Research Article

Marian Sauter*, Maximilian Stefani, Wolfgang Mack

Equal Quality for Online and Lab Data: A Direct Comparison from Two Dual-Task Paradigms

https://doi.org/10.1515/psych-2022-0003
received July 28, 2021; accepted January 3, 2022.

Abstract: Conducting behavioral experiments online has become more prevalent recently. Still, there is reluctance to embrace the possibilities this technology has to offer. So far, only simple tasks have been replicated in an online setting. In order to investigate whether collecting online also leads to high quality data in demanding tasks, we directly compared data collected in the lab with data collected online from a demanding dual-task paradigm and a psychological refractory period paradigm. In Experiment 1, we recruited from local pools, online and offline; in Experiment 2, we collected lab data from our local pool and online data from a remote commercial participant platform. We found that all relevant effects were replicated in the lab and online settings; effect sizes were similar. Additionally, most response time distributions were even statistically equivalent when comparing online and lab data. Thus, online effect sizes and variances can be comparable to lab-based data. Online studies are time-efficient and recruiting an online sample instead or on top of a laboratory sample should be considered for basic behavioral research. This can serve an important role in the generalizability and replicability of findings in the cognitive and behavioral sciences.

Keywords: online experiments; web-based experiments; dual-task.

1 Introduction

Nowadays, researchers who conduct research with questionnaires would hardly think of collecting them with pen and paper. What was still common 10 years ago has now been largely replaced by computer technology and the internet, because these make it possible to finally collect huge samples and reach people from all countries and population strata. The cost-benefit analysis is extremely positive (Birnbaum & Birnbaum, 2000). But for psychological experiments, we still think of laboratory-based measurements as the gold standard. The predominant concern with running experiments online is poor data quality resulting from the largely uncontrolled technical setup and surroundings, which is seen as troublesome especially for response time sensitive paradigms and small effects (Sauter et al., 2020). There are still very few online studies in cognitive psychology even if the precision, accuracy, and popularity rise (Anwyl-Irvine, Dalmaijer, et al., 2020; Bridges et al., 2020). To alleviate this concern, prior studies have shown that online data can be reliable and comparable to lab data in various cognitive tasks (Arechar et al., 2018; Crump et al., 2013; Semmelmann & Weigelt, 2017). One of the first larger online replication approaches included stroop, switching, flanker, simon, posner cuing, attentional blink, subliminal priming, and category learning tasks (Crump et al., 2013). The authors found that most effects (but not all: e.g. subliminal priming) could be well replicated using a sample recruited from Amazon Mechanical Turk. The authors further found that 'random responding' was almost never an issue and they offer specific recommendations for web-based experiments. However, since they did not directly compare online-data with lab-data, it is uncertain, whether the issues in replicating well-known effects stems from the wider online sample or the web-based experiment technology. Semmelmann and Weigelt (2017) also replicated well-known psychological experiments (i.e. reaction time tasks, stroop tasks, flanker tasks or priming tasks) either in a classical laboratory, online, or online

***Corresponding author: Marian Sauter,** Universität der Bundeswehr München, Allgemeine Psychologie; Ulm University, General Psychology, E-mail: marian.sauter@uni-ulm.de
Maximilian Stefani, Wolfgang Mack, Universität der Bundeswehr München, Allgemeine Psychologie

in the laboratory setting. Except for the priming task, they were able to replicate all experiments in the three different settings. Semmelmann and Weigelt (2017) argued that the power of the priming task was very weak due to a high rate of exclusions. On the one hand, this may have been because they were able to replicate other effects that were less than 50 ms, but on the other hand, priming tasks are very time sensitive (as effect sizes are typically very small) and they missed accurate timing in their experiment. Additionally, Barnhoorn et al. (2015) showed that time-sensitive response time tasks also work online. Using also a JavaScript-based environment, the masked-priming task could be replicated and thus effects below 50 ms could be demonstrated. Direct comparisons of typical lab-based student recruitment and recruitment in the wild' are sparse and limited to one-dimensional tasks (Birnbaum, 2000; Germine et al., 2012; Reimers & Stewart, 2015; Semmelmann & Weigelt, 2017) and generalization across psychophysical tasks is necessary to build a holistic picture of data quality in online experiments. This is now easily possible due to the wide range of available tools that have been developed in recent years.

The technical implementation of online based studies has been studied since the popularity of the internet (see Musch & Reips, 2000, for an overview). While 10 years ago people tried to implement online experiments with small self-programmed platforms (Keller et al., 2009; T. W. Schubert et al., 2013), today there are popular platforms with large communities. These range from classic lab-based software like PsychoPy (Peirce et al., 2019) or OpenSesame (Mathôt et al., 2012) with corresponding online extensions to pure browser-based solutions like lab.js (Henninger et al., 2019) or Gorilla (Anwyl-Irvine, Massonnié, et al., 2020), see Sauter et al. (2020) for an overview. However, they all have in common that they are based on HTML5 / JavaScript, are thus executed locally in the browser and do not depend on a fast internet connection during execution. Bridges et al. (2020) demonstrated that if sub-millisecond accuracy is not critical, pretty much all platforms are suitable for acquiring cognitive experiments with visual and auditory stimuli.

