sp open
swiss psychology open

How to Run Behavioural Experiments Online: Best Practice Suggestions for Cognitive Psychology and Neuroscience

RESEARCH

ubiquity press

NATHAN GAGNÉ
LÉON FRANZEN
*Author affiliations can be found in the back matter of this article

ABSTRACT

The combination of a replication crisis, the global COVID-19 pandemic in 2020, and recent technological advances, have accelerated the on-going transition of research in cognitive psychology and neuroscience to the online realm. When participants cannot be tested in-person, data of acceptable quality can still be collected online. While online research offers many advantages, numerous pitfalls may hinder researchers in addressing their questions appropriately, potentially resulting in unusable data and misleading conclusions. Here, we present an overview of the costs and benefits of conducting online studies in cognitive psychology and neuroscience, coupled with detailed best practice suggestions that span the range from initial study design to the final interpretation of data. These suggestions offer a critical look at issues regarding recruitment of typical and (sub)clinical samples, their comparison, and the importance of context-dependency in each part of a study. We illustrate our suggestions by means of a fictional online study, applicable to traditional paradigms such as research on working memory with a control and treatment group.

CORRESPONDING AUTHOR:
Léon Franzen
University of Lübeck, DE
leon.franzen@mail.com

KEYWORDS:
online research; best practice; psychology; neuroscience; cognition; dyslexia

TO CITE THIS ARTICLE:
Gagné, N., & Franzen, L. (2023). How to Run Behavioural Experiments Online: Best Practice Suggestions for Cognitive Psychology and Neuroscience. *Swiss Psychology Open,* 3(1): 1, pp. 1–21. DOI: https://doi.org/10.5334/spo.34

INTRODUCTION

In the midst of a global pandemic, experimental research has exceeded the realm of physical space through online experimentation. In a change that started years ago, a large-scale transition was long overdue. The increasing shift to online experiments in cognitive psychology and neuroscience was enabled by increasing technological innovations; shifting the focus to both the costs and benefits attributed to online experiments. The 2020 COVID-19 pandemic forced most researchers to transition online almost overnight, yet many of these researchers possess limited experience with this new delivery method, and several challenges await them.

With careful implementation, the benefits attributed to online research present the potential to address some of the current issues in the fields of cognitive psychology and neuroscience. In recent years, researchers in these fields have found themselves in a dire replication crisis (Baker & Penny, 2016; Ioannidis, 2005; Ioannidis et al., 2014; Makel et al., 2012). The failure to replicate findings of previous work has been a growing trend (Sharpe & Poets, 2020), as reproducibility rates of published findings in the field of psychology are estimated to only range between 35% and 75% (Artner et al., 2020; Camerer et al., 2018; Etz & Vandekerckhove, 2016). In fact, more than half of researchers have tried and failed to reproduce their own studies (Baker & Penny, 2016). These numbers do not build trust in reported findings within academia and the public eye (Sauter et al., 2022). Causes for low replication rates had been attributed to the complex nature of reproducing experimental methodology, including problems with statistical power, selective and small convenience sampling, engaging in questionable research practices, publication bias, and high costs associated with running a study repeatedly with a sufficient sample size (Button et al., 2013; Holcombe, 2020; Ioannidis et al., 2014; John et al., 2012; Munafò et al., 2017; Nosek et al., 2012; Rosenthal, 1979; Simmons et al., 2011). However, the replication crisis constitutes only one example of a general problem that can be addressed with online research. Other aspects include greater online accessibility for recruitment, as well as quicker and cheaper training of budding researchers for instance. Nonetheless, online designs are not a blanket solution and must be thoughtfully implemented.

Until the start of the COVID-19 pandemic, data validity and reliability had only been investigated for selected leading platforms, such as Amazon's Mechanical Turk (MTurk; e.g., Berinsky et al., 2012; Chandler et al., 2014; Crump et al., 2013), and groups of experiments (e.g., psychophysics: de Leeuw & Motz, 2016; perception: Woods et al., 2015). However, the sudden shift encouraged publications addressing the generic implementation of online experiments in accordance with the current technological possibilities (Grootswagers, 2020; Sauter et al., 2020). These constitute a great starting point, but a more nuanced take on many practical and experimental aspects of online research remains absent from the literature. Recent work found no statistical difference between online and in-person testing (Sauter et al., 2022), whereas other work reveals a small yet acceptable loss of data quality when comparing online testing to in-person testing (Uittenhove et al., 2022). Notably, the latter authors place a large emphasis on sampling by suggesting that "who" is tested is more important than "how" they are tested. Specifically, participants from the online platform Prolific (Palan & Schitter, 2018) were found to be more indistinguishable to students than MTurk participants. In fact, Prolific consistently provided higher data quality than MTurk (Peer et al., 2021). These findings underline the importance of deliberately choosing the platform and participants for successful studies.

The present work provides an overview of the costs and benefits of online experimentation, paired with some best practice suggestions spanning the entire study cycle from the design to interpreting the results—with a focus on the sampling of specific populations. To this end, we use a fictional example of a study recruiting a specific sub-clinical dyslexia sample, as a practical illustration that can be generalized to most difficult-to-recruit populations. The recruitment of these populations has a long history of proving difficult and being biased (Blythe et al., 2018).

OVERVIEW OF COSTS AND BENEFITS OF ONLINE RESEARCH

Online research provides a wealth of opportunities to improve psychological studies and to address some of the postulated pitfalls (henceforth, costs) that have previously deterred researchers from using this delivery method (see costs listed in Table 1). Often methodological problems with online research have appeared as major obstacles in the eyes of many researchers. Specifically, **costs of online research** may include reduced control of the testing environment and possibility for intervention, especially in a mass online testing setting. The physical absence of researchers in the online realm prevents certain types of direct intervention, if issues arise during the experiment, and precludes the mitigation of extraneous distractors during testing. However, this limitation only applies if participants are being tested in a mass online setting. Intervention on an individual level, such as organising an online call during testing, remains a possibility, specifically for research with special populations. These extraneous distracting variables present an opportunity to confound results when a study is conducted outside of the laboratory. Distractors, and unforeseen issues may also add noise to the data. For example, participants may start answering at random part-way through the experiment due to a lack of motivation or attention, or they may even start cheating on the task by screen grabbing.

COSTS
▪ Limited control of testing environment, higher risk of distractions
▪ Limited possibility for intervention during testing in mass online testing, physical absence of researchers
▪ Noisy data
▪ Increased lack of motivation and attention due to extended general computer time usage
▪ Higher dropout rates
▪ Compensation of cognitive or perceptual differences by expensive hardware
▪ Varying computer processing capabilities, timing inaccuracies for brief stimulus display
▪ Greater potential for cheating (providing invalid data) and participant fraud (pretending to be someone one is not)
▪ Potentially greater temptation to pay participants non-adequately, no in-person interaction
▪ If a bug is present and data was not collected in batches, much data needs to be discarded
▪ Reliable access to online studies required (internet access and proper equipment)
▪ Sampling of non-naïve participants (particularly via platforms)
BENEFITS
▪ No physical presence needed
▪ Time reduction for experimenter(s)' testing time commitment – without supervision requirements
▪ Rapid data collection (in parallel)
▪ Easily collect more data for larger sample sizes
▪ Increased possibilities for recruitment, generally more accessible for most people (depending on socioeconomic background)
▪ Collect more representative and heterogenous samples (depending on sampling method)
▪ Increased autonomy for participants about the time and location of participation
▪ Reduced social pressure and feelings of obligation to finish a study
▪ Access for more trainees to run their own study earlier in their career
▪ No lab equipment and space needed; all-in-one solutions provided by platforms
▪ Potential for increased data anonymity (e.g., for special populations)
▪ Reduced equipment and research costs

Table 1 List of costs and benefits of online studies. To facilitate evaluation of the importance and consequences of items, costs and benefits indicate a potential trade-off any experimenter would engage in when conducting research online.

The physical presence of the researcher in an in-person setting not only adds social pressure to perform well on the experiment, but the social interaction itself may also offer added motivation to participate. As such, it is conceivable that participants who are looking at a screen for the entirety of an online experiment, in the comfort of their own home, may feel more unmotivated and get easily distracted, as it may feel less personal and purposeful than completing a study in-person. Participant dropout rates are higher online compared to offline (Yetano & Royo, 2017). In fact, a dropout rate of 20% is not uncommon online (Peer et al., 2021), potentially due to a lack of social pressure to complete a study and type of motivation (Jun et al., 2017).

Expensive hardware may also be used to compensate for cognitive and perceptual differences to a certain extent. For example, without the appropriate screen calibration checks in place, a larger screen size may allow for faster and better detection of a stimulus in a reaction time task. Additionally, participants' computers may have varying processing capacities which could come to influence millisecond timing accuracy in brief stimulus presentation tasks.

The increased anonymity provided by online experiments gives participants a certain level of protection against fraudulent behaviour, such as claiming they meet the eligibility criteria when they fail to do so in reality. The temptation to pay participants non-adequately may also be greater in the online realm, given the lack of physical and social contact with participants. Additionally, if a bug is present, then much data needs to be discarded, unlike in offline research where participants are tested individually, and the bug could be eliminated without having affected hundreds of datasets in a matter of minutes or hours. Online studies also require both participants and researchers to have access to a computer and stable Internet connection, rendering it more difficult to reach poorer populations. Since not everyone has reliable access to the Internet and technology and is given access to online platforms such as MTurk or Prolific, not everyone may be reached. Lastly, when sampling from common online platforms, there may be a higher likelihood of sampling from non-naïve participants. Individuals registered on common online platforms may already be well-versed in research, as these platforms offer unlimited opportunities for participants to familiarize themselves with common study designs. This concern is heavily task-dependent, and more impactful for social and learning tasks. The

amount of time spent on a platform and whether this is one's main source of income factors into naivety and data quality concerns as well (Peer et al., 2021).

Recruitment strategies and sampling methods also remain important components to increasing an experiment's ecological validity. Many sampling biases in experimental research are often a product of convenience sampling, which involves drawing a sample from a population that is close and easy to reach (Saunders et al., 2019). This type of sampling method is often useful for pilot studies, but psychological research has become increasingly reliant on convenience samples of undergraduate students from western universities, resulting in a misrepresentation of the true population (Masuda et al., 2020). Henrich et al. (2010) grouped these issues with the growing reliance on such samples in psychological research in the descriptive term WEIRD (Westernized, Educated, Industrialized, Rich, and Democratic). These individuals tend to be undergraduate students who are taking courses in psychology or neuroscience and have previously been exposed to psychological research. WEIRD samples in university participant pools allow researchers to conduct their experiments at a low cost and with limited recruitment effort. Students are already familiar with the experimental process and only receive course credit as compensation, which results in an easy and low-cost form of recruitment (Masuda et al., 2020), but places an inherent limitation on the generalisability of results.

