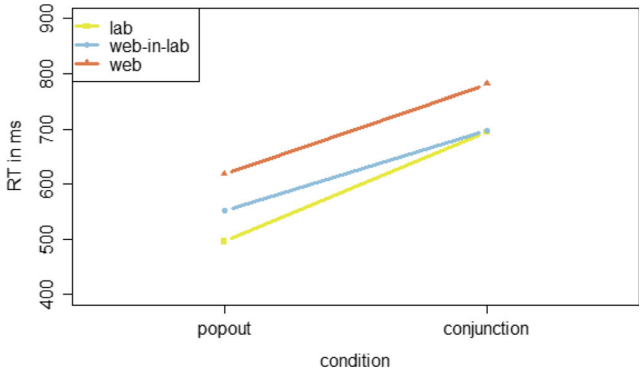

Fig. 5 Search setting × condition interaction. The *graph* shows reaction time (RT) per condition per setting. The expected search effect can easily be seen in each setting by an increase in RT from popout to conjunction search. Still, the unexpected setting × condition interaction can be explained through the lab data, which shows a steeper increase over conditions than the other two settings

combinations of setting and condition differed significantly, $p < .02$, $d > 0.87$.

To investigate the condition × items interaction, we performed an ANOVA for each setting. In lab, there was a significant main effect for condition, $F(1, 42) = 434.60$, $p < .001$, items, $F(1, 42) = 373.30$, $p < .001$, and condition × items interaction, $F(1, 42) = 228.10$, $p < .001$. For the web-in-lab, the same held true for condition, $F(1, 38) = 483.70$, $p < .001$, items, $F(1, 38) = 415.00$, $p < .001$, and condition × items, $F(1, 38) = 219.20$, $p < .001$. Also the web data showed main effects for condition, $F(1, 40) = 369.9$, $p < .001$, items, $F(1, 40) = 215.4$, $p < .001$, and condition × items, $F(1, 40) = 163.30$, $p < .001$.

Overall, popout RT increased by 40 ms when increasing the display size from 4 ($M = 534$ ms, $SD = 74.31$) to 16 ($M = 574$ ms, $SD = 88.28$) items, $d = -0.50$. Conjunction RT on the other hand increased by 214 ms from four ($M = 617$ ms, $SD = 78.22$) to 16 ($M = 831$ ms, $SD = 130.89$) items, $d = -1.98$. Averaged over conditions, lab data ($M = 620$ ms, $SD = 158.65$ ms) is 32 ms faster than web-in-lab data ($M = 652$ ms, $SD = 136.42$ ms), $d = -0.22$, which in turn is 72 ms faster than web data ($M = 724$ ms, $SD = 162.20$ ms), $d = -0.54$, totaling 104 ms between lab and web, $d = -0.69$. For further details, see Fig. 6.

With regard to the ER data, a mixed ANOVA showed a significant effect for condition, $F(1, 120) = 121.16$, $p < .001$, and number of items, $F(1, 120) = 28.19$, $p < .001$), but not for the between-subject factor setting, $F(2, 120) = 0.49$, $p = .61$. It revealed a significant interaction of setting × items, $F(2, 120) = 4.73$, $p = .01$, a significant interaction of condition × number of items, $F(1, 120) = 87.80$, $p < .001$, and a significant three-way interaction of setting × condition × items, $F(2, 120) = 4.69$, $p = .01$. The interaction setting × condition was not significant, $F(2, 120) = 0.42$, $p = .66$.

In short, the expected offset for web technology and online experimentation was indicated, the experimental effect – an interaction between number of items and condition – was found and there was no main effect in ERs over settings. These results support our three overarching hypotheses.

Experiment 6: masked priming task

In priming experiments, participants are given a cue regarding which kind of target might appear. The prime could refer to the position or target type and is usually a very short presentation before the actual target (Meyer & Schvaneveldt, 1971). For example, priming a side of the screen where the target will appear usually leads to faster RTs when the target appears where it was primed (congruent condition), compared to slower RTs in incongruent cases (e.g., target appears on the opposite side the prime indicated). The general priming paradigm can be extended by using a mask before and/or after the prime. The task of the mask is to avoid conscious processing of the prime even further: Not only is it presented very shortly, but additionally it is immediately masked, which is an attempt to prohibit any episodic memory traces (Forster & Davis, 1984).

We included this task in our investigation as it is a very delicate paradigm that relies on very short and exact stimulus presentation times to work appropriately. Many priming paradigms present the prime around 50 ms, which needs to be performed very accurately if one wants to reproduce the main effects. Having this experiment as part of our comparison we were testing the upper limits of presentation and recording accuracy through web technology used in our web-in-lab and web settings. An earlier study by Crump et al. (2013) was not able to replicate the main effects of a masked priming paradigm from an in-lab study (Eimer & Schlaghecken, 2002) through using web technology. Thus, we wanted to investigate further whether we can identify at which point the replication fails: could it be the change of technology or the recording at home?

We expected an advantage of the non-primed side in RT in early prime durations (16, 32 ms) that changes to an advantage for the primed side in late prime durations (80, 96 ms) as reported by Forster and Davis (1984). The latter part was replicated by Crump et al. (2013), whereas the former was not. Additionally, as usual, we expected an RT offset for web technology, but no differences in ER rates over settings.

