Journal of Memory and Language 134 (2024) 104472

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Moving experimental psychology online: How to obtain high quality data when we can't see our participants

Jennifer M. Rodd

Department of Experimental Psychology, Psychology and Language Sciences, University College London, 26 Bedford Way, London WC1H 0AP, UK

ARTICLE INFO

Keywords:
Web-based experiments
Online recruitment
Data quality
Online research
Remote testing
Cognitive psychology
Crowdsourcing

ABSTRACT

The past 10 years have seen rapid growth of online (web-based) data collection across the behavioural sciences. Despite the many important contributions of such studies, some researchers have concerns about the reduction in experimental control when research moves outside of laboratory conditions. This paper provides an accessible overview of the issues that can adversely affect data quality in online experiments, with particular focus on cognitive studies of memory and language. I provide checklists for researchers setting up such experiments to help improve data quality. These recommendations focus on three key aspects of experimental design: the technology choices made by researchers and participants, participant recruitment methods, and the performance of participants during experiments. I argue that ensuring high data quality for online experiments requires significant effort prior to data collection to maintain the credibility of our rapidly expanding evidence base. With such safeguards in place, online experiments will continue to provide important, paradigm-changing opportunities across the behavioural sciences.

Introduction

In 2020 the Coronavirus pandemic led to many researchers closing up their labs and working from home. For many researchers, the next two years saw significant restrictions on their ability to allow volunteers into the lab to participate in research. Fortunately, the preceding 10 years had seen rapid advances in the software tools and services needed to set up experimental tasks online (e.g., Anwyl-Irvine et al., 2020; de Leeuw, 2015) and to recruit online from well-regulated pools of remote, paid participants (e.g., Crump et al., 2013; Palan & Schitter, 2018). These advances allowed those of us who had already embraced online data collection to continue to run experiments. In addition, many lab-based researchers, who may have been sceptical about online data collection, made the switch to online experiments.

This rapid shift to online research has not been without problems. Many researchers have well-founded concerns about the reduction in experimental control that arises when we move research outside of laboratory conditions and can no longer directly observe our participants. Now that researchers have the option to return to in-person testing, behavioural scientists face important choices as to when (and if) to continue with remote, web-based data collection. These choices have important consequences for the quality and cost-effectiveness of our research, as well as for our decisions about which experimental paradigms to prioritise. Many excellent reviews of the methodological issues that arise for online experiments already exist (e.g., Chandler & Shapiro, 2016; Gosling & Mason, 2015; Sauter et al., 2020; Stewart et al., 2017; Woods et al., 2015). Recent studies have also provided extensive data about the characteristics of the experiments that are currently being conducted online (Tomczak et al., 2023). The aim of this paper is not to provide a comprehensive review of the rapidly developing technical aspects of online data collection, but instead to review the use of online recruitment from the perspective of experimental design and data quality, with particular focus on cognitive studies of memory and language. I provide a set of checklists focused on experimental design considerations that will allow researchers to further improve the quality of the data they collect online. I focus here primarily on studies with adult participants. (See Chuey et al., 2021, 2022; Kominsky et al., 2021; W. Li et al., 2022; Zaadnoordijk et al., 2021; Zaadnoordijk & Cusack, 2022 for recent reviews of online research with infants and children.)

To preview these recommendations, I will caution researchers against viewing online experiments as a 'quick fix', a simple matter of translating tasks from lab-based to browser-based software, or as necessarily being cheaper, easier or more efficient than in-person testing. Ensuring appropriate data quality for online experiments requires significant effort prior to data collection, but is vital to maintain

E-mail address: j.rodd@ucl.ac.uk.

https://doi.org/10.1016/j.jml.2023.104472
Received 14 April 2023; Received in revised form 6 September 2023; Accepted 10 October 2023
Available online 21 November 2023

the credibility of data obtained using online methods. With such safeguards in place, online experiments will continue to provide important, paradigm-changing opportunities for researchers.

Importantly, many of the issues highlighted in this review as being potential causes of reduced data quality for online experiments (e.g., participant inattentiveness) also apply to lab-based experiments. In these cases, some of the potential solutions should simply be considered good experimental practice, regardless of the experimental setting.

Why would we want to do our experiments online?

Web-based data collection methods provide exciting new opportunities for behavioural research that would not be feasible using in-person recruitment. For example, my earliest venture into online data collection was a mass participation experiment conducted in collaboration with BBC Radio 4 that was only possible due to the emerging online data collection tools (Rodd et al., 2016; Experiment 1). Participants were recruited for this priming experiment from across the United Kingdom via a live radio programme. The lexical primes were embedded within the radio show such that listeners were exposed to them in a highly naturalistic manner. At the end of the programme they were invited to take part in an online task that measured the impact of these primes on their subsequent lexical processing. This experiment was then followed up by two online experiments in which we specifically recruited participants with idiosyncratic linguistic experiences via their participation in recreational rowing to test how their experiences had modified their representations of words that are used differently within a rowing context (e.g., using "catch" to refer to the time where their oar enters the water; Rodd et al., 2016; Experiments 3 & 4). These niche participants would have been difficult to recruit in sufficient numbers using conventional lab-based testing.

Some of the most ambitious and innovative research in the fields of memory and language over the past 10 years have only been possible because the researchers have adopted a large-scale 'citizen science' approach that has capitalised on the power of online data collection to take research outside of our laboratories and collect large, diverse datasets (W. Li et al., 2022). Recent years have seen the publication of several very large online studies that include participant numbers that were previously inconceivable. The online game Sea Hero Quest (Coughlan et al., 2019; Coutrot et al., 2019; Spiers et al., 2023) provided data about the spatial navigation skills of *4.3 million* individuals from around the globe. Such large-scale experiments have provided important, novel insights into key theoretical debates concerning critical periods (Chen & Hartshorne, 2021), and the effects of age and gender on cognitive skills (Erb et al., 2023; Spiers et al., 2023). Of particular interest to researchers interested in how language is learned and processes is the opportunity to open up their research to large groups of participants who speak a wide range of different languages beyond those spoken in the researchers' current location (Pavlick et al., 2014).

These studies illustrate the exciting potential of online recruitment methods to dramatically expand the scope of our research both in terms of the number and the diversity of participants. However, it is also the case that many online experiments use more conventional experimental designs that could straightforwardly have been conducted in the lab. In such cases the switch to online testing has likely been made for pragmatic reasons, such as cost or convenience. For example, in pre-pandemic times a PhD student in my research group submitted a thesis comprising ten behavioural experiments using conventional psycholinguistic paradigms with data from 986 participants (Betts, 2018) – a feat that was only possible because (i) the pool of potential online participants is far larger than typical local participant pools, and (ii) the researcher time needed to test large groups of participants is greatly reduced. Therefore, even for traditional cognitive experiments, online recruitment allows us to collect larger datasets extremely quickly (Peer et al., 2017). This improved efficiency is a legitimate reason for shifting experiments online as it facilitates the highly desirable scaling up of our lab-based paradigms to larger sample sizes (Hartshorne et al., 2019). There is a growing consensus that published studies in the psychological sciences have been systematically underpowered, and that studies with low statistical power are more likely to report effects that cannot be replicated by other researchers (Brysbaert, 2019; Szucs & Ioannidis, 2017; Fraley & Vazire, 2014). Fortunately, the last 10 years have seen some improvements in typical sample sizes, which may in part be due to increased use of online recruitment (Fraley et al., 2022; Sassenberg & Ditrich, 2019). This view is supported by recent data showing that of the 1 million online participants tested using the Gorilla platform in 2022, 61% were in small studies ($N < 250$), 15% were in medium studies (251–500 N) and 24% were part of large studies ($N > 500$; Tomczak et al., 2023).

In addition, many current platforms for collecting data online make it relatively straightforward to share everything that is needed for other researchers to collect additional data in an exact replication or modified version of the original experiment in an extremely time-efficient manner. Thus the switch to online recruitment may provide, at least in theory, one part of the solution to improving the reproducibility of our evidence base (Aarts et al., 2015; Chambers, 2017; Munafò et al., 2017).

A final factor that is facilitating the shift towards online experiments is the recent flurry of exciting developments in the range of tools available to help researchers conduct a wide variety of types of experiments. For instance, the year 2022 alone saw the publication of tools to collect eye-tracking data from infants in their homes using an online webcam-linked eye tracker (OWLET; Werchan et al., 2022), adaptive methods for estimating psychometric function parameters in online experiments (jsQuestPlus; Kuroki & Pronk, 2022), methods for collecting speeded overt pronunciation responses in language production tasks (Vogt et al., 2022), mouse-tracking methods that provide an alternative to eye-tracking in order to track participants' spatial attention (MouseView.js; Anwyl-Irvine et al., 2022), tools to support highly interactive coalition formation experiments (The Online Coalition Game; Wissink et al., 2022), methods to allow researchers using auditory studies to standardize sound level adjustments, detect stereo/mono playback, and assess lower frequency limits (Headphone and Loudspeaker Test (HALT); Wycisk et al., 2022), a toolbox for facilitating the collection of a range of ratings data for auditory material (Donhauser & Klein, 2022), and methods for measuring sensorimotor synchronization (SMS) in online experiments using the built-in microphone and speakers of standard laptop computers (Anglada-Tort et al., 2022). While this flurry of methods development is hugely exciting for those of us using online data collection methods, it is perhaps timely that we pause to reflect on the rigour of our approach to psychological experimentation.

Can we trust the data from online experiments?

The history of online data collection dates back to the mid-1990s when the introduction of *Forms* (or "fill-out forms") first allowed readers of web documents to send responses back to the server. (See Musch & Reips, 2000 for a comprehensive review of the early days of online data collection and the relevant technological developments.) Initially, web-based data collection within the psychological sciences was focused on text-based surveys and questionnaires that would more typically have been administered by means of face-to-face interviews, postal mail, or telephone calls (see Gosling et al., 2004), but detailed guidance on how researchers could use web-based data collection across a range of methodologies dates back to at least 1996 (Hewson et al., 1996). Large-scale studies that recruited many more participants than were typical in lab-based studies quickly followed. For example, Nosek et al., (2002) reported the results of 600,000 implicit attitude and stereotype tests obtained in an 19 month period between 1998 and 2000, with over 150,000 datasets collected in a 5-day period following national television coverage in 2000.

Despite this early progress, many researchers remained sceptical in the extent to which results from online experiments should be given the

same credibility as lab-based studies, especially for paradigms that rely on precise timing of either stimulus presentation or response recording (Chetverikov & Upravitelev, 2016). These concerns began to be alleviated by a set of important studies comparing data collected online to the findings from existing lab-based studies. Germine et al., (2012) reported the performance of visitors to their TestMyBrain.org website on a range of cognitive/perceptual tasks known to be sensitive to individual differences in healthy participants (e.g., face memory, emotion perception, visual memory, episodic memory, working memory). In some cases these tasks rely on presentation of complex visual stimuli, use brief presentation timing and, in the case of the memory tests, depend on participant honesty and are potentially susceptible to cheating. Although their self-selecting group of participants performed better in some tasks and more poorly in others compared with the participants in previously published lab-based studies, the internal reliability of their online data was at least as good as the lab data on all tests. Other studies replicated key within-participant findings from a range of classic paradigms online. For example, Crump et al., (2013) replicated a number of classic cognitive psychology paradigms that required millisecond timing for response time collection (Stroop, task-switching, flanker task, Simon task) as well as paradigms for which precise control was needed over stimulus presentation due to the relatively short stimulus presentation times (visual cueing, attentional blink, masked priming). Similar reassurance came from studies of performance on linguistic tasks such as acceptability judgements (Sprouse, 2011) and self-paced reading (Enochson & Culbertson, 2015). Since then, researchers have extensively explored the validity and reliability of online data collection methods (e.g., see Stark et al., 2022 for recent demonstration of a classic semantic interference effect in word production as measured by key press responses; Angele et al., 2022 for a replication of classic masked priming effects; Ratcliff & Hendrickson, 2021 for exploration of lexical decision data). Importantly, several studies directly compared the data from the same task when it was performed in the lab using standardized equipment or in participants' own homes with their own hardware. These studies have been broadly positive in their conclusions. For example in their comprehensive review of the literature Hartshorne et al., (2019; Appendix C) concluded that "internet volunteers comply with instructions and answer truthfully at rates matching or exceeding lab-based subjects, resulting in data with similar psychometric validity".

However, these demonstrations of successful online experiments should not, of course, be taken as evidence that the data from *all* online experiments can be trusted. First, not all paradigms can be transferred online with such ease. For example, the results from a recent web-based implementation of the visual-world paradigm, which collected eye-movement data online (Slim & Hartsuiker, 2022) showed that while the spatial accuracy of the data was sufficient to distinguish looks across the four quadrants of the computer screen, there was an unexplained delay of roughly 300 ms in the time course of the eye movements. In addition, researchers should ensure that the particular experimental platform that they are used has been appropriately validated, particularly with respect to timing accuracy.

In summary, while the increasing number of demonstrations of successful online experiments gives us cause for general optimism about web-based data collection, care should still be taken when setting up individual experiments to ensure that we maintain high standards with respect to data quality. Broadly speaking, preserving high data quality during online testing requires us to carefully consider three key issues that might have somewhat different considerations for online experiments compared with lab-based tasks: technology, recruitment and participant performance. I review these three issues in turn, making specific recommendations for researchers.

Technology

When shifting to online data collection, researchers must carefully consider the technology that supports their research (Anwyl-Irvine et al., 2021). Not only must researchers make careful decisions about what experimental software to use to communicate with participants (usually via the internet), but they must also consider the impact of relying on participants' own software (e.g., web-browser) and hardware (e.g., laptop, headphones). In some cases, experimenters choose to send out specialised equipment to participants in advance of data collection, either to ensure equivalent quality of stimulus presentation (e.g., headphones), or because their study requires equipment that they could not reasonably expect participants to have access to (e.g., sleep monitoring devices). Here I focus on the more common situation where researchers rely on participants to provide all the hardware themselves (See Table 1 for a summary of recommendations).

