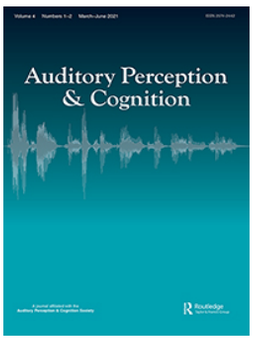

Auditory Perception & Cognition

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rpac20

Online Data Collection in Auditory Perception and Cognition Research: Recruitment, Testing, Data Quality and Ethical Considerations

Tuomas Eerola, James Armitage, Nadine Lavan & Sarah Knight

To cite this article: Tuomas Eerola, James Armitage, Nadine Lavan & Sarah Knight (2021): Online Data Collection in Auditory Perception and Cognition Research: Recruitment, Testing, Data Quality and Ethical Considerations, Auditory Perception & Cognition, DOI: 10.1080/25742442.2021.2007718

To link to this article: https://doi.org/10.1080/25742442.2021.2007718

Published online: 01 Dec 2021.

AUDITORY PERCEPTION & COGNITION
https://doi.org/10.1080/25742442.2021.2007718

Routledge
Taylor & Francis Group

Online Data Collection in Auditory Perception and Cognition Research: Recruitment, Testing, Data Quality and Ethical Considerations

Tuomas Eerola [a], James Armitage [a], Nadine Lavan [b] and Sarah Knight [c]

[a]Department of Music, Durham University, Durham, UK; [b]Department of Biological and Experimental Psychology, School of Biological and Behavioural Sciences, Queen Mary University of London, London, UK; [c]Department of Psychology, University of York, York, UK

ABSTRACT

Online studies using recruitment services (such as Prolific or Amazon's MTurk) and online testing platforms (such as Gorilla or PsyToolkit) are becoming increasingly common in psychological science. Although auditory disciplines have been slower to adopt these methods, uptake is rapidly increasing in auditory perception and cognition research. Utilizing online data collection and recruitment presents several advantages to researchers in terms of the speed of research and the range of target demographics available compared to either traditional lab studies or web-based recruitment via traditional means. Online platforms and recruitment services also present a set of technical and ethical challenges owing to the fact that the people completing experiments are working with their own devices outside the lab. This article discusses the potential technical and ethical implications of online studies, including both recruitment services and online testing platforms, with specific reference to auditory perception and cognition research. Rates of remuneration, sampling characteristics, anonymity, quality control, and ethics are all discussed with respect to these approaches. We also provide proposals for how researchers can ensure that online research meets present-day ethical and technical guidelines as well as research transparency standards

ARTICLE HISTORY
Received 16 April 2021
Accepted 2 November 2021

KEYWORDS
Online; platforms; recruitment services; experiment; ethics; auditory; crowdsourcing

Empirical research in auditory disciplines has historically taken place in highly controlled lab settings. In these settings, researchers have been able to control the exact environment under which participants hear and respond to auditory stimuli, for instance, volume, sound quality, distance from speakers or screens, and minimized or completely removed ambient noise and ambient sound distortions and reflections. During the last decade researchers have begun to collect auditory data online. In part, this change in data collection procedure has been driven by necessity (i.e., the Covid-19 pandemic), but it is also a consequence of increased access to Internet-enabled technologies and the improvements in web-based data collection software. Online data collection technologies

CONTACT Tuomas Eerola tuomas.eerola@durham.ac.uk Department of Music, Durham University, Palace Green, Durham DH1 3RL, UK

This article has been republished with minor changes. These changes do not impact the academic content of the article.

can be (and are) used with samples recruited in a traditional way, such as psychology undergraduates meeting course requirements, volunteers from the general public or specifically targeted populations. However, the research community in cognitive sciences, social sciences, economics and AI research is increasingly turning to recruitment services to target large numbers of participants quickly and efficiently (Buhrmester et al., 2018; Stewart et al., 2017) and running behavioral studies online has become relative common (Grootswagers, 2020). Moreover, there has also been an increase in studies utilizing so-called gamification approaches, where game-like elements (competition among online gamers, smooth visual feedback, etc.) are brought to online data collection to increase participant enjoyment and engagement with the research task (Nacke & Deterding, 2017).