Methods

Before the experiment started, the participant was instructed to indicate the direction of a stimulus that followed a short mask through pressing the corresponding arrow key. To get used to the fast-paced paradigm, participants took part in a 48 random trial training block with feedback. It was followed by six

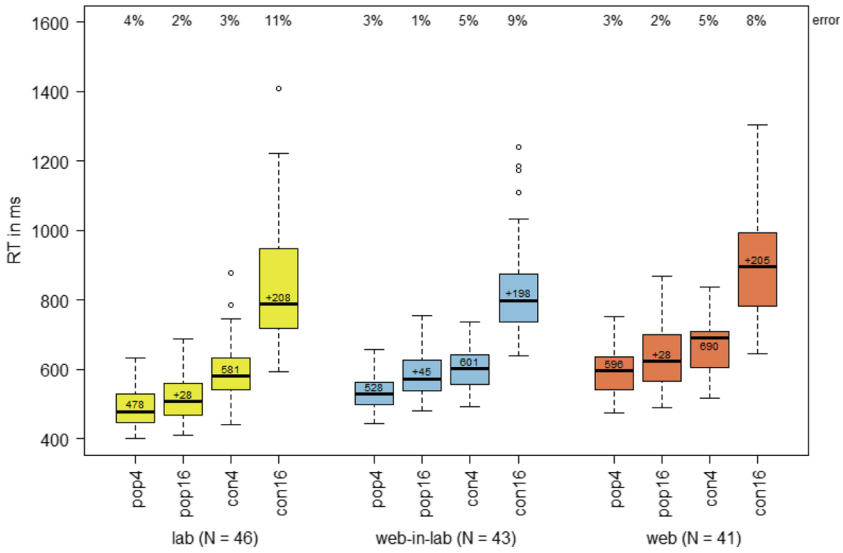

Fig. 6 Visual search results. Reaction time (RT) data from the visual search task per setting (lab, web-in-lab, web) and condition ("pop" for popout, "con" for conjunction with number of items). Outliers are displayed as *circles*, while ERs are shown at the top of the graph

experimental blocks with 48 trials each. Every trial started with a 500-ms fixation cross, followed by a 16-, 32-, 48-, 64-, 80-, or 96-ms prime (e.g., "<<"), a 100-ms mask ("##"), a 50-ms blank, and a 100-ms target (e.g. "<<"). The possible conditions were congruent (e.g., "<<" following a "<<" prime) or incongruent (e.g. "<<" following a ">>" prime). The participants had 3,000 ms to identify the target direction through the arrow keys ("left" or "right") before the trial ended. Each symbol was a black, 40 px font print on a gray background. All combinations of primed side, condition, and prime duration were equiprobable and randomly selected.

Results

After removal of extreme outlier trials, we excluded seven participants in the lab, nine in the web-in-lab, and 21 in the web setting due to high ERs (SD = 3). No additional participants were excluded because of high RTs (SD = 1.5). Another subject had to be removed from the web data because he/she did not complete the experiment. This yielded a total of 43, 39, and 21 data sets for lab, web-in-lab, and web, respectively.

In RT data, a violation of sphericity was indicated by Mauchly's test, $\chi^2(65) = 41.53$, $p < .001$, for the interactions condition × prime duration and setting × condition × duration, and thus the degrees of freedom were corrected through the Greenhouse-Geisser method ($\varepsilon = 0.93$). Main effects were identified through a mixed-design ANOVA for the between-subjects factor setting, $F(2, 100) = 25.96$, $p < .001$, and within-subjects factors condition, $F(1, 115) = 6.99$, $p = .001$, and prime duration, $F(4.65, 465) = 34.68$, $p < .001$. There was a significant interaction between condition and prime duration, $F(5, 500) = 11.28$, $p < .001$, $\eta^2 = 0.007$ and between setting × condition × duration, $F(10, 500) = 2.68$, $p = .003$, $\eta^2 = 0.003$.

Following up on the main effect condition, a post-hoc t-test revealed a significant difference ($p = .007$, $d = -0.13$) between congruent (M = 439 ms, SD = 49.22) and incongruent (M = 446 ms, SD = 55.14) items with incongruent being slower than congruent. Regarding the main effect settings, we found a significant difference between lab (M = 411 ms, SD = 40.49) and web-in-lab (M = 452 ms, SD = 35.17), $p < .001$, $d = -1.10$, and lab and web (M = 487 ms, SD = 53.13), $p < .001$, $d = -1.70$, but not for web-in-lab and web, $p = .007$, $d = -0.83$, Bonferroni corrected. For details, see Fig. 7.

To follow up on the condition × prime duration interaction, we calculated Bonferroni corrected t-tests between primed and unprimed trials for each prime duration at each setting. For the web-in-lab data, the 96-ms prime duration yielded a significant difference between congruent and incongruent trials, $t(38) = -4.54$, $p < .001$, $d = -0.63$, while all other combinations were not significant, $p > .05$.

For ER data, Mauchly's test was not violated. A mixed ANOVA indicated significant effects for condition, $F(1, 100) = 24.40$, $p < .001$, setting, $F(2, 100) = 3.10$, $p < .05$, a significant interaction of setting × duration, $F(10, 500) = 2.49$, $p < .01$, condition × duration, $F(5, 500) = 12.63$, $p < .001$, and setting × condition × duration, $F(10, 500) = 2.07$, $p < .03$. There was no significant main effect for duration, $F(5, 500) = 1.34$, $p = .25$, or the setting × condition interaction, $F(2, 100) = 1.09$, $p = .34$.