The evident convenience associated with WEIRD samples has often left researchers reluctant to explore new ways to expand their sampling efforts. However, online research provides an alternative delivery method that has the potential to counteract this reluctance by allowing for greater data collection for a similar budget. To increase sampling efforts without much effort and recruit a wider variety of participants, one may use one of the many existing platform solutions. If the appropriate sampling method is paired with a complementary delivery system (online vs offline), there is potential for capturing a more heterogenous sample. While online platforms and large-scale initiatives simplify recruitment of larger samples, it is crucial to investigate who are the standard users of platforms, such as Amazon's MechanicalTurk (MTurk; www.mturk.com), Prolific Academic (www.prolific.co; Palan & Schitter, 2018), OpenLab (www.open-lab.online), and university participant pools, among others, before recruiting through them (Chandler et al., 2014; Rodd, 2019; for details on implementation and platforms, see Grootswagers, 2020; Sauter et al., 2020). Here, we do not aim to describe populations on these platforms in much detail, as this has been done elsewhere (e.g., Berinsky et al., 2012; Levay et al., 2016; Walters et al., 2018; Woods et al., 2015). Specifically for MTurk, the interested reader can draw back on several analyses of participants and task performance in the context of online tasks (Berinsky et al., 2012; Casler et al., 2013; Chandler et al., 2014; Crump et al., 2013; Goodman et al., 2013; Hauser & Schwarz, 2016; Levay et al., 2016; Mason & Suri, 2012; Paolacci et al., 2010; Shapiro et al., 2013; Sprouse, 2011; Walters et al., 2018). The prevalence of self-reported clinical conditions in these samples matches that seen in the general population, but can also surpass that observed in laboratory studies, making crowdsourcing a viable way to conduct research on clinical populations (Gillan & Daw, 2016; Shapiro et al., 2013; van Stolk-Cooke et al., 2018). Another investigation in the context of political science shows that users of MTurk are more representative of the US population than in-person convenience samples (Berinsky et al., 2012). While MTurk can provide a representative US sample, other platforms such as Prolific allow to collect data from representative samples (US, UK, Germany, etc.) based on census data (stratified by age, sex, and ethnicity). Reports also show that results obtained from more diverse MTurk samples are almost indistinguishable from those collected in laboratory settings, as many well-known laboratory effects replicate (Casler et al., 2013; Crump et al., 2013; Goodman et al., 2013; Sprouse, 2011). This also holds true for Prolific (Sauter et al., 2022). Nonetheless, researchers must be wary of undesired effects. These include a lack of seriousness that can occur when a specific type of user (i.e., younger men) is dominant on a platform (Downs et al., 2010). Additionally, socioeconomic differences may prevent reaching everyone equally, even in industrialized countries. Although this list of potential costs of online experimentation appears lengthy, many of the presented costs have the potential to be mitigated in these studies, given best practice adaptations. We will illustrate these adaptations in a fictional study on dyslexia.

Some of the adaptations are solutions to issues associated with the replication crisis that can be implemented with relative ease. Increased possibility for recruitment is the **most frequently mentioned benefit of online studies** (see benefits listed in Table 1). Online experimentation attracts a wider audience that would otherwise be difficult to reach in-person. This benefit results in the important ability to efficiently collect large amounts of data as a function of greater accessibility brought on by the online realm, particularly in industrialized countries with wide-spread access to technological devices and reliable internet. Though, socioeconomic differences may still affect an equal reach. Online samples, therefore, generally provide higher accessibility to more representative samples. This benefit contributes to reducing the use of selective and small convenience samples and increases statistical power, which has positive implications for the replication crisis.

Another benefit of online studies stems from the recent availability of all-in-one platforms offering experiment building, presentation, recruitment, and distribution capabilities. Platforms that offer integrated

features for building, hosting, and recruiting include Inquisit Web, Labvanced, and Testable (Sauter et al., 2020). These integrated online environments can result in less monetary and time investments on the researchers' end when it comes to data collection. However, the same or more time may need to be dedicated to building and setting up an experiment. Therefore, having a good workflow is important to maximize efficiency, as we will discuss later on. Online platforms also allow for rapid data collection that requires less lab equipment and space. For example, no experimental consumables need to be purchased or costly lab space booked (for discussions, see Grootswagers, 2020; Mathôt & March, 2021; Sauter et al., 2020). The combination of these benefits also provides access for trainees to run their own online experiments earlier on in their careers when funds can be limited.

The absence of the researcher also reduces any social pressures and feelings of obligation to finish a study that may be present in offline research, and increases the potential for data anonymity. Participants may also experience greater autonomy with regards to the time and location of their participation, as there is no dependence on the experimenter and the laboratory's availability. Hence, online research could become advantageous for both single studies and entire research fields.

Although researchers may accept some of the costs to achieve increased sample sizes that may result in increased statistical power, online research is not the one-size-fits-all solution to all problems. Only if study objectives and online methods are aligned, the online experiment can quickly become a useful and trustworthy tool that is able to replicate (Crump et al., 2013; Sauter et al., 2022), extend, and generalise findings from laboratory experiments. Hence, it is indispensable to tailor suggestions for successful online studies to the specific context of a study and its research questions. Otherwise, well-intended generic blueprints bear the potential to be counterproductive by leading the avid trainee astray. For example, the generic recommendation to provide participants with feedback in each trial (Crump et al., 2013) may help to avoid missed trials but risks adding an unintended, confounding learning component to a perceptual decision-making experiment. Thereby, it could fail to address the research question altogether. To facilitate the transition to online experiments, this paper provides suggestions for future online studies, focusing on context-dependent leveraging of the opportunities of online studies in cognitive psychology and neuroscience research.

ONLINE RESEARCH SUGGESTIONS

In this section, we introduce best practice suggestions generally applicable to traditional paradigms in the field of cognitive psychology and demonstrate their application using a fictional online study (for an overview, see Figure 2). These suggestions equally apply to behavioural data forming a part of neuroscience studies, which may be complemented or followed up with neuroscientific methods in an offline setting. While the outlined suggestions cover many aspects involved in online studies and their organisational structure, they are not intended to be all-encompassing.

The proposed fictional (i.e., simulation) study investigates whether adult dyslexia is associated with visuo-spatial working memory deficits (i.e., accuracy and reaction time) in an adapted version of a Sternberg task (Sternberg, 1966) using stimuli from consumer psychology (Henderson & Cote, 1998). Such a research question can be answered with behavioural data collected online and fuses a cognitive research question with the investigation of a hard-to-recruit sample. This fictional example fits the current zeitgeist as its experimental paradigm could also be run in the laboratory, but some aspects, such as recruitment from a specific population, pandemic-related limitations of in-person interaction, and cost efficiency, would benefit crucially from the online delivery method.

At the very beginning, when designing a study, the specific objectives and research questions of the study are important to consider in determining whether running the study online would address the research question(s) appropriately. For example, if physiological measurements, such as those obtained from electroencephalography (EEG), functional magnetic resonance imaging (fMRI) or galvanic skin response devices are a key component of the study, answering questions regarding those data by means of an online study alone would be impossible. If the study also asks a question about behavioural performance, however, the experimenters could decide to run this part online. This online option could work as a pilot for a subsequent lab experiment including physiological and neuroscientific measures. Vice versa, if experimenters get to run the study with a smaller sample size in the lab first, those pilot data collected under controlled circumstances can be used to establish data screening thresholds for the future analysis of the online data. However, it is not recommended to estimate power based on effect sizes from small sample pilot studies, as this could lead to overall biased estimations (Albers & Lakens, 2018). Depending on the task, information from meta-analyses, previous published research, or data simulations may be used as appropriate for power computations. In this way, online and offline studies can be complementary and benefit each other, but carefully considering the research question(s) that can be addressed with each delivery method is key. Such complementary methodology also has the advantage of avoiding a reductionist bias in neuroscience (Krakauer et al., 2017).

WORKFLOW AND ORGANISATION

An efficient online workflow is important for success, which is discussed at length elsewhere (Grootswagers, 2020; Mathôt & March, 2021; Sauter et al., 2020). The hypothetical workflow of the presented simulated experiment would include the creation of our experiment using the Builder or python-based Coder of PsychoPy3 (Peirce et al., 2019) before exporting a preliminary form of the experiment to the experimental platform Pavlovia.org with the click of a button. This transfer translates the Python-based PsychoPy code into JavaScript code and creates a repository on GitLab to host the translated code. The experiment would then be linked to this new repository and Pavlovia at the same time. It could also be updated directly from within a PsychoPy file, if the JavaScript code on GitLab has not yet been modified manually. Should one want to implement custom or advanced aspects, the easily accessible JavaScript code upon which the experiment is based can be modified directly. Other experiment builders are also available, such as Gorilla (Anwyl-Irvine et al., 2020), LabVanced (www.Labvanced.com), InquisitWeb (www.millisecond.com), PsyToolkit (Stoet, 2010, 2017), OpenSesame/OSWeb (Mathôt et al., 2012), Testable (www.testable.org; Sauter et al., 2020), and FindingFive (www.findingfive.com). Here, we do not intend to give specific recommendations. Instead, we encourage researchers to examine the technical capabilities and licensing costs of each platform closely before committing.

Participants may be recruited through a university participant pool (e.g., SONA) or advertisements sharing a link to a questionnaire hosted on an online survey platform, such as Qualtrics (www.qualtrics.com), SurveyMonkey (www.surveymonkey.com) or SoSci Survey (www.soscisurvey.de). Written consent, demographics and other relevant information can be collected online using a questionnaire on Qualtrics. Some software including SONA allows for the automatic generation of ID codes and porting them from one software to the next. In the case of the SONA platform, it generates a random ID code that can be automatically forwarded to Qualtrics and subsequently to Pavlovia. Here, it is important to choose this code carefully for it to work across platforms and coding languages. The Pavlovia experiment can then be opened in the participants' default browser and start the automatic download of the experimental stimuli to run the study locally via the browser. Upon completion of the task, a poststudy questionnaire and a thank you message can also be displayed.

Importantly, to avoid participants spending a long time wrongly doing the task and the researchers having to award credit or pay participants for unusable data, checks should be performed in the background and problems may be presented as error messages before the end of the experiment, and may even lead to automatic experiment abortion. It is crucial to pilot your workflow using multiple machines, browsers, and participants extensively before starting any data collection.

Lastly, having good lab organisation is an often-underrated factor of successful online studies—and in-person experiments alike. Acquiring research assistants, ensuring having access to all software and platforms, implementing a system of on-boarding or training on these platforms and all aspects involved in the study, and keeping track of administrative tasks such as emails, subject IDs, and credits should all be independent of all individuals to mitigate the impact of the absence of one crucial person or researchers moving to different labs. This is especially important considering many younger trainees (e.g., undergraduate thesis students) getting the opportunity to run an online study but moving to a different programme shortly after its completion. Figure 2 summarises our suggestions for best practices in online research.

EXPERIMENTAL DESIGN

To examine the research question in our fictional study, we used a **mixed (within-between-subjects) experimental design**. That is, comparing two memory conditions and two experimental groups. The initial step includes a power analysis. Since we generate data for our fictional example, we take a simulation approach, which evaluates the power to detect a potential effect by means of repeated linear mixed-effects modelling (tutorial available: DeBruine & Barr, 2021). This approach's flexibility allows for running simulated power and sensitivity analyses for various experimental designs including their specific factors. Our assumptions were based on data from a recent working memory dyslexia study (for details, see R markdown script on the OSF; Franzen et al., 2022). For comparison, an a priori power analysis for a traditional repeated measures ANOVA with a within- and between-subject interaction resulted in a required total sample size of 68 participants (two predictors, Cohen's f effect size = 0.225, alpha = 0.05, power = 95%; Faul et al., 2007). We used this number as a stopping criterion for the data simulation process. Online research gains strengths from collecting as much data as possible and effects are often small. However, collecting large amounts of data will highly depend on a trade-off of statistical, resource, and time considerations. In real-life studies, the results of a power analysis are important guiding principles to avoid introducing data collection biases, which have been contributors to the replication crisis.