A group of paradigms suitable for approaching a generalization is multi-tasking. Classical multi-tasking (or rather: dual-tasking) experiments are simple in their setup but demanding and high measurement accuracy is important – similar to priming tasks. There are the classical dual-task paradigm (DT) in which either two extremely simple tasks are presented at exactly the same time (Ruthruff et al., 2001) or the refractory period paradigm (PRP), in which the two tasks are presented shortly after one another with varying second stimulus onset asynchronies (see review, Pashler, 1994). In the DT paradigm, the participants are instructed not to prefer one of the two tasks, and in most cases, they are free to decide in which order to perform the tasks. Without training, this inevitably leads to an increase in response time in the task that was performed second. By comparing the tasks in the dual-task and in the single-task condition, it is now possible to calculate costs that differ depending on stimulus and response pairing and training status. In the PRP paradigm, participants are instructed to always respond in the order in which the stimuli were presented. The stimuli are presented in a time-shifted manner (stimulus onset asynchrony or SOA), which can cause an overlap of the processes of both tasks. However, the SOA only affects the second task to be processed (RT2), the smaller the SOA (e.g. at 16 ms or 50 ms) the higher is RT2. With large SOAs (e.g. 1000 ms) there is no more overlapping of the processes and RT2 is as fast as if the task had been set as a single-task.

In the present study, we conducted two experiments to investigate whether dual-task effects can reliably be shown in an uncontrolled online setting. In Experiment 1, we compared dual-task costs in both the classic dual-task paradigm as well as the PRP paradigm for participants recruited in the lab versus participants recruited online, while we advertised predominantly among the same participant pools for both methods. In Experiment 2, we compared dual-task costs and variances in the PRP paradigm for new lab-based participants versus participants recruited online through a separate commercial participant pool (prolific.co). Note that for both of the experiments, we used the same experimental script for the lab-based recruitment and the online recruitment so that our comparison is not confounded by technological differences in the experimental software.

So overall, we were expecting two things, (1) regardless of the setting (in-lab vs. online), both the dual-task costs and the PRP effect can be observed, (2) despite high demands on timing, we can observe the highest response times for RT2 with a SOA of 16 ms (PRP effect) in the in-lab (hereinafter referred to as lab) and online setting.

2 Experiment 1

2.1 Methods

2.1.1 Participants

We excluded all participants who responded to less than 85% of all trials correctly (13 participants - 12 online, 1 offline - in the PRP condition, accuracy range: 65% to 84%, and zero participants in the DT condition), because we assumed that they did not understand the task instructions. In the final data, 127 participants took part in the dual-task condition (15 offline, 112 online) and 113 participants took part in the PRP condition. Across both conditions, 236 individual participants took part (median age: 23, range: 18-54; 164 male, 85 female, 1 diverse), while participants in the lab participated in both conditions and online in only one condition. Participants in the lab conditions were all students of the Bundeswehr University Munich (median age: 23, range: 20-29). Participants in the online condition were not selected from a specific participant pool (median age: 23, range: 18-54). The study was advertised among students of the Bundeswehr University Munich and across social media. All participants indicated that they have normal or corrected-to-normal vision. They provided informed consent and received course credit (lab) or no compensation (online) for their participation. This study was carried out in accordance with the recommendations of the Universität der Bundeswehr München and the Deutsche Forschungsgesellschaft. Strict Covid-19 hygiene protocols were in place. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

2.1.2 Setup

The experiment was programmed using OpenSesame version 3.3.8 with Python 3.7.6 and the OSWeb extension (version 1.3.13) with JavaScript ES6 (Mathôt et al., 2012). Psycho was used as backend and the resolution was set to 1280 px x 720 px. The experiment was hosted locally on servers at the university using JATOS as the participant management software (Lange et al., 2015).

Online. The only requirement was that the experiment was conducted on a PC or laptop with a proper keyboard. Apart from that, there were no restrictions specified with regards to the participants' hardware. The stimulus size was not scaled according to screen size. This means, that the stimuli were of equal size for all participants in terms of physical properties – but not in terms of visual angle, as we do not know how far the participants were sitting from the screen.

Lab. The visual stimuli were displayed on an EIZO® color monitor with a screen diagonal of 27 inches and a frame rate of 144 Hz at a resolution of 3840 × 2160 pixels. The experiment was started in the Firefox Browser (version 86.0+build3+0ubuntu0.20.04.1) on PCs with Ubuntu 20.04.2 LTS (64-bit).

2.1.3 Stimuli and Procedure

We contrasted data from traditional lab-based sources and unconstrained online sources in two short implementations of classical dual-task paradigms. In particular, online participants had to complete either a task in the psychological refractory period (PRP) paradigm or in the dual-task (DT) paradigm (see Figure 1 for the task progression of a single trial), whereas lab participants completed both paradigms (counterbalanced order). In the DT condition, participants had to first learn two basic tasks in 2x8 training trials: (1) a green (Hex: #88D18A) or blue (Hex: #20639D) disk (height: 32 px; width: 128 px) appeared centrally on the screen and participants had to press the "g" or "b" key accordingly using their left hand index- and middle finger (color task); (2) a disk (Hex: #999999; radius = 50 px) appeared on the screen either left or right of a central fixation cross and participants had to indicate this by pressing the "left arrow" or "right arrow" key with their right-hand index or middle-finger respectively (location task). After they did eight training trials in these tasks individually in single-task blocks, they were combined in a dual-task. In dual-task blocks, dual-task trials and single-task trials were mixed. This means that participants either had to respond to a gray disk appearing left or