Over conditions, participants in the lab condition had an ER of 2.93 % (SD = 2.12 %), in web-in-lab 3.11 % (SD = 2.01 %), and in web 1.89 % (SD = 1.35 %).

In short, there was an RT offset between lab data and the other settings, but no significant differences in ERs between the settings. Regarding the task specific effects, we were not able to reproduce the full priming effect: Only in the 96-ms condition did we see significantly

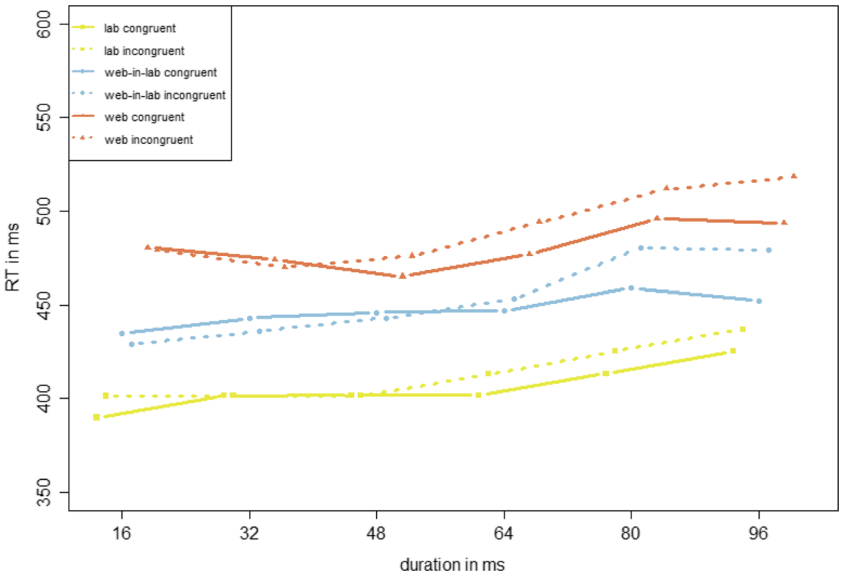

Fig. 7 Priming task results. The *dashed line* depicts the reaction time (RT) in ms for the incongruent condition for each setting (lab, web-in-lab, web), whereas the solid line represents the congruent condition. The primes had durations of between 16 and 96 ms. The expected pattern would have a higher RT for congruent trials during the prime durations of 16 and 32 ms, inverting the advantage at the durations 80 and 96 ms to higher RTs for incongruent trials

faster RTs for congruent than incongruent trials, but only for web-in-lab data, not for the other settings.

Experiment 7: attentional blink task

The attentional blink (AB) task shows that humans have a short attentional gap for identifying a second target after identifying a first target stimulus when presented within a rapid serial visual presentation (RSVP). Usually participants are presented with a train of uppercase letters, each presented for about 100 ms. One of the letters has a different feature (e.g., white color instead of black) to the others, called Target1 (T1). The target has to be memorized by the participant, while the rest of the stimulus train keeps flashing on the screen. Another target, T2, might follow in the range of 100–1,000 ms after T1. The accuracy of identifying the presence of T2 is usually highly impaired when presented about 200–400 ms after T1 (Raymond, Shapiro, & Arnell, 1992; Shapiro, Raymond, & Arnell, 1997) – an effect called attentional blink. Participants in this experiment are reporting identification of T1 (e.g., typing the letter that was white), and whether T2 appeared during the trial. Through this approach, participants need to focus on the appearance of T1 and consciously process it, instead of trying to just report whether T2 was present.

Overall, due to its robust nature in a rapid presentation environment, we found this paradigm as fitting to test our three settings. It needs consistent quick rendering of the necessary letters on screen, whereas any inaccuracies might interfere with the paradigm. This is specifically interesting when looking at the difference between lab and web technology (web-in-lab, web) experiments, as rapid presentation of stimuli is the main requirement on software in this case.

In contrast to the other experiments, AB relies on ER as an identifier for its task-specific effect. Usually a maximum of ER is found when T2 is presented about 200–400 ms after T1, which recedes as more time between T1 and T2 is given.

As the usual timing offset is supposed to be due to inaccuracies in web technology, we also expect a lower RT for web technology experiments, despite not being of relevance in the paradigm itself. Still, the ER rates should not differ between the different settings, just between conditions.

Methods

Before the start of the experiment, participants were instructed that they will see a white letter that they need to identify. Additionally they were informed that there might be a black "X," which they should identify as well. Each trial started with a fixation cross for 500 ms. It was followed by a train of 16–24 different letters from the alphabet (excluding "X"), with N-8 being the position of T1. The distractors and T2 were black size 60 px font, while T1 was white. There was a 50 % probability that T2 (black letter "X") was shown within the last eight items of the train. Each letter was shown for 100 ms centered on a gray background. The first response the participant had to give was identifying the white letter through a corresponding press on the keyboard, while the second response was about the presence of an "X" in the train (pressing 0 indicated absent, 1 present). In total, four blocks of 20 trials each were recorded, preceded by ten training trials with feedback.

Results

Following earlier studies, only trials in which T1 was correctly identified were considered in the analysis (Crump et al., 2013). Subjects with a T1 position error result of 100 % were removed from analysis (lab: 3, web-in-lab: 4), as the task was to primarily identify T1. No participants were excluded because of RT, as it was not a task-relevant measurement. Based on ER (SD = 3), five data sets were excluded. One was removed from web-in-lab, while four participants were removed from web, whic yielded 45, 43, and 39 data sets in lab, web-in-lab, and web, respectively. All further analysis was performed on T2 data.