Incorporating a **within-subjects comparison of conditions** has been recommended as a safeguard to decrease both participant and technical noise by making the data less dependent on differences between participant set-ups (i.e., excluding technical noise; Anwyl-Irvine et al., 2021; Bridges et al., 2020; Pronk et al., 2020; Rodd, 2019). Our design included a three-item working

memory condition, which could serve as a familiar reference condition in the analysis by calculating the difference between the target and reference condition. A between-participant component could then investigate potential differences of two groups of independent participants between relevant conditions.

Specific to our fictional study design is a sample comprised of **two participant groups** that are captured by a **between-subjects factor**. Like in an ideal scenario, the simulated groups were of equal size. Here, when recruiting from hard-to-find populations, online studies can be particularly beneficial by providing locally unbound access to more potential participants. We would expect to be able to collect a much larger sample by conducting the study online rather than in the laboratory, especially assuming circumstances surrounding data collection during a pandemic.

However, **participants on online platforms may be non-naïve** resulting in a form of dependency of data points, as some participants are more likely to take part in similar tasks (Chandler et al., 2014; Chandler & Paolacci, 2017; Meyers et al., 2020; for an empirical evaluation, see Woike, 2019). To circumvent issues arising from certain task types being frequently offered on a given platform (e.g., economic game paradigms), researchers might opt for less traditional paradigms. An alternative could also be to switch the stimulus type of a traditional paradigm if this fits with the proposed research question and objectives. For example, an n-back task could be presented using letters or symbols rather than numbers. The issue of naivety may be less prominent for speeded reaction time tasks, as having a previous understanding of the instructions, may not result in improved performance if the stimuli are different. This point would be especially relevant for implicit learning tasks, where the absence of experience is critical to the paradigm. Lastly, avoiding a predominantly western sample may still be difficult, as about ⅔ of the recently active users on Prolific Academic were born in the UK, US, Germany or Canada (accessed 28/07/2021) but efforts in this direction should be made regardless.

All inclusion/exclusion criteria should be established prior to data analysis to avoid introducing biases (Chandler et al., 2014; Munafò et al., 2017; Nosek et al., 2018), checking the amount of time participants spent on reading instructions, using instructional manipulation/comprehension checks (Oppenheimer et al., 2009), integrating timed attention/catch trials to weed out inattentive participants, and analysing data that was flagged for exclusion separately or including this fact as a moderator variable in the analysis (Chandler et al., 2014). These suggestions can be implemented for experimental research and questionnaires alike. Ideally, a study would perform participant screening and exclusion before collecting actual experimental data by using one of the many tools built into the aforementioned recruitment platforms. For example, catch trials would then represent a complementary check of the data. These attention checks are meant to determine whether participants are paying attention and completing the study in accordance with the instructions, and could mitigate the lack of motivation and attention present in the online realm. In fact, probing participants' attention using catch trials is generally a useful strategy, as results revealed that inattention may be even larger in the lab. One study found that 61% of participants failed attention checks in the lab compared to only 5% of MTurk participants (Hauser & Schwarz, 2016). We also suggest testing participants in small batches, as an alternative, to prevent having to discard large amounts of data at once, if a bug is present in the code. In between batches, a pre-registered script assessing the data quality and checking for bugs only should be run.

An ever-existing risk in an online environment is the **potential for participant fraud**—particularly when recruiting rare populations (Chandler & Paolacci, 2017). While fraudulent behaviour cannot be entirely excluded, it is important to verify as much as possible that participants really are who they say they are and are prevented from participating repeatedly (Rodd, 2019). Here, recruiting through a trustworthy official platform is likely to mitigate potential fraud due to in-built screening and monitoring tools. Besides demographics, system information, and ID numbers, one may also include checks throughout the study by asking about similar information several times using different wording. As an example, to ensure that a participant is truly an individual with dyslexia, they may be asked "have you been officially diagnosed with dyslexia?". However, later in the study they could also be asked "have you experienced much difficulty with reading or spelling at some point during your life?". The responses to these questions may help solidify the participant's identity as an individual with dyslexia.

As in previous studies, we suggest **safeguarding against including participants in the control group who might be part of a special population**. For instance, participants who are suspected to encounter dyslexia-related issues can be filtered by means of the Adult Dyslexia Checklist (Franzen, 2018; Franzen et al., 2021; Smythe & Everatt, 2001; Stark et al., 2022). This checklist is a self-report questionnaire that can be easily administered and scored online without much effort on the experimenters' side. Once implemented in the software of choice, it can easily be reused in multiple studies using the same software. This checklist adds another level of safety when recruiting participants fitting specific group in-/exclusion criteria online. We used a conservative cut-off score of 40 as an upper cut-off and exclusion criterion from the control group (Figure 1b). Online studies could also be run in two parts, the first part for screening participants, and the second only for those with valid information that meets the eligibility criteria.

We simulated accuracy and reaction times as dependent performance measures from individuals with and without dyslexia. The **use of a speeded task** in an online design is recommended, as fast reaction times may avoid cheating and serve as a supplementary screening measure by recording hints of a lack of attention (Rodd, 2019). We implemented sensible but challenging time limits, particularly for the memory decision (i.e., 3 seconds; Figure 1a). These are fast enough to keep participants attentive and discourage "screen grabbing" on memory tasks, while not being too fast either making the task too difficult or to avoid participants dropping out. Researchers may also review the relevant literature and pilot ahead of time to determine an appropriate response deadline, depending on the testing population. For example, while A/B perceptual decision-making frequently uses a 1.5 second response deadline, the deadline for the same task has been extended to 3 seconds for patients with psychosis. Participants were required to use a physical keyboard and button responses to avoid cursor moving times or touch screen inaccuracies to confound our results. Nevertheless, using this set-up, one has to keep in mind the standard polling rate of keyboard devices of 125 Hz (i.e., one sample every 8 ms) when interpreting the precision of logged reaction times (Anwyl-Irvine et al., 2020).

Equally, the monitor's frame rate has important implications for the **stimulus presentation duration** as it constrains the rate at which stimuli can be presented, and in turn their presentation duration. We suggest recording the frame rate of all participants to get an idea of potential stimulus presentation timing differences. 60 frames per second (also termed Hz) is the standard for most laptop

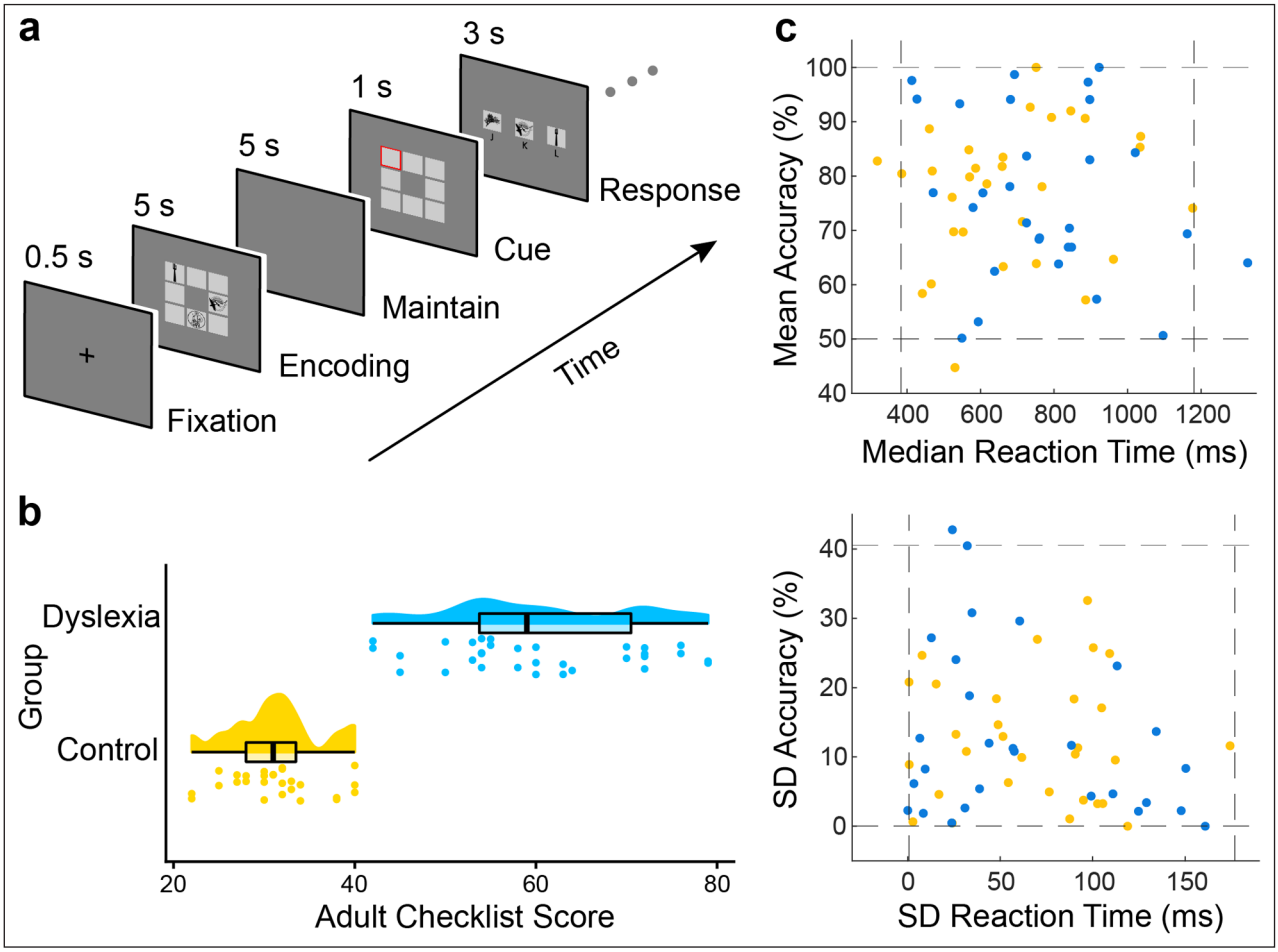

Figure 1 Experimental design, dyslexia scores, and data screening visualisations. a) Schematic of the fictional paradigm and trial sequence. First, participants saw an encoding period in which they encoded the location of various stimuli of the same type presented in different locations. A total of eight fixed locations were available on each trial and either three or six locations were filled with items. Then, a spatial retrieval cue was followed by a decision screen presenting three different stimuli of the same type. Participants were instructed to respond as quickly as possible using their physical keyboard. Further experimental details are available from the study's Open Science Framework repository (Franzen et al., 2022). **b)** Raincloud plots (Allen et al., 2019) of the simulated dyslexia checklist scores that served as screening tool after the removal of excluded participants. Dyslexia data is depicted in blue colour, while the yellow colour indicates data of the control group. Overlaid boxplots show the median, upper and lower quartile. A maximum score of 40 was used to delineate between participants included in the control group and others without an official dyslexia diagnosis who were excluded from further analyses and this plot. **c)** Scatterplots of accuracy as a function of reaction time across all conditions (top: measures of central tendency; bottom standard deviations). One mean value per participant computed across mean accuracy or median reaction times of both working memory conditions. Colours indicate groups. Blue dots depict single-participant values of the dyslexia group, whereas yellow dots depict values of the control group. Dashed lines indicate the lower and upper bounds of the 95% confidence interval for all participants included in the analyses.

screens and laboratories, which means that a new stimulus can be presented every 16.6 ms and its multiples. Most platforms exhibit a positive delay, whereby they present a stimulus longer than intended (Anwyl-Irvine et al., 2021). Therefore, achieving exact presentation of short stimulus durations of 50 ms or less is difficult to impossible to guarantee online and should be avoided for this reason (Grootswagers, 2020). However, timing concerns for longer stimulus presentation durations on modern browsers, and with the optimal operating system/browser pairing, have been alleviated in recent years (Anwyl-Irvine et al., 2021; Bridges et al., 2020; Gallant & Libben, 2019; Mathôt & March, 2021; Pronk et al., 2020). Timing accuracy of visual displays and reaction times tend to be best when chromium-based browsers are being used, independent of the operating system (e.g., Google Chrome; Anwyl-Irvine et al., 2021). But substantial differences between the browser/presentation software/operating system combinations exist and need to be considered in the study design (Bridges et al., 2020).