Reliance on web-based communication

Most remote experiments require participants to access the experimental task via a web-browser. This approach avoids the need for participants to download bespoke software or apps, which can be perceived by participants to be time-consuming and to have additional security risks compared with the web-browser approach. A consequence of delivering experiments via web-browsers is that information about the task and participants' responses is usually transferred to/from each participant at the time of testing. Given that many researchers have little or no web development experience, most researchers choose to create their browser-based experiments using specialised software that provides a JavaScript programming framework and/or a graphical user interface (GUI). These experiment-building tools vary on a range of dimensions such as (1) how much time/expertise is needed to get started, (2) how much flexibility and control it offers the researcher, (3) available features, (4) cost, (5) compatibility with open-science practices, and (6) long-term maintenance and sustainability. In addition to creating experimental tasks that can be run in a web browser, researchers will need a web server to host their experiments and store the data. Again, there are a range of hosting options that vary along the same dimensions described above, as well as considerations such as data protection. It is up to individual researchers and labs to decide what factors are most important, and where they sit on the customization-flexibility spectrum, from fully-serviced and GUI-based tools (e.g., Gorilla Experiment Builder, https://www.gorilla.sc; Anwyl-Irvine et al., 2020; Qualtrics, qualtrics.co.uk) to highly-customised and often programming-heavy solutions, e.g. The Music Lab (https://www.themusiclab.org/), MIT's Lookit platform (https://lookit.mit.edu/). Several options are available that fall somewhere in the middle of this spectrum (e.g., jsPsych, lab.js, PsychoJS, OSWeb, Pavlovia, Cognition.run, JATOS, Inquisit, Pushkin, PCIbex; See Hartshorne et al., 2019 for an extremely useful overview of the functionality of these different options). A final important factor that researchers must explore when choosing between these options is the requirements of their local ethics or data protection regulations with respect to where and how data from participants is stored. This is of particular concern if they are collecting sensitive personal information or making sound/video recordings of participants.

Regardless of the exact set up of any web-based experiment, participants must be able to access the task via an appropriate internet connection. Fortunately, most internet connections are sufficiently fast and reliable to support this approach, although care may need to be taken when testing on mobile devices or when targeting participants from geographical locations or demographic groups where internet connections may be more variable. Most well-used experimental platforms include sufficient data buffering to ensure that transient fluctuations or reductions in connections don't directly impact on the timing of stimulus presentation or response recording. At the start of an experiment, the information needed to deliver the task is typically loaded into the browser's (limited) memory cache to avoid subsequent loading delays during the experiment. Using this approach, slow connections may cause longer initial loading times, and it is usually good practice to have

Table 1
Checklist for researchers: Dealing with variability in participants' technology.

Source of Variability	Potential Solutions				
	Specify Requirements During Recruitment	Obtain Information about Participants' Tech before Task	Obtain Information about Participants' Performance before Task	Piloting of Task	Post-Experiment Participant Feedback
Participant Hardware					
Main Device	Laptop/desktop vs. tablet vs. phone	Via experimental software Tick-box confirmation			
Audio Peripherals	Speakers vs. headphones Headphone type (e.g., in-ear vs. on-ear vs. over-ear)	Headphone screen task[1] Tick-box confirmation	Audio discrimination task	Compare screening task performance to self-report	Check they heard all stimuli clearly
Visual Peripherals	Minimum screen size	Viewing distance or physical size screen task[2]	Visual discrimination task	Compare screening task performance to self-report	Check they saw all stimuli clearly
Input Devices	Keyboard, touchscreen, mouse, microphone	Via experimental software Tick-box confirmation			
Participant Software					
Web-browser	Browser identity (Chrome, Safari, Firefox, Edge)	Via experimental software		Pilot on all common browsers	
Operating system	Windows vs. Mac	Via experimental software		Pilot on all common OSs	
Web connection	Specify need for stable connection	Via experimental software		Ask participants about delays/glitches. Check trial timings on (simulated) slow connections	Ask participants about unexpected glitches/ delays

Note that the importance of these different components will vary across tasks, stimuli and participant cohorts. 1. Milne et al., (2021), 2. Li et al., (2020).

a time-out if the experiment hasn't fully loaded after a certain duration. Similarly, results are often sent back during the experiment to avoid data loss if the participant doesn't finish, and this can be done in the background at any time, without interrupting the flow or timing of the experiment.

All the above means that for most online experiments that are run using well validated software, issues to do with data transfer are dealt with 'behind the scenes' without the need for careful management by the researcher. For most online tasks, this is not something that places significant constraints on our experimental design choices or impacts on data quality. However, all experiments should be routinely piloted on a slow connection to better understand any consequences that connectivity issues may have for participants' experience of the task. Note that many browsers can simulate this slow connection as part of their developer tools, often referred to as 'network throttling' (e.g., https://developer.chrome.com/docs/devtools/network/reference/#throttling). Second, it is useful to monitor how a particular experiment runs in practice by including questions in the debrief that would reveal any issues that might arise due to connectivity issues, e.g., slow loading times, unexpected delays between trials. Importantly, this qualitative information from participants should be supplemented by quantitative information. Most experimental software now provides detailed trial-by-trial timing information. This data can be reviewed to ensure that our experiments have run with the expected timings. Taken together, careful checking and reporting of this important timing information can reassure both experimenters and other researchers that an experiment has run as expected without significant glitches.

Reliance on participants' hardware and software

Experimental psychologists are accustomed to constructing a bespoke lab-based experimental setup, choosing carefully from a wide range of hardware, software, services, and programming languages, and then collecting data from all participants via this standardised setup. In contrast, online experiments are usually conducted on participants' own devices – sometimes referred to as 'commodity devices', a term which refers to devices/components that are relatively inexpensive, widely available and more or less interchangeable with other hardware of its type. Unfortunately, the idiosyncratic combinations of the hardware and software that are used by individual participants will differ widely in terms of their specifications. This variability can affect the primary device being used (i.e., their computer, laptop, tablet, phone) and their internet browsers (e.g., Chrome, Firefox, Safari, Edge; see Anwyl-Irvine et al., 2021 for an analysis of the equipment current online participants are likely to use), as well as any peripherals being used to present stimuli (e.g., headphones, speakers, monitors), or record responses (e.g., keyboards, touch screens; Pronk et al., 2020).

Recommendations: Minimum tech requirements

Prior to data collection researchers should specify the (experiment-specific) tech requirements that must be met by all participants. The level of restriction will depend critically on the nature of the experimental paradigm and on the experimental aims.

Perhaps the most salient concern to researchers when deciding what tech is considered adequate relates to the precision/reliability of timing information with respect to both stimulus presentation and response measurement. While modern web browsers have the capacity for millisecond timing both in terms of stimulus presentation and response recording (see Lukács & Gartus, 2022), the precision and reliability of the timing that is achieved in practice is inevitably limited by participants' hardware/software. Researchers may choose to restrict participants' tech in order to improve the precision or consistency of this timing information. While it is often suitable for participants to complete survey tasks (for which response time data are not critical) on a phone or tablet, many reaction time tasks are limited to laptop or desktop computers. In addition for experiments where timing precision is critical researchers may place additional restrictions in terms of operating system or web browser. Two major recent studies have provided a valuable evidence base about the timing precision that is achieved by a range of experimental combinations (Anwyl-Irvine et al., 2021; Bridges et al., 2020). The findings from these studies are broadly positive in terms of the overall levels of performance: timing accuracy is generally good and there are no major outliers in terms of particular hardware/software combinations. They do however report some variability across OS/browser combinations. Anwyl-Irvine et al., (2021)

report that, at the time of testing, the most common combination (Windows + Chrome) was the best performing combination in terms of the accuracy and precision of display and response timing. Interestingly, Anwyl-Irvine et al., (2021) report that the choice of experimental platform contributes greater variance than the participants' choice of device. Bridges et al. (2020) specifically highlight the variability they observed in audio-visual synchrony, which they report to be the least precise aspect of the browser-based experiments. Unfortunately, as noted by the authors of these papers, while these studies provide useful snapshots of the state of play at a given point in time, platforms, browsers, and operating systems evolve rapidly making it difficult for researchers to assess the likely impact of updates that have occurred since these data were collected. This need for caution when interpreting the relative merits of different set-ups based on published performance data is emphasised by the presence of inconsistencies in the timing lags reported for particular hardware/software combinations between these two papers, which were published within a relatively short window of time.

The second area in which researchers often wish to restrict participants' tech relates to the peripherals used to present stimuli, either auditory or visual. Broadly speaking, researchers can take two approaches to imposing restrictions on participants' tech. They can either specify the characteristics of the tech itself (e.g., "must use headphones") or can focus on participants' performance on tasks that are designed to reveal important information about their experience with the stimuli (e.g., ability to make fine-grained audio discriminations). Once a researcher has decided what aspects of a participant's set up are most relevant to their current experiment, they must consider how they will obtain information about these relevant characteristics. There are a range of options, which vary in the effort required on the part of the researcher. First, important information about participants' hardware/software is often reliably and automatically provided by the experimental software being used (e.g., phone vs. tablet vs. computer; web-browser identity). In the absence of such data, researchers may elect to simply rely on participants to honestly and accurately report the relevant information – this can be particularly time efficient for relatively 'low-stakes' issues where the researcher judges that any misinformation might lead to additional noise in the data, but is unlikely to introduce any systematic bias. If this approach is not judged to be adequate then they can build into their experimental procedures a screening task that can identify salient characteristics of participants' experimental set ups. For example, researchers may include a simple alternative choice task to assess whether participants can adequately hear/see stimuli that are similar to those to be used in later tasks. Finally, researchers may choose to incorporate one of the increasing number of sophisticated screening tasks that are available. For example, it is now relatively standard for studies that use auditory stimuli to include tasks that can reliably reveal whether or not participants are wearing headphones (e.g., Milne et al., 2021). It is also possible to measure a participant's viewing distance in the web browser, for example by detecting their blind spot (e.g., Q. Li et al., 2020; see Brascamp, 2021 for review). Note that in some cases, researchers can set up their experiment in a way that flexibly adapts in response to this information about participants' equipment, for example, by varying the relative size of visual stimuli in response to feedback from participants about absolute size of a standard stimulus (e.g., a credit card) on the screen (Q. Li et al., 2020).

Ideally, we would always exclude participants with unsuitable tech *before* the start of the experiment by indicating our requirements at sign up. In such cases I would strongly recommend adding a tick box section near the start of the experiment asking them to confirm that they meet these requirements as sign-up instructions are not always read carefully. If this is not possible because we need to embed screening questions or screening tasks, it is usually preferable to include these elements as early in the experiment as possible. Note that it is not always possible to reject participants who fail screening tasks at this point in order to avoid paying unsuitable participants as this may contravene the rules of local ethics boards or the relevant crowdsourcing recruitment panel. In addition, it is sometimes necessary to collect data from participants with unsuitable tech if our screening tasks are too complex to allow instantaneous exclusion of participants based on performance. (See the section below on preregistration of exclusion criteria for guidance on these situations where data must be excluded *after* data collection.)

Finally, in some cases, researchers may be uncertain about whether particular tech choices will influence participants' performance in important ways. In these cases, researchers can face difficult decisions that balance their desire to collect high quality data with constraints related to recruitment. For example, it may be that allowing participants to complete the tasks on a tablet may reduce data quality compared to when a laptop/desktop computer is required, but this compromise may be worthwhile if it allows recruitment of larger or more diverse sets of participants. In such cases, where exclusion criteria are not clear cut, the best approach may be to record all potentially relevant information about participants' tech setup and then explore the impact of these differences in the analysis stage.

General considerations: The tricky cases

The above sections highlight some of the many issues that arise for online data collection that relate directly to the technological requirements of such experiments. This array of potential issues may seem somewhat daunting, especially for researchers embarking on online experiments for the first time. Fortunately, in the case of many relatively straightforward experimental paradigms many of these considerations are unlikely to directly impact on the quality of the data that is collected, especially when experiments are delivered using tools that have been well validated. Unfortunately, it is not always immediately clear to an individual researcher what might constitute a 'straightforward experimental paradigm'. In other words, it is not always clear whether an individual researcher can safely use an off-the-shelf solution or whether they need to dig more deeply into the relevant tech issues. One helpful approach may be for researchers to carefully consider whether the variability that is inevitably introduced by participants' hardware/software choices can safely be considered to be 'random noise' that can largely be mitigated by larger sample sizes.

One clear case where this assumption is likely to *not* be safe is in between-subject designs that explore differences between individuals (or groups of individuals). For example, studies that explore how factors such as age or education might modulate performance must consider that participants who are younger, more affluent or more highly educated may have more recent/expensive tech that might systematically impact their data, particularly on tests requiring fast reactions or fine motor movements (Passell et al., 2021). In these cases, the systematic differences that exist across devices (e.g., in display/response latencies; Nicosia et al., 2022) can potentially lead to significant, but artefactual, performance differences between groups who differ systematically in their technology (see Hartshorne et al., 2019 for discussion). This could, for example, result in an apparent decline in performance with age in any sample where older participants own devices with slower latencies. The presence of such a relationship between the quality of a participant's equipment and an experimental variable of interest is not straightforward to address. For example, simply restricting access to those participants who meet some pre-specified baseline technological requirements would likely result in sampling biases that might interact in complex ways with variables of interest. At the very least, researchers using such designs should systematically record any potentially important tech information for inclusion in analyses that attempt to disentangle effects of their variables of interest from confounding differences in participants' tech.