Investigating the RT data, Mauchly's test for sphericity showed a violation for the within factor T2 position, $\chi^2(35) = 243.17$, $p < .001$, and setting × position interaction, thus a Greenhouser-Geiser correction was performed ($\varepsilon = 0.69$). A mixed ANOVA did not yield any significant effects: Neither the main effects for setting, $F(2, 124) = 2.04$, $p = .13$, nor position of T2, $F(5.52, 684.48) = 1.65$, $p = .11$, nor the interaction between setting × position, $F(11.04, 684.48) = 1.13$, $p = .32$, were significant. In total, lab data were 162 ms (M = 492 ms, SD = 410.54 ms) faster than web-in-lab (M = 654 ms, SD = 477.58 ms), while web (M = 595 ms, SD = 396.20 ms) data were 59 ms lower than web-in-lab.

Regarding ERs, the assumption of sphericity was also violated as indicated by Mauchly's test, $\chi^2(35) = 180.43$, $p < .001$, therefore degrees of freedom were corrected through the Greenhouse-Geisser method ($\varepsilon = 0.67$). Main effects were identified through a mixed-design ANOVA for the between-subjects factor setting, $F(2, 124) = 6.43$, $p = .002$, and within-subjects factor T2 position, $F(5.36, 664.63) = 90.99$, $p < .001$. No significant interaction was found for setting × T2 position, $F(10.72, 664.63) = 1.23$, $p = .24$.

Following up on the main effect of setting, post-hoc Bonferroni corrected t-tests indicated a significant difference between the settings web-in-lab and web, $p < .002$, $d = 0.86$, but not for the other combinations ($p > .14$, $d < 0.42$). Overall, lab data exhibited an ER of 39 % (SD = 35 %), web-in-lab 46 % (SD = 34 %), and web 34 % (SD = 32 %).

Following up on the main effect of T2 position, we performed Bonferroni corrected post-hoc t-tests for positions 1, 2, 3, and 4 (peak of AB effect). We found a significant difference between positions 1 and 2, $t(126) = -6.14$, $p < .001$, $d = -0.62$, and 3 and 4, $t(126) = 5.71$, $p < .001$, $d = 0.56$, but not between positions 2 and 3, $t(126) = -1.45$, $p = .45$, $d = -0.14$; a clear resemblance of the AB around 200 and 300 ms (see Fig. 8 for details).

In short, unlike in other experiments, we did not find an RT offset, but the ER data showed significant effects for setting (web more accurate than web-in-lab) and T2 position (AB effect).

Analysis summary

All results are summarized in Fig. 9. First, we note whether the task-specific effects of each paradigm have been reproduced (top row). Second, we show the RT offset between settings through the differently colored bars. Lastly, ERs are depicted as small black dots, thereby showing that the four experiments that rely on RT as a metric do not differ in ERs. In short, we found the effects of four of our five paradigms well reproduced, while the last experiment (priming) was not replicated in either setting. Still, each experiment showed significant differences in RTs between settings, with in-lab being the fastest, followed by web-in-lab and web. In an attempt to further quantify the first finding post-hoc, we calculated the Bayes factors (see Jarosz & Wiley (2014) for an overview) for each experiment (JASP 0.7.5.6, www.jasp-stats.org). Briefly, Bayes factor analysis compares two hypotheses with regard to how well they predict given data. The resulting evidential value indicates the empirical data as more probable under one of the two hypotheses. In our case, we found very strong evidence that the data were more likely to occur under the model including an effect of setting, rather than without it in RT tasks (Stroop, Flanker, Search, Priming), all $\log(BF_{10}) > 12.04$, while the attentional blink data that relied on ERs lacked the data for conclusions, $\log(BF_{10}) = 1.94$. Regarding the task-specific effects, we also found very strong evidence, all $\log(BF_{10}) > 98.45$, in all experiments, except for priming, $\log(BF_{10}) = 3.74$. See Table 1 for details.

With regard to the RT offset between settings, we tried to identify additional factors that might account for the differences over settings. First, we used the performance task in the web technology-based studies as a first indicator on whether there is a difference between browser types. The data was log-transformed due to a violation of the assumption of equality of variances (Levene's test, $F = 5.48$, $p = .03$). A Mann-Whitney test, due to violations of normality (Shapiro-Wilk test, $W = 0.70$, $p < .001$, for "Firefox" and $W = 0.79$, $p = .004$, for "Chrome"), showed a significant difference in performance, i.e. just the deployment of stimuli in regard of temporal accuracy, over browsers with the browser Chrome (N = 14, M = 9.18, SD = 5.45) being slower than the browser Firefox (N = 25, M = 5.80, SD = 3.15) in our web participants, $U = 265.5$, $p = .008$, $d = 0.43$. Yet, Bayes analysis did not yield concise evidence for either model, $\log(BF_{10}) = -0.49$. The browsers Safari (N = 3, M = 2.09, SD = 1.56) and Internet Explorer (N = 1, M = 9.87) were not incorporated into this analysis due to the low number of participants using these browsers. Based on these results – that browser type might influence performance of timing accuracy – we investigated two further points. First, we calculated a correlation between the results of the performance task and the RTs averaged over all experiments to see whether a general indication of display inaccuracy can be indicative of higher RTs. It yielded a non-