Generally, **timing accuracy** in online studies is affected by the variety of set-ups and varies slightly more than in the lab across all measurements (for more details, see Bridges et al., 2020). For example, the accuracy of reaction times may be adversely affected if certain operating system/browser pairings are used (for details, see Anwyl-Irvine et al., 2021; Bridges et al., 2020). Laptops offer more recording precision and are preferable over touch screen devices. In contrast, research questions less reliant on speeded answers or interested in using push notifications may alternatively opt to use smart phones and tablet devices. Thus, we suggest recording all possible timings regardless of the allowed testing devices to check whether participants stayed attentive and on task.

Counterbalancing can be applied, as normal experimental design principles become even more important online to avoid even small order effects becoming relevant due to larger sample sizes (DePuy & Berger, 2014). Optimally, a Latin square can be used to determine the counterbalancing order of conditions. Counterbalancing of the block or trial order could also be performed using two or more separate versions of the experiment to avoid unequal dropout rates imbalancing groups during automatic, participant-id-dependent assignment. Some platforms offer built-in quota management or randomisation modes. An alternative method to structured counterbalancing would be full randomisation of stimuli. Here, the order of blocks and trials within blocks gets randomised to mitigate the impact of their order altogether. Counterbalancing and randomisation are supposed to be common practice in experimental psychology online and in the laboratory (Kingdom & Prins, 2016). Particularly, when participants can be randomly assigned to two groups independent of their characteristics in between-subject designs (Altman & Bland, 1999).

To **equalise visual stimuli appearance** across a variety of screens, we suggest adapting it based on a box that the participant needs to manually adapt to the size of a common reference object (e.g., a credit card) at the start of the experiment. This method has been implemented for some of the platforms including Psychtoolbox (Psychtoolbox Team, 2021), PsychoPy/PsychoJS/Pavlovia (Carter, 2021), Gorilla (Gorilla Team, 2021), and Labvanced (Labvanced Team, 2022). Additionally, we did not allow participants to use their phones or tablets to complete the study. Most online delivery platforms have the option to select the specific types of devices allowed for a particular study before publishing it. This may add additional safety regarding consistent stimulus presentation. A mixture of screen size, refresh rate, and resolution may serve as quick a posteriori proxy of stimulus size effects and help to rule these hardware parameters out as potential confounds of performance. Screen size and resolution are easily accessible measures that most platforms provide, but other options could also be used. Researchers can correlate these proxies with behavioural performance to get an indication. The combination of all these proxies (common reference objects, screen size, screen refresh rate, and screen resolution) would be most informative.

To ensure full attention on the task and reduce the likelihood of distractions, we suggest **forcing the browser in full screen mode**. Researchers may consider pausing the experiment as soon as exiting the full screen mode is detected by the software. This mode reduces, but does not eliminate, the possibility of distractions, since notifications often need to be disabled manually and extended or second monitors are not affected by the full screen mode.

The absent possibility for the experimenter to intervene in a mass online testing setting calls for improved study designs that make experiments "dummy proof". This places particular emphasis on the use of **clear and concise instructions and substantial practice**. Instructions may include a step-by-step guide. Their minimum presentation time should be fixed to increase the likelihood of them being read by the participant without simply moving on. Visual instructions may be accompanied by a read-out audio file for increased comprehension and accessibility. However, this supplementary audio would require participants to turn up the volume on their device. A technical check needs to be implemented by the experimenter. Headphone checks could also be implemented to ensure participants are wearing them when the task contains auditory material. In turn, this addition may result in more time costs during the implementation but may be worth the trade-off depending on the tested population. For example, individuals with dyslexia are likely to benefit from this audio-visual presentation format, while fast readers may be distracted by a slower audio. Researchers can also

consider instructional manipulation checks, which follow a format similar to other questions or instructions in the study but ask participants to click on a non-intuitive part of the display (Oppenheimer et al., 2009). If time allows, participants may be quizzed about the nature of the study after reading the instructions (Woods et al., 2015) and a post study questionnaire can be used as well.

Subjecting participants to **practice trials with intuitive feedback** at the beginning of the experiment to ensure proper understanding of the task requirements is already standard in behavioural studies. However, its exact implementation becomes crucial online. This practice should consist of at least three trials per experimental condition (a first encounter and two practice trials for performance evaluation). Depending on the study, the practice could also include many more trials and a staircase procedure for achieving a first performance plateau. Participants should also receive feedback on whether their response was correct, incorrect, or too slow (in a speeded task). Intuitive colouring of the feedback itself, such as correct in green, incorrect in red, and too slow in blue may facilitate learning the task. Experimenters should consider repeating the practice, if a valid response (correct or incorrect) was given in fewer than 51% of the trials for performance evaluation. This general threshold can be applied to any paradigm measuring accuracy. A threshold level of 51% allows for the distinction between mere chance level of responding yet is not too high as to eliminate participants who have understood the instructions but are low performing (51–70%). These thresholds are subject to change based on a study's specific paradigm, its experimental conditions, and prior evidence in the literature, as other researchers set their accuracy threshold at 85% for instance (Sauter et al., 2022). Participants could also be allowed to repeat the practice multiple times if they want to or feel that they need to, however, a limit of repetitions (3–5 times) should be imposed to reduce the possibility of confounding practice effects. These effects can occur quickly (Bartels et al., 2010; Benedict & Zgaljardic, 1998; Calamia et al., 2012). If participants are allowed to repeat the practice as often as they wish, repeating the practice could be more tied to personality type or motivation, as it is conceivable that rather unconfident or highly perfectionistic individuals opt for repeating the practice more often than others, even though they understand the task perfectly well. If a participant is systematically failing the practice (i.e., > 3–5 times) and has failed both attempts at the comprehension checks prior to starting the practice, they should be prevented from continuing the study.

In our fictional study, we designated an **overall study duration** of 30 minutes as ideal and 45 minutes as maximum; to reduce dropout rates based on the assumption that participants are at higher risk of distraction and have shorter attention spans in an online setting. Nonetheless, 45 minutes can be considered rather long, as 70% of respondents in a recent survey indicated to prefer studies with a duration of less than 30 minutes (Sauter et al., 2020). This study duration refers to a single session. If the experiment were to consist of multiple sessions, separated by days or hours, each session would ideally be a maximum of 45 minutes or much less. Depending on the total number of sessions, to limit fatigue effects in a particular session, the duration of each session should be equally long. With increasing duration of an experiment, the dropout rate increases, as participants become less attentive, motivated, and occasionally discouraged. However, increasing the monetary incentive structure to a pay slightly above minimum wage only mitigates this issue somewhat (Crump et al., 2013). Researchers could also consider implementing a game-style reward structure to mitigate participant dropout. An example would be to give block-by-block feedback, which may motivate participants to continue the experiment and track their performance along the way. Nevertheless, shorter studies have increasing appeal for participants (Sauter et al., 2020) and should in turn benefit researchers.

When it comes to keeping it short, researchers always face a context-dependent trade-off between statistical power per condition, often represented by the trial count, and the length and complexity of the experiment. For instance, to perform modelling of reaction times with linear mixed-effects models, one requires enough sampling units (i.e., participants and trials) to increase the likelihood of these models' conversion and thereby credible results. Although devising rules of thumb is difficult, 40 participants and 40 different stimuli are considered a good starting point in cognitive psychology (Meteyard & Davies, 2020). Importantly, the sampling unit count depends mainly on the research design (e.g., within or between-subject/item factors) and researchers' planned analyses. Here, we emphasise that this is highly specific to each study and a simulation approach is recommended to determine the power and number of sampling units for a specific design and all its factors.

In terms of **procedural design recommendations** for online research, behaviours leading to irreproducible research (Button et al., 2013; Ioannidis, 2005, 2008; Ioannidis et al., 2014; Munafò et al., 2017; Simmons et al., 2011) can be partly avoided by pre-registering an unmodifiable analysis plan consisting of research questions, hypotheses, exclusion criteria, a priori power calculations, and target sample size on the Open Science Framework for instance (OSF; www.osf.io; Foster & Deardorff, 2017; Nosek, 2015; Nosek et al., 2018). This does not mean that the door for exploratory or unplanned analyses is closed, which could be added as an "unplanned protocol deviations" section in the final report (Camerer et al., 2018). Pre-registration templates exist (van den Akker et al., 2021) and data, analyses, and materials can also be made easily accessible and

reproducible by sharing these in hybrid formats, such as markdown (Artner et al., 2020). These formats merge code, results, and text. They can be stored via cloud-based solutions including the OSF, GitHub or Databrary (Gilmore et al., 2018; for a guide on transparent research steps, see Klein et al., 2018). Due to the lack of hypotheses and fictional nature, the presented example simulation study was not pre-registered. However, in a truly experimental setting, all aspects of the fictional study's procedures would have been pre-registered, and authors may even consider submitting for a registered report.

Overall, asking participants to find a quiet environment, exit all non-essential programmes, close all other browser tabs and applications, and empty their browser's cache can all help the study to run smoothly and provide the best possible timing accuracy. Instructions may rather be extensive, and their visual form complemented by an auditory read out. Attention and comprehension checks are recommended. Participants could be guided through a sample trial step-by-step. We also suggest complementing experimental testing with questionnaire items (after the experiment) to collect self-report data on experienced noise and distractions in the participant's environment, on what may have gone wrong, and give participants a chance to provide feedback that is often collected informally in the lab.

DATA SCREENING AND ANALYSIS

Once the aforementioned experimental design considerations have been implemented and data collected, researchers face the challenging task to evaluate whether their data could give rise to valid empirical results. Therefore, the implementation of robust data screening measures is of utmost importance. Since control over a participant's setting, their technical equipment and understanding of the task is more limited online than in the laboratory, increased noise in a dataset should be expected. This gives even more importance to the screening of data collected online as opposed to a laboratory. The data screening procedure has to be able to 1) identify and quantify this noise and 2) lead to clear decisions on whether a dataset should be kept or excluded from all analyses.

When dealing with empirical data, the first data screening step consists of **identifying the number of participants that did not fully partake**. The number of completed experimental trials lends itself as a good proxy to evaluate whether sufficient trials have been recorded for analysis and whether sheer experimental duration might have been the reason for participants dropping out. This threshold should be based on the requirements of the statistical analyses, such as power and sensitivity, and set a priori to avoid issues that have contributed to the replication crisis (Button et al., 2013; Ioannidis, 2005, 2008; Ioannidis et al., 2014; Munafò et al., 2017; Simmons et al., 2011). In our fictional study, we simulated power for a linear mixed-effects model as a function of both the number of participants and stimuli. This simulation allows for determining the critical number of trials (i.e., stimuli) necessary for a reliable minimum of 80% power, given a sample size and assumptions about the effect. This threshold should be determined for every study. In the simulation, it resulted in a sample of 70 participants for detecting a group difference (between-subjects factor) in response times when using 40 or more stimuli. Repeated participation attempts can be prevented using the settings in most experiment manager software. For example, the platform Prolific automatically tracks participants' completed, rejected, and aborted attempts, providing a full report that includes participant ID numbers and session durations. We suggest excluding data of quickly aborted and repeated sessions from all analyses, unless this number is part of the research question.