A second case where it is unsafe to treat variability in participants' tech as random noise is those experiments where the conclusions that a researcher wishes to draw from their data relies on highly accurate

information about the precision of stimulus presentation. For example the inferences that are made from studies of masked priming can rely on the exact absolute duration of stimulus presentation (Van den Bussche et al., 2009). In such cases, researchers need to ensure that they can rely on timing information with a similar level of confidence to in-lab studies (Angele et al., 2022; Barnhoorn et al., 2014).

Recruitment

One of the most appealing aspects of online data collection is the ability to collect data from participants anywhere in the world at the click of a button without them having to travel to our lab spaces. Researchers have a number of different options available to them in terms of how these participants are recruited. These choices have important consequences for the quality of the data that we collect, and for the demographic diversity of the participants that we recruit. The following section will first provide an overview of the different approaches to online recruitment, before considering some of the important factors that researchers should take into account when deciding on their recruitment strategy (See Table 2 for a summary of these recommendations).

Recruitment approaches

Researchers must choose between several general approaches to online participant recruitment. First, they can continue to recruit in exactly the same manner that they use for their lab-based studies, making use of locally organised participant panels or local students who participate in experiments for course credit. This local approach to recruitment can have advantages such as increased in-person vetting procedures, and in the case where participants are psychology students can form a valuable part of their research methods training. However, in most cases researchers are keen to recruit participants from outside their locality usually using one of two approaches. First, researchers may use *indirect recruitment* where they stay one step removed from their participants and subcontract recruitment to a crowdsourcing platform (or local participant pool). Alternatively, researchers can take a more *direct* approach of contacting remote participants directly via social media or other networks.

Indirect recruitment via crowdsourcing platforms

The most common approach to online recruitment is to rely on crowdsourcing platforms that give researchers near instantaneous access to large pools of paid participants. A dominant force since it was launched in 2005 has been Amazon Mechanical Turk (MTurk; https://www.mturk.com; Paolacci & Chandler, 2014). This platform was not specifically designed for researchers – it was set up as a generic online crowdsourcing service to allow anonymous online individuals (known as workers or Turkers) to receive payment for completing web-based tasks (known as HITs: human intelligence tasks). HITs can be offered by commercial organisations, researchers or other individuals (known as requesters). In the case of academic research, these HITs (i.e., surveys or experiments) are typically hosted on external websites. A second widely used platform, Prolific (https://www.prolific.co, formerly known as Prolific Academic) was set up by academic researchers in 2014, primarily in response to geographical limitations of MTurk and concerns about the limited assurances about data quality and pre-screening provided by other alternatives (Palan & Schitter, 2018; Peer et al., 2017). Several other similar platforms are also available (e.g., Testable Minds, https://www.testable.org/minds); Cloud Research, https://www.cloudresearch.com/; Crowdworks, https://crowdworks.jp/).

These participant pools have several key characteristics that make them appealing to researchers. First, the platform is responsible for finding individuals who are willing to complete tasks, and maintaining a database of these individuals. Second, the platforms' payment systems typically allow researchers to make a single lump sum payment to the platform rather than setting up their own bespoke payment systems to reward individual participants. Third, these platforms provide, to differing degrees, the ability to select participants with particular characteristics. For example, within its standard commission rate, Prolific currently includes basic level screening for a relatively wide range of factors (age, sex, gender identity, nationality, country of birth/residence,[1] language background, ethnicity, employment/educational status, political/religious affiliation, sexual orientation, handedness, marital status and socioeconomic status). They also allow researchers to screen on the basis of participants' approval rating (i.e. how well did participants do in past studies). Finally, in some cases these platforms can provide participant samples that are carefully curated according to demographic factors. For example, for additional payment Prolific currently offers representative samples for UK and US populations, based on recent census data about age, sex and ethnicity. Of course, these benefits come at a price – MTurk currently charges a commission rate of 40%, while Prolific's standard commission rate for academic researchers is currently 33% (https://www.prolific.co/researchers#pricing). It is up to individual researchers to determine whether this constitutes value for money when compared with other recruitment approaches, which will load differently on staff time and may produce different levels of data quality. (See Peer et al., 2022 for recent comparison of data quality across different recruitment platforms.)

Several studies have explored the characteristics of participants recruited via these platforms. Of particular interest for language researchers, Pavlick et al., (2014) highlighted the linguistic diversity of the participants who are available, reporting that a sample of 5,043 bilingual MTurk workers from across 106 countries included individuals with 95 different native language. US workers alone reported 61 different native languages. On the basis of data from an online translation task they concluded that at least 13 languages provided sufficiently large populations of participants who gave quick and accurate translation responses: Dutch, French, German, Gujarati, Italian, Kannada, Malayalam, Portuguese, Romanian, Serbian, Spanish, Tagalog, and Telugu). Given the increasing awareness of the importance of studying human language across the widest possible range of languages, this relatively easy access to linguistically diverse participant pools is extremely exciting.

Direct recruitment via social media or existing networks

The alternative to using established crowdsourcing platforms is for researchers to directly recruit their own participants. For example, by posting a link to their experiment on social media and asking followers/friends to participate and share with their networks. Or researchers may share their experiment with existing networks (e.g., local schools or campaign groups). This direct approach can be attractive due to its low cost, apparent ease, and its potential to collect extremely large datasets if the experiment is widely shared. In such cases, researchers have the choice as to whether or not to directly reward their participants for their time, either by direct payment or other reward (e.g., charitable donations).

Over the past 15 years, several highly successful large-scale studies have been run in this way to explore important theoretical questions within cognitive psychology. Halberda et al. (2012) investigated numerical intuitions and their relation to students' performance in school mathematics across the lifespan from more than 10,000 participants.

[1] Note that using recruitment platforms such as Prolific may be particularly beneficial to researchers for whom demographic factors such as geographic locations are of particular importance. IP addresses are not a reliable way of determining someone's geographic location, so if this information is important to your experimental aims then it's best to use a participant recruitment service like Prolific, because they are better able to verify participants' country of residence.

Table 2
Checklist for researchers: Recruitment Issues.

Potential Issues	Experimental Considerations	Potential Solutions		
		Indirect Recruitment (e.g., MTurk, Prolific)	Direct Recruitment (e.g., via Social Media)	All Recruitment Methods
Inaccurate Demographic Information[1]	Is information critical to research question? (e.g., language background as dependent variable) Are participants likely to lie to gain access to task? (e.g., financial reward, restricted access)	Check wording of platform's screening questionnaires		Speeded tasks to verify key demographics (e.g., vocab tests) Ensure terms are consistently understood. (e.g., "native language") Repeat key questions post-experiment (payment guaranteed)
Non-naivety (i.e., Super Workers)	How might experience with similar tasks impact/bias performance? (e.g., surprise elements, practice effects, priming)	Avoid restricting by high approval rating Exclude participants from earlier, related experiments	Target recruitment at non-traditional pools (e.g. avoid academic networks)	Questions about prior task experience in debrief (for exclusion or use in analysis)
High Attrition	Is experiment particularly long, dull or difficult?	Consider restricting to participants with high approval rating (although see non-naivety issue)		Honestly describe task at sign-up Ask participants to explicitly confirm their availability for duration before task begins Emphasise importance of research
Selective Attrition	Between-participants or individual differences design? Might dropout differ across conditions? (e.g., more difficult condition, time of day preference) Might dropout differ across groups? (e.g., younger, lower performing)	Increase motivation to complete (e.g., higher payment, completion bonus)	Increase motivation to complete (e.g., emphasize benefit to society)	Warm-up task such that dropout occurs *before* condition allocation Emphasise length of task at sign-up Collect partial data from incomplete participants Report attrition by condition/ group Consider impact of asymmetric drop out on conclusions Participant partnership (Table 3)

1. Specific concerns about bots (i.e., automatic survey-takers) are best dealt with by careful consideration of participant task performance (See Table 4).

Brysbaert et al. (2016) tested the vocabulary knowledge of 221,268 individuals who were each shown a random list of 67 words (and 33 non-words) selected from a list of 61,800 dictionary entries that they believe includes "the vast majority of reasonably known English words" (https://vocabulary.ugent.be/). Similarly, Guasch et al., (2022) recruited more than 200,000 native speakers of Catalan for an online visual lexical decision task using a social media campaign supported by radio interviews and newspaper articles. They report that 'word of mouth' sharing via social media was responsible for the majority of their recruitment success. Hartshorne, Tenenbaum & Pinker (2018; see also Chen & Hartshorne, 2021; Hartshorne & Germine, 2015) set up a grammar quiz that went viral, allowing them to recruit 669,498 participants for an experiment that explored the role of critical periods in language development by disentangling participants' age at first exposure to English from both their current age and their number of years' experience. Finally, in what is perhaps the current pinnacle of online cognitive research, the online game Sea Hero Quest (https://glitchers.com/project/sea-hero-quest/) represents a successful collaboration between researchers and commercial game/web developers (Coughlan et al., 2019; Coutrot et al., 2019; Spiers et al., 2023). The high quality free-to-download game provided data about the spatial navigation skills of 4.3 million self-selecting players from around the globe who provided over 117 years of total game play without need for participant payment.

Unfortunately, these hugely successful studies remain exceptional – for many researchers direct recruitment can be frustrating and unpredictable, and can result in recruitment of highly atypical participants if sharing takes place primarily through academic networks. In my view, successful direct-recruitment experiments where participants are not directly paid require researchers to plan a sophisticated recruitment strategy that maximises the likelihood that their experiment will be widely accessed and shared. Any such strategy requires that researchers carefully consider participants' motivation to take part in their experiment. Not only can this help to ensure that participants are recruited quickly and efficiently, but that the people who take part are well motivated to complete tasks with appropriate levels of care and attention.

Participant motivation

Participants take part in our experiments for a variety of reasons. These usually tap into one (or more) of the following sources of motivation: financial reward, altruism, knowledge seeking, and

entertainment. Careful consideration of these factors is essential for experiments where participants are *not* paid, but is worth considering for all experiments. Even if participants are being directly paid for their time, this payment is unlikely to be the sole factor that drives their decisions about whether or not to participate (Göritz, 2014), and other factors will likely impact on the care they take while completing individual tasks.

Altruism

Online experiments may succeed in recruiting large numbers of highly motivated participants by tapping into altruistic motivation by being framed as 'citizen science' projects in which participants are making an important contribution to research that has potential societal impact (W. Li et al., 2022). The social importance of the research is often highlighted to enhance participants' sense that they are donating their time to a good cause. Explicit partnerships with well-established charities can help legitimise this link with the general public (see Sea Hero Quest's link with Alzheimer's Research UK).

A somewhat more targeted approach to recruitment that also taps into altruistic motivation is to use social media as a gateway to specific community networks. For example, researchers may focus on participants with particular clinical diagnoses or demographic characteristics (e.g., second language learners, young children, twins). Researchers may have strong existing links with the relevant communities (e.g., schools or clinical services) that allow participants to be recruited either directly on social media or by asking relevant contacts (e.g., teachers, clinicians) to pass on information about their online experiments to members of the target population. Existing relationships can be important in order to establish the credibility of the research team and to convince participants that they are giving up their time in a manner that will be of (long-term) benefit to their community. Ideas of reciprocity can also be important here, with researchers giving up their time to provide workshops or community events that strengthen the sense that they are part of a shared communal effort to address issues that are of concern within a particular section of the community.

This altruistic motivation that can drive participants to contribute to online research is relevant, to a lesser degree, when researchers share their experiment with their social media followers/friends with a plea to "help out my student" or similar. This approach is less likely to succeed in gathering large datasets: their immediate friends/colleagues may choose to help out, but this desire to help is less likely to extend beyond their immediate circle, which of course can be problematic in terms of participant demographics – the friends/followers of academic researchers are unlikely to be typical of the wider population. In general, I advise against this approach to data collection, which is unlikely to provide large numbers of suitable participants, especially when experiments are circulated amongst populations that may already be saturated by similar previous requests.

Knowledge seeking

Experiments that are widely shared on social media often motivate participants by promising to provide them some information in return for their time. Such experiments may be modelled on the ubiquitous social media quizzes used by marketing agencies to create audience engagement by promising to provide some information either about themselves (e.g., insights into personality/intelligence) or with more generic interesting facts. Sites such as https://www.labinthewild.org (Reinecke & Gajos, 2015) and testmybrain.org (Germine et al., 2012) are examples of experimental platforms in which all experiments provide personalized results about how participants' performance or preferences compare to others.

The use of personalised feedback to motivate participants can be enhanced by allowing participants to share their performance scores on social media (Hartshorne et al., 2019). One successful example of this approach is the lexical decision task used by Brysbaert et al. (2016) who motivated participants' to take part by providing an estimate of the proportion of English words that each participant likely knows, and how this compares to other participants. Participants were allowed to take the test as often as they liked, with a different subset of words being tested on each session. They tested 221,268 individuals, some of whom took the vocabulary test more than 100 times. Note that this approach of feeding back details of participants' performance needs to be treated carefully from an ethical point of view, taking into account the potential impact for participants who perform poorly, especially if children or vulnerable adults may be included in the sample.

Entertainment

Successful large-scale online experiments have often been carefully designed to be highly engaging and entertaining: participants are more likely to give up their time for free if they are participating in a fun, gamified task (Long et al., 2023). The most successful example of this is the Sea Hero Quest game mentioned above (Coughlan et al., 2019; Coutrot et al., 2019). This was built by game developers to such a high standard that participants willingly gave up many hours of their time to play. Of course this approach is not always feasible: not all experimental paradigms are susceptible to being fully gamified, and researchers usually don't have the skills or budget to invest in sophisticated game design. Fortunately tasks that more closely resemble classic experimental paradigms (e.g., lexical decision) can succeed in recruiting large numbers of participants if they are kept short and accessible and participants are highly motivated by the altruistic or knowledge-seeking motivations described above (Brysbaert et al., 2016).