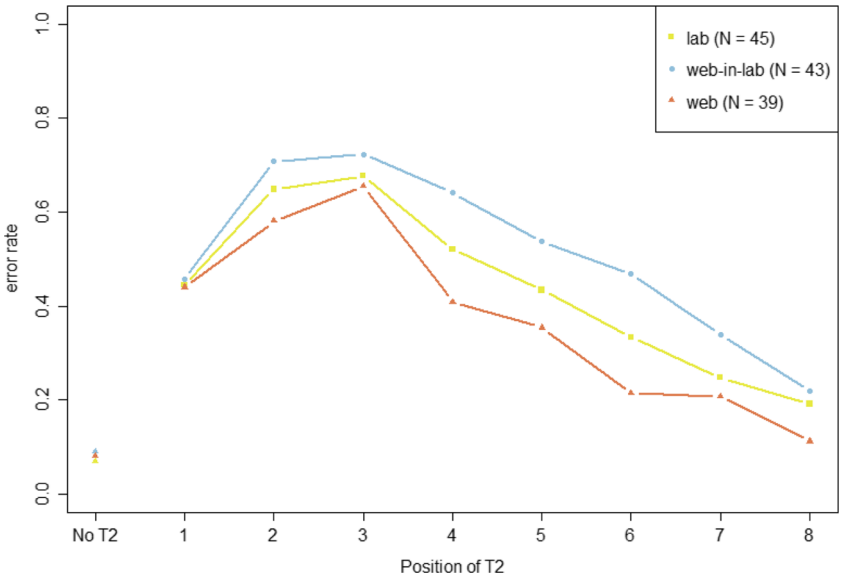

Fig. 8 Attentional blink results. Error rates of the attentional blink paradigm per setting (lab, web-in-lab, web) and position of T2 relative to T1. Attentional blink impairments are usually increasing recognition errors about 200–400 ms after T1 (position 2 to 4 in our case)

significant result, Pearson's $r(41) = -.16$, $p = .31$. Second, we ran a Mann-Whitney test (due to a violation of the normality assumption for Chrome, Shapiro-Wilk test, $W = 0.81$, $p = .006$) of the average RT (over all experiments and conditions) per browser. The results indicated a significant speed advantage in RT time for the browser Chrome ($M = 559$ ms, $SD = 50.52$) compared to Firefox ($M = 667$ ms, $SD = 55.99$), $U = 25.00$, $p < .001$, $d = 1.93$, which was very strongly supported by Bayes factor analysis, $\log(BF_{10}) = 9.18$. In short, while internal display accuracy measurements did not indicate conclusive results, there seems to be an influence of browser type on the speed of RTs.

Discussion

This study used a three step approach to identify potential differences between classical psychophysical in-lab measurements and online experiments. We conducted five well-established experiments coupled with two performance indicators to see whether the change of technology itself (lab to web-in-lab) or the change of environment (web-in-lab to web) has a specific influence on psychophysical measurements, especially RT. For four of the five main experiments (Stroop, Flanker, visual search, and attentional blink) we were able to replicate the main effects in each setting. The fifth and very timing sensitive experiment (priming) was not replicated in any setting, i.e. not even using classical methodology. These results argue for web technologies and online conduction to be in line with classical in-lab data recording.

To complement this picture, we also found our other two overarching hypotheses confirmed. First, we did find a coherent additive offset of web technology on RTs, which had an additional increase when conducted online. Second, ERs, which are not timing-dependent, were not different over settings, thereby further arguing for comparable data quality over all settings. Additionally, post-hoc analysis indicated that

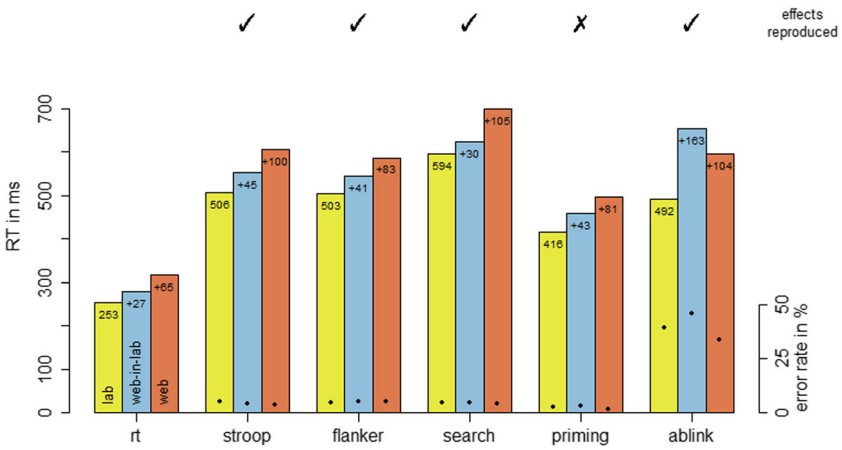

Fig. 9 Summary of results. The graph depicts the main results of our study. Reaction time (RT) offsets can be seen by comparing each colored bar with another setting, thereby showing the additive offset for web-in-lab and web data. Task-specific results – whether the effects of each paradigm were reproduced in each setting – are summed at the top. That there are no differences in ER between settings, except for attentional blink, can be seen from the black dots for each setting