In a second step, the **double-checking of crucial experimental variables** at independent time points is essential and should be considered good practice. Participants should only be excluded in accordance with hard exclusion criteria and/or repeated failing of attention and/or comprehension checks, which were obtained at different timepoints throughout the experiment. The platform Prolific provides useful recommendations on how to integrate attention and comprehension checks within their platform but can also be regarded as general advice (Prolific Team, 2022). Specifically, they allow two different types of attentional checks. The first type is an **Instructional Manipulation Check (IMC)**, where participants are asked to answer a question in a very specific way, and thus, researchers can determine whether participants were paying attention or not, based on their answer. The second method is to **integrate a nonsensical item** into a survey, where only one objectively logical answer makes sense. These checks may take the form of a multiple-choice question. To ensure participants have understood critical information about the experiment, Prolific also recommends implementing valid comprehension checks right at the beginning of the study alongside the instructions (i.e., before the training). As per Prolific guidelines, participants should be given at least two attempts for comprehension checks, and if they fail both attempts, they should be asked to return to their submission (for more information, see Prolific's Attention and Comprehension Check Policy; Prolific Team, 2022).

Third, one should **clean all trials without a valid response**. For example, in perceptual decision-making all trials without a decision before the limits of the response deadline are commonly discarded (Franzen et al., 2020), because a decision is simply absent and should not be evaluated as incorrect for this reason. However, criteria for the validity of trials may vary by research field and question. In a design where missed responses could be of interest, such as inhibited responses in the go/no-go

task, these trials may be retained for analysis, as they present valuable information for answering the research question.

Additional data screening measures that are specific to the collected behavioural data (e.g., accuracy and reaction times) can be applied. First, as an indicator of the ability to understand and perform the task, we suggest screening the simulated data for **mean accuracy levels** that fall below 50% on the rather easy three-load (low-WM load) condition. As mentioned previously, a threshold of 51% is a soft recommendation that should allow for the differentiation between mere chance and low performance. Specific to the fictional study, a three-digit working memory load should not be a problem for anyone, unless they are affected by a more general cognitive impairment (Vogel et al., 2001), and can thus be used as a baseline. To finish this step, report the number of participants removed due to accuracy concerns.

Next, we suggest **screening reaction times** for trials considered too fast for any valid decision-making, memory or other relevant processes having been completed and indicated via button press(es). For instance, given that the lower bound of object processing of a single object was found to be around 100 ms (Bieniek et al., 2016) and participants saw three choice options, we would remove all trials with reaction times faster than 200 ms. This threshold is highly study specific.

In general, it is good practice to **base data screening decisions**, reliability estimates, and analysis of results **on robust statistical evidence** (Field & Wilcox, 2017; Parsons et al., 2019). When computing averages, robust measures of central tendency, such as the median, trimmed or winsorized means should be considered (Rousselet et al., 2017; Wilcox & Rousselet, 2018). These measures are best suited for continuous variables. We also suggest to exclude participants whose data represents outliers based on robust measures of variation in a univariate or multivariate distribution including the median absolute deviation (MAD) or minimum covariance determinant (MCD; Leys et al., 2019). Similarly, the combination of frequent fast responses without much variation and poor performance suggests mere guessing. This would be indicated by a pattern of fast median reaction time, small median absolute deviation in reaction times, and low mean accuracy (< 5th percentile of the group distributions). Researchers can first explore visually whether such a pattern might exist by plotting mean/SD accuracy vs mean/SD reaction times (Figure 1c).

Since many studies in cognitive psychology aim to generalise results to a population of participants and/or stimuli, we suggest applying multilevel analysis including random effects, as is done in linear mixed-effects modelling (for a tutorial, see DeBruine & Barr, 2021). Alternatively, one may use Bayesian mixed modelling, which allows for quantifying evidence for both the alternative and the null hypothesis (Kruschke, 2014; Kruschke & Liddell, 2018). Modern computing power equally allows for extrapolation from one's data by means of bootstrapping or permutation approaches (i.e., sampling at random with or without replacement, respectively; Efron, 1979; Wilcox & Rousselet, 2018). Researchers should ensure reproducibility of their analysis and avoid inconsistencies by implementing all data screening procedures in custom scripts in open source languages available to everyone such as RStudio (RStudio Team, 2021).

Taken together, in our fictional example, applying these data screening measures resulted in the exclusion of six participants. The remaining data shows that all average values are within reasonable boundaries, with only a few participants from both groups exhibiting more extreme performance (Figure 1c).

DISCUSSION

The opportunities for future online research are manifold and provide some exciting possibilities but the devil is in the detail. A combination of specifically tailored research questions and experimental designs paired with online attention checks and rigorous data screening is required for success.

With increased cost efficiency and feeling of anonymity, online experiments could help in the recruitment of larger and locally independent samples or specific populations and demographics that are otherwise hard to reach. Lower costs would allow undergraduate students to run scientifically meaningful independent research projects, demonstrating the type of access the online delivery method provides to motivated trainees, even in times of in-person lab closures. Particularly in these times, the lack of required physical presence in labs—a benefit of online research—was complimentary to the COVID-19 social distancing measures and facilitated cost efficient access to data. Thereby, it allowed trainees to fulfil their degree requirements.

Increased anonymity may be a particularly important advantage for individuals with learning disabilities who may not feel comfortable disclosing their diagnosis and standing physically present in front of an experimenter who they may even know. Equally, the hurdle to quit an experiment is much reduced by simply closing one's browser tab instead of having to walk out of a laboratory. This is of particular relevance for studies conducted physically in the university context where students participate in their professors' experiments for course credit, and may feel obligated to finish a study due to implicit social pressure associated with this relationship. Hence, especially the recruitment of participants with specific diagnoses (e.g., dyslexia; for psychiatric conditions, see Gillan & Daw, 2016) may benefit widely from shifting experimental research in cognitive psychology and neuroscience online—if done thoughtfully.

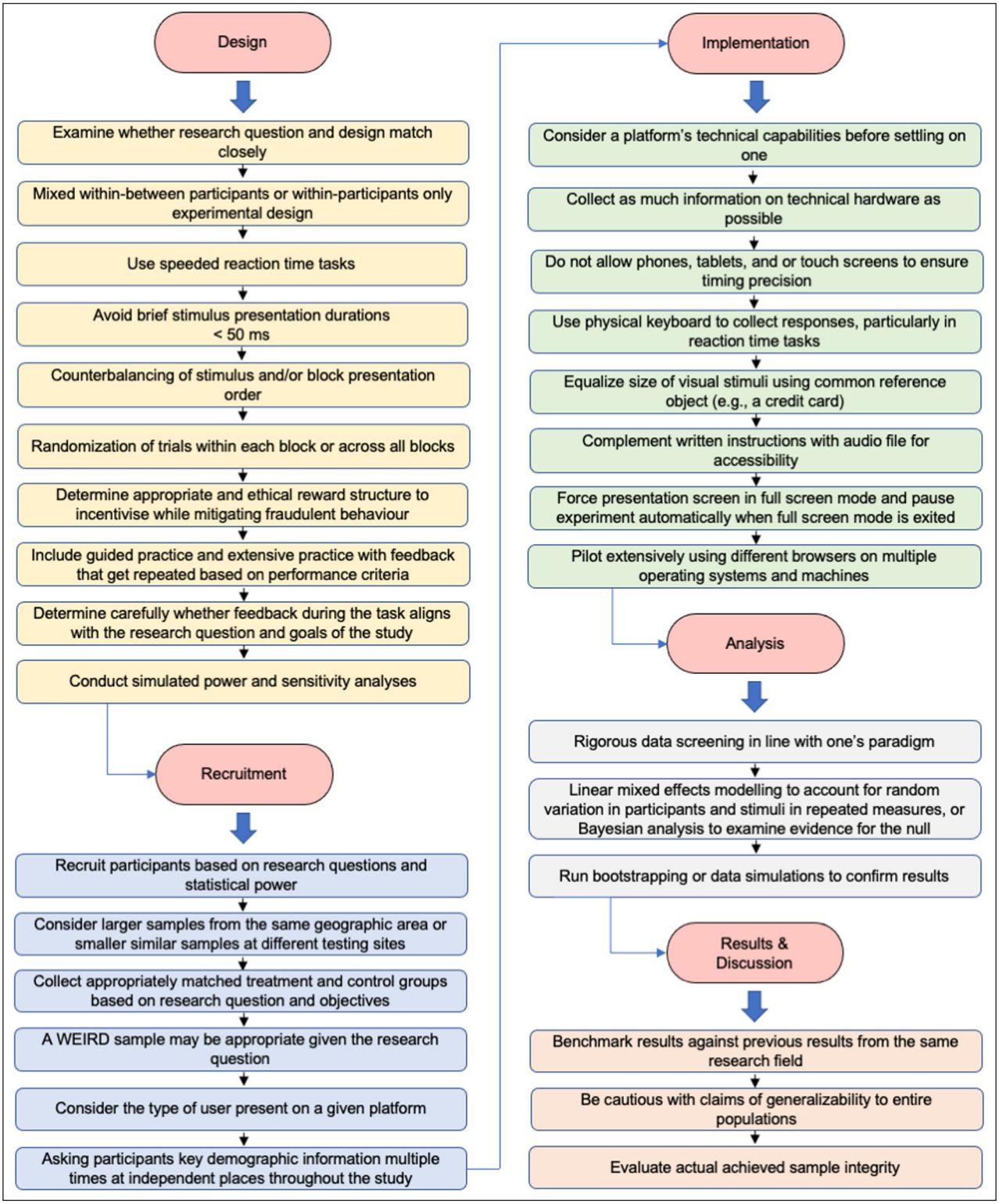

Figure 2 Overview of suggestions for online research by study stage. Flowchart following the workflow of experimental studies in cognitive psychology and neuroscience.

However, increased anonymity comes at the cost of lower experimenter control online. Some level of control could be regained in a proctored setting with online intervention and supervision, such as organising a call where both participant and experimenter are online during testing. The decision to add researcher supervision online would also depend on the type of study design, such as generic paradigms that are built and ready to run independently, as opposed to specific protocol administrations that require additional supervision (i.e., neuropsychological test administration). Additionally, the level of intervention could also be based on the type of population, such as older adults requiring more supervision and assistance in the online realm.

Recruiting populations with dyslexia and other specific diagnoses or traits has a long history of proving difficult,

especially in a university setting. The nature of dyslexia and the predominant focus of the education system on reading and writing mean that individuals with medium to severe cases of dyslexia may not be captured when recruiting at the university level, as many have most likely not progressed to this level (Warnke, 1999). Therefore, generalising lab-based results to the general population of individuals with dyslexia can become problematic, particularly if based on a WEIRD sample. Here, extended sampling of individuals from a variety of socioeconomic backgrounds would be a step towards more generalisability. Collecting a sample diverse in socioeconomic background, age, country of origin, etc. is not automatic and needs to keep the distribution of access to the internet and technology among other variables in mind (Lourenco & Tasimi, 2020).