Here we examine the issues relating to online *implementation* of studies via online platforms and the *recruitment* of participants via recruitment services, with a specific focus on auditory research. Implementation covers techniques, web protocols, online hosting, and tasks and controls related to collecting good quality data online. Recruitment of participants refers to the way participants are contacted, either via traditional recruitment methods (participant pools, social media, etc.) or through recruitment services (such as Prolific.co or Amazon's MTurk). Together, studies which are both implemented online and also utilize recruitment services are often called *crowdsourcing* studies (Stewart et al., 2017). Lastly, gamification, which is a related but different style of recruitment and implementation of online study, will be briefly presented as the approach shares many of the considerations of all online studies in the auditory domain.

In this paper, we first introduce the typical tasks utilized in online auditory research, and present some common online platforms for hosting and organizing online experiments. We move on to summarize the main recruitment services, and then proceed to review the pros and cons of online studies, online studies with recruitment services, and finally cover online studies using a gamified approach. Overall, the elements of the online studies can be visualized as a process cycle (see Figure 1) where the new elements beyond the participants, researcher, and the experiment consist of an *online platform* that hosts and runs the experiment, and a *recruitment service* – or alternatively traditional recruitment – that brings the participants to the experiment. As discussed further below, the design of the research can be pre-registered, and the data obtained from the experiment can be shared according to Open Data initiatives in behavioral sciences (Shrout & Rodgers, 2018) as any empirical research, but in the text we address the specific issues of transparency unique to online studies.

Types of Online Tasks in Auditory Research

To orient the reader to the challenges of online research, we first outline four types of tasks (rating, forced-choice, production, and stimulus manipulation tasks) reported in online studies in the auditory domain.[1]

In *rating tasks* participants are asked to make decisions based on subjective qualities of stimuli in a continuous manner. Examples range from the perception of social traits from voices (Lavan, Mileva et al., 2021, 2021), to rating the consonance of intervals and chords (Lahdelma & Eerola, 2020), evaluating the "song-like-ness" of the speech-to-song illusion (Tierney et al., 2021), or rating the recognition of spoken words (Slote & Strand, 2016). In

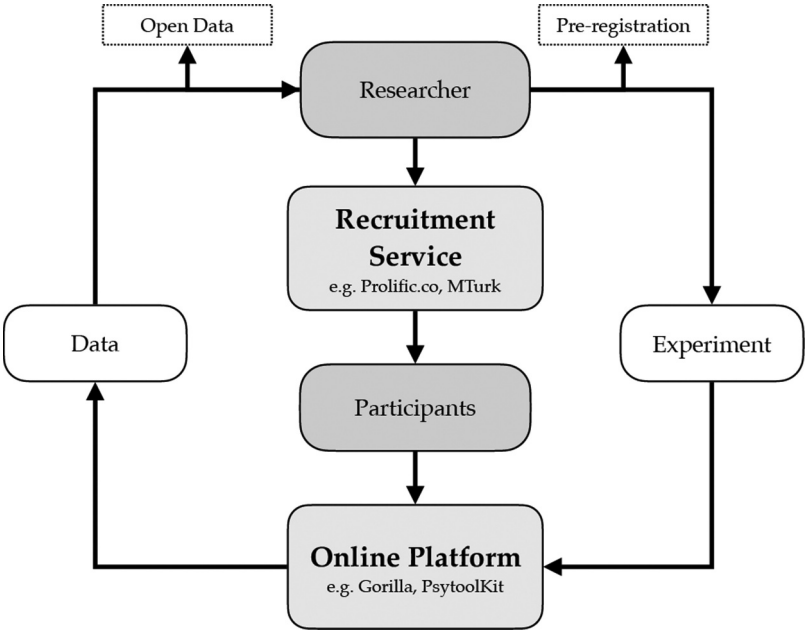

Figure 1. Diagram to illustrate the elements of an online experiment with a recruitment service and online platform.

certain cases, these tasks can be implemented as indirect measures, where for instance, economic games can be used to assess the impact of perceived social traits from voices (Knight et al., 2021).