A hurdle that can sometimes prevent mass participation is the inclusion of lengthy consenting/instructions phases before the task begins. There can be large differences in what the ethics panels at different institutions require, and of course the nature of the experiment is key here in terms of the potential for any distress or harm resulting from participating in the experiment and the degree of personal information that is required. But it is worth considering how these stages of the experiment can be streamlined so that participants obtain all necessary information in an accessible and time efficient manner.

Summary

While the direct-recruitment approach has been highly successful in a small number of cases (Brysbaert et al., 2016; Coughlan et al., 2019; Coutrot et al., 2019; Guasch et al., 2022; Hartshorne et al., 2018), the success of these studies is due, in part, to the care taken by researchers to consider the motivation of their participants – participants in all these cases *wanted* to participate, and were willing to give up their time for free, most likely because of a combination of the factors discussed above. In cases where researchers wish to take this approach to psychological research, there is much that can be learned from the broader 'citizen science' literature. For example, interviews with participants in a conservation-focused citizen science project (Rotman et al., 2012) indicated that initial participation was primarily driven by a desire to take part in studies that would interest and educate them (i.e., knowledge seeking), but that their continued participation over longer periods of time was influenced by a wider set of factors such as whether they felt their contributions would be worthwhile (i.e., altruism). This study also indicated that providing appropriate feedback about the project and recognising the contributions of participants was key to maintaining motivation. There is also evidence that the relative contributions of these factors to participants' motivation also varies across populations (see Li et al., (2018) for specific discussion of what motivates older adults and people with disabilities to take part in online studies).

In general, I would strongly caution researchers against expecting 'data for free' on (relatively dull) standard experimental paradigms where participants have no strong intrinsic motivation to participate. For more traditional cognitive psychology tasks where participants are not driven to participate by altruism, knowledge seeking or entertainment it is, in my opinion, more appropriate (and time-efficient) to reward online participants by payment, perhaps recruited via

established crowd-sourcing platforms.

Finally, it is worth considering that in cases where participants are highly motivated they may be will to provide highly valuable input to our research beyond completion of our tasks. Oliveira et al. (2017) analyzed open ended comments from 8,288 volunteers who took part in online experiments on the experiment platform Lab in the Wild (https://www.labinthewild.org/). They found that some participants were highly motivated to contribute to the research projects - making detailed and highly appropriate suggestions for how the research project could be improved or extended. These findings suggest an opportunity to involve volunteer participants more broadly into our research as 'citizen scientists' that is currently largely untapped.

Participant diversity, naivety and the superworker problem

One clear, and perhaps unavoidable, limitation of traditional lab-based testing is its overreliance on unrepresentative, non-diverse undergraduate populations who are routinely oversampled due to their easy availability for on-campus testing. There is now strong evidence that participant pools such as MTurk are more representative of the U.S. population than typical in-person convenience samples (e.g., Berinsky et al., 2012; Woods et al., 2015). Indeed this desire to sample the population more broadly has been a driving force in many researchers' choice to switch to online recruitment. (See Henrich et al., 2010; Rad et al., 2018 for broader discussion of overreliance on Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations.)

But online recruitment does not automatically ensure recruitment of diverse pools of participants. It is still worth giving thought as to how our particular recruitment strategy might result in over (or under) sampling of particular participant demographics (De Man et al., 2021). In the early days of online data collection, when internet access was not widespread, researchers were concerned that the web-based recruitment might oversample "techies", or "social rejects and loners with no other outlet for social contact" (Gosling et al., 2004). Although the increased prevalence of internet access has largely alleviated these specific concerns, it is almost certainly the case that we continue to oversample particular groups within society. The extent to which this is problematic for any given study will depend both on its aims and recruitment strategy. While much is known about the typical demographic characteristics of participants on large crowdsourcing platforms such as MTurk (see Chandler & Shapiro, 2016 for review; see https://www.mturk-tracker.com for up-to-date pool characteristics), when researchers use more bespoke recruitment strategies, such as those that rely on citizen science approaches, they will need to consider how this choice will impact on their likely pool of participants, and perhaps require them to collect extensive demographic information.

Perhaps the most important concern about the characteristics of participants who are recruited online, particularly from large crowdsourcing platforms such as MTurk, is that researchers may have become overly reliant on small cohorts of 'superworkers' who participate in large numbers of studies to create a steady stream of personal income (see Robinson et al., 2019 for comprehensive review). The impact of this problem is difficult to quantify. While it is clear that repeated testing on the *same* task can reduce effect sizes (Chandler et al., 2015), the consequences of repeated participation across *different* tasks is unclear. In addition, the prevalence of 'superworkers' on different platforms is also unclear. Crowdsourcing platforms provide variable levels of information about the size of their participant pool. Prolific recently advertised having more than 130,000 active participants (https://prolific.co/prolific-vs-mturk). While Amazon advertises more than 500,000 registered workers the number of these workers who are currently active is likely to be considerably smaller. Different approaches have been used to estimate the size of the MTurk active worker pool and have provided widely different answers to this question. Robinson et al., (2019) base their estimate on a large dataset acquired by TurkPrime, an independent company that facilitates access to the MTurk database (Litman et al., 2017). Based on the subset of MTurk participants that were based in the US and were recruited via TurkPrime between 2016 and 2019 they estimated that there were *at least* 80,000 to 85,000 active US MTurk workers per year, and that within any given year more than half of these workers were new to the platform, with approximately 4,600 new workers joining the pool each month (Robinson et al., 2019). This estimate is higher than some previous suggestions that "the average lab has access to less than 10,000 workers in any three-month span" (Stewart et al., 2015), but is more closely in line with other estimates (Difallah et al., 2018).

While these relatively high estimates for the number of active workers may give some reassurance, it does not give a full picture of the superworker problem. Importantly, Robinson et al.'s (2019) analysis indicates that a relatively small subset of MTurk workers complete a disproportionate proportion of experiments: over a 3-year period just 5.7% of the US workers made up close to half of the participants each year. Robinson et al., (2019) suggest that a primary cause of the superworker problem is the way that researchers choose to recruit participants: in order to improve the quality of their data they restrict access to their experiment to workers with an established reputation as 'good workers' by, for example requiring participants to have a 95% approval rating and have previously completed at least 100 HITs. Robinson et al., (2019) suggest that these selection practices routinely lock out close to 35% of less experienced MTurk workers and argue that this practice of restricting studies to the same group of experienced workers exacerbates problems of non-naivety.

Robinson et al., (2019) explored the impact of such recruitment restrictions on worker naivety and data quality. They compared performance of a group of participants with "standard" worker qualifications (at least 95% approval rating; more than 100 HITs) to an "inexperienced" group (less than 50 HITs) and an "open" group that was open to all workers. Results show that the "standard" cohort was highly similar to the "open" sample in terms of their approval ratings and number of previous HITS, and that they performed similarly on a wide range of measures of data quality (attention checks, internal consistency, experimental effect sizes). Of particular interest are the findings that the "inexperienced" group, who reported less prior exposure to experimental manipulations than workers in the standard or open samples, also performed well in terms of data quality measures. These results suggest that including relatively naive workers in a study will have minimal impact on the outcome of any individual study, while allowing new individuals to more easily refresh the pool of workers. Robinson et al., (2019) suggest that while targeting highly experienced participants may be beneficial for experiments that are particularly long or complex, or include multiple testing sessions, for many routine experiments there is no compelling reason to exclude inexperienced workers. Other studies provide mixed evidence on this issue: while Peer et al., (2014) found that MTurk participants with high approval ratings (>95%) were significantly more likely to pass attention check questions and produced data with higher reliability values on standardised personality questionnaires, Peer et al., (2022) found that approval rating did *not* predict data quality.

Importantly, as with other aspects of experimental design, the issue of superworkers should be considered in the context of a researcher's specific experimental tasks and aim: the extent to which superworkers should be considered problematic will differ widely across tasks, depending on how repeated exposure to similar experiments might influence participants' behaviour due to either their familiarity with the tasks themselves (e.g., practice/boredom effects) or to insights gained either during the experiment itself or during post-experiment debriefings that might influence their future behaviour. For example, lack of naivety may be of particular concern in designs that include some element of deception – experienced participants may, for example, be more likely to suspect the presence of a 'surprise' memory test at the end of an experiment, or to spot connections between apparently unconnected elements of an experiment. While it is often possible to exclude

participants who have recently performed highly similar experiments from *within* a research group, we usually have no information about participants' other experimental experience. In cases where the conclusions we wish to draw from our data can potentially be undermined by participants' prior experience, it may be useful to explicitly ask participants about this experience, either with respect to particular experimental tasks or with experimental research more generally. Although this approach will never allow us to be certain about exactly what other experiments individual participants have completed, exploratory analyses may reveal whether, at a group level, reported experience has a substantial influence on performance.

Reliability of participant demographics

When we don't have the opportunity to meet our participants in person it can be difficult to check even the most basic demographic information. The most extreme version of this concern is the worry that some of our data may be coming from bots, also known as automatic survey-takers, or fraudsters. In most cases, at least for typical experimental psychology paradigms, bots can be detected using the range of tools discussed below to assess the quality of participant data, for example by looking carefully at the time taken to complete different task components as well as trial-by-trial reaction time data (e.g., Storozuk et al., 2020; Teitcher et al., 2015). But even setting aside concerns about bots, it can be difficult to verify basic information from genuine participants, such as their age or whether they have already taken part in your experiment. The extent to which we should be concerned about issues to do with participant identity will, as with most aspects of experimental design, be experiment specific. For example, when comparing monolingual to bilingual participants, we must be confident that our participants do indeed fulfil our language background requirements in order to be confident in any inferences that we draw from their data. This is particularly important in the case of any null findings in order to ensure that the absence of a group difference doesn't simply reflect our failure to reliably classify participants. For other experiments, it may be particularly important to be certain that we have accurate information about participants' ages, either because this is a variable of interest, or because for ethical reasons we need to be confident that we are not testing inappropriately young participants. Finally in most experiments we want to avoid resampling participants, especially for designs that contain surprise elements such as an unexpected memory test, or where the aim of the experiment becomes clear (either explicitly or implicitly) towards the end of the procedure.

Many of the crowdsourcing platforms that provide access to large pools of participants, implement various forms of participant screening such that they give the impression that at the click of a button we can ensure that all our participants meet some pre-specified demographic requirements, such as age or language background. I *strongly* caution against relying too heavily on this out-sourcing of this important aspect of data quality and would recommend additional within-experiment recording of participants' demographic information, especially when this information is key to the aims of the research (See Pavlick et al., 2014 for discussion related to reporting of language proficiency).

When deciding how much to trust (or not) the demographic information provided by participants (either via the recruitment platform or from our own questionnaires) it is important to consider *why* this information may be inaccurate. One possibility is that participants are explicitly and deliberately misrepresenting themselves. To assess the likelihood of this possibility we should carefully consider how participants' behaviour may be influenced by the reward structure set up by your recruitment approach. For example, relatively high rates of payments might incentivise participants to misrepresent themselves in order to meet your recruitment criteria, for example by lying about their age or language background (Pavlick et al., 2014), or about whether they have previously taken part. Similarly, if your experiment is perceived as fun or has been widely circulated on social media, but has a minimum age requirement (e.g., due to ethical considerations) then some younger participants may give a false age in order to access your experiment. Finally, in the case where you explicitly provide feedback on performance, or some other insight that may be of interest to participants, they may choose to take part repeatedly in order to better their score or evaluate the consistency of the information that you provide them. In cases where such incentive structures exist, greater care should be taken to verify participants' information. In contrast, in the absence of such incentives, it seems unlikely that significant numbers of participants would choose to misrepresent their identity.

In addition to deliberate misrepresentation, demographic information can be misleading because participants fail to adequately understand what is being asked of them. This can be particularly problematic when it comes to information about participants' language background, where researchers may need to be precise in specifying what they mean by 'native language' or 'bilingual', which may be used inconsistently or without sufficient explanation on the screening forms of the recruitment platforms. At the very least, researchers for whom this information is critical should ensure that they know exactly what precise wording was used by the platform's screening forms. Ideally, this information should be supplemented by questionnaires embedded within the task in order to obtain the precise information that the researcher needs. Note that this can cause (minor) difficulties when the information provided directly by the participants does not match the information provided by the platform. Specifically, some platforms will not permit you to use your internal data to exclude participants from payment on the grounds that they don't meet your conditions for participation. But of course this doesn't prevent you from excluding data from your analysis based on additional within-experiment screening. When including multiple-choice demographic questionnaires experimenters should include clear definitions for any terms that may be interpreted differently by participants. When using open-ended questions it is helpful to give examples of typical answers to help participants easily understand what is required. Particular care and sensitivity is needed when asking questions about participants' sex, gender or gender identity.

Finally, in cases where this demographic information is critical to your experimental aims, I'd recommend including speeded tasks that can help to confirm information provided in demographic questionnaires (e.g., speeded vocab test to confirm language background; questions about local culture to confirm residency information).

Selective attrition

One final recruitment issue that can have potentially devastating consequences for the ability of any given experiment to address the intended research question is selective attrition, i.e., non-randomness in participant dropout. Note that I refer here only to dropout that occurs *after* participants have started the experiment. In particular, I focus on the critical case where participants may choose to drop out after they have been assigned to an experimental condition, and where participants are more or less likely to dropout depending on which condition they have been assigned to. The (related) issue of how participants' decisions about whether or not to sign up for an experiment in the first place can affect the demographic representativeness of participant samples is dealt with in the earlier section on Participant Recruitment.

Historically, the issue of selective attrition (also known as 'attrition bias') has been more carefully considered in the context of clinical trials, epidemiological studies and long-term longitudinal studies, where significant numbers of participants are likely to withdraw from the study over time (see Graham, 2009 for overview). In contrast, single-session cognitive studies have historically largely ignored the potential pitfalls of selective attrition due to their relatively low dropout rates. Researchers have typically assumed that any impact of non-random dropout are likely to be minimal.