This issue equally applies to individuals with psychiatric disorders. These individuals would usually not be part of university subject pools and require even better data protection standards. For example, patients with psychotic disorders may be outpatients at a university clinic or affiliated hospital facilitating recruitment through collaborations but requiring pro-active, in-person recruitment due to high levels of distrust. If researchers are not able to establish or benefit from those in-person collaborations and would like to recruit online, extending the research question to other (model) populations/aspects would be a desirable option. In the case of schizophrenia, one such population are individuals on the schizotypy personality spectrum. According to the continuum hypothesis of psychosis, schizotypy can be regarded as a subclinical model of schizophrenia among the normal population (Ettinger et al., 2014; Giakoumaki, 2012), featuring subclinical symptoms of psychosis, but does not need to lead to a clinically diagnosable state (Kwapil & Barrantes-Vidal, 2015; Nelson et al., 2013; Raine et al., 1994; Siddi et al., 2017). Recruiting from the normal (i.e., non-clinical) population comes with the advantages of being able to recruit through standard platforms, assess relevant traits with standardised questionnaires instead of requiring an official diagnosis, and often avoid medication-related confounds. Further, investigating a dimensionally expressed trait in the normal population using a continuum (correlational) approach avoids the need for a well-matched control group.

To recruit hard-to-reach populations, other recruitment attempts could include the use of mailing lists, listservs, online forums or Facebook (sponsored) posts, as has been done for recruiting infant participants (Brouillard & Byers-Heinlein, 2019b, 2019a). Having one's study featured in relevant newspaper or pop-science articles on the topic presents another opportunity. However, as with different platforms, there may be expected differences between samples solely based on the recruitment source and technique. In using these methods, it is important to attempt to recruit outside of the researchers' own social media networks, since such recruitment may not increase the diversity of the sample—leading instead to homophily (Aiello et al., 2012; Sheskin et al., 2020). Some of the popular platforms for managing data collection, such as MTurk or Prolific, provide access to quite large populations from around the world. Most importantly, the characteristics, intentions, and motivations of these samples must align with the study's objectives.

The need for an appropriately matched control group is a crucial aspect of research with specific populations, which intends to compare groups. Its appropriateness is highly context-dependent, as a control group with limited years of education may be matched for age but is likely a mismatch on a cognitive task (e.g., reading age) when compared to a group of individuals with dyslexia taking part in higher education. In these cases, a WEIRD sample (Henrich et al., 2010) may also be appropriate, particularly if sampling of a population taking part in higher education is important for the research question. Another reason might be that the budget for advertising the study is constraint, as is often the case in student research. In these cases, matched WEIRD samples are important for scientific validity of the comparison but care needs to be exercised regarding the generalisability of results. In the case of our fictional study, comparing a diverse dyslexia sample from all walks of life to undergraduate students on a challenging cognitive task would have rendered this group comparison not meaningful. It would have been more beneficial to compare groups similar in certain demographic characteristics, raising awareness for the limitations and skewness in the interpretation of the results. Hence, WEIRD samples need to be well-justified but should not be categorically condemned.

One suggestion for increasing ecological validity, however, is to collect similar, but more, samples in independent locations. In other words, running the identical study at several universities in different cities, provinces or even countries. Thereby, cross-cultural questions could be addressed (Knoeferle et al., 2010; Woods et al., 2013). Selecting specifically matching control groups is equally key for those studies run at multiple sites. Networks facilitating such collaboration exist, such as the Psychology Science Accelerator (www.psysciacc.org; Forscher et al., 2021), and have their roots in the promising open science movement that seeks to counteract the replication crisis by promoting transparency, openness, and replication (Klein et al., 2018; Munafò et al., 2017; Nosek, 2015; Nosek et al., 2012, 2018; Simmons et al., 2011). However, these initiatives can come with a separate set of challenges regarding their organisation and logistics.

Another aspect, under the researcher's control, that can help increase ecological validity by minimising the risk of participant fraud and ethical concerns, is a study's reward structure. The reward structure of an experiment must be in line with payment amounts offered in laboratory settings to be ethical, while not being too large to avoid participants providing false information and/or providing random data simply to gain access to the monetary

reward. Arbitrarily increasing monetary incentives much above the local minimum wage is unlikely to have a positive effect. Compared to payment below minimum wage, it has been shown to affect only the speed at which a sample and data is being collected as well as reducing dropout rates, but importantly, it did not affect experimental results (Crump et al., 2013). In accordance with Woods and colleagues (2015), we suggest an ethical and reasonable rate based on the local minimum wage (e.g., 20 cents EUR per minute). Variable bonus incentives could also be introduced for completing certain aspects or the entire task successfully (Chandler et al., 2014). Lastly, researchers also need to consider the likelihood of potential participants on a platform using participation as their main source of income, as this may affect data quality (Peer et al., 2021).

Besides participant recruitment and sample characteristics, technological capabilities relevant to the experimental design are another main aspect underlying successful online studies. These become much more important in the online realm, as their variability increases with every participant bringing their own set-up to the study. To minimise unwanted and negative effects, researchers are well-advised to include hardware requirements in their experiment description and recruitment filters. Checking whether these have been fulfilled when starting the experiment and enforcing their compliance is a must, as it allows for better standardisation of procedures and results. In this respect, the necessity of extensive piloting of an experiment cannot be emphasised enough. Piloting the workflow should be carried out using multiple machines, browsers, and participants. This should include an evaluation of the instructions and accuracy of technical and other screening checks upon which the experiment should get aborted, if a mismatch is detected. Researchers need to keep in mind that extensive piloting takes time, potentially even more than in the lab.

Very recently, technological innovations have given rise to the possibility of webcam-based eye-tracking. This follows on from the introduction of more and more portable eye-trackers to the market since 2015. The possibility of collecting eye-tracking data in online studies using a laptop's in-built or a USB webcam is a promising prospect and could benefit cognitive research majorly, as eye movements provide a mechanism to examine cognitive and physiological processes simultaneously. Current sampling rates are often restricted to 30 Hz (one data point every ~33 ms), which is still low compared to high-end lab-based eye-tracking systems often achieving 1000 Hz (one data point per millisecond). This maximum rate depends on the specifications of the webcam, processing power of the computer, current computational load (e.g., number of open browser tabs), and the eye-tracking algorithm itself. It is important to keep in mind that the maximum rate may not always be achieved by all systems and at all times, as the algorithm is often run on the participant's local machine. Therefore, it is likely sufficient for experiments expecting longer and more steady fixations of at least 150 ms, as some samples may be skipped or not accurately collected. This type of research is often conducted in consumer psychology or sustained attention paradigms. Hence, the sampling rate remains a caveat that requires careful consideration of its usefulness for a given study.

Accuracy of some webcam-based solutions is estimated to be around 1 degree of visual angle by the developers (GazeRecorder, 2021), comparable to lab-based systems. Occasional recalibration throughout an experiment can be useful, since accuracy may decrease over time (Pronk, 2021). Factors in the participants' environment, such as the general lighting conditions, can affect tracking performance as well. As accuracy often decreases towards the edges of a display, focusing the stimulus presentation around the centre of the display would be good practice. At the time of writing, webcam-based eye-tracking is available on platforms including PsychoPy/Pavlovia, Labvanced, and Gorilla. Most use the WebGazer.js library (Papoutsaki et al., 2016) and require a mouse/cursor to perform the calibration. Taken together, if used purposefully, webcam-based eye-tracking and automated video analysis have great potential for adult and developmental research (Chouinard et al., 2019), as one of the first physiological measures to be reliably collected online.

CONCLUSION

As the transition to online research proves difficult to achieve based on a simple blueprint, the presented suggestions aim to provide a starting point. They are intended to guide one's critical thinking about experimental design considerations without claiming to be all-encompassing. Crucial questions as one begins to design any online study are: What is the goal of my study, is the online delivery method appropriate and sufficient, and are all the measures needed to answer my research question accurately collectable online? It is also important to consider worst-case scenarios with regards to the experimental design, participants, and technology, and to think of ways to mitigate these issues beforehand. Our fictional study illustrates these suggestions in practice. Often the benefits outweigh the costs, as the future of research is heading towards technological innovation, and the COVID-19 pandemic offered many a first opportunity at trying to leverage the benefits of online research. As increased environmental changes and biological hazards may result in an uncertain future with regards to global pandemics (Beyer et al., 2021), making the transition to online experimentation sooner rather than later, could prove to be more advantageous for research teams in many different settings and research fields in the long run. Thus, whether more researchers adopt this method should simply be a matter of time. The key factor is how it is being done.

TRANSPARENCY STATEMENT

Regarding the presented fictional example study, we reported how we determined the sample size and stopping criterion. We reported all experimental conditions and variables. We reported all data exclusion and outlier criteria and whether these were determined before or during the data analysis.

DATA ACCESSIBILITY STATEMENT

Data and code is accessible from the experimental study's Open Science Framework repository: https://osf.io/dyn5t/ (doi: 10.17605/OSF.IO/DYN5T).

ACKNOWLEDGEMENTS

We would like to thank Bianca Grohmann, Malte Wöstmann, Aaron P. Johnson, and Jonas Obleser for their feedback on earlier drafts of this manuscript. Further, we wish to acknowledge all the suggestions and feedback that were collected in response to a tweet by LF on the topic.

FUNDING INFORMATION

This research was supported by a Fonds de Recherche du Québec – Société et Culture team grant (196369), and a Horizon Postdoctoral Fellowship awarded to LF by Concordia University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Both authors contributed equally to most stages of this research, while Léon Franzen also acted as senior, supervising author. Specifically, authors contributed in the following way: Conceptualization: Nathan Gagné and Léon Franzen. Data curation: Nathan Gagné and Léon Franzen. Formal analysis: Nathan Gagné and Léon Franzen. Funding acquisition: Léon Franzen. Investigation: Nathan Gagné and Léon Franzen. Methodology: Nathan Gagné and Léon Franzen. Project administration: Léon Franzen. Resources: Léon Franzen. Software: Nathan Gagné and Léon Franzen. Supervision: Léon Franzen. Validation: Nathan Gagné and Léon Franzen. Visualization: Nathan Gagné and Léon Franzen. Writing – original draft: Nathan Gagné and Léon Franzen. Writing – review & editing: Nathan Gagné and Léon Franzen.