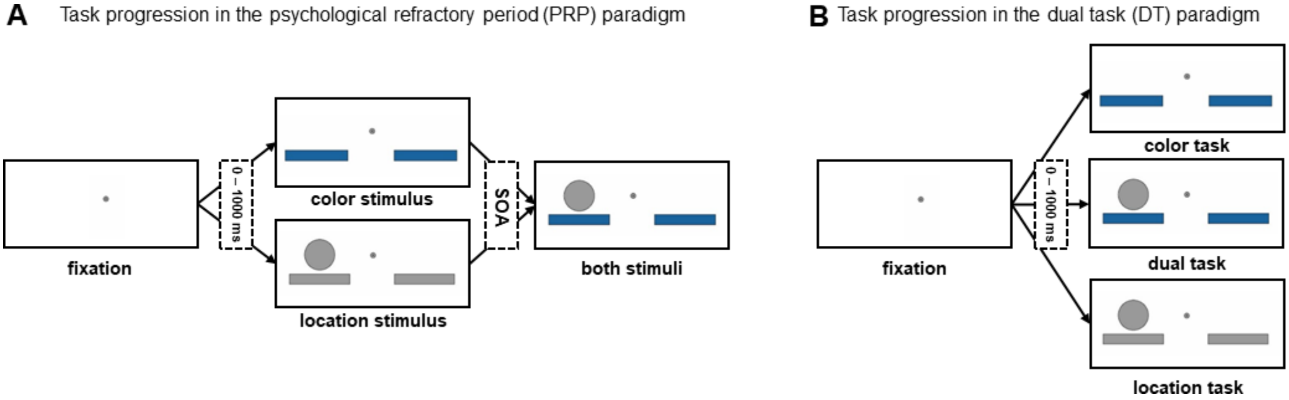
A Task progression in the psychological refractory period (PRP) paradigm
fixation
0 – 1000 ms
color stimulus
location stimulus
SOA
both stimuli
B Task progression in the dual task (DT) paradigm
fixation
0 – 1000 ms
color task
dual task
location task

Figure 1: Depiction of the task progression in **A.** the psychological refractory (PRP) paradigm and **B.** the dual task (DT) paradigm. In the PRP paradigm, a fixation dot is shown for a random time between 0 and 1000 ms, then the first stimulus appears and stays on the screen for a specified time (SOA) until the second stimulus appears. In the DT paradigm, a fixation dot is shown for a random time between 0 and 1000 ms and then either the color stimulus, location stimulus or both stimuli are shown at the same time.

right of the central fixation (location single-task), or a green or blue disk that appeared at the central fixation (color single-task) or a disk that appeared left or right of the central fixation cross and was either blue or green (dual-task).

In the DT condition, a trial started with a fixation dot that was shown for a random time between 0 ms and 1000 ms. Immediately after, either only one stimulus appeared (single-task trial) or both stimuli appeared (dual-task trial) and participants were tasked to indicate the respective responses. Participants completed 4 blocks à 32 trials (16 dual-task, 8 location single-task, 8 color single-task trials).

In the PRP condition, stimuli were identical to the DT condition. A trial started with a fixation dot that was shown for a random time between 0 ms and 1000 ms. Immediately after, the first stimulus appeared (counterbalanced which task appeared first). the second stimulus followed after a specifically set stimulus onset asynchrony (SOA) of 16ms, 133ms, 500ms or 1000ms (note that they cannot be controlled don to the millisecond in an online setting due to various monitors' refresh rates). Participants completed 4 blocks à 32 trials. In contrast to the DT condition, there were no single-task trials intermixed.

In both conditions, the stimuli stayed on the screen until all required responses were indicated. Then a fixation dot appeared. If the responses were correct, the fixation dot stayed on the screen for a random time between 1000 ms and 2000 ms. If the responses were incorrect, the fixation dot stayed on the screen for a random time between 2000 ms and 4000 ms (in order to discourage errors). Then the next trial started. Participants received feedback about the average response time and accuracies after each block.

2.2 Data analysis

All response time calculations are reported as means of individual participant mean values (± standard deviation). If not otherwise specified, in the subsequent analyses, we used Welch's t-tests for between-group comparisons (i.e. online vs. lab) and paired-samples t-tests for within-group comparisons. Prefacing the hypotheses tests, we excluded all extreme outlier trials according to the boxplot method (above the third quantile plus three times the inter-quartile-range and below the first quantile minus three times the inter-quartile range). In the dual-task data, we then looked at when first and second responses were indicated and contrasted it for online and lab data. We then investigated the critical dual-task costs by means of t-tests. For the PRP data, we calculated a repeated-measures ANOVA with the between-subject factors method (online vs. lab) and within-subject factor SOA (16 ms, 133 ms, 500 ms, 1000 ms), followed by t-tests contrasting online and lab data in the individual SOA conditions. In both dual task and PRP experiments, when there was no significant difference in the between-group t-tests, we proceeded with an equivalence analysis to reveal whether the two effects (i.e. for online data vs. lab data) are *practically* equivalent. The rationale behind this analysis is to find a suitable method to replace an existing method when it offers practically equivalent effects (e.g. in medical research it can be used to judge whether a new and cheaper treatment results in equivalent therapeutic effects for the patients). In our case, we want to see whether the more time-efficient and (often) less expensive method of testing dual-task paradigms in the lab can be replaced be online testing, while resulting in similar effects. The analysis follows the established TOST (two-one-sided-*t*-tests) method (Daniël Lakens, 2017). Essentially, two equivalence bounds (upper and lower) are specified based on the smallest effect size of interest (e.g., $d = 0.4$). Then a confidence interval is calculated around the observed effect size. All effects more extreme than the equivalence bounds are rejected, when their confidence interval overlaps with the equivalence bounds. We used the R TOSTER package (Daniel Lakens, 2018).

2.3 Results

2.3.1 Dual-task (DT) condition

Before beginning the analysis, we investigated for outlier trials based on the task types (single vs. dual) for the individual participants. Using the boxplot method, we excluded extreme outliers (2.5%).

Response times. Overall, participants in the DT condition gave their first response within 720 ± 106 ms (lab), respectively, 738 ± 557 ms (online), $t(113.54) = -0.3$, $p = .763$, $d = 0.03$, 95% CI [-135 ms, 99 ms]. They indicated their second response within 953 ± 161 ms (lab), respectively, 912 ± 230 ms (online); $t(22.5) = 0.89$, $p = .385$, $d_z = 0.19$, 95% CI [-56 ms, 139 ms]. To proceed, dual-task RT costs were calculated by subtracting single-task response times from dual-task response times for each participant and task (color vs. location) separately. We compared the variances using Levene's test, which was not significant ($p = .636$). There was no significant difference in dual-task costs between lab (229 ms) and online (90 ms) data; $t(119.06) = 1.27$, $p = .205$, $d_z = 0.13$, 95% CI [-77 ms, 354 ms]. The equivalence test was non-significant, t(125) = -1.053, p = 0.147, given equivalence bounds of -447 and 447 (on a raw scale) and an alpha of 0.05. Based on the equivalence test and the null-hypothesis test combined, we can conclude that the observed effect is statistically not different from zero and statistically not equivalent to zero.