Unfortunately, dropout rates in *online* experiments can be dramatically higher than for in-person experiments. Zhou & Fishbach (2016)

Table 3
Building a positive, collaborative relationship with participants.

Recruitment
Set appropriate payment rates
Honest descriptions of duration and nature of task (especially if long/dull)
Before Experiment
Reduce anxiety by ensuring instructions and expectations about performance are clear
Reduce anxiety by allowing access to instructions after practice trials
During Experiment
Avoid unnecessary trick questions, also known as Instructional Manipulation Checks[1], that may be perceived as distrustful or antagonistic
End of the Experiment
Thank participants and explain why your research is important
Feedback on participant performance (if appropriate and ethics permits)
Direct to relevant podcasts, blogs etc.
Debrief participants on their experience with the tasks, and any other thoughts
Ask participants for suggestions for future similar experiments
Reply promptly to queries
Avoid rejecting participants' payment claims due to poor performance

1. Oppenheimer et al., 2009.

compared participant dropout in a set of 88 online social psychology studies conducted via MTurk with 82 similar single-session lab studies. Dropout rates were high in the online experiments: virtually all (99%) had non-zero dropout rates, and over 20% had dropout rates above 30%. In contrast for the lab-based studies 96% had no dropout, and the highest dropout rate was 4.7%. Zhou & Fishbach (2016) suggest that this difference reflects a combination of the higher sunk costs associated with in-person experiments (e.g., arranging the appointment, travel time etc.), as well as the social costs of withdrawing part way through (e.g., awkwardness, embarrassment). The prevalence of distractions and interruptions in online participants' environments may also play a role. More recent data from the Gorilla Experiment Builder platform suggests that, for studies conducted in 2022, only 32.5% of the participants who were recruited did not complete the whole study successfully, either because they dropped out or were excluded by the researcher at some stage during the study (Tomczak et al., 2023).

Importantly, Zhou and Fishbach (2016) noted that researchers may often be unaware of the extent of participant dropout because some online software may only upload participant data to the server once the experiment is complete, so researchers may not have access to information about how and when additional participants may have dropped out. Some software (e.g., Qualtrics) requires researchers to explicitly alter default parameters to ensure that incomplete responses are included. Other software (e.g., jsPsych) is intrinsically agnostic to when data should be downloaded, and gives researchers the freedom to configure this aspect of their tasks as they see fit. Even in such cases, problems can arise when researchers make use of existing tutorials or shared code that may use a one-shot data upload model, perhaps without realising that this choice has been made without their consideration. Indeed when Zhou and Fishbach (2016) replicated six published social psychology experiments they found overall dropout rates exceeded 30% on all six tasks (31.9% to 51% across studies), which was in stark contrast to the original studies which reported *no* information about participant attrition.

It seems clear that dropout rates for online experiments are both high and potentially under-reported. But does this matter? For experiments with carefully controlled within-participant designs where no inferences are being made about the participants being representative of the broader population, it may be relatively safe to assume that dropout can safely be viewed as an irritation that can be fixed by recruiting additional participants, and that is unlikely to systematically distort the effects of the key experimental manipulations. In contrast, selective dropout in between-participant designs can be potentially catastrophic. These designs rely on the assumption of random assignment of participants to different experimental conditions to ensure that any between-condition differences are unlikely to reflect pre-existing differences in the groups of participants that were assigned to these conditions. Most existing experimental paradigms make it relatively straightforward to achieve random assignment at the *start* of an experiment by randomly allocating each participant to a particular condition. However, in cases where the likelihood of any given participant completing the experiment depends on the condition to which they were assigned, this assumption will no longer hold. This can arise, for example, because one condition has been set up to be longer, less interesting, or more difficult. In such cases, it is likely that the characteristics of the participants who complete each condition will differ, for example with the more 'difficult' conditions ending up with participants who are, on average, more motivated or conscientious, or who are more skilled at the particular task being performed. As with many aspects of experimental design, the impact of selective attrition can be highly experiment-specific. For example, experiments that include a time-of-day manipulation (e.g., sleep studies) may find that participants are more likely to comply if they are randomly assigned to same-day condition compared with an over-night condition. Selective attrition can result either in false positive or false negative results, which might appear highly replicable if the observed patterns of selective attrition are consistent across experiments. Indeed, Zhou & Fishbach (2016) demonstrated that, in an extreme case, selective attrition can result in highly counterintuitive 'reverse' effects. They compared participants' ratings of the perceived difficulty of recalling either many (12) or few (4) happy events from the past 12 months. Counterintuitively, recall was, on average, *less* effortful in the 'many' condition, which they attribute to selective attrition, suggesting that the relatively small subset of participants (31%) who persevered with this challenging condition had self-selected to be individuals for whom recalling happy memories was relatively easy.

A range of possible solutions are available to the selective attrition problem. First, it is advisable, where possible, to reduce overall rates of attrition, such that the impact of these issues is minimised. In addition, researchers can boost participants' motivation to complete the task by, for example, increasing payments or highlighting the importance of complete datasets. In particular, for multi-session experiments it can be advisable to include a well-publicised 'completion bonus' where participants receive a disproportionately large payment for completing the last component. A complementary approach is to accept that some level of attrition is inevitable but to try and ensure that it happens *before* participants are assigned to a particular experimental condition by, for example, warning participants more clearly about aspects of the study that might lead them to withdraw part way through. Researchers should be honest that parts of the task will be difficult or tedious, and should ask participants to explicitly confirm that they will complete the experiment

despite these issues.

Regardless of the inclusion of such strategies, researchers must be transparent when reporting participant dropout. First they should ensure that the experimental software is recording all partial data. Not only will this make any selective attrition visible, but it can provide useful insights into when and why participants are withdrawing, such that this can be addressed in future experiments. Second, researchers should ensure that details of any dropout are reported clearly, broken down by experimental condition, so that readers can consider the potential impact of selective attrition.

General considerations

Although online experiments inevitably result in increased separation between a researcher and their participants, I'd encourage researchers to try and overcome this perceived distance and make explicit attempts to build a collaborative relationship with their participants by treating them as a valued part of the research team (See Table 3). Although it is likely that each individual researcher may only test any given online participant on a single occasion, the way that we treat our participants will impact not only the data we collect during that session, but their willingness to participate in our colleagues' future experiments and the time and attentiveness that they give to any such experiments.

Participant performance

Even if you have done everything that you can to ensure that participants are using appropriate tech (hardware and software) and that your recruitment strategy has provided a pool of suitable participants, things can still go awry once any individual participant begins the experiment. The following sections review a number of (related) issues that can arise during the experiment, such that participants fail to perform your task in the way you had intended, and thereby reduce (or even destroy) your data quality. In most cases these are exacerbated versions of issues that arise when participants complete tasks in the lab. I summarise some of the steps that you can take to minimise the impact of such issues in Table 4.

Table 4
Checklist for researchers: Issues affecting participant performance.

Potential Issue	Experimental Considerations	Potential Solutions				
		Piloting	Sign-up & Instructions	During Tasks	Post-Task Debrief	Data Analysis
Misunderstanding of Instructions	Is the task complex? Might participants have completed similar tasks? Will performance data always show if they misunderstand task?	Open ended question: "What did you understand your task to be"	Avoid jargon (e.g., "lexical decision", "response times") Video instructions for complex tasks Dynamic/annotated instructions Practice block, with feedback and chance to reread instructions Instruction check question(s) before task	Reminders of button assignments (e.g., "press X for yes")	Open ended question: "What did you understand your task to be"	Check time taken to read instructions
Participants finding task stressful	Are low levels of performance expected?	Open ended question about experience	Emphasize expectations in instructions		Open ended question about their experience	
Sustained Attention During Task	Would inattentiveness add noise or introduce systematic bias?	Evaluate changes in performance across time	Honest estimate of task duration Emphasise the importance of research Tick-box to confirm time available and quiet environment	Easy filler trials Instructional Manipulation Check[1] Optional breaks	Report disruptive interruptions	Overall duration of all task components Trial-by-trial RTs
Explicit Cheating	Are answers available online? (e.g., vocab test) Can note-taking aid performance (e.g., memory test) Any incentive to cheat? (e.g., exclusion from future studies)		Emphasize expectations: "we don't expect perfect performance" Emphasize importance of genuine data to research	Short trial time-outs Record all RTs	Opportunity to report cheating (payment guaranteed)	Confirm well-established data patterns (e.g., serial-position effects)

Note that the importance of these different aspects will vary widely across experiment. Not all measures are needed for all experiments.
1. Oppenheimer et al., (2009).

(Mis)understanding of task instructions

For data to provide an appropriate test of our experimental hypotheses we need to be confident that all participants have understood the task instructions. We need to be sure that they have not missed key elements of how they should approach our task (Oppenheimer et al., 2009). Even minor misunderstandings can (at best) add to the noise in our data signal requiring us to expend additional resources testing more participants, or (at worst) add systematic bias that can compromise any conclusions that we wish to draw. In addition, a lack of confidence as to whether they are 'doing it right' may increase participant anxiety leading to variability in performance and a more negative experience. This concern is not unique to online experiments, but is exacerbated compared with in-person testing where participants usually have an opportunity to ask questions prior to starting the experiment, and experimenters are more likely to notice uncertainties or anxieties. Online participants may also be less experienced with standard experimental tasks than typical in-person undergraduate participants.

Importantly, it is often surprisingly difficult to determine from task data alone whether participants have failed to interpret the instructions in the way that we intended. Such misunderstandings can result in relatively subtle shifts in behaviour that may not be immediately apparent. For example, if a participant is explicitly instructed to remember stimuli in order to be able to later recall them, poor performance can arise either because of a failure to notice that they were supposed to try and remember the items, or from genuinely poor memory ability, or from low levels of attention during the learning/recall periods. Similarly, if asked to rate stimuli on a scale (e.g., semantic relatedness), participants might be rating on some other factor that is sufficiently similar to our dimension of interest that their misunderstanding is not immediately clear from their data.

Unfortunately, even when great care is taken to make our instructions clear, confusion can arise either because participants rush through reading the instructions or because of pre-existing expectations about what they think their task should be, perhaps based on experience with other similar tasks. I recommend a two-pronged approach to ensuring adequate understanding of instructions: do everything you can when setting up your experiment to ensure that all participants understand your instructions, and then build in safeguards to ensure that you can identify and exclude any participants who have misunderstood these instructions.

The primary challenge when optimising task instructions is to overcome our own familiarity with the tasks and ensure that they are completely clear to participants with no previous knowledge of our paradigm. We need to avoid any jargon that may be unclear to naïve participants (e.g., "lexical decision"; "grammaticality"). Avoiding overly long instructions, which are more likely to be skim read, is also advisable. In the case of complex multi-component tasks that unfold over time or when recruiting participants that may find written instructions challenging (e.g., children, poor comprehenders), it may be worth using video demonstrations to indicate what the task will look like and how/when they should respond. It can also be beneficial to make instructions dynamic. For example, instructions can be divided into smaller chunks that are presented to participants in sequence in response to button clicks, or pop ups can be used to annotate a screen shot of an experimental trial in order to highlight any crucial elements. In-person piloting of instructions with naïve participants can be highly beneficial.

The inclusion of practice trials (often with explicit feedback) is routinely used in most experiments, but often these are set up in a way that assumes that participants will easily pass the practice trials and may not take account of how best to assist participants who remain confused after these trials. Participants should have the opportunity to revisit the instructions between any practice block and the start of the experimental items. In addition, for any complex tasks where there is significant possibility of misunderstandings, I recommend including a short questionnaire at the end of the instruction phase which tests knowledge of the task, and which requires participants who answer any of these questions incorrectly to read the instruction again before starting the task (see Crump et al., 2013 for evidence of the effectiveness of this approach). It can also be useful, after data collection, to check the time taken by participants to read instructions in order to identify any participants who skimmed (or did not read) this information. Finally, some researchers have found it useful to ask participants at the end of the experiment to describe in their own words what they 'understood their task to be'. This can provide insights into whether their interpretation of the task, perhaps in relatively subtle ways, may have shifted during the experiment, and help to understand any unusual participant performance. This information can also be used to guide our decisions about removing participants from our analyses if they have misunderstood the mechanics of the task.

We should also be clear about any expectations that we might have about the level of performance that we expect. For example, vocabulary and working memory tasks routinely include difficult items to avoid ceiling effects. In such cases it is important to tell participants that we don't expect perfect performance in order to alleviate any potential stress and to avoid participants giving up or looking things up online when things get unexpectedly difficult. In particular, in experiments that use the staircase method to adjust item difficulty to a participant's performance threshold (Cornsweet, 1962) we should emphasise that perfect performance is impossible. Participants who usually perform well on experimental tasks can be surprisingly disconcerted when they find that the task difficulty is higher than expected.

Participant attentiveness

A key concern amongst researchers is whether online participants will devote sufficient focused attention to their task. This concern has been extensively explored within the field of survey design over many years, with the theory of *satisficing* providing a useful framework for exploring the factors that can influence the choices that participants make about the cognitive resources that they devote to our tasks (Krosnick, 1991; see Roberts et al., 2019 for recent review; see Couper, 2011 for broader historical perspective on the influence of mode of delivery on survey data collection).

There are several reasons that participant attentiveness may be lower for online experiments compared with lab-based tasks. In contrast to the relatively sterile environment of a psychological laboratory, online testing permits experimental settings that may be noisy and distracting. The potential scale of this issue is illustrated by Clifford and Jerit (2014): when 435 undergraduate students were randomly assigned to complete a political survey either online or in the lab, the online group reported higher rates of phone use (21% vs. 1%), internet browsing (11% vs. 1%) and talking to another person (21% vs. 2%). The online group also reported relatively high rates of watching TV (14%) and listening to music (20%). Interestingly, responses to relatively tricky catch trials showed no differences between the two groups, suggesting that these distractions had little impact on their ability to complete a survey. However, this conclusion will likely *not* apply to more demanding speeded tasks: reaction time measures are inevitably more strongly impacted by external distractions.