AUTHOR AFFILIATIONS

Nathan Gagné orcid.org/0000-0003-2232-9408
Concordia University, Montreal, CA
Léon Franzen orcid.org/0000-0003-2277-5408
University of Lübeck, DE

REFERENCES

Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., & **Menczer, F.** (2012). Friendship prediction and homophily in social media. *ACM Transactions on the Web*, *6*(2). DOI: https://doi.org/10.1145/2180861.2180866

Albers, C., & **Lakens, D.** (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, *74*, 187–195. DOI: https://doi.org/10.1016/j.jesp.2017.09.004

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., & **Kievit, R.** (2019). Raincloud plots: a multi-platform tool for robust data visualization [version 1; peer review: 2 approved]. *Wellcome Open Res*, *4*(63). DOI: https://doi.org/10.12688/wellcomeopenres.15191.1

Altman, D. G., & **Bland, J. M.** (1999). Treatment allocation in controlled trials: why randomise? *BMJ*, *318*(7192), 1209–1209. DOI: https://doi.org/10.1136/bmj.318.7192.1209

Anwyl-Irvine, A. L., Dalmaijer, E. S., Hodges, N., & **Evershed, J. K.** (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*, *53*(4), 1407–1425. DOI: 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. DOI: https://doi.org/10.3758/s13428-019-01237-x

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & **Vanpaemel, W.** (2020). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*, *26*(5), 527–546. DOI: https://doi.org/10.1037/met0000365

Baker, M., & **Penny, D.** (2016). Is there a reproducibility crisis in science? *Nature*, 452–454. DOI: https://doi.org/10.1038/d41586-019-00067-3

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., & **Ehrenreich, H.** (2010). Practice effects in healthy adults: A longitudinal study on frequent repetitive cognitive testing. *BMC Neuroscience*, *11*. DOI: https://doi.org/10.1186/1471-2202-11-118

Benedict, R. H. B., & **Zgaljardic, D. J.** (1998). Practice effects during repeated administrations of memory tests with

and without alternate forms. *Journal of Clinical and Experimental Neuropsychology*, *20*(3), 339–352. DOI: https://doi.org/10.1076/jcen.20.3.339.822

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. DOI: https://doi.org/10.1093/pan/mpr057

Beyer, R. M., Manica, A., & **Mora, C.** (2021). Shifts in global bat diversity suggest a possible role of climate change in the emergence of SARS-CoV-1 and SARS-CoV-2. *Science of the Total Environment*, *767*, 145413. DOI: https://doi.org/10.1016/j.scitotenv.2021.145413

Bieniek, M. M., Bennett, P. J., Sekuler, A. B., & **Rousselet, G. A.** (2016). A robust and representative lower bound on object processing speed in humans. *European Journal of Neuroscience*, *44*(2), 1804–1814. DOI: https://doi.org/10.1111/ejn.13100

Blythe, H. I., Kirkby, J. A., & **Liversedge, S. P.** (2018). Comments on: "What is developmental dyslexia?" brain sci. 2018, 8, 26. the relationship between eye movements and reading difficulties. *Brain Sciences*, *8*(6). DOI: https://doi.org/10.3390/brainsci8060100

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). DOI: https://doi.org/10.7717/peerj.9414

Brouillard, M., & **Byers-Heinlein, K.** (2019a). *Recruiting hard-to-find participants using Facebook sponsored posts*. DOI: https://doi.org/10.17605/OSF.IO/9BCKN

Brouillard, M., & **Byers-Heinlein, K.** (2019b). *Recruiting infant participants using Facebook sponsored posts*. DOI: https://doi.org/10.17605/OSF.IO/9BCKN

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & **Munafò, M. R.** (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. DOI: https://doi.org/10.1038/nrn3475

Calamia, M., Markon, K., & **Tranel, D.** (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *Clinical Neuropsychologist*, *26*(4), 543–570. DOI: https://doi.org/10.1080/13854046.2012.680913

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. DOI: https://doi.org/10.1038/s41562-018-0399-z

Carter, W. L. (2021). *ScreenScale*. DOI: https://doi.org/10.17605/OSF.IO/8FHQK

Casler, K., Bickel, L., & **Hackett, E.** (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156–2160. DOI: https://doi.org/10.1016/j.chb.2013.05.009

Chandler, J. J., Mueller, P., & **Paolacci, G.** (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130. DOI: https://doi.org/10.3758/s13428-013-0365-7

Chandler, J. J., & **Paolacci, G.** (2017). Lie for a Dime: When Most Prescreening Responses Are Honest but Most Study Participants Are Impostors. *Social Psychological and Personality Science*, *8*(5), 500–508. DOI: https://doi.org/10.1177/1948550617698203

Chouinard, B., Scott, K., & **Cusack, R.** (2019). Using automatic face analysis to score infant behaviour from video collected online. *Infant Behavior and Development*, *54*, 1–12. DOI: https://doi.org/10.1016/j.infbeh.2018.11.004

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

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. DOI: https://doi.org/10.3758/s13428-015-0567-2

DeBruine, L. M., & **Barr, D. J.** (2021). Understanding Mixed-Effects Models Through Data Simulation. *Advances in Methods and Practices in Psychological Science*, *4*(1). DOI: https://doi.org/10.1177/2515245920965119

DePuy, V., & **Berger, V. W.** (2014). Counterbalancing. *Wiley StatRef: Statistics Reference Online*. DOI: https://doi.org/https://doi.org/10.1002/9781118445112.stat06195

Downs, J. S., Holbrook, M. B., Sheng, S., & **Cranor, L. F.** (2010). Are your participants gaming the system? Screening mechanical Turk workers. *Conference on Human Factors in Computing Systems – Proceedings*, *4*, 2399–2402. DOI: https://doi.org/10.1145/1753326.1753688

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, *7*(1), 1–26. DOI: https://doi.org/10.1214/aos/1176344552

Ettinger, U., Meyhöfer, I., Steffens, M., Wagner, M., & **Koutsouleris, N.** (2014). Genetics, Cognition, and Neurobiology of Schizotypal Personality: A Review of the Overlap with Schizophrenia. *Frontiers in Psychiatry*, *5*, 1–16. DOI: https://doi.org/10.3389/fpsyt.2014.00018

Etz, A., & **Vandekerckhove, J.** (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE*, *11*(2), 1–12. DOI: https://doi.org/10.1371/journal.pone.0149794

Faul, F., Erdefelder, E., Lang, A.-G., & **Buchner, A.** (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Journal of Materials and Environmental Science*, *39*(2), 175–191. DOI: https://doi.org/10.3758/BF03193146

Field, A. P., & **Wilcox, R. R.** (2017). Robust statistical methods: A primer for clinical psychology and experimental

psychopathology researchers. *Behaviour Research and Therapy*, *98*, 19–38. DOI: https://doi.org/10.1016/j.brat.2017.05.013

Forscher, P., Wagenmakers, E.-J., Coles, N. A., Silan, M. A., Dutra, N. B., Basnight-Brown, D., & **IJzerman, H.** (2021). *A Manifesto for Big Team Science*.

Foster, E. D., & **Deardorff, A.** (2017). Open Science Framework (OSF). *Journal of the Medical Library Association*, *105*(2). DOI: https://doi.org/10.5195/jmla.2017.88

Franzen, L. (2018). *Neural and visual correlates of perceptual decision making in adult dyslexia* (Issue December) [University of Glasgow]. https://theses.gla.ac.uk/71950/

Franzen, L., Delis, I., Sousa, G. De, Kayser, C., & **Philiastides, M. G.** (2020). Auditory information enhances post-sensory visual evidence during rapid multisensory decision-making. *Nature Communications*, *11*, 5440. DOI: https://doi.org/10.1038/s41467-020-19306-7

Franzen, L., Gagné, N., Johnson, A. P., & **Grohmann, B.** (2022). *Behavioral markers of visuo-spatial working memory load in adult dyslexia*. Open Science Framework. DOI: https://doi.org/10.17605/OSF.IO/DYN5T

Franzen, L., Stark, Z., & **Johnson, A. P.** (2021). Individuals with dyslexia use a different visual sampling strategy to read text. *Scientific Reports*, *11*, 6449. DOI: https://doi.org/10.1038/s41598-021-84945-9

Gallant, J., & **Libben, G.** (2019). No lab, no problem: Designing lexical comprehension and production experiments using PsychoPy3. *The Mental Lexicon*, *14*(1), 152–168. DOI: https://doi.org/10.1075/ml.00002.gal

GazeRecorder. (2021). *Gaze flow*. https://gazerecorder.com/gazeflow/.

Giakoumaki, S. G. (2012). Cognitive and prepulse inhibition deficits in psychometrically high schizotypal subjects in the general population: Relevance to schizophrenia research. *Journal of the International Neuropsychological Society*, *18*(4), 643–656. DOI: https://doi.org/10.1017/S135561771200029X

Gillan, C. M., & **Daw, N. D.** (2016). Taking Psychiatry Research Online. *Neuron*, *91*(1), 19–23. DOI: https://doi.org/10.1016/j.neuron.2016.06.002

Gilmore, R. O., Kennedy, J. L., & **Adolph, K. E.** (2018). Practical Solutions for Sharing Data and Materials From Psychological Research. *Advances in Methods and Practices in Psychological Science*, *1*(1), 121–130. DOI: https://doi.org/10.1177/2515245917746500

Goodman, J. K., Cryder, C. E., & **Cheema, A.** (2013). Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224. DOI: https://doi.org/10.1002/bdm.1753

Gorilla Team. (2021). *Gorilla Screen Calibration*. https://support.gorilla.sc/support/reference/task-builder-zones#eyetracking

Grootswagers, T. (2020). A primer on running human behavioural experiments online. *Behavior Research Methods*, *52*(6), 2283–2286. DOI: https://doi.org/10.3758/s13428-020-01395-3

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. DOI: https://doi.org/10.3758/s13428-015-0578-z

Henderson, P. W., & **Cote, J. A.** (1998). Guidelines for selecting or modifying logos. *Journal of Marketing*, *62*(2), 14–30. DOI: https://doi.org/10.1177/002224299806200202

Henrich, J., Heine, S. J., & **Norenzayan, A.** (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. DOI: https://doi.org/10.1017/S0140525X0999152X

Holcombe, A. (2020). *The reproducibility crisis*. DOI: https://doi.org/osf.io/r4wpt/

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. DOI: https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648. DOI: https://doi.org/10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & **David, S. P.** (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*(5), 235–241. DOI: https://doi.org/10.1016/j.tics.2014.02.010

John, L. K., Loewenstein, G., & **Prelec, D.** (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. DOI: https://doi.org/10.1177/0956797611430953

Jun, E., Hsieh, G., & **Reinecke, K.** (2017). Types of motivation affect study selection, attention, and dropouts in online experiments. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 1–15. DOI: https://doi.org/10.1145/3134691

Kingdom, F., & **Prins, N.** (2016). *Psychophysics* (2nd ed.). Academic Press.

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Mohr, A. H., Jzerman, H. I., Nilsonne, G., Vanpaemel, W., & **Frank, M. C.** (2018). A practical guide for transparency in psychological science. *Collabra: Psychology*, *4*(1), 1–15. DOI: https://doi.org/10.1525/collabra.158

Knoeferle, K. M., Woods, A., Käppler, F., & **Spence, C.** (2010). That Sounds Sweet: Using Cross-Modal Correspondences to Communicate Gustatory Attributes. *Psychology & Marketing*, *30*(6), 461–469. DOI: https://doi.org/10.1002/mar

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & **Poeppel, D.** (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron*, *93*(3), 480–490. DOI: https://doi.org/10.1016/j.neuron.2016.12.041

Kruschke, J. K. (2014). *Doing Bayesian Data Analysis*. Academic Press.