Table 1 Overview of p-testing results and Bayes factor analysis

Experiment	Factor	p	log(BF_{10})	Factor	p	log(BF_{10})
Stroop	Setting	p < .001	21.89	Condition	p < .001	154.75
Flanker	Setting	p < .001	17.31	Condition	p < .001	139.42
Search	Setting	p < .001	12.04	Condition	p < .001	119.51
	Items	p < .001	55.88	Condition * Items	p < .001	378.01
Priming	Setting	p < .001	16.19	Condition	p = .001	**3.74**
	Duration	p < .001	98.45	Condition * Duration	p < .001	108.53
Att. Blink	Setting	p = .002	**1.94**	Position	p < .001	164.36

This table shows the results of p-hypothesis-testing and the Bayes factor analysis for all relevant main effects and interactions. Cells marked in bold leave the results indecisive, as p-hypothesis testing points towards a significant difference, while Bayes factor analysis does no support or reject this indication

browser type might have an influence on the RT measurement accuracy.

Reaction time offsets

As our first overarching hypothesis we expected higher absolute RTs in web technology experiments, as shown in studies before (de Leeuw & Motz, 2015; Reimers & Stewart, 2015). On average, we identified a significantly slower RT of 37 ms (range 27–45 ms, *SD* = 8.14) when comparing web-in-lab to lab data (see Fig. 9, colored bars). When conducting the experiments at home, a total offset of 87 ms (range 65–105 ms, *SD* = 16.04) was found compared to lab data. In earlier studies, RT offsets of 60 ms (Reimers & Stewart, 2015), 43 ms (Schubert et al., 2013), 25 ms (de Leeuw & Motz, 2015), and 27 + 83 ms (Hilbig, 2015) have been found. Thus, our results fit in with earlier results.

How does this additional offset emerge? As we were keeping all other factors identical between lab and web-in-lab, except for the change of experimentation software, we can safely assume that this difference is actually due to inaccuracies in the browser engine. This is reasonable, as specific software like Matlab's psychtoolbox are optimized for stimulus presentation (e.g., through screen update synchronization) and response recording, whereas web technology is not. This can be seen from the initial performance measurements where we found a significant increase of 6 ms in presentation timing. Already at this minimal load with no keyboard input or similar computationally expensive operations we found timing offset. Coupled with results from earlier studies (Barnhoorn et al., 2014; Reimers & Stewart, 2015), which showed a presentation timing offset of around 15–20 ms, inaccuracies might be attributed to the insensitivity of JavaScript.

Performing the experiments online added another 50 ms. These reasons are manifold but can be attributed to three categories: hardware, software, and environment. Taking the experiment from the lab to the home of students, onto their personal computers and laptops, changes the setting from our well-configured in-lab equipment to systems, which are not optimized for psychophysical testing. This can be due to older hardware, but could also be due to an overloaded system with many programs installed. Additionally, despite being asked to close other programs, participants might have been running additional software in the background while conducting the experiment, which can slow down performance as well. However, our additional correlation analysis did not indicate that there is a relationship between performance of the system and RT accuracy of the tasks. This allows the inference that the additional offset to conduct experiments online might be due to the changed experimental environment or a speed-accuracy tradeoff (see the section on ERs below).

Replication of task-specific effects

For the first four experiments (Stroop, Flanker, visual search, and attentional blink) the main task-specific effects could be replicated in each setting (lab, web-in-lab, web) through p-hypothesis testing and Bayesian factor analyses. These results confirm our second overarching hypothesis: Independent from the higher cognitive involvement and computational requirements of the tasks (e.g., attentional blink shows over a dozen stimuli compared to a singular item in the Stroop task), our results indicate that web technology is able to replicate well-known psychological effects.

These replications allow the inference of two statements: First, despite having a timing offset (see above), web technology is accurate enough to record RTs that differentiate between the conditions. One has to keep in mind that the duration of each task was below 10 min. This yielded a range of 24–72 trials per condition, a rather low number even for behavioral tasks. On top of that we worked with task-specific effects well below 50 ms, while still replicating them. Therefore the influence of differences of timing accuracies seems to be rather marginal. In line with Reimers and Stewart (2015), we attribute the feasibility to our within-

subjects designs, which obviously counteracts potential performance-related influences.

Second, considering that there was just a short contact without further supervision with the subjects who participated from home (web), environmental effects do not seem to influence the main task-specific effects significantly. Despite having an additional offset for online studies, the results still have been replicated through the experiments. Additionally, in the only task that solely relied on ERs (attentional blink), not lab but web data were the most accurate. This result argues against the common preconception that participants who are taking part from home are less attentive and more sloppy in their answers, which is the point of debate in the scientific community (Germine et al., 2012; Gosling et al., 2000).

The only task that was not fully replicated in either setting was the priming paradigm. Only web-in-lab and web data achieved a partial replication of faster RTs when being cued to the stimulus side at a prime duration of 96 ms. While existing literature shows contradicting results for the replication of this paradigm (Barnhoorn et al., 2014; Crump et al., 2013), we would argue that the combination of a few trials (24 per condition), very short presentation times (16–96 ms) and a high rate of exclusion (25 % of subjects) reduced the power of the data. This is also the case for the classical in-lab data, which did not produce coherent results. Bayesian factor analysis supported our assumption by indicating neither a difference nor no difference as being more probable.