In addition to the external distraction factors that may be more prevalent for online participants, the physical presence of a human researcher for lab-based experiments may serve as a behavioural nudge to remind them to maintain their attention (e.g., see Oppenheimer et al., 2009 for evidence of increased attentiveness for proctored surveys). In addition, lab-based participants have often made a substantial time/travel commitment when deciding to participate, which may act to increase their perception of the importance of their data and their sense of commitment to a real life researcher.

It is therefore important to build safeguards into our data collection and analysis procedures to deal with this issue. As with the approach to experimental instruction set out above, these safeguards come in two

forms: we should set up experimental procedures that maximise participant engagement (e.g., short, entertaining experiments that emphasise the importance of the data), while also ensuring that we can adequately identify any participants who have not adequately engaged. As was discussed with respect to variability in participant hardware/software, it is important to consider whether, for any particular experiment, participants' inattentiveness can be treated as an additional source of random variability that might be largely overcome by increasing the number of participants. This is less problematic than when lack of attention can result in the introduction of systematic bias into the dataset that could undermine the conclusions that we wish to draw from our data.

There are some very general safeguards that we should employ. First, we should not take for granted that our participants share our assumptions about what constitutes an appropriate experimental environment. Participants can be unaware that we are often interested in the very precise details of the timings of their responses and so can underestimate the impact of distractions on the usefulness of their data. I recommend including questions in the set-up stage of the experiment where participants are asked to explicitly confirm that they are in a suitable, quiet environment where distractions are unlikely (e.g., that they don't have any other windows open on their computer and have turned their phone to silent). It can sometimes be difficult to set reasonable expectations for participants – for example for short tasks it might be appropriate to ask them to turn off phone notifications, but participants are unlikely to agree to this for longer experiments. Similarly, it can be beneficial to ask participants to report at the end of the task if their performance was adversely affected by any interruptions. As with all such debriefing, we should encourage honesty by stating that participants' responses will not influence their payment.

Some researchers advocate embedding tricky, unexpected instructions, often referred to as 'Instructional Manipulation Checks' within either the initial task instructions or throughout the tasks to identify any participants who are not paying sufficient attention. For example, Oppenheimer et al., (2009) embedded an instruction to ignore a highly salient task demand (to press a large red button marked 'continue') and instead to click on the title of the screen. They report that a large proportion of participants (46%) failed this check, and that removing these participants from the sample improved the quality of the data. Perhaps surprisingly, Hauser & Schwarz (2016) found evidence that MTurk participants were more attentive to such instruction checks than college students. Peer et al., (2014) reported that high-reputation MTurk workers rarely failed such checks and that their inclusion only improved data quality for low-reputation MTurk workers. However there are concerns that such 'gotcha' questions don't just *measure* participants' attention, but instead actively change how participants approach tasks. By implicitly teaching participants that there is "more than meets the eye" to the tasks they are being set, we may encourage them to try and avoid future traps by approaching subsequent trials with a more analytic or reflective mode of thinking than may be desired (Hauser & Schwarz, 2015). More broadly, there is a concern that 'trap questions' can undermine the collaborative relationship between researcher and participants by highlighting our distrust of them. My view is that we should be extremely cautious when introducing such measures. My preference, for many experiment types, is instead to include occasional very *easy* trials on which we can confidently expect near-ceiling performance in our target population, such that we can exclude any participants who fail on these trials. For example in multiple choice vocabulary tests we included fillers with extremely low age-of-acquisition.

It is also sometimes possible to record additional informative data about participants' behaviour during the task that can provide useful insights into their multi-tasking behaviour. For example, jsPsych can provide information about whether the user has clicked on other non-experiment windows, or has exited full screen mode during the experiment (https://www.jspsych.org/7.3/overview/record-browser-interactions/).

Finally, as noted earlier, many highly successful online experiments have 'gamified' their tasks. Not only does this help with participant recruitment, but by making our experiments more entertaining and engaging we inevitably make it more likely that participants, from a wide range of backgrounds, will devote significant sustained attention to our tasks (Long et al., 2023). Similarly, increasing participants' motivation by highlighting the value of their data to particular societal issues that may be important to them (see above) can help encourage high levels of sustained attention.

Explicit cheating

For some experiments we need to be concerned about explicit cheating, such as taking notes in a memory task or searching online during tests of crystallised knowledge (e.g., vocabulary). A demonstration that we cannot always take participants' honesty for granted come from a survey of political knowledge conducted by Clifford and Jerit (2014) who compared results with a lab-based cohort and found clear evidence that the online participants specifically boosted their scores on those factual questions that could have been looked up online.

Fortunately, many of the tasks that are routinely used in cognitive psychology involve making speeded responses where there is insufficient time for participants to make use of external sources of information. In addition, many tasks that are typically delivered in a non-speeded manner (e.g., vocabulary tests) can be modified to a speeded format to either prevent cheating or make any such cheating evident from the response time data. For memory tasks, it is harder to completely rule out the possibility that participants may have cheated. Options include honesty checks that the end of the experiment, making clear that participants' payment will not be adversely affected, and including checks of participants' data for well-established patterns that might be absent if participants have not relied on their own memory (e.g., recency/primacy effects, serial position effects). These safeguards will necessarily be highly paradigm specific and will need careful piloting, perhaps in the lab where cheating can more easily be prevented.

As with the earlier discussion of how to avoid participants misrepresenting their age or language background, it is always worth thinking carefully about the reward structures that might lead participants to cheat. In particular, if your recruitment platform excludes participants with poor performance from future experiments (or if participants *perceive* this to be the case), this may incentivise cheating as participants aim to avoid being excluded from future earning opportunities. Thinking carefully about such reward structures can allow us to adjust task instructions to alleviate any such concerns and emphasise the importance, from our perspective, that their data provides an accurate reflection of their abilities and our expectation that we do *not* expect perfect performance. However, it is important to note that for some experiments, it will simply be impossible to be entirely sure that participants haven't cheated, and if this assurance is critical to the experimental aim then such experiments will have to be conducted in person.

Finally, our assumptions about what constitutes 'cheating' may well not align with our participants' assumptions. Participants may previously have taken part in data-entry style MTurk tasks for which looking up information online or taking notes is entirely acceptable. If there are particular behaviours that we want to discourage we should explicitly state these in the instructions, perhaps explaining why this is important for our research.

General advice

The above list of potential pitfalls and solutions is not intended to be exhaustive. Indeed a recurring theme in this article is that many of the issues that can reduce data quality are highly specific to particular paradigms. Therefore, it is important that researchers maximise their

Table 5
Reasons to Exclude Data from Analysis: Checklist.

	Reasons to Exclude	Examples
Technology (See Table 1)	Failure to meet baseline tech requirements	• Experimental software indicated used phone/tablet not laptop/desktop • Failed headphone or audio quality check • Self-reported not using headphones • Self-reported not hearing stimuli clearly • Failed visual acuity check • Evidence of variable trial-to-trial event timing • Self-reported glitches in timing
Recruitment (See Table 2)	Failure to meet demographic requirements Evidence of inaccurate demographic information Evidence of experience with similar experiments	• Self-reported native language, age, or country of residence • Failed vocabulary test aimed to confirm native language • Answered questions about demographics inconsistently • Self-reported having completed similar experiments
Participant Performance (See Table 4)	Evidence that instructions were misunderstood Evidence of low/inconsistent attentiveness Evidence of cheating	• Inaccurate description of task instructions at debrief • Self-reported ignoring aspect of instructions at debrief • Short time reading instructions • Low performance on easy filler trials • Low performance on instructional manipulation checks • Low/variable performance on main task • Long total experiment duration • Long delays between or within tasks • Unfeasibly fast response or reading times • Absence of well-established data patterns (e.g. serial order effects, frequency effects) • Self-reported distractions/interruptions/multi-tasking at debrief • Self-reported cheating

opportunities to discover any unexpected, idiosyncratic issues that have affected their specific experiment. I therefore make two very general recommendations.

First I strongly advocate that all online tasks be set up to collect data about how long participants take to complete all its constituent elements. Even for questionnaire-type elements of an experimental procedure where reaction-time data is unlikely to be formally analysed, it is important to get a sense of how long participants are taking to (i) read the instructions and (ii) respond to each item. It is also helpful to get a general sense of when/if they are taking unscheduled breaks. This data can provide important insights into how participants are approaching your task. In addition, trial-by-trial reaction time data (both durations and variance) can provide a wonderfully rich source of information about participants' performance across the duration of the experiment. As set out in the later section of preregistration, it is important to think carefully about the likely timings of participants' performance (at the levels of both task and trials) *before* data collection, setting out as precisely as possible our expectations about how attentive participants should proceed through our task(s).

Second, the inclusion of open ended questions at the end of the experiment can be highly beneficial and give important insights into participant behaviour. Specifically, I recommend routinely asking participants whether anything unexpected happened during the experiment, or to report any issues that they think may have impacted their performance. Additional helpful information can be obtained by indicating to participants that you plan to conduct similar experiments in the future and to ask for their suggestions for improvements. In part these questions aim to recreate the kinds of informal interactions with participants that happen after in-person testing sessions that tend to build up researcher knowledge about how their tasks operate from a participant's perspective. They also help to foster a sense that your participants are playing an active role in your research community and that their views and opinions are valued (Table 3).

Preregistering exclusion criteria

No matter how careful you have been when setting up an experiment, it is inevitable that some data will need to be excluded from the analyses that address your experimental hypotheses. Data should not be excluded because it is in some general sense 'low quality' but because it has some very specific characteristic that makes it inappropriate with respect to your specific research question. The key to safeguarding data quality is to specify very precisely the conditions under which data can legitimately be excluded, and the reasons for these exclusions. In general, the reasons for excluding data map onto the three areas of concern described above: we typically exclude participants if (i) their technology does not meet our requirements, (ii) they do not meet our demographic requirements or (iii) they performed the task inappropriately (see Table 5 for an overview/checklist).

These decisions about data exclusions should be specific to the particular experimental paradigm(s) being used and to the particular inferences that a researcher might wish to draw from their data. While there may be some elements that become standardised across a set of similar experiments, it is important to consider these decisions in the context of the aims of the current experiment, and the extent to which any apparently problematic data might (or might not) undermine the specific conclusions that we might want to be able to draw on the basis of our analyses.

Decisions about exactly what data will be excluded from our analyses should, so far as is possible and practical, be made *in advance* of data collection. This is important for two related reasons. First, the process of specifying the different ways in which data might be 'problematic', will often lead to changes in the experimental procedure to ensure that we have sufficient information on which to make fully justified decisions about data exclusion. In other words, by making these decisions in advance of data collection, we run *better* experiments and obtain a better understanding of our participants' data. Second, this pre-emptive approach allows us to formally preregister our exclusion criteria. As with other forms of preregistration, this increases the extent to which readers will be able to trust the outcomes from reported analyses (Nosek

et al., 2018, 2019). As has been discussed extensively elsewhere (e.g., Munafò et al., 2017), typical analysis pipelines require researchers to make a large number of decisions, and this flexibility opens the door to systematic bias towards making choices that lead to statistically significant results. Formal preregistration reassures readers that decisions about data exclusion were made without knowledge of the observed data and thereby increases confidence that key findings are not the consequence of selective reporting or 'cherry picking' of those subsets of the initial dataset that are most neatly consistent with the researchers' predictions. This is of particular importance for online studies where rates of participant exclusion are often higher than comparable lab-based studies (see above).

Importantly, any decisions about whether data from individual participants should be included in an analysis should be separate from decisions about whether these participants should be paid. The latter decisions will (rightly) be governed by ethical considerations. I usually advocate paying *all* participants who participate in your experiments regardless of any indications that they may not have fully engaged with your task. It is extremely difficult, if not impossible, to distinguish participants who are capable of performing your task but have *chosen* to be inattentive from participants who, despite being part of your intended demographic sample found your task challenging, for example due to reduced comprehension, memory, or attention skills. This can be particularly problematic when translating a lab-based task where participants may be high-performing undergraduate students, to online recruitment approaches that may sample the distribution of cognitive skills more broadly.

The following section outlines a general, systematic approach to developing data exclusion criteria *prior* to data collection.

Stage 1: Specify experiment-specific data quality concerns

The first stage in developing data-quality exclusion criteria is to specify, in as much detail as possible, the primary data quality concerns for your specific experiment. These concerns may be based on pilot data, existing published studies, or on intuitions about participants' behaviour. Feedback from previous experiments can play a key role here, especially if participants had an opportunity to answer open-ended questions about their experience with the tasks. These discussions should be guided by our experimental aims, such that we focus primarily on data quality issues that could potentially compromise our ability to answer our primary research question(s).

As set out above the three areas where we might expect data collected remotely to be more problematic compared to more conventional lab-based approaches are:

- Technology (e.g., sound volume/quality, use of headphones, appropriate visual display, suitable internet connectivity etc.; Table 1)
- Participant identity (i.e., demographic details; Table 2)
- Participant behaviour (e.g., understanding of instructions, inattentive/variable performance, explicit cheating; Table 3)

Although these three areas of concerns will likely significantly affect *all* behavioural experiments, their likely impact will differ considerably across different experiments. For example, a researcher may be more concerned about the veracity of participants' demographic information if they are setting up an experiment designed to test for between-group differences (e.g., monolingual vs. bilingual; older vs. younger adults) compared with a within-participant design that includes participants with a broad range of demographics. Similarly, concerns about cheating would be particularly salient to a researcher using a working memory task where the answers could easily be written down and such note-taking would invalidate their results. It is therefore critical to be specific about how these issues might undermine the trustworthiness of the data from your particular paradigm, and to specify how these different issues could potentially undermine your ability to draw appropriate inferences from your data. It is also important to consider exactly how these concerns might impact your data. For example, as described above, it is important to consider whether participant inattentiveness will likely introduce random noise or systematic bias. By specifying these concerns *before* you focus on developing your exclusion criteria you are less likely to overlook a key aspect of data quality that could potentially undermine your overall experimental aims. In particular, specifying the 'worst-case scenario' at this point can focus your efforts on making sure that this outcome is avoided. And in the case that this outcome cannot be appropriately mitigated you might need to return to more conventional lab-based approaches, where participants and their behaviour can be more closely observed, or at least starting off your experimental journey in the lab in order to better understand participant performance before shifting recruitment online.