Error rates. Errors were infrequent overall (lab: 3.1 ± 3.2 %, online: 3.4 ± 2.9 %). Dual-task error costs were calculated by subtracting single-task rates from dual-task rates for each participant and task (color vs. location) separately. In terms of error rates, dual-task costs were not visible, lab: 1.3 ± 3.2 %, online: -0.1 ± 3.0 %, $t(17.46) = 1.66$, $p = .114$, $d = 0.48$, 95% CI [-0.4 %, 3.3 %].

2.3.2 PRP condition

Response times. Again, we first excluded extreme outlier trials using the boxplot method (2.9%). All response time calculations are reported as means of individual participant mean values. Overall, participants gave their first response in 748 ± 258 ms (lab), or 790 ± 241 ms (online) and their second response in 755 ± 203 ms (lab) or 718 ± 253 ms (online). We calculated a repeated-measures ANOVA on the second response times with the between-subject factors method (online vs. lab) and within-subject factor SOA (16 ms, 133 ms, 500 ms, 1000 ms). For an overview, see Figure 3. Only the main effect SOA was significant ($F(3, 363) = 719.28$, $p < .001$, details see Table A1). We also calculated *t*-tests in order to reveal whether the PRP effect was consistently shown in all SOA-comparisons for online and lab data. It was significant for all comparisons (all *ps* < .0001)

Since there was no main effect of method, we went on to do an equivalent analysis for each of the SOA conditions separately (alpha = 0.05). Levene's test indicated that we can assume equal variance for all comparisons (all *ps* > .095). Equivalence bounds were chosen based on a Cohen's $d = 0.4$ for all comparisons. *For the 16-ms-SOA condition*, the equivalence test was non-significant, $t(121) = -0.936$, $p = 0.176$, given equivalence bounds of -81 ms and 81 ms. *For the 133-ms-SOA condition*, the equivalence test was non-significant, $t(121) = -0.806$, $p = 0.211$, given equivalence bounds of -79 ms and 79 ms. *For the 500-ms-SOA condition*, the equivalence test was non-significant, $t(121) = -0.636$, $p = 0.263$, given equivalence bounds of -69 ms and 69 ms. *For the 1000-ms-SOA condition*, the equivalence test was non-significant, $t(22.89) = -0.0297$, $p = 0.488$, given equivalence bounds of -49 ms and 49 ms. As with the DT results, there is still uncertainty regarding the cross-method comparison results for the PRP effect.

Error rates. We calculated error rates by separately calculating the rates for the first and second response and then averaging them. Similar to the DT condition, error rates were comparable and error infrequent: Participants responded

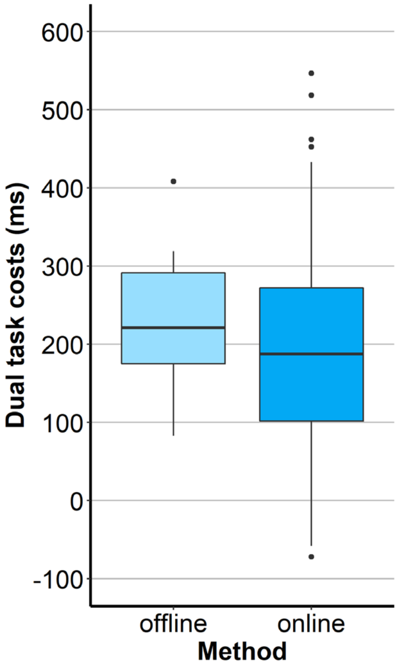

Figure 2: Dual-task costs for the DT task as a function of the method (lab vs online). Colored boxes mark the 25th to 75th percentile range of the group data; whiskers extend to 1.5 times the inter-quartile range. Black dots show the outlier participants' mean RTs.

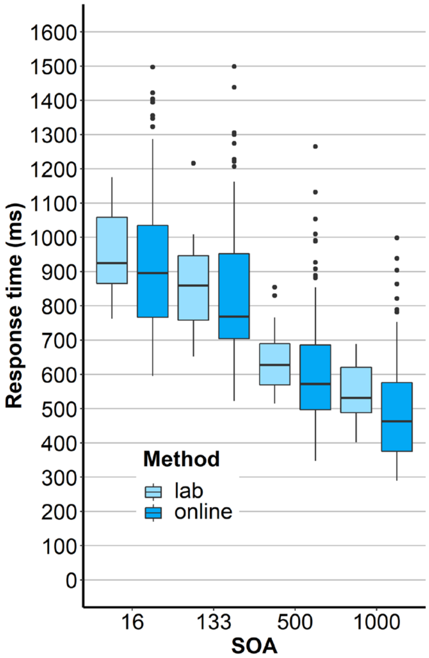

Figure 3: Response times for the PRP task as a function of the SOA.). Colored boxes mark the 25th to 75th percentile range of the method data (light blue: lab; blue: online); whiskers extend to 1.5 times the inter-quartile range. Black dots show the outlier participants' mean RTs.

wrongly in 4 ± 3 % (lab) or 3 ± 3 % of trials (online). As with RTs, we calculated the same repeated-measures ANOVA as for response times with the factors method and SOA to check whether errors depended on condition. The main effect method was not significant: $F(1, 121) = 1.19$, $p = .277$, $\eta^2_p = 0.01$; The main effect SOA was not significant: $F(3, 363) = 1.64$, $p = .18$, $\eta^2_p = 0.01$; The interaction method X SOA was not significant: $F(3, 363) = 2.52$, $p = .057$, $\eta^2_p = 0.02$.