In *forced-choice tasks* the participant is exposed to different audio stimulus comparisons and asked to choose between the alternatives, for instance, classifying genre with different levels of noise (Kell et al., 2018), detecting a melody among variants (Woods & McDermott, 2018), or recognizing syllables from audio-video presentation (Brown et al., 2018). Similarly, studies have examined how listeners learn to recognize people from voices (Lavan, Knight et al., 2019; Lavan, Knight, McGettigan et al., 2019), speech intelligibility (Yoho et al., 2019) or how the perception of familiar and unfamiliar voices may differ (Kanber et al., 2021; Lavan, Kreitewolf et al., 2021). Finally, studies have explored whether and how listeners can incorporate voices into their self-concept (Payne et al., 2021). One interesting variant of a discrimination task is the computerized adaptive beat alignment test (Harrison & Müllensiefen, 2018), where a participant is presented with a beep track and a musical track at the same time. The key variable is the temporal match between the beep track and the true beat locations, which are varied in an adaptive fashion during the experiment. In some forced choice tasks, the decision timing information can be used to gain insights about lexical or semantic decisions related to the auditory signal (Armitage et al., 2021; Escudero et al., 2021; Lahdelma et al., 2020; Slote & Strand, 2016).

Production tasks can consist of singing, tapping, or altering parameters of sound/music. A good example is tapping with an external stimulus, which can be done in an online setting utilizing the built-in microphone and speakers of a standard laptop

computer to a high degree of temporal accuracy (~2 ms latency and jitter) using REPP (Rhythm ExPeriment Platform) (Anglada-Tort et al., 2021). For a broad review of the timing capacities and accuracy of online platforms, see Bridges et al. (2020). Similarly, responses to tasks can be collected via singing in online experiments that also utilize the microphone input (Pfordresher & Demorest, 2020, 2021). Other examples of production tasks run in an online setting include a study where participants read sentences alone and in synchrony with another speaker in an online setting (Bradshaw & McGettigan, 2021), and a set-up in which adults had to try to learn novel words and were then required to verbally produce the words at test (James et al., 2020). Participants can also type their responses on a keyboard, which has been used to study speech perception (Guang et al., 2020; Heffner et al., 2017).

Stimulus manipulation tasks may employ sorting and arranging stimuli or controlling the parameters of the stimuli. For instance, Lavan and her colleagues (Lavan, Burston et al., 2019, 2019; Lavan et al., 2020; Njie et al., 2021) asked listeners to sort a number of voice recordings by perceived identities using a drag and drop interface. An example of auditory research where the participants controlled the creation of the stimuli according to specific criteria is the work by Harrison and others (Harrison et al., 2020). In their experiment, participants were asked to alter expressive parameters such as duration, pitch, and intensity of spoken prosody to create emotional expressions conveying one of three emotions. They also applied a similar methodology to enable participants to create pleasant musical chords by allowing them to modify the frequency of the intervals on a continuous domain. It is likely that there will be an increase in future in flexible and creative tasks that utilize visual layouts and new kinds of assessment and production tasks.

In summary, there is already considerable diversity in the approaches, techniques and methods utilized in online studies involving audio stimuli. However, most of the studies that we cite here are from the pre-pandemic era and it is reasonable to expect a significant growth in the number of online studies within the next few years. Many labs and research groups have turned their attention to online opportunities because traditional lab experiments have been paused at the point of writing for more than 21 months during the periods of lockdowns or social distancing. It will be exciting to see how this current expansion of online studies will develop more robust and accurate ways of capturing our engagement with auditory stimuli and how these might shape empirical data collection practices – including lab-based behavioral experiments – in the near future.

Online Testing Platforms

There are numerous services that offer full online testing capabilities. Most of these have recently been reviewed in detail by Sauter and his colleagues (Sauter et al., 2020). The services can be divided into *integrated services*, which contain an experimenter builder and host the live experiment data, such as *Gorilla.sc* (Anwyl-Irvine, Massonnié et al., 2020), *Labvanced* (Finger et al., 2017), *Testable* (Rezlescu et al., 2020), *JATOS* (Lange et al., 2015) and *Pavlovia*[2]; and those that are mainly experiment builders but may also have capabilities to manage other aspects of the data such as *jsPsych, lab.js, OpenSesame, PsychoPy* (Peirce, 2007), *Qualtrics* and *PsyToolkit* (Stoet, 2010). We note here that "hosting the live data" refers to the actions whereby the service records, stores and

