A primer on running human behavioural experiments online

Tijl Grootswagers*

*School of Psychology, University of Sydney, NSW, 2006, Australia, tijl.grootswagers@sydney.edu.au

Abstract

Moving from the lab to an online environment opens up enormous potential to collect behavioural data from thousands of participants with the click of a button. However, getting the first online experiment running requires familiarisation with a number of new tools and terminologies. There exist a number of tutorials and hands-on guides that can facilitate this process, but these are often tailored to one specific online platform. The aim of this paper is to give a broad introduction to the world of online testing. This will provide a high-level understanding of the infrastructure before diving into specific details with more in-depth tutorials. Becoming familiar with these tools allows moving from hypothesis to experimental data within hours.

Introduction

Lightning fast internet speeds and significant technological improvements have made it possible to perform complex experiments within a modern web-browser. It is becoming increasingly popular to combine browser-based experiments with recruiting participants on platforms such as Amazon's Mechanical Turk (MTurk) or Prolific Academic (Palan & Schitter, 2018). There are several reasons why researchers opt for online instead of lab-based testing. The first is efficiency. The recruitment platforms (e.g., MTurk) have access to large numbers of participants, allowing to test many (thousands) of participants simultaneously, which would not be possible in a lab-based setting. They are also not restricted to office hours or teaching schedules, and do not require an on-campus presence for participants or researchers (Note this document was written during the COVID-19 pandemic). Secondly, participants from the online platforms are a better reflection of the general population than the undergraduate students who typically participate in experiments on campus (Berinsky et al., 2012). Finally, online experiments are more economical[1], because there is no need to spend time recruiting, scheduling, and testing participants.

Our lab has had an overwhelmingly positive experience with running online studies (Grootswagers et al., in press, 2017, 2018). While early days involved extensive JavaScript programming for relatively simple online studies, recent advancements have made it much easier to get complex studies up and running (Anwyl-Irvine, Massonnié, et al., 2020; Barnhoorn et al., 2014; De Leeuw, 2015; Henninger et al., 2019; Peirce et al., 2019). These generally come with associated tutorials and hands-on guides, but these are often specific to a single platform or method. Therefore, it can be a challenge to get familiar with the infrastructure, tools, and terminology, especially when starting out from scratch. This document aims to facilitate this process by introducing the basics to online testing. It is intended to serve as a high-level overview, and guide the reader to relevant in-depth literature, reviews, and tutorials.

[1] There has been discussion about online studies being exploitative but the experimenter can pay participants a fair compensation in accordance with institutional ethics review boards (c.f. Crump et al., 2013; Mason & Suri, 2012; Shank, 2016)

The basics

The core infrastructure needed for online experiments consists of: (1) a browser-based experiment (2) a server to host the experiment, and (3) a participant recruitment tool. Figure 1 illustrates the general infrastructure and workflow for online experiments. Experiments are programmed to run in a browser and hosted on a server. Participants are recruited from online marketplaces and perform the task on their local machine. The data is uploaded to the hosting server where the experimenter can collect the results.