2.4 Discussion

In the first experiment, we compared whether commonly shown dual-task effects were present in both online and classical lab recruitment methods. In the dual-task paradigm, we found that the dual-task costs were pretty much comparable. A clear average cost could be found in dual-tasks when compared to their single-task equivalents. As can be seen in Figure 2, in the online recruitment, the effect was solid across most participants, with the effect size for the 25th to 75th percentile easily above 100 ms. In the PRP paradigm, we found that the second responses were dependent on the SOA (as was expected) for both online and lab data. This means, that for both online and lab experiments we fully replicated the commonly found dual-task effects.

Variances are numerically higher across participants recruited online and there were some participants who did not show a dual-task effect. This might be the dominant reason why the lab vs online effects were not to be found statistically equivalent. Note that higher variance was expected due to the unconstraint nature of the recruitment and environment participants were situated in while participating in the experiment. Additionally, a wider demographic was attracted (participants were slightly older online and the range was wider) and we recruited almost eight times more participants in the online condition compared to the lab condition. For a just comparison of variances, sample sizes in both conditions should be comparable as well. Arguably, especially sample size in the lab experiment was likely too low for the equivalence test (Rusticus & Lovato, 2014). It might additionally be, that reliability of the data suffered due to the comparably low amount of trials for each participant. While typical dual-task studies have around 400-1000 trials in one session (e.g. Hazeltine et al., 2006; T. Schubert et al., 2008; Strobach et al., 2018), the present experiment prioritized faster completion time and was therefore left with only 128 trials.

It can be argued that the online condition was not a “true in the wild” online condition as participants were recruited from similar participants’ pool in both conditions (i.e. advertised mostly in circles of psychology students in the same University). So, while it could be shown that effects are reliable across the two recruitment methods, participants’ demographics were likely still quite homogenous, especially in their socio-economic status. In order to allow for generalizations across true internet-based samples, a demographic beyond local samples needs to be recruited. Overall, sizable dual-task and PRP effects could be shown in both the online and lab condition, but the effects were not found to be equivalent.

3 Experiment 2

In the first experiment, we showed dual-task and PRP effects in both an online and lab environment. However, the effects were not found to be equivalent. In the second experiment, sample sizes will be more closely aligned between online and lab condition (increased power for equivalence tests), the duration of the experiment is to be increased (higher within-subject reliability) and online recruitment is not based on the University’s student sample but on an unrelated globally-accessible online participant pool (Prolific.co). Only[1] the PRP paradigm will be reproduced, as it is more sensible to reveal differential effects because it comprises of five individual SOA-conditions (i.e. five comparisons) instead of only one comparison.

3.1 Methods

3.1.1 Participants

In a first step, we excluded all participants with incomplete data (2 participants) and who responded to less than 85% of all experimental trials correctly (6 participants, accuracy range: 76% to 84%), because we assumed that they did not fully internalize the task instructions. All excluded participants were from the online condition. In the final data set, 94 participants (51 lab, 43 online) were analyzed for this experiment (median age: 23, range: 18-35; 57 male, 38 female). Participants in the lab condition (median age: 23, range: 20-30) were students of the Bundeswehr University Munich. Participants in the online condition (median age: 25, range: 18-35) had to be between 18 and 35 years old, have student status, and speak German as a first language to remain comparable to the lab condition. All participants indicated that they have normal or corrected-to-normal vision. This study was carried out in accordance with the recommendations of the Universität der Bundeswehr München and the Deutsche Forschungsgemeinschaft. Strict Covid-19 hygiene protocols were in place. All subjects gave written informed consent in accordance with the Declaration of Helsinki. They received course credit (lab) or get 3.75 £ (online) for their participation.

3.1.2 Setup

The setup was the same as in Experiment 1.

3.1.3 Stimuli and Procedure

The stimuli and general procedure were exactly the same as for the PRP paradigm from Experiment 1, except that Experiment 2 consisted of 40 practice trials in a single-task block and 10 dual-task blocks à 32 experimental trials (as opposed to 4 blocks in Experiment 1). Thus, Experiment 2 was longer compared to Experiment 1: it took around 30 minutes.

[1] This trade-off was unfortunately necessary due to limited access to the on-site laboratory during the Covid-19 pandemic.

3.2 Results

Data analysis was the same as in Experiment 1, if not described otherwise. Before doing the response time and error analyses, we investigated the data for outliers based on each participants' performance in each SOA condition. We excluded all extreme outlier trials (2.3%) identified by the boxplot method (all trials with response times +- three times the inter-quartile range).

3.2.1 Comparison of online and in-lab data

All response time calculations are reported as means of individual participant mean values (± standard deviation). Overall, participants gave their first response in 809 ± 173 ms (lab), or 886 ± 263 ms (online) and their second response in 788 ± 275 ms (lab) or 804 ± 286 ms (online). We first went on to test the normality assumption with the Shapiro-Wilk test for all combinations of our factors method (online vs lab) and SOA (16 ms, 133 ms, 500 ms, 1000 ms). All tests were significant (*p*s < .03), so normality assumption was not met, which will be factored into the interpretation. We also tested variance homogeneity with Mauchly's test of sphericity, which was significant (*p* < .001). We went on to calculate a repeated-measures ANOVA with the between-subject factor method and the within-subject factor SOA. Only the main effect SOA was significant (details see Table 1, means see Table 2). As we did not observe an interaction, there are reasonable indications, that generally, the patterns of results for online and lab data were similar.

Table 1: Results of the repeated-measures ANOVA with factors method (online vs lab) and SOA (16 ms, 133 ms 500 ms, 1000 ms).