possibly displays the choices made by the participants during the experiment, and does not refer to the long-term data sharing solutions where the full data is deposited to an open access repository (e.g., to OSF, Github, etc.). It is also worth pointing out that most of the experiment builders (e.g., Psychopy, Gorilla, jsPsych, OpenSesame, Labvanced, and Psytoolkit) can run locally as well, which may well be desirable in situations requiring lab facilities, specialist interfaces (monitor, response device, or linking the responses with psychophysiology or neural responses), or in situations where the highest possible timing accuracy is needed. Here we focus on the platforms that have been featured in multiple auditory studies such as Gorilla, PsyToolkit, Qualtrics, and PsychoPy (see Table 1).

Qualtrics is a survey platform that supports JavaScript and html5 techniques. The platform interface for participants supports 75 languages, and provides good programmable control over the presentation orders (block and item randomizations) and allows the answers and the block choices to interact with responses. Qualtrics also supports links to recruitment services and have their own specific service called Qualtrics Panel. They are the most expensive in comparison to other services highlighted here, but universities often have institute- or department-wide subscriptions to the service. While Qualtrics handles complex survey and questionnaire designs and basic presentation of text, images, and limited audio and video, the specific needs of auditory research require an additional layer of coding in JavaScript. For instance, collecting timing responses in relation to audio or adjusting the audio presentation options requires controlling these parameters with JavaScript, which also requires specialized competence on how to integrate the code with the custom Qualtrics JavaScript Form Engine. For this reason, Qualtrics is not an optimal platform to develop complex, custom auditory experiments, although it would not be impossible.

Gorilla is an integrated platform that has an easy-to-use experiment design interface (task builder, questionnaire builder and experiment builder tools). It supports a wide variety of techniques, has tools for detecting bots, and supports recruitment services (Anwyl-Irvine, Massonnié et al., 2020). It also achieved excellent precision (under 3 ms) in response time tasks although suffers from variability in the audiovisual synchrony in a direct comparison of the timing capacities of five online platforms (Bridges et al., 2020) (more about these in the section *Online testing in auditory research*). Gorilla also offers a rich set of experiment and task libraries that can be used in building new experiments. Gorilla charges users for 'tokens,' each enabling users to collect and download one data set. Users can either buy batches of tokens, or commit to lab/team/department subscriptions that may have different pricing strategies. Creating complex experiments and tasks is relatively easy with the task builder that relies on a graphical user interface. As such,

Table 1. Summary of the popular online testing platforms in auditory studies.

Platform	Hosted	Fee	Example	Task
Qualtrics	Within	$300/month	Lahdelma & Eerola, 2020	Consonance rating
Gorilla	Within	$1/participant	Lavan, Knight et al., 2019	Voice recognition
PsyToolkit	Within	Free	Armitage et al., 2021	Auditory & word priming
PsychoPy	Pavlovia	$145/month	Escudero et al., 2021	Auditory & visual priming

Gorilla offers more flexibility in producing iterative, trial-based experiments than e.g., Qualtrics can easily provide. Additional functionality can be added to tasks by using snippets of JavaScript in the task builder. These snippets can be relatively easily implemented by those with only limited experience of JavaScript, while experts can set up entire tasks and experiments purely through code.

Psytoolkit is a free platform that contains experiment builders for experiments and surveys, and a large roster of example experiments and surveys. It also hosts the running of experiments and stores the data (Stoet, 2010). It is free for noncommercial use and geared for academic studies, especially for those using visual stimuli, but has good support for audio as well. It also supports integration of recruitment services in the data collection routines. There is no visual user interface and the surveys and experiments are programmed through a custom syntax. However, the examples and the help files are organized in a way that allow users to master the syntax after some period of study and consequently create sophisticated experiments with feedback options.

PsychoPy is another set of free tools for creating advanced behavioral experiments that can be run online (Gallant & Libben, 2019). It has a wide selection of techniques and experiment libraries. Online versions of studies are hosted at Pavlovia.org, but the integration of PsychoPy and Pavlovia is seamless. Whereas PsychoPy is free, hosting experiments in Pavlovia has a yearly cost ($2,053 or £1,500) or per participant cost (measured in credits). Pavlovia supports popular online techniques such as jsPsych and lab.js, which are high-level JavaScript libraries designed for creating behavioral studies with minimal programming experience. Like other online study platforms, it also supports integration of recruitment services in the data collection routines. The graphical user interface in the experiment builder makes this a suitable entry level tool, but competence in Python is useful to develop experiments that go beyond presenting auditory stimuli (e.g., to precisely control the timing of the visual and auditory stimuli).