Kruschke, J. K., & **Liddell, T. M.** (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin and Review*, *25*(1), 178–206. DOI: https://doi.org/10.3758/s13423-016-1221-4

Kwapil, T. R., & **Barrantes-Vidal, N.** (2015). Schizotypy: Looking back and moving forward. *Schizophrenia Bulletin*, *41*(2), S366–S373. DOI: https://doi.org/10.1093/schbul/sbu186

Labvanced Team. (2022). *Labvanced Eye-tracking Guide*. Scicovery GmbH. https://www.labvanced.com/docs/guide/eyetracking/

Levay, K. E., Freese, J., & **Druckman, J. N.** (2016). The Demographic and Political Composition of Mechanical Turk Samples. *SAGE Open*, *6*(1). DOI: https://doi.org/10.1177/2158244016636433

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & **Ley, C.** (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology*, *32*(1), 1–10. DOI: https://doi.org/10.5334/irsp.289

Lourenco, S. F., & **Tasimi, A.** (2020). No Participant Left Behind: Conducting Science During COVID-19. *Trends in Cognitive Sciences*, *24*(8), 583–584. DOI: https://doi.org/10.1016/j.tics.2020.05.003

Makel, M. C., Plucker, J. A., & **Hegarty, B.** (2012). Replications in Psychology Research: How Often Do They Really Occur? *Perspectives on Psychological Science*, *7*(6), 537–542. DOI: https://doi.org/10.1177/1745691612460688

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

Masuda, T., Batdorj, B., & **Senzaki, S.** (2020). Culture and Attention: Future Directions to Expand Research Beyond the Geographical Regions of WEIRD Cultures. *Frontiers in Psychology*, *11*, 1394. DOI: https://doi.org/10.3389/fpsyg.2020.01394

Mathôt, S., & **March, J.** (2021). Conducting linguistic experiments online with OpenSesame and OSWeb. *PsyArXiv*. DOI: https://doi.org/10.31234/osf.io/wnryc

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. DOI: https://doi.org/10.3758/s13428-011-0168-7

Meteyard, L., & **Davies, R. A. I.** (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*. DOI: https://doi.org/10.1016/j.jml.2020.104092

Meyers, E. A., Walker, A. C., Fugelsang, J. A., & **Koehler, D. J.** (2020). Reducing the number of non-naïve participants in Mechanical Turk samples. *Methods in Psychology*, *3*, 100032. DOI: https://doi.org/10.1016/j.metip.2020.100032

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie Du Sert, N., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & **Ioannidis, J. P. A.** (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021. DOI: https://doi.org/10.1038/s41562-016-0021

Nelson, M. T., Seal, M. L., Pantelis, C., & **Phillips, L. J.** (2013). Evidence of a dimensional relationship between schizotypy and schizophrenia: A systematic review. *Neuroscience and Biobehavioral Reviews*, *37*(3), 317–327. DOI: https://doi.org/10.1016/j.neubiorev.2013.01.004

Nosek, B. A. (2015). Promoting an open research culture: The TOP guidelines. *Science*, *348*(6242), 1422–1425. DOI: https://doi.org/10.1126/science.aab2374

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & **Mellor, D. T.** (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2600–2606. DOI: https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., Spies, J. R., & **Motyl, M.** (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. DOI: https://doi.org/10.1177/1745691612459058

Oppenheimer, D. M., Meyvis, T., & **Davidenko, N.** (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. DOI: https://doi.org/10.1016/j.jesp.2009.03.009

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

Paolacci, G., Chandler, J., & **Ipeirotis, P. G.** (2010). Running experiments on Amazon mechanical turk. *Judgment and Decision Making*, *5*(5), 411–419.

Papoutsaki, A., Daskalova, N., Sangkloy, P., Huang, J., Laskey, J., & **Hays, J.** (2016). WebGazer: Scalable webcam eye tracking using user interactions. *IJCAI International Joint Conference on Artificial Intelligence*, 3839–3845. DOI: https://doi.org/10.1145/2702613.2702627

Parsons, S., Kruijt, A.-W., & **Fox, E.** (2019). Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395. DOI: https://doi.org/10.1177/2515245919879695

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & **Damer, E.** (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, *54*(4), 1643–1662. DOI: https://doi.org/10.3758/s13428-021-01694-3

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. DOI: https://doi.org/10.3758/s13428-018-01193-y

Prolific Team. (2022). *Prolific's Attention and Comprehension Check Policy*. https://researcher-help.prolific.co/hc/en-gb/articles/360009223553-Prolific-s-Attention-and-Comprehension-Check-Policy

Pronk, T. (2021). *Demo Eye Tracking 2*. https://gitlab.pavlovia.org/tpronk/demo_eye_tracking2

Pronk, T., Wiers, R. W., Molenkamp, B., & **Murre, J.** (2020). Mental chronometry in the pocket? Timing accuracy of

web applications on touchscreen and keyboard devices. *Behavior Research Methods*, *52*(3), 1371–1382. DOI: https://doi.org/10.3758/s13428-019-01321-2

Psychtoolbox Team. (2021). *Psychtoolbox MeasureDpi*. http://psychtoolbox.org/docs/MeasureDpi

Raine, A., Lencz, T., Scerbo, A., & **Kim, D.** (1994). Disorganized Features of Schizotypal Personality. *Schizophrenia Bulletin*, *20*(1), 191–201. DOI: https://doi.org/10.1093/schbul/20.1.191

Rodd, J. (2019). *How to maintain data quality when you can't see your participants*. Psychological Science. https://www.psychologicalscience.org/observer/how-to-maintain-data-quality-when-you-cant-see-your-participants

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. DOI: https://doi.org/10.1037/0033-2909.86.3.638

Rousselet, G. A., Pernet, C. R., & **Wilcox, R. R.** (2017). Beyond differences in means: Robust graphical methods to compare two groups in neuroscience. *European Journal of Neuroscience*, *46*(2), 1738–1748. DOI: https://doi.org/10.1111/ejn.13610

RStudio Team. (2021). *RStudio: Integrated development for R*.

Saunders, M. N. K., Lewis, P., & **Thornhill, A.** (2019). "Research Methods for Business Students" Chapter 4: Understanding research philosophy and approaches to theory development. In *Research Methods for Business Students* (8th ed., pp. 128–171). Pearson Education. www.pearson.com/uk

Sauter, M., Draschkow, D., & **Mack, W.** (2020). Building, hosting and recruiting: A brief introduction to running behavioral experiments online. *Brain Sciences*, *10*(4), 1–11. DOI: https://doi.org/10.3390/brainsci10040251

Sauter, M., Stefani, M., & **Mack, W.** (2022). Equal Quality for Online and Lab Data: A Direct Comparison from Two Dual-Task Paradigms. *Open Psychology*, *4*(1), 47–59. DOI: https://doi.org/10.1515/psych-2022-0003

Shapiro, D. N., Chandler, J., & **Mueller, P. A.** (2013). Using mechanical turk to study clinical populations. *Clinical Psychological Science*, *1*(2), 213–220. DOI: https://doi.org/10.1177/2167702612469015

Sharpe, D., & **Poets, S.** (2020). Meta-analysis as a response to the replication crisis. *Canadian Psychology*, *61*(4), 377–387. DOI: https://doi.org/10.1037/cap0000215

Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., Fei-Fei, L., Keil, F. C., Gweon, H., Tenenbaum, J. B., Jara-Ettinger, J., Adolph, K. E., Rhodes, M., Frank, M. C., Mehr, S. A., & **Schulz, L.** (2020). Online Developmental Science to Foster Innovation, Access, and Impact. *Trends in Cognitive Sciences*, *24*(9), 675–678. DOI: https://doi.org/10.1016/j.tics.2020.06.004

Siddi, S., Petretto, D. R., & **Preti, A.** (2017). Neuropsychological correlates of schizotypy: a systematic review and meta-analysis of cross-sectional studies. *Cognitive Neuropsychiatry*, *22*(3), 186–212. DOI: https://doi.org/10.1080/13546805.2017.1299702

Simmons, J. P., Nelson, L. D., & **Simonsohn, U.** (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. DOI: https://doi.org/10.1177/0956797611417632

Smythe, I., & **Everatt, J.** (2001). *Adult Dyslexia Checklist*. http://www.itcarlow.ie/public/userfiles/files/Adult-Checklist.pdf

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*(1), 155–167. DOI: https://doi.org/10.3758/s13428-010-0039-7

Stark, Z., Franzen, L., & **Johnson, A. P.** (2022). Insights from a dyslexia simulation font: Can we simulate reading struggles of individuals with dyslexia? *Dyslexia*, *28*(2), 228–243. DOI: https://doi.org/10.1002/dys.1704

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, *153*(3736), 652–654. DOI: https://doi.org/10.1126/science.153.3736.652

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods*, *42*(4), 1096–1104. DOI: https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments. *Teaching of Psychology*, *44*(1), 24–31. DOI: https://doi.org/10.1177/0098628316677643

Uittenhove, K., Jeanneret, S., & **Vergauwe, E.** (2022). From lab-based to web-based behavioural research: Who you test is more important than how you test. *PsyArXiv*, February 16. DOI: https://doi.org/10.31234/osf.io/uy4kb

van den Akker, O. R., Weston, S., Campbell, L., Chopik, B., Damian, R., Davis-Kean, P., Hall, A., Kosie, J., Kruse, E., Olsen, J., Ritchie, S., Valentine, K., Van 't Veer, A., & **Bakker, M.** (2021). Preregistration of secondary data analysis: A template and tutorial. *Meta-Psychology*, *5*. DOI: https://doi.org/10.15626/MP.2020.2625

van Stolk-Cooke, K., Brown, A., Maheux, A., Parent, J., Forehand, R., & **Price, M.** (2018). Crowdsourcing Trauma: Psychopathology in a Trauma-Exposed Sample Recruited via Mechanical Turk. *Journal of Traumatic Stress*, *31*(4), 549–557. DOI: https://doi.org/10.1002/jts.22303

Vogel, E. K., Woodman, G. F., & **Luck, S. J.** (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(1), 92–114. DOI: https://doi.org/10.1037/0096-1523.27.1.92

Walters, K., Christakis, D. A., & **Wright, D. R.** (2018). Are Mechanical Turk worker samples representative of health status and health behaviors in the U.S.? *PLoS ONE*, *13*(6), e0198835. DOI: https://doi.org/10.1371/journal.pone.0198835

Warnke, A. (1999). Reading and spelling disorders: Clinical features and causes. *European Child & Adolescent Psychiatry*, *8*(S3), S002–S012. DOI: https://doi.org/10.1007/PL00010689

Wilcox, R. R., & **Rousselet, G. A.** (2018). A Guide to Robust Statistical Methods in Neuroscience. *Current Protocols in Neuroscience*, *82*(1), 8–42. DOI: https://doi.org/10.1002/cpns.41

Woike, J. K. (2019). Upon Repeated Reflection: Consequences of Frequent Exposure to the Cognitive Reflection Test for Mechanical Turk Participants. *Frontiers in Psychology*, *10*, 2646. DOI: https://doi.org/10.3389/fpsyg.2019.02646

Woods, A. T., Spence, C., Butcher, N., & **Deroy, O.** (2013). Fast lemons and sour boulders: Testing crossmodal correspondences using an internet-based testing methodology. *I-Perception*, 4(6), 365–379. DOI: https://doi.org/10.1068/i0586

Woods, A. T., Velasco, C., Levitan, C. A., Wan, X., & **Spence, C.** (2015). Conducting perception research over the internet: a tutorial review. *PeerJ*, *3*, e1058. DOI: https://doi.org/10.7717/peerj.1058

Yetano, A., & **Royo, S.** (2017). Keeping Citizens Engaged: A Comparison Between Online and Offline Participants. *Administration and Society*, *49*(3), 394–422. DOI: https://doi.org/10.1177/0095399715581625

PEER REVIEW COMMENTS

Swiss Psychology Open has blind peer review, which is unblinded upon article acceptance. The editorial history of this article can be downloaded here:

- **PR File 1.** Peer Review History. DOI: https://doi.org/10.5334/spo.34.pr1

TO CITE THIS ARTICLE:
Gagné, N., & Franzen, L. (2023). How to Run Behavioural Experiments Online: Best Practice Suggestions for Cognitive Psychology and Neuroscience. *Swiss Psychology Open,* 3(1): 1, pp. 1–21. DOI: https://doi.org/10.5334/spo.34

Submitted: 23 January 2022 **Accepted:** 23 December 2022 **Published:** 04 January 2023

Swiss Psychology Open is a peer-reviewed open access journal published by Ubiquity Press.