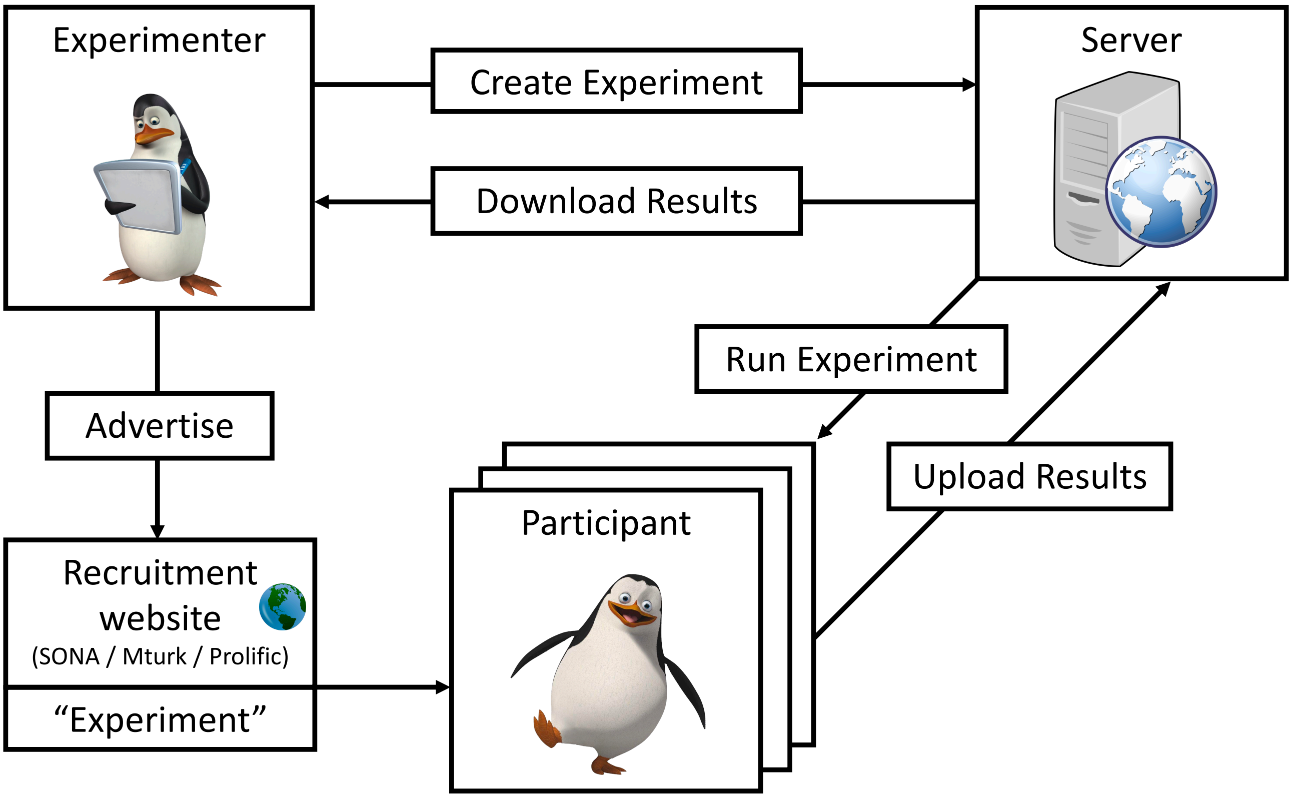

Figure 1. Infrastructure model for online experiments.

Creating the experiment

The experiment needs to be able to run in a web-browser (e.g., Safari, Google Chrome, Internet Explorer). It therefore needs to be programmed in a browser-compatible programming language (e.g., JavaScript, PhP). The most popular language for online experiments is JavaScript, and there exist several JavaScript modules (e.g., JsPsych, PsychoJS, GorillaJS, Lab.js) tailored to behavioural experiments. The libraries provide a number of high-level functions to facilitate experiment-specifics, such as presenting stimuli, control timing, randomisation, and collecting responses. Some are accompanied by graphical interfaces that allow creating experiments without the need for any programming. For example, the Psychopy builder (Peirce et al., 2019) can export the experiment as JavaScript code.

Hosting the experiment

The experiment needs to be accessible to the world. This involves *hosting* the experiment code, stimuli, and libraries on a server. This allows a participant to access the experiment code from their web browser. The experiment then runs in the browser on the participants computer. The participant completes the experiment, and the script sends the participants experimental data back to the server. This means that the server should be able receive and store the experiment data. Several hosting tools exist that are specifically aimed at collecting behavioural data online, such as pavlovia or gorilla. Alternatively, experiments can be hosted using cloud services (e.g., Google or Amazon) but this requires a more hands-on approach.

Recruiting participants

The final step is to recruit participants. What is needed for this is a marketplace (on the web) where participants can view and sign up for experiments. When they decide to participate, they get the link (URL) to the experiment server and complete the task. Examples of such marketplaces are SONA systems (often used for undergraduate testing at universities), MTurk, or Prolific (Palan & Schitter, 2018). To be able to give participants compensation (e.g., course credits or payment) for their participation, online experiments often display a unique code that participants can enter in the recruitment system so the experimenter can verify their participation. It is useful to note the time zone of the participants, for example, MTurk workers (based in the US) will be more likely to be online and see the experiment if it is posted during their daytime. The recruitment systems will have the option to specify how many participants are needed, and some provide additional screening criteria. When all participants have completed the experiment, the researcher can simply download the data from the server and start analysing.

Frequently asked questions

The basic infrastructure needed for online testing is not overly complex, as described in the previous section. In addition, the available infrastructure has improved significantly in recent years with the development of more sophisticated hosting solutions and programming libraries. Once familiar with these powerful tools, it is extremely easy to go from hypothesis to experimental data within hours. The remainder of this paper will cover a number of frequently asked questions with regards to online testing.

How good are the data?

Several studies have compared data from online markets to data collected in the lab (Barnhoorn et al., 2014; Crump et al., 2013; de Leeuw & Motz, 2016; Simcox & Fiez, 2014; Zwaan & Pecher, 2012), with overall positive results. Tutorials and reviews have suggested that online experiment data is generally better when experiments are short, pay well, are fun, and have clear instructions. It is good to keep in mind that participants from online marketplaces (e.g., MTurk) are not as familiar with psychology experiments compared to undergraduate students. Therefore, it is essential to make very clear instructions and sometimes include a number of practice trials to ensure they understand the task.

How good is the timing?

Despite the progress in web-based technology, stimulus and response timing will be less reliable than the commercial equipment used in the lab. In general, latencies and variabilities are higher in web-based compared to lab-environments. Several studies have assessed the quality of timing in online studies, with encouraging results (Anwyl-Irvine, Dalmaijer, et al., 2020; Bridges et al., 2020; Pronk et al., 2019; Reimers & Stewart, 2015). An online evaluation of a masked priming experiment showed that very short stimulus durations (i.e., under 50ms) can be problematic (but see Barnhoorn et al., 2014), and other classic experimental psychology paradigms that rely on reaction times (e.g., Stroop, flanker, and Simon tasks) were successfully replicated (Crump et al., 2013).