Predictor	df_{Num}	df_{Den}	*Epsilon*	*F*	*p*	η^2_g
Method	1.00	92.00		0.17	.680	.00
SOA	1.64	150.70	0.55	707.60	.000	.51
Method x SOA	1.64	150.70	0.55	0.90	.392	.00

Note. df_{Num} indicates degrees of freedom numerator. df_{Den} indicates degrees of freedom denominator. Epsilon indicates Greenhouse-Geisser multiplier for degrees of freedom, *p*-values and degrees of freedom in the table incorporate this correction. η^2_g indicates generalized eta-squared.

In order to investigate, whether the two methods led in fact to similar results, we further conducted post-hoc paired *t*-tests and equivalence tests for each of the SOAs separately. Levene's tests revealed that equal variances can be assumed for all but the 1000-ms-SOA condition[2]. Following Simonsohn (2015)'s suggestion, the equivalence bound was set to the effect size Experiment 1 had 33% power to detect (this results in *d* = 0.42). Alpha was set to 0.05. *For the 16-ms-SOA condition*, the equivalence test was significant, *t*(92) = -1.918, *p* = 0.0291, given equivalence bounds of -93 ms and 93. The null hypothesis test was non-significant, *t*(92) = 0.111, *p* = 0.912. *For the 133-ms-SOA condition*, the equivalence test was significant, *t*(92) = 1.803, *p* = 0.0374, given equivalence bounds of -94 ms and 94 ms. The null hypothesis test was non-significant, *t*(92) = -0.226, *p* = 0.822. *For the 500-ms-SOA condition*, the equivalence test was non-significant, *t*(92) = -0.708, *p* = 0.481, given equivalence bounds of -77 ms and 77 ms. The null hypothesis test was non-significant, *t*(92) = -0.708, *p* = 0.481. *For the 1000-ms-SOA condition*, the equivalence test was non-significant, *t*(68.29) = 1.052, *p* = 0.148, given equivalence bounds of -65 ms and 65 ms. The null hypothesis test was non-significant, *t*(68.29) = -0.940, *p* = 0.350.

[2] To investigate whether the wider age range for online participants is a prime determinant for the increased standard deviation in the 1000-ms-SOA condition, we first correlated the within-subject SD with age in Experiment 2 and found a significant negative correlation, Pearson's r = -.18, t(358) = -3.4008, p < .001. For a more straightforward interpretation, we followed this up with a linear regression only for the 1000-ms-SOA condition which resulted in a regression coefficient of -10.8 for age. This means that for each year a participant got older, the SD decreased by 10.8 ms. Given that the mean age is slightly higher for online (24.4 years) than for offline data (23.0 years), age does not seem to be a prime determinant of SD.

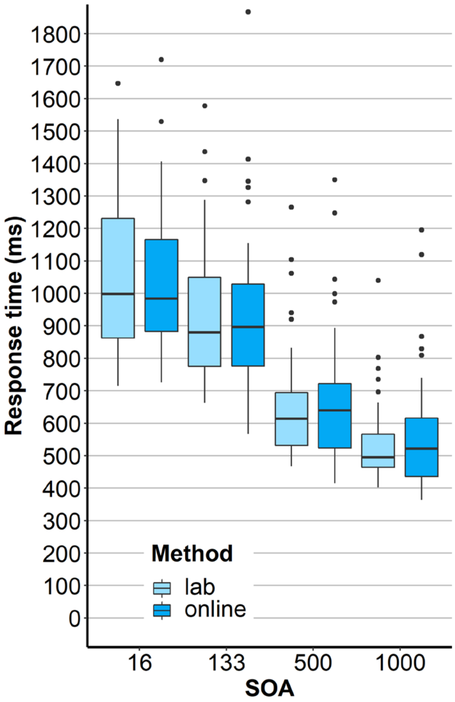

Figure 4: Means and standard deviations for the second response time as a function of a 4 (SOA) X 2 (method) design. Colored boxes mark the 25th to 75th percentile range of the method data (light blue: lab; blue: online); whiskers extend to 1.5 times the interquartile range. Black dots show the outlier participants' mean RTs.

Table 2: Means and standard deviations for the second response time as a function of a 4 (SOA) X 2 (method) design.

method: lab			
soa	*M*	*M* 95% CI [LL, UL]	*SD*
16	1042	[981, 1104]	219
133	930	[872, 989]	208
500	649	[601, 696]	168
1000	532	[499, 564]	116
method: online			
soa	*M*	*M* 95% CI [LL, UL]	*SD*
16	1038	[969, 1106]	224
133	941	[867, 1014]	240
500	675	[613, 738]	202
1000	562	[506, 619]	184

Note. *M* and *SD* represent mean and standard deviation, respectively. *LL* and *UL* indicate the lower and upper limits of the 95% confidence interval for the mean, respectively. The confidence interval is a plausible range of population means that could have created a sample mean (Cumming, 2014).

Overall, based on the equivalence tests and the null-hypothesis test combined, we can conclude that the observed effect is never statistically different from zero. It is statistically equivalent to zero for the 16-ms and 133-ms condition.

3.2.2 PRP effect comparisons

Lastly, we calculated *t*-tests in order to reveal whether the PRP effect was consistently shown in all comparisons for online and lab data. The results can be seen in Table A2.

4 Discussion

In the second experiment, we compared whether commonly shown dual-task effects (specifically: psychological refractory period effects) were present in both online and classical lab recruitment methods. We recruited an “standard” lab-based student sample and compared this to a more diverse online sample recruited through a globally available recruitment platform. We found that the second responses were dependent on the SOA (as was expected), with all SOA conditions significantly differing from all other SOA conditions for both online and lab data. This means, we fully replicated the commonly found PRP effects both in-lab and online.