Stage 2: Design study-specific exclusion criteria

Once you have a clear idea of the potential data quality issues that are particularly worrisome for your particular experiment, the next stage is to ensure that your experimental procedure contains the necessary questionnaires, tasks or other elements that will allow you to identify (and then exclude) problematic data.

In some cases, the data from the tasks that you have developed to test your hypotheses may themselves provide useful exclusion criteria. For example, unusually high error rates, as well as slow or highly variable reaction times on your primary task, might strongly indicate that individual participants have either not understood the instructions or were not paying appropriate attention. Indeed, we would anticipate that most experiments would contain some exclusion criteria of this kind. But for most experiments, it is often beneficial to introduce new elements to the experimental procedure to improve our ability to reliably identify problematic data, and to help us better understand why particular participants may be showing unusually poor or variable performance.

One relatively straight forward approach can be to introduce additional trials into your existing paradigm. The most common form of this approach is to introduce trials that are designed to be sufficiently straightforward that you can reasonably expect all participants to respond correctly as long as they are attending to the task. Similarly, for some tasks it may be appropriate to ask the same questions twice, perhaps in a slightly different manner, to check for consistency of responding – inconsistent responses might indicate a lack of attention or deliberate misrepresentation (e.g., demographic information).

The final, and most time consuming, approach to ensuring you have sufficient information to be able to appropriately exclude problematic data sets is to introduce entirely new tasks or questionnaires into your experimental procedure in order to confirm some key information about your participants or their behaviour (e.g., vocabulary tests). This approach can be particularly worthwhile for those potential issues that could have particularly catastrophic consequences for the interpretation of your data.

Stage 3: Piloting of exclusion criteria

For many experiments, particularly those using new tasks or applying familiar tasks to new populations, it can be highly beneficial to pilot any planned exclusion criteria prior to data collection. For example, you may wish to reassure yourself that a pre-set level of performance on a screening task won't inadvertently exclude large numbers of eligible participants due to an overly strict performance threshold. In some cases it is optimal to pilot using highly trusted participants (e.g., researchers from outside your research team) who you have reason to expect to meet all your inclusion criteria and who are likely to engage appropriately with your tasks. This approach may provide particularly informative insights as to the feel of the experiment from a participant's view. However, in other cases it may be more appropriate to pilot using

participants who have similar demographics to those who will participate in the main experiment. This approach will allow you to more appropriately characterise the likely distribution of performance on your critical measures, and so may be more helpful in ensuring that your criteria are not overly strict. A final approach that can prove helpful is to run pilot studies face-to-face in the lab. This can potentially provide the best of both the two previous alternatives - it allows for recruitment of a more appropriate sample of participants while still providing an opportunity to closely observe task performance and conduct an extensive debrief that may uncover unexpected issues from the participants' point of view. As with other experimental design choices, decisions about whether and how to pilot your experiment will depend on many different factors that are specific to your current situation. In general, although it is easy to justify skipping this stage in order to proceed more quickly to data collection, this step has the potential to minimise the risk that you end up either excluding very large numbers of participants from the main experiment or have to significantly diverge from your pre-registered exclusion criteria (see below). Such pilot data will also allow you to increase the specificity of your pre-registered exclusion criteria (see below).

Finally, open-ended questions at the end of pilot experiments allow participants to report issues with your current procedure that you had not anticipated and can result in significant improvements to experiments. These questions are typically phrased in terms of asking participants to reflect on their own experience of our tasks, but I'd also advocate asking them to look forward and make suggestions for how we might improve future studies using these methods (Table 4; see Oliveira et al., 2017 for evidence that participants can provide highly varied, actionable feedback).

Stage 4: Preregister study-specific exclusion criteria

Once you have devised a set of exclusion criteria that you are confident will allow you to restrict your analysis to data from appropriate participants with appropriate tech, who have not cheated and have appropriately understood your instructions and attended to your task, these exclusion criteria should then be preregistered in a publicly accessible time-stamped preregistration repository such as the Open Science Framework (https://osf.io/) and AsPredicted (https://AsPredicted.org/). These criteria should be specified in as much detail as is possible. For tasks that have been extensively piloted or used in prior studies it can be optimal to express these criteria in absolute terms (e.g., minimum proportion correct, maximum number of time-outs), but for newer tasks or for familiar tasks being used with new populations, participant performance will be somewhat less certain so it may be necessary to specify these requirements using more general procedures that take into account the observed distribution in the current dataset (e.g., specifying cut-offs in units of variance). Preregistration can be viewed as a continuum - researchers make choices about how specific their preregistration document should be based on their level of certainty about the general characteristics of their to-be-collected data. These choices will come with consequences - the higher the level of detail in the preregistration document, the higher the level of reassurance that their readers will be given about the trustworthiness of their conclusions (Nosek et al., 2018, 2019). In my view there are very few experiments where it is *not* worth preregistering our exclusion criteria – even for relatively new tasks where we have significant uncertainty about participants' performance, it is usually worth specifying what we can, even if these criteria subsequently prove to be incomplete or insufficient.

Stage 5: Review exclusion criteria after data collection

For familiar paradigms that have been used multiple times within a particular population these exclusion criteria will likely perform as expected such that once they have been applied your final data set will only include participants from your desired population, who have performed your task with appropriate levels of attentiveness and without cheating. Equally importantly, you will be reassured that you have not overenthusiastically excluded participants in a manner that may have biased your results by excluding participants from within your target population who simply found your task more difficult than other participants. However, in some cases, initial data quality checks will reveal that the exclusion criteria were inappropriate or insufficient. This may arise because you had not foreseen a particular way in which participants might perform your task that is clearly inappropriate. For example, in an early experiment using an auditory word association task where participants were instructed to type in the first word that came to their mind in response to each target word we discovered a handful of participants had simply typed back the target word. It is also reasonably likely that your open ended questions at the end of the experiment may also elicit unexpected responses that lead you to decide that a participant's data may be unreliable. For example, a participant once reported treating an experiment as a social game performed collaboratively with friends. Of course, such cases will often already have been excluded on the basis of specific preregistered data checks (i.e., poor or variable task performance), but the possibility remains that you end up with participants who meet your preregistered inclusion criteria, but whose data is clearly inappropriate. In such cases, as with all aspects of preregistration it is completely acceptable to diverge from your preregistration in a clear, transparent manner setting out your reasons for making additional exclusions. (See Nosek et al., 2019; "preregistration is a plan not a prison"; https://cos.io/blog/preregistration-plan-not-prison/.) In the case where such issues affect substantial numbers of participants, it may be necessary to report analyses both with and without these troublesome participants in order to (hopefully) reassure readers that your decision to diverge from your preregistration is not responsible for a dramatic change to your findings.

Conclusions

Maintaining high experimental standards for online experiments takes time and careful thought. In the absence of face-to-face contact with our participants, we need to persuade ourselves (and our peers) that the data that arrived as if by magic via the click of a button is sufficiently trustworthy that it can adequately help to answer important research questions.

This paper reviews, in a non-exhaustive manner, some of the key challenges faced by researchers running online experiments related to (i) technology, (ii) participant recruitment, and (iii) participant performance. In all cases, I recommend a two-pronged approach to maximise data quality. First, I outline some of the many possible steps we can take when setting up our experiments to maximise data quality. Clearly, not all these measures will be needed for all experiments – we should carefully target our experimental interventions in the way that is most beneficial in the context of our current experimental aims and methods. Second, I argue that regardless of how much care is taken when setting up an experiment, it is inevitable that a non-zero proportion of the data that we collect should be excluded from analysis. I encourage the preregistration of exclusion criteria to allow us to reliably and appropriately identify any data that should be legitimately excluded from our analyses.

Finally, I reiterate that many of the data quality issues that arise when collecting data online are amplified versions of issues that also arise in the lab. Many of the lessons learned about to how to improve the data quality of individual online experiments should therefore be transferred back into the lab to improve our in-person research. For instance, the suggestions for improving participants' attentiveness and their comprehension of instructions are highly relevant to lab-based experiments. In addition, the lessons learned when considering sources of between-participant variability (e.g., participants' choices of software and hardware) can help us to understand between-lab

differences that result from *researchers'* technology choices. By being ambitious for *all* our experiments in terms of their experimental rigour and data quality, we can improve the validity and reliability of the data that we collect and enhance the quality of our science.

CRediT authorship contribution statement

Jennifer M. Rodd: Conceptualization, Writing – original draft, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgement

This work was funded by a grant from the UK Economic and Social Research Council (ES/S009752/1). I thank Becky Gilbert for her advice, suggestions and wisdom.

References

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., Babel, M., Bahník, Š., Baranski, E., Barnett-Cowan, M., Bartmess, E., Beer, J., Bell, R., Bentley, H., Beyan, L., Binion, G., Borsboom, D., Bosch, A., Bosco, F. A., … Zuni, K. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://10.1126/science.aac4716.

Angele, B., Baciero, A., Gómez, P., & Perea, M. (2022). Does online masked priming pass the test? The effects of prime exposure duration on masked identity priming. *Behavior Research Methods*. https://10.3758/s13428-021-01742-y.

Anglada-Tort, M., Harrison, P. M. C., & Jacoby, N. (2022). REPP: A robust cross-platform solution for online sensorimotor synchronization experiments. *Behavior Research Methods, 54*(5), 2271–2285. https://doi.org/10.3758/s13428-021-01722-2

Anwyl-Irvine, A. L., Armstrong, T., & Dalmaijer, E. S. (2022). MouseView.js: Reliable and valid attention tracking in web-based experiments using a cursor-directed aperture. *Behavior Research Methods*, *54*(4), 1663–1687. https://10.3758/s13428-021-01703-5.

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. https://doi.org/10.3758/s13428-020-01501-5

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

Barnhoorn, J. S., Haasnoot, E., Bocanegra, B. R., & van Steenbergen, H. (2014). QRTEngine: An easy solution for running online reaction time experiments using Qualtrics. *Behavior Research Methods, 47*(4), 918–929. https://doi.org/10.3758/s13428-014-0530-7

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

Betts, H. N. (2018). Retuning lexical-semantic representations on the basis of recent experience [Doctoral, UCL (University College London)]. In *Doctoral thesis, UCL (University College London)*. (pp. 1–232). https://discovery.ucl.ac.uk/id/eprint/10049908/.

Brascamp, J. W. (2021). Controlling the spatial dimensions of visual stimuli in online experiments. *Journal of Vision, 21*(8), 19. https://doi.org/10.1167/jov.21.8.19

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

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition, 2*(1), 16. https://doi.org/10.5334/joc.72

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant's Age. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01116

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice* (pp. xii, 274). Princeton University Press. https://10.1515/9781400884940.

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using Nonnaive Participants Can Reduce Effect Sizes. *Psychological Science, 26*(7), 1131–1139. https://doi.org/10.1177/0956797615585115

Chandler, J., & Shapiro, D. (2016). Conducting Clinical Research Using Crowdsourced Convenience Samples. *Annual Review of Clinical Psychology, 12*(1), 53–81. https://doi.org/10.1146/annurev-clinpsy-021815-093623

Chen, T., & Hartshorne, J. K. (2021). More evidence from over 1.1 million subjects that the critical period for syntax closes in late adolescence. *Cognition, 214*, 104706. https://10.1016/j.cognition.2021.104706.

Chetverikov, A., & Upravitelev, P. (2016). Online versus offline: The Web as a medium for response time data collection. *Behavior Research Methods, 48*(3), 1086–1099. https://doi.org/10.3758/s13428-015-0632-x

Chuey, A., Asaba, M., Bridgers, S., Carrillo, B., Dietz, G., Garcia, T., Leonard, J. A., Liu, S., Merrick, M., Radwan, S., Stegall, J., Velez, N., Woo, B., Wu, Y., Zhou, X. J., Frank, M. C., & Gweon, H. (2021). Moderated Online Data-Collection for Developmental Research: Methods and Replications. *Frontiers in Psychology, 12*, Article 734398. https://doi.org/10.3389/fpsyg.2021.734398

Chuey, A., Boyce, V., Cao, A., & Frank, M. C. (2022). *Conducting developmental research online vs. in-person: A meta-analysis*. PsyArXiv. https://10.31234/osf.io/qc6fw.