Recruitment Services (Crowdsourcing)

When online studies are carried out with the help of recruitment services, we usually talk about crowdsourcing (Bhatti et al., 2020). This combination has the potential to be transformative for certain types of research questions and tasks within cognitive sciences (Stewart et al., 2017), and data sciences. In auditory studies, using recruitment services in

Table 2. Summary of popular recruitment services.

Service	Participants	Median Wage[a]	Auditory-related[b]	Citations
Mechanical Turk (US, 2005)	250,000[c]	$10.20	167	86,100
Appen[d]	1,000,000[e]	$1.85	5	6,850
Prolific.co (UK, 2015)	153,308	$6.91	2	4,610
Qualtrics Panel (US, 2018)	Unknown	Unknown	37	1,980

[a]According to http://faircrowd.work although the median wages do not include the proportion of nonpayment reported by participants, which are 60% for Mechanical Turk, 11% for Appen, 29% for Prolific.co according to http://faircrowd.work, see Section about *ethical considerations* for a full discussion.
[b]Web of Science search terms with auditory|audio|music|speech+recruitment service.
[c]Though 7K are estimated to be active each year (Stewart et al., 2015).
[d]Formerly known as Crowdflower.
[e]Unsubstantiated.

conjunction with online studies is relatively common. With the rise of recruitment services such as Amazon's *Mechanical Turk, Prolific.co, Clickworker* or *Appen*, there are new opportunities which are already used by many scholars working on auditory issues. Table 2 pools together the summary details such as the size of the recruitment pool, the median hourly salary, citations, and citation frequency of audio-related studies for each of four popular services. We note, however, that the landscape of the available services may shift rapidly as there are new services such as *Sojump*, the Chinese recruitment service, and *CloudResearch* (formerly known as *TurkPrime*) and *Prime Panels* associated with CloudResearch, which provide additional ways of using Mechanical Turk with advanced customization options.

Although there are papers already available that summarize several of these recruitment services, particularly Mechanical Turk (Chandler et al., 2019; Grootswagers, 2020), we nonetheless provide a short overview of the currently available popular recruitment services and their relation to auditory studies.

Amazon Mechanical Turk (hereafter MTurk) offers a large online workforce who can complete experiments and surveys at a very competitive price (Henrich et al., 2010). MTurk (and the alternative platforms discussed below) has proven to be more representative of the population than the usual lab or survey samples (Behrend et al., 2011). Participants using these services tend to be more reliable than typical online surveys or lab participants (Daniël Lakens, 2014), and even more attentive than lab participants (Hauser & Schwarz, 2016). Researchers can specify a number of criteria from the participants, including their past performance in tasks. Over the last ten years, large-scale studies with MTurk participants have successfully replicated classic studies in psychology (Berinsky et al., 2012), political science (Goodman et al., 2013), and economics (Mullinix et al., 2015). However, not all reports about MTurk are positive; several scholars have noted that the participants in MTurk are no longer naïve participants (Chandler et al., 2019). Overall, however, MTurk has been a popular recruitment service among researchers, such that there are numerous studies involving auditory stimuli and participants recruited from MTurk (Aljanaki et al., 2017; Harrison et al., 2020; Howe & Lee, 2021; e.g., Schmidtke et al., 2018; Speck et al., 2011).

Prolific.co is another popular service that offers a large and flexible participant pool. Prolific offers a particular advantage to the academic community in that it is specifically designed as a platform for recruiting participants for research studies (Palan & Schitter, 2018). Prolific allows researchers to pre-screen participants for a range of variables that are pertinent to empirical auditory research, such as musical training, hearing loss, native language, handedness as well as linguistic and other demographic variables. Similarly to MTurk, Prolific allows researchers to include participants based on their past performance on the platform (approval rate, registration date, participation in different types of studies, number of completed studies, etc.). There are options to allow Prolific to offer representative samples for an additional fee or to balance certain demographic aspects of the samples. There have been a considerable number of studies involving auditory stimuli or music conducted with participants recruited from Prolific.co, including dozens of online experiments by the authors. Our studies have addressed consonance and dissonance of chords and interval using self-reports (Athanasopoulos et al., 2021; Lahdelma et al., 2021; e.g., Lahdelma & Eerola, 2020), auditory priming (Armitage & Eerola, 2020;

Armitage et al., 2021; Lahdelma et al., 2020), perception of social traits from voices (Knight et al., 2021; Lavan, Mileva et al., 2021, 2021), or various voice identity perception tasks (Kanber et al., 2021; Lavan, Knight et al., 2019).