Variances were not found to differ and the SOA-specific RT distributions can be considered equivalent between online and lab data (exception: 1000-ms-SOA condition). Additionally, all post-hoc *t*-tests revealed significant differences in RTs between the various SOA conditions We therefore argue that we consistently showed PRP effects in both online and in-lab data. This is especially meaningful as the online sample was recruited through the platform Prolific, which mostly represents people untrained in cognitive psychology tasks.

4.1 General Discussion

In the present study, we investigated whether online experiments serve as a good expansion of classically lab-based dual-task research. In particular, we contrasted effect sizes and variances in online and lab-based experiment-conduction for a dual-task paradigm (E1) and psychological refractory paradigm (E1 and E2). We showed that effect sizes are practically indifferent between online and lab data, even when the same number of participants is recruited. And while the distribution of most of the participants (i.e. 25th to 75th percentile) seems to be similar for online and lab data, there are arguably more outlier participants (> 1.5 IQR) in the online dataset. This was to be expected as generally, the environment is less restricted in online settings, both in terms of control on the surroundings as well as the wider demographic that is recruited.

We can state three important implications about online vs. lab experiments. First, the within-subject variance was not found to differ between online and lab data (as demonstrated by Levene tests), except for the 1000-ms-SOA condition (variance seems higher for online data). We believe that this condition could have behaved differently because it is the only condition that allowed for significant time to respond to the first task, before the second stimulus – 1000 ms after the first stimulus - was even shown on the screen. This means it was the only condition in which the two sub-tasks were clearly de-coupled, i.e. it was easily possible to respond to the first task but not the second task. Second, we observed significant differences between all SOA conditions. This means that timing posed little to no difficulties. Although the variance was largest in the conditions with a SOA of 16 ms and 133 ms, this was to be expected since most participants are only equipped with screens with a refresh rate of 60 Hz at home. This means that especially with a SOA of 16 ms it could happen that the second stimulus was presented one frame too late, i.e. at 34 ms. However, the results indicate that this not necessarily disrupts effects and shows that effects smaller than 50 ms can also be shown online. Third, online experiments do not necessarily suffer from data quality issues, even in longer (i.e. 30min) studies, at least if the experiment is instructed and designed appropriately and participants are paid. Incorporating block-wise and trial-wise feedback may have been helpful as well.

We had to exclude some participants in the online condition, because they did not match our accuracy criterion of 85%-correct. Exclusions and dropouts are quite common for online experiments (Sauter et al., 2020), possibly, as task instructions cannot be easily tailored to individual participants. Perhaps, our instructions were not clear for everyone. Some researchers claim that motivation might be lower in online experiments (a fear, which is often not warranted, Clifford & Jerit, 2014; Hauser & Schwarz, 2016), so some participants will press random keys just to ‘get through’ and get paid. However, as the accuracies ranged from 65% to 84% for the participants who were excluded, this is clearly not the case in the present study. Chance performance would have resulted in a 25%-correct rate, as a trial was considered incorrect when one or both responses were incorrect. These participants might have prioritized speed over accuracy. In any case, for online experiments, it is valuable to take a priori measures for ensuring data quality (Kees et al., 2017; Sauter et al., 2020).

Typical behavioral studies in cognitive psychology make wide clams about the generalizability of their results (at least implicitly) but are really only based on a very niche WEIRD (western, educated, industrialized, rich, democratic) demographic, i.e. young adult psychology students. Our study has used a sampling method for the online data that is accessible to researchers across the globe. This allows other researchers to easily replicate the results if they choose to do so, because not only the experimental protocol is available, also the same demographic is targetable. In addition, it would even be easily possible to lift the limits on occupation status (we only recruited students), first language and age in order to allow for a much farther generalization of the results than can be achieved in any lab setting. Given the powerful tools for running behavioral experiments online that are available nowadays (Sauter et al., 2020), such a sampling strategy should be adopted for all behavioral experiments in which it is ethically and technically feasible. In particular, we argue that all researchers conducting simple behavioral experiments in cognitive psychology should consider recruiting an online sample in addition to their classical lab-based sample allowing them to make claims of generalizability beyond their typical niche demographic. This can help restore confidence in basic research among researchers and the public, which has been damaged by the replication crisis (Wingen et al., 2020).

Overall, in the present study we showed, that data in dual-task paradigms conducted online, amount of errors, response times, effect sizes and variances can be comparable to lab-based data, especially in the psychological refractory period paradigm. As online studies are efficiently conducted, they do not tax researchers’ resources too much and we argue that considering to recruit an online sample in addition to or instead of a laboratory sample should become a standard for basic behavioral research.

Author contributions: M.S. = Marian Sauter, M.St. = Maximilian Sefani, WM = Wolfgang Mack. Conceptualization: M.S. (lead) and M.St.; Data curation: M.S. and M.St.; Formal analysis: M.S. and M.St.; Funding acquisition: W.M.; Investigation: M.S. and M.St. (lead); Methodology: M.S. (lead) and M.St.; Project administration: M.S.; Resources: W.M.; Software: M.S. (lead) and M.St.; Validation: M.S. (lead) and M.St.; Visualization: M.S.; Writing – original draft: M.S.; Writing - review & editing: M.S. (lead), M.St. and W.M.;

Financial statement: This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projektnummer PA 2947/1-1 (to S.P.).

Conflict of interest: Authors report no conflict of interest.

Ethical statement: The protocol of this study was approved by the ethics committee of the Universität der Bundeswehr München. All subjects gave written informed consent in accordance with the Declaration of Helsinki and the German Psychological Society (DGPs).