To tackle this issue, researchers could employ longer experimental durations (e.g. a 1-hour long experiment instead of a few minutes) with a higher amount of trials to be able to pinpoint the effects more efficiently (as one has to consider that Crump et al.'s data was collected within 15 min through 572 trials, Barnhoorn et al.'s data in 30 min through 576 trials, and we only employed 336 trials in under 10 min). Introducing a suitable singular experiment of a much longer duration in our case would have contradicted the intention of having online publishable experiments, which can be performed quickly, as this is assumed to be one of the main factors of participation of unpaid volunteers. While many scientists rely on paid participation services like Amazon Turk, we tried to keep our tasks within a realistic time for free participation from "wild" participants, especially under the light of recent discussions of validity and generalization of AMT participants (e.g., Stewart et al., 2015).

On the other hand, an important point to note is that we used the *setTimeout()* method, as Crump et al. did, and they also could not replicate the results. In contrast, Barnhoorn et al. used the *requestAnimationFrame()* method, which synchronizes with the refresh rate of the screen, and were able to replicate the paradigm. This leads to the question of whether there is a quantifiable advantage of using one method over the other and should be investigated further.

Error rates

Our third overarching hypothesis assumed if the experiments are replicable and the timing offset is solely due to inaccuracies in the JavaScript engine, other measurements should be coherent over all three settings. In line with this argument, the three RT-based experiments that we replicated (Stroop, Flanker, visual search) did not show a difference in ER data (see Fig. 9, black dots). Through the isolation of non-timing metrics, we can argue that data quality is equal over settings, while assigning the RT differences to the JavaScript engine itself. Despite not replicating the priming task, the ER from this data was also in line with this argument. With the attentional blink task on the other hand, a significant difference of ERs over settings was found. More precisely, web data were significantly more accurate than web-in-lab data. In contrast, a common preconception would be that the in-lab participants (lab and web-in-lab) would be more likely to answer more correctly, due to the lab setting and the presence of an experimenter. Taken together, we can infer that the data quality of online participants is the same as that of in-lab participants, independent of surrounding and other variables.

Future steps

While the results of our study argue for using web technology and the Internet as a research method, we think scientists need to continue to dissect the different parts that might influence data collection when conducting online experimentation. Technology-wise, there is still a lack of knowledge on which of the many ways to present a stimulus is the most time accurate through web technology. Not only can we vary the programming language (JavaScript, Java, Flash), but also the timing methods within one language (*setTimeout()*, *requestAnimationFrame()*, see Barnhoorn et al. (2014) for an example) and stimulus presentation method for each of these timings (z-index, visibility, display, libraries like jQuery vs. native JavaScript). Here, trade-offs with regard to availability (Java might be more accurate than JavaScript, but would need to be downloaded and installed, which would defeat the purpose of large-scale online experiments) and computational load (e.g., using large libraries to ease the display of stimuli) need to be considered carefully. Additionally, our analysis of browser types indicated an effect of the browser that is used. Having different browsers producing different speeds within the same experiment argues for possible differences in timing accuracies between browser manufacturers, which in turn might introduce another, very important layer of variance. Overall, right now we have not yet identified a single optimal way to use web technology in psychophysics and therefore need to remain attentive to effects on presentation and input recording accuracy.

Another step in this experimental line is to publish the array of tasks for the broader public to participate. Here we would enter the "wild" part of the internet by changing the population compared to our web setting. Studies up to now argued that volunteers from the internet (Germine et al., 2012) as well as paid participants through merchant systems like AMT (Buhrmester, Kwang, & Gosling, 2011) are contributing valid and usable data, yet there is not much evidence on whether there is a difference between types of participants and, if so, which ones. An interesting approach would be to use earlier ideas from questionnaire research (Göritz, 2006) and simply test different incentives for the same task. Not only would this allow identification of the effectiveness of different recruiting approaches, but also show whether data quality varies with the origin and motivation of participants too.

Summary

Summing up, our results extend the current state of knowledge of online experimentation in psychology by providing an extensive array of experiments that single out the potential differences when changing from a classical in-lab to an online setting. We used directly comparable experimental setups to reproduce five well-known experiments through a single investigation, thereby keeping the settings experimentally identical. In this process we were able to coherently replicate the different task-specific effects through web technology. The resulting argument, that web technology has reached a stage that allows for online experimentation, was confirmed by additional metrics like ERs, which either were consistent over settings or even exhibited a slight advantage for online conducted data. A reaction timing offset for each change (technology and environment) was found and seems to be due to inherent factors of the JavaScript engine, possibly varying between browser types, but details need further studies to confirm our assumptions. Taken together, all three findings are strong arguments that online experimentation through the use of web technology can be a near to equivalent substitute for classical in-lab data acquisition in psychophysical experimentation.

Acknowledgments We would like to thank Marisa Nordt for valuable discussions regarding analysis and Astrid Hönekopp, Tobias Meißner, Helen Prüfer, Katharina Sommer, and Ricarda Weiland for help in collecting the data. All code, raw data and analysis files can be found at The Open Science Framework (http://osf.io/qzy2g).