What are the limitations?

Online experiments only work for some stimulus modalities. While the online approach is well suited for experiments consisting of visual stimuli and keyboard or mouse responses (but see previous question on timing), other paradigms are harder or impossible to move online. For example, studies requiring auditory stimuli are possible (Cooke et al., 2011; Gibson et al., 2011; Schnoebelen & Kuperman, 2010; Slote & Strand, 2016), but may necessitate a more extensive set-up procedure, such as procedures to make sure the participants set-up works. Presenting stimuli in other modalities, such as tactile or olfactory stimuli, are impossible to achieve in an online environment.

A second limitation is the lack of experimental control. For example, while participants screen size is reported by the browser, there is no way to know the participants distance to screen. It is therefore impossible to control the exact visual angle of stimuli, which can be a limiting factor for some experiments. It is also hard to test whether participants are paying attention to the experiment. A common approach is to exclude participants based on their performance on catch-trials (Mason & Suri, 2012). Still, there can be a large amount of variability in attention amongst online participants and they could be distracted by other sources while performing experiments, such as listening to radio, looking at their phone, or watching their children.

Conclusion

Online experiments offer large-scale participant testing in a short time and are cheaper to run than their lab-based counterparts. They can be a suitable option for many research questions but have some limitations in the amount of experimental control. This manuscript has provided a high-level overview of the infrastructure. For more in-depth reading, the reader is referred to the more specialised tutorials and reviews cited above. The JavaScript experiment libraries (e.g., JsPsych, PsychoJS, GorillaJS, Lab.js) also have associated hands-on tutorials and contain many examples of classic cognitive science experiments, which are a good place to start with programming the online experiment.

References

Anwyl-Irvine, A. L., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2020). *Online Timing Accuracy and Precision: A comparison of platforms, browsers, and participant's devices* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/jfeca

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

Barnhoorn, J. S., Haasnoot, E., Bocanegra, B. R., & Steenbergen, H. (2014). 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

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. https://doi.org/10.1093/pan/mpr057

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* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/d6nu5

Cooke, M., Barker, J., Lecumberri, M. L. G., & Wasilewski, K. (2011). Crowdsourcing for word recognition in noise. *Twelfth Annual Conference of the International Speech Communication Association.*

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. https://doi.org/10.1371/journal.pone.0057410

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12.

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

Gibson, E., Piantadosi, S., & Fedorenko, K. (2011). Using Mechanical Turk to Obtain and Analyze English Acceptability Judgments. *Language and Linguistics Compass*, *5*(8), 509–524. https://doi.org/10.1111/j.1749-818X.2011.00295.x

Grootswagers, T., Cichy, R. M., & Carlson, T. A. (2018). Finding decodable information that can be read out in behaviour. *NeuroImage*, *179*, 252–262. https://doi.org/10.1016/j.neuroimage.2018.06.022

Grootswagers, T., Kennedy, B. L., Most, S. B., & Carlson, T. A. (in press). Neural signatures of dynamic emotion constructs in the human brain. *Neuropsychologia*. https://doi.org/10.1016/j.neuropsychologia.2017.10.016

Grootswagers, T., Ritchie, J. B., Wardle, S. G., Heathcote, A., & Carlson, T. A. (2017). Asymmetric Compression of Representational Space for Object Animacy Categorization under Degraded Viewing Conditions. *Journal of Cognitive Neuroscience*, *29*(12), 1995–2010. https://doi.org/10.1162/jocn_a_01177

Henninger, F., Shevchenko, Y., Mertens, U., Kieslich, P. J., & Hilbig, B. E. (2019). *lab.js: A free, open, online experiment builder*. Zenodo. https://doi.org/10.5281/zenodo.2775942

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23. https://doi.org/10.3758/s13428-011-0124-6

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

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

Pronk, T., Wiers, R. W., Molenkamp, B., & Murre, J. (2019). Mental chronometry in the pocket? Timing accuracy of web applications on touchscreen and keyboard devices. *Behavior Research Methods*. https://doi.org/10.3758/s13428-019-01321-2

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

Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research1. *PSIHOLOGIJA*, *43*(4), 441–464.

Shank, D. B. (2016). Using Crowdsourcing Websites for Sociological Research: The Case of Amazon Mechanical Turk. *The American Sociologist*, *47*(1), 47–55. https://doi.org/10.1007/s12108-015-9266-9

Simcox, T., & Fiez, J. A. (2014). Collecting response times using Amazon Mechanical Turk and Adobe Flash. *Behavior Research Methods*, *46*(1), 95–111. https://doi.org/10.3758/s13428-013-0345-y

Slote, J., & Strand, J. F. (2016). Conducting spoken word recognition research online: Validation and a new timing method. *Behavior Research Methods*, *48*(2), 553–566. https://doi.org/10.3758/s13428-015-0599-7

Zwaan, R. A., & Pecher, D. (2012). Revisiting Mental Simulation in Language Comprehension: Six Replication Attempts. *PLOS ONE*, *7*(12), e51382. https://doi.org/10.1371/journal.pone.0051382