References

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

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

Arechar, A. A., Gächter, S., & Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics*, *21*(1), 99–131. https://doi.org/10.1007/s10683-017-9527-2

Barnhoorn, J. S., Haasnoot, E., Bocanegra, B. R., & van Steenbergen, H. (2015). Qrtengine: An easy solution for running online reaction time experiments using Qualtrics. *Behavior Research Methods*, *47*(4), 918–929. https://doi.org/10.3758/s13428-014-0530-7

Birnbaum, M. H. (2000). Introduction to psychological experiments on the internet. In M. H. Birnbaum & M. O. Birnbaum (Eds.), *Psychological Experiments on the Internet* (pp. XV–XX). Elsevier. https://doi.org/10.1016/B978-012099980-4/50001-0

Birnbaum, M. H., & Birnbaum, M. O. (Eds.). (2000). *Psychological Experiments on the Internet*. Elsevier.

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

Clifford, S., & Jerit, J. (2014). Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory and Online Studies. *Journal of Experimental Political Science*, *1*(2), 120–131. https://doi.org/10.1017/xps.2014.5

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.

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), 847–857. https://doi.org/10.3758/s13423-012-0296-9

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407.

Hazeltine, E., Ruthruff, E., & Remington, R. W. (2006). The role of input and output modality pairings in dual-task performance: Evidence for content-dependent central interference. *Cognitive Psychology*, *52*(4), 291–345. https://doi.org/10.1016/j.cogpsych.2005.11.001

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2019). lab.js: A free, open, online study builder. Advance online publication. https://doi.org/10.31234/osf.io/fqr49

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An Analysis of Data Quality: Professional Panels, Student Subject Pools, and Amazon's Mechanical Turk. *Journal of Advertising*, *46*(1), 141–155. https://doi.org/10.1080/00913367.2016.1269304

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

Lakens, D [Daniel]. (2018). *Package 'TOSTER'*. https://cran.microsoft.com/snapshot/2018-07-03/web/packages/toster/toster.pdf

Lakens, D [Daniël] (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362.

Lange, K., Kühn, S., & Filevich, E. (2015). “Just Another Tool for Online Studies” (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies. *PloS One*, *10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). Opensesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Musch, J., & Reips, U.-D. (2000). A Brief History of Web Experimenting. In M. H. Birnbaum & M. O. Birnbaum (Eds.), *Psychological Experiments on the Internet* (pp. 61–87). Elsevier. https://doi.org/10.1016/B978-012099980-4/50004-6

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, *116*(2), 220–244. https://doi.org/10.1037//0033-2909.116.2.220

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). Psychopy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

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

Rusticus, S. A., & Lovato, C. Y. (2014). *Impact of Sample Size and Variability on the Power and Type I Error Rates of Equivalence Tests: A Simulation Study.* https://doi.org/10.7275/4S9M-4E81

Ruthruff, E., Pashler, H., & Klaassen, A. (2001). Processing bottlenecks in dual-task performance: Structural limitation or strategic postponement? *Psychonomic Bulletin & Review*, *8*(1), 73–80. https://doi.org/10.3758/BF03196141

Sauter, M., Draschkow, D., & Mack, W. (2020). Building, Hosting and Recruiting: A Brief Introduction to Running Behavioral Experiments Online. *Brain Sciences*, *10*(4). https://doi.org/10.3390/brainsci10040251

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. https://doi.org/10.1371/journal.pone.0067769

Schubert, T., Fischer, R., & Stelzel, C. (2008). Response activation in overlapping tasks and the response-selection bottleneck. *Journal of Experimental Psychology. Human Perception and Performance*, *34*(2), 376–397. https://doi.org/10.1037/0096-1523.34.2.376

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

Simonsohn, U. (2015). Small Telescopes:Detectability and the Evaluation of Replication Results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Strobach, T., Hendrich, E., Kübler, S., Müller, H., & Schubert, T. (2018). Processing order in dual-task situations: The “first-come, first-served” principle and the impact of task order instructions. *Attention, Perception & Psychophysics*, *80*(7), 1785–1803. https://doi.org/10.3758/s13414-018-1541-8

Wingen, T., Berkessel, J. B., & Englich, B. (2020). No Replication, No Trust? How Low Replicability Influences Trust in Psychology. *Social Psychological and Personality Science*, *11*(4), 454–463. https://doi.org/10.1177/1948550619877412

Appendix

Table A1: Repeated-measures ANOVA with the factors method and SOA in Experiment 1.

Predictor	df_{Num}	df_{Den}	*Epsilon*	SS_{Num}	SS_{Den}	*F*	*p*	η^2_g
(Intercept)	1.00	121.00		256642870.54	13199695.48	2352.61	.000	.94
method	1.00	121.00		66638.16	13199695.48	0.61	.436	.00
soa	2.04	247.41	0.68	14548409.92	2447382.68	719.28	.000	.48
method x soa	2.04	247.41	0.68	2856.96	2447382.68	0.14	.873	.00

Note. df_{Num} indicates degrees of freedom numerator. df_{Den} indicates degrees of freedom denominator. Epsilon indicates Greenhouse-Geisser multiplier for degrees of freedom, *p*-values and degrees of freedom in the table incorporate this correction. SS_{Num} indicates sum of squares numerator. SS_{Den} indicates sum of squares denominator. η^2_g indicates generalized eta-squared.

Table A2: Post-hoc t-tests for all comparisons of method and SOA combinations in Experiment 2. p shows unadjusted and p.adj. shows Holm-adjusted p-values.

method	SOA1	SOA2	n1	n2	*p*	*p*. adj
lab	16	133	51	51	0,002	0,003
	16	500	51	51	0,000	0,000
	16	1000	51	51	0,000	0,000
	133	500	51	51	0,000	0,000
	133	1000	51	51	0,000	0,000
	500	1000	51	51	0,001	0,003
online	16	133	43	43	0,037	0,037
	16	500	43	43	0,000	0,000
	16	1000	43	43	0,000	0,000
	133	500	43	43	0,000	0,000
	133	1000	43	43	0,000	0,000
	500	1000	43	43	0,015	0,030