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

Cornsweet, T. N. (1962). The Staircase-Method in Psychophysics. *The American Journal of Psychology, 75*(3), 485–491. https://doi.org/10.2307/1419876

Coughlan, G., Coutrot, A., Khondoker, M., Minihane, A.-M., Spiers, H., & Hornberger, M. (2019). Toward personalized cognitive diagnostics of at-genetic-risk Alzheimer's disease. *Proceedings of the National Academy of Sciences, 116*(19), 9285–9292. https://doi.org/10.1073/pnas.1901600116

Couper, M. P. (2011). The Future of Modes of Data Collection. *Public Opinion Quarterly, 75*(5), 889–908. https://doi.org/10.1093/poq/nfr046

Coutrot, A., Schmidt, S., Coutrot, L., Pittman, J., Hong, L., Wiener, J. M., Hölscher, C., Dalton, R. C., Hornberger, M., & Spiers, H. J. (2019). Virtual navigation tested on a mobile app is predictive of real-world wayfinding navigation performance. *PLoS One1, 14*(3), e0213272. https://doi.org/10.1371/journal.pone.0213272

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS One1, 8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

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

De Man, J., Campbell, L., Tabana, H., & Wouters, E. (2021). The pandemic of online research in times of COVID-19. *BMJ Open, 11*(2), e043866. https://doi.org/10.1136/bmjopen-2020-043866

Difallah, D., Filatova, E., & Ipeirotis, P. (2018). Demographics and Dynamics of Mechanical Turk Workers. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, 135–143*. https://doi.org/10.1145/3159652.3159661

Donhauser, P. W., & Klein, D. (2022). Audio-Tokens: A toolbox for rating, sorting and comparing audio samples in the browser. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01803-w

Enochson, K., & Culbertson, J. (2015). Collecting Psycholinguistic Response Time Data Using Amazon Mechanical Turk. *PLoS One1, 10*(3), e0116946. https://doi.org/10.1371/journal.pone.0116946

Erb, C. D., Germine, L., & Hartshorne, J. K. (2023). Cognitive control across the lifespan: Congruency effects reveal divergent developmental trajectories. *Journal of Experimental Psychology: General*, No Pagination Specified-No Pagination Specified. https://10.1037/xge0001429.

Fraley, R. C., Chong, J. Y., Baacke, K. A., Greco, A. J., Guan, H., & Vazire, S. (2022). Journal N-Pact Factors From 2011 to 2019: Evaluating the Quality of Social/Personality Journals With Respect to Sample Size and Statistical Power. *Advances in Methods and Practices in Psychological Science, 5*(4). https://10.1177/25152459221120217.

Fraley, R. C., & Vazire, S. (2014). The N-Pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power. *PLoS One1, 9*(10), e109019. https://doi.org/10.1371/journal.pone.0109019

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin and Review, 19*(5), 847–857. https://doi.org/10.3758/s13423-012-0296-9

Göritz, A. S. (2014). Determinants of the starting rate and the completion rate in online panel studies1. In *Online Panel Research* (pp. 154–170). John Wiley & Sons, Ltd. https://10.1002/9781118763520.ch7.

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annu. Rev. Psychol., 66*, 877–902. https://doi.org/10.1146/annurev-psych-010814-015321

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *The American Psychologist, 59*(2), 93–104. https://doi.org/10.1037/0003-066X.59.2.93

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Guasch, M., Boada, R., Duñabeitia, J. A., & Ferré, P. (2022). Prevalence norms for 40,777 Catalan words: An online megastudy of vocabulary size. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01959-5

Halberda, J., Ly, R., Wilmer, J. B., Naiman, D. Q., & Germine, L. (2012). Number sense across the lifespan as revealed by a massive Internet-based sample. *Proceedings of the*

National Academy of Sciences of the United States of America, 109(28), 11116–11120. https://doi.org/10.1073/pnas.1200196109
Hartshorne, J. K., de Leeuw, J. R., Goodman, N. D., Jennings, M., & O'Donnell, T. J. (2019). A thousand studies for the price of one: Accelerating psychological science with Pushkin. *Behavior Research Methods, 51*(4), 1782–1803. https://doi.org/10.3758/s13428-018-1155-z
Hartshorne, J. K., & Germine, L. T. (2015). When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the lifespan. *Psychological Science, 26*(4), 433–443. https://doi.org/10.1177/0956797614567339
Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition, 177*, 263–277. https://doi.org/10.1016/j.cognition.2018.04.007
Hauser, D. J., & Schwarz, N. (2015). It's a Trap! Instructional Manipulation Checks Prompt Systematic Thinking on "Tricky" Tasks. *SAGE Open*, *5*(2), 2158244015584617. https://10.1177/2158244015584617.
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. https://doi.org/10.3758/s13428-015-0578-z
Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X
Hewson, C. M., Laurent, D., & Vogel, C. M. (1996). Proper methodologies for psychological and sociological studies conducted via the Internet. *Behavior Research Methods, Instruments, & Computers, 28*(2), 186–191. https://doi.org/10.3758/BF03204763
Kominsky, J. F., Begus, K., Bass, I., Colantonio, J., Leonard, J. A., Mackey, A. P., & Bonawitz, E. (2021). Organizing the Methodological Toolbox: Lessons Learned From Implementing Developmental Methods Online. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.702710
Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213–236. https://doi.org/10.1002/acp.2350050305
Kuroki, D., & Pronk, T. (2022). jsQuestPlus: A JavaScript implementation of the QUEST+ method for estimating psychometric function parameters in online experiments. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01948-8
Li, Q., Gajos, K. Z., & Reinecke, K. (2018). Volunteer-Based Online Studies With Older Adults and People with Disabilities. *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, 229–241. https://10.1145/3234695.3236360.
Li, Q., Joo, S. J., Yeatman, J. D., & Reinecke, K. (2020). Controlling for Participants' Viewing Distance in Large-Scale, Psychophysical Online Experiments Using a Virtual Chinrest. *Scientific Reports*, *10*(1), Article 1. https://10.1038/s41598-019-57204-1.
Li, W., Germine, L. T., Mehr, S. A., Srinivasan, M., & Hartshorne, J. (2022). Developmental psychologists should adopt citizen science to improve generalization and reproducibility. *Infant and Child Development*, e2348. https://10.1002/icd.2348.
Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z
Long, B., Simson, J., Buxó-Lugo, A., Watson, D. G., & Mehr, S. A. (2023). How games can make behavioural science better. *Nature, 613*(7944), 433–436. https://doi.org/10.1038/d41586-023-00065-6
Lukács, G., & Gartus, A. (2022). Precise display time measurement in JavaScript for web-based experiments. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01835-2
Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods, 53*(4), 1551–1562. https://doi.org/10.3758/s13428-020-01514-0
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*(1), 0021. https://doi.org/10.1038/s41562-016-0021
Musch, J., & Reips, U.-D. (2000). Chapter 3—A Brief History of Web Experimenting. In M. H. Birnbaum (Ed.), *Psychological Experiments on the Internet* (pp. 61–87). Academic Press. https://10.1016/B978-012099980-4/50004-6.
Nicosia, J., Wang, B., Aschenbrenner, A. J., Sliwinski, M. J., Yabiku, S. T., Roque, N. A., Germine, L. T., Bateman, R. J., Morris, J. C., & Hassenstab, J. (2022). To BYOD or not: Are device latencies important for bring-your-own-device (BYOD) smartphone cognitive testing? *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01925-1
Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice, 6*(1), 101–115. https://doi.org/10.1037/1089-2699.6.1.101
Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., & van 't Veer, A. E., & Vazire, S.. (2019). Preregistration Is Hard And Worthwhile. *Trends in Cognitive Sciences, 23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009
Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114
Oliveira, N., Jun, E., & Reinecke, K. (2017). Citizen Science Opportunities in Volunteer-Based Online Experiments. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 6800–6812). https://doi.org/10.1145/3025453.3025473
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. 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. https://doi.org/10.1016/j.jbef.2017.12.004
Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science, 23*(3), 184–188. https://doi.org/10.1177/0963721414531598
Passell, E., Strong, R. W., Rutter, L. A., Kim, H., Scheuer, L., Martini, P., Grinspoon, L., & Germine, L. (2021). Cognitive test scores vary with choice of personal digital device. *Behavior Research Methods, 53*(6), 2544–2557. https://doi.org/10.3758/s13428-021-01597-3
Pavlick, E., Post, M., Irvine, A., Kachaev, D., & Callison-Burch, C. (2014). The Language Demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics, 2*, 79–92. https://doi.org/10.1162/tacl_a_00167
Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology, 70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006
Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods, 54*(4), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3
Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods, 46*(4), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y
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. https://doi.org/10.3758/s13428-019-01321-2
Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences, 115*(45), 11401–11405. https://doi.org/10.1073/pnas.1721165115
Ratcliff, R., & Hendrickson, A. T. (2021). Do data from mechanical Turk subjects replicate accuracy, response time, and diffusion modeling results? *Behavior Research Methods, 53*(6), 2302–2325. https://doi.org/10.3758/s13428-021-01573-x
Reinecke, K., & Gajos, K. Z. (2015). LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 1364–1378. https://10.1145/2675133.2675246.
Roberts, C., Gilbert, E., Allum, N., & Eisner, L. (2019). Research Synthesis: Satisficing in Surveys: A Systematic Review of the Literature. *Public Opinion Quarterly, 83*(3), 598–626. https://doi.org/10.1093/poq/nfz035
Robinson, J., Rosenzweig, C., Moss, A. J., & Litman, L. (2019). Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool. *PLoS One1, 14*(12), e0226394. https://doi.org/10.1371/journal.pone.0226394
Rodd, J. M., Cai, Z. G., Betts, H. N., Hanby, B., Hutchinson, C., & Adler, A. (2016). The impact of recent and long-term experience on access to word meanings: Evidence from large-scale internet-based experiments. *Journal of Memory and Language, 87*, 16–37. https://doi.org/10.1016/j.jml.2015.10.006
Rotman, D., Preece, J., Hammock, J., Procita, K., Hansen, D., Parr, C., Lewis, D., & Jacobs, D. (2012). Dynamic changes in motivation in collaborative citizen-science projects. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (pp. 217–226). https://doi.org/10.1145/2145204.2145238
Sassenberg, K., & Ditrich, L. (2019). Research in Social Psychology Changed Between 2011 and 2016: Larger Sample Sizes, More Self-Report Measures, and More Online Studies. *Advances in Methods and Practices in Psychological Science, 2*(2), 107–114. https://doi.org/10.1177/2515245919838781
Sauter, M., Draschkow, D., & Mack, W. (2020). Building, Hosting and Recruiting: A Brief Introduction to Running Behavioral Experiments Online. *Brain Sciences, 10*(4), Article 4. https://doi.org/10.3390/brainsci10040251
Slim, M. S., & Hartsuiker, R. J. (2022). Moving visual world experiments online? A web-based replication of Dijkgraaf, Hartsuiker, and Duyck (2017) using PCIbex and WebGazer.js. *Behavior Research Methods*. https://10.3758/s13428-022-01989-z.
Spiers, H. J., Coutrot, A., & Hornberger, M. (2023). Explaining World-Wide Variation in Navigation Ability from Millions of People: Citizen Science Project Sea Hero Quest. *Topics in Cognitive Science, 15*(1), 120–138. https://doi.org/10.1111/tops.12590
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. https://doi.org/10.3758/s13428-010-0039-7
Stark, K., van Scherpenberg, C., Obrig, H., & Abdel Rahman, R. (2022). Web-based language production experiments: Semantic interference assessment is robust for spoken and typed response modalities. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01768-2
Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing Samples in Cognitive Science. *Trends in Cognitive Sciences, 21*(10), 736–748. https://doi.org/10.1016/j.tics.2017.06.007
Stewart, N., Ungemach, C., Harris, A. J. L., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making, 10*(5), 13. https://doi.org/10.1017/S1930297500005611
Storozuk, A., Ashley, M., Delage, V., & Maloney, E. A. (2020). Got Bots? Practical Recommendations to Protect Online Survey Data from Bot Attacks. *The Quantitative Methods for Psychology, 16*(5), 472–481. https://doi.org/10.20982/tqmp.16.5.p472
Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology, 15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

Teitcher, J. E. F., Bockting, W. O., Bauermeister, J. A., Hoefer, C. J., Miner, M. H., & Klitzman, R. L. (2015). Detecting, Preventing, and Responding to "Fraudsters" in Internet Research: Ethics and Tradeoffs. *Journal of Law, Medicine & Ethics, 43*(1), 116–133. https://doi.org/10.1111/jlme.12200

Tomczak, J., Gordon, A., Adams, J., Pickering, J. S., Hodges, N., & Evershed, J. K. (2023). What over 1,000,000 participants tell us about online research protocols. *Frontiers in Human Neuroscience, 17*, 1228365. https://doi.org/10.3389/fnhum.2023.1228365

Van den Bussche, E., Van den Noortgate, W., & Reynvoet, B. (2009). Mechanisms of masked priming: A meta-analysis. *Psychological Bulletin, 135*, 452–477. https://doi.org/10.1037/a0015329

Vogt, A., Hauber, R., Kuhlen, A. K., & Rahman, R. A. (2022). Internet-based language production research with overt articulation: Proof of concept, challenges, and practical advice. *Behavior Research Methods, 54*(4), 1954–1975. https://doi.org/10.3758/s13428-021-01686-3

Werchan, D. M., Thomason, M. E., & Brito, N. H. (2022). OWLET: An automated, open-source method for infant gaze tracking using smartphone and webcam recordings. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01962-w

Wissink, J., van Beest, I., Pronk, T., & van de Ven, N. (2022). The Online Coalition Game: A tool for online interactive coalition formation research. *Behavior Research Methods, 54*(3), 1078–1091. https://doi.org/10.3758/s13428-021-01591-9

Woods, A. T., Velasco, C., Levitan, C. A., Wan, X., & Spence, C. (2015). Conducting perception research over the internet: A tutorial review. *PeerJ, 2015*(7). https://doi.org/10.7717/peerj.1058

Wycisk, Y., Kopiez, R., Bergner, J., Sander, K., Preihs, S., Peissig, J., & Platz, F. (2022). The Headphone and Loudspeaker Test – Part I: Suggestions for controlling characteristics of playback devices in internet experiments. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01859-8

Zaadnoordijk, L., Buckler, H., Cusack, R., Tsuji, S., & Bergmann, C. (2021). A Global Perspective on Testing Infants Online: Introducing ManyBabies-AtHome. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.703234.

Zaadnoordijk, L., & Cusack, R. (2022). Online testing in developmental science: A guide to design and implementation. *Advances in Child Development and Behavior, 62*, 93–125. https://doi.org/10.1016/bs.acdb.2022.01.002

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology, 111*, 493–504. https://doi.org/10.1037/pspa0000056