References

Barnhoorn, J. S., Haasnoot, E., Bocanegra, B. R., & van Steenbergen, H. (2014). QRTEngine: An easy solution for running online reaction time experiments using qualtrics. *Behavior Research Methods*. doi:10.3758/s13428-014-0530-7

Birnbaum, M. H. (2000). Introduction to psychological experiments on the internet. In M. H. Birnbaum (Ed.), *Psychological Experiments on the Internet* (pp. xv – xx). Academic Press. doi: 10.1016/B978-012099980-4/50001-0

Birnbaum, M. H. (2004). Human research and data collection via the internet. *Annual Review of Psychology, 55,* 803–32. doi:10.1146/annurev.psych.55.090902.141601

Brainard, D. (1997). The psychophysics toolbox. *Spatial Vision*, 433–436. doi: 10.1163/156856897X00357

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. doi:10.1177/1745691610393980

Cohen, J., Collins, R., Darkes, J., & Gwartney, D. (2007). A league of their own: demographics, motivations and patterns of use of 1,955 male adult non-medical anabolic steroid users in the United States. *Journal of the International Society of Sports Nutrition, 4*(1), 12. doi:10.1186/1550-2783-4-12

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*(3), e57410. doi:10.1371/journal.pone.0057410

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*, (2014). doi: 10.3758/s13428-015-0567-2

Eimer, M., & Schlaghecken, F. (2002). Links between conscious awareness and response inhibition: Evidence from masked priming. *Psychonomic Bulletin & Review, 9*(3), 514–520. doi:10.3758/BF03196307

Elze, T., & Tanner, T. G. (2012). Temporal properties of liquid crystal displays: Implications for vision science experiments. *PloS One, 7*(9), e44048. doi:10.1371/journal.pone.0044048

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics, 16*(1), 143–149.

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. doi:10.1037/0278-7393.10.4.680

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*(5), 847–57. doi:10.3758/s13423-012-0296-9

Göritz, A. S. (2006). *Incentives in Web Studies : Methodological Issues and a Review, 1*(1), 58–70.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2000). Should we trust web-based studies? a comparative analysis of six preconceptions about internet questionnaires. *The American Psychologist, 59*(2), 93–104. doi:10.1037/0003-066X.59.2.93

Stewart, N., Ungemach, C., Harris, A. J. L., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making, 10*(5), 479–491. doi:10.1017/CBO9781107415324.004

Hilbig, B. E. (2015). Reaction time effects in lab- versus Web-based research: Experimental evidence. *Behavior Research Methods*. doi:10.3758/s13428-015-0678-9

Jarosz, A. F., & Wiley, J. (2014). What are the odds? a practical guide to computing and reporting Bayes Factors. *The Journal of Problem Solving, 7,* 2–9. doi:10.7771/1932-6246.1167

Joinson, A. N. (2001). Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology, 31*(2), 177–192. doi:10.1002/ejsp.36

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109*(2), 163–203. doi:10.1037//0033-2909.109.2.163

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*(2), 227–234. doi:10.1037/h0031564

Neath, I., Earle, A., Hallett, D., & Surprenant, A. M. (2011). Response time accuracy in Apple Macintosh computers. *Behavior Research Methods, 43*(2), 353–62. doi:10.3758/s13428-011-0069-9

Plant, R. R., & Turner, G. (2009). Millisecond precision psychological research in a world of commodity computers: New hardware, new problems? *Behavior Research Methods, 41*(3), 598–614. doi:10.3758/BRM.41.3.598

Ramsey, J. D., & Morrissey, S. J. (1978). Isodecrement curves for task performance in hot environments. *Applied Ergonomics, 9*(2), 66–72. doi:10.1016/0003-6870(78)90150-3

Raymond, J., Shapiro, K., & Arnell, K. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal of Experimental ….* Retrieved from http://psycnet.apa.org/journals/xhp/18/3/849/

Reimers, S. (2007). The BBC internet study: General methodology. *Archives of Sexual Behavior, 36*(2), 147–61. doi:10.1007/s10508-006-9143-2

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*(2), 309–327. doi:10.3758/s13428-014-0471-1

Reips, U.-D. (2000). The web experiment: Advantages, disadvantages, and solutions. *Psychology Experiments on the Internet*, 89–117.

Sanders, A. F., & Lamers, J. M. (2002). The Eriksen flanker effect revisited. *Acta Psychologica, 109,* 41–56.

Schmidt, W. C. (2001). Presentation accuracy of Web animation methods. *Behavior Research Methods, Instruments, & Computers : A Journal of the Psychonomic Society, Inc, 33*(2), 187–200. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11447672

Schubert, T. W., Murteira, C., Collins, E. C., & Lopes, D. (2013). ScriptingRT: A software library for collecting response latencies in online studies of cognition. *PloS One, 8*(6), e67769. doi:10.1371/journal.pone.0067769

Shapiro, K. L., Raymond, J. E., & Arnell, K. M. (1997). The attentional blink. *Trends in Cognitive Sciences*. doi:10.1016/S1364-6613(97)01094-2

Skitka, L. J., & Sargis, E. G. (2006). The internet as psychological laboratory. *Annual Review of Psychology, 57,* 529–55. doi:10.1146/annurev.psych.57.102904.190048

Stewart, N., Ungemach, C., Harris, A. J. L., Bartels, D.M., Newell, B. R., Paolacci, G., et al. (2015). The average labo-ratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making, 10*(5), 479–491. doi:10.1017/CBO9781107415324.004

Stroop, J. R. (1935). Stroop color word test. *Journal of Experimental Physiology* (18), 643–662. doi: 10.1007/978-0-387-79948-3

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record, 58,* 475–482.

Windes, J. D. (1968). Reaction time for numerical coding and naming of numerals. *Journal of Experimental Psychology, 78*(2), 318–22. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/5722448

Wolfe, J. M. (1998). Visual search. *Attention, Perception, & Psychophysics, 20,* 13–73. doi:10.1016/j.tics.2010.12.001