Appen is the third popular provider of recruitment services to researchers, which was known until 2019 as *Crowdflower*. Like MTurk, Appen provides a large and highly flexible pool of participants who can complete a range of tasks that include but are not limited to participating in academic research. Peer et al. (2017) contend that Prolific and Crowdflower provide higher quality data than MTurk, despite the latter's prominence as a research tool. While Appen is therefore another viable recruitment service that can be used for research, so far there are to our knowledge no auditory or music studies that used Appen as a source of participants.

Qualtrics Panel provides customized recruitment services in the form of a panel of participants, which is negotiated with the company, and the pricing is related to the length of the survey and specificity of the sample required. Unfortunately the details of the panel participants (how many, what are the typical fees and so on) are not readily available, but several studies have compared Qualtrics Panel to other recruitment services in terms of the quality of the data and representativeness of the samples to nationally representative samples in the US (Zack et al., 2019) and in India (Boas et al., 2020). The results suggest that Qualtrics Panel may offer slightly more representative data than MTurk with specified criteria, and generally the notion of representative sample should be used with caution when recruiting through these services, especially outside the US. We have only identified one study relating to auditory stimuli and music that has used Qualtrics Panel as the recruitment service (Jakubowski et al., 2021).

As a basic visualization of the increase in prevalence in crowdsourcing in auditory and music research over the last 15 years, Figure 2 shows the number of results returned by Google Scholar searches for the keywords "Music" or "Auditory" coupled with each of the recruitment services, "MTurk," "Prolific.co" and "Crowdflower" for each year from 2005 to 2020 (MTurk, Prolific.co and Crowdflower were launched in 2005, 2015 and 2007 respectively).

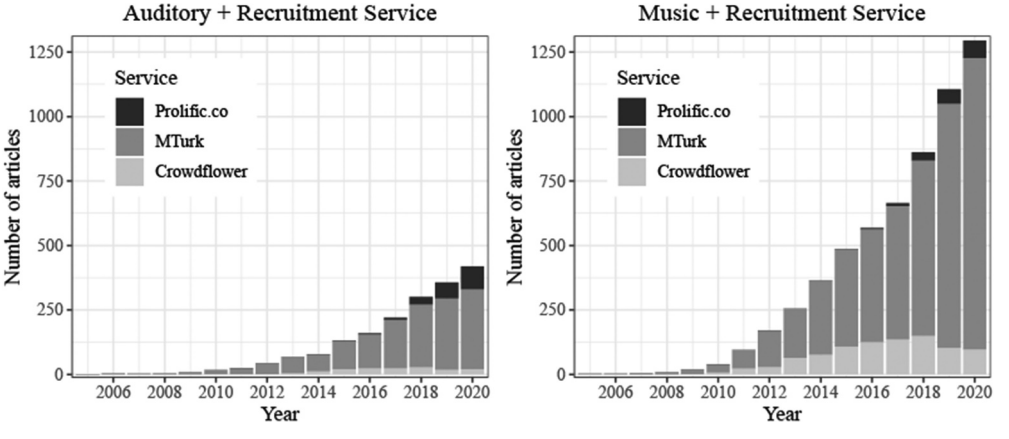

Figure 2. Frequency of the top three recruitment services in published studies since 2005 according to Google Scholar results for "Auditory" or "Music" + recruitment service.

In addition to scholars capitalizing on recruitment services, there have been initiatives to build complete "virtual labs" or meta-services that function as a one-stop shop for online study needs. Such a service would provide the full framework which handles every aspect of the online research, from recruitment that taps into recruitment services (if needed), to implementing any online task and rendering the interface to the participant via a browser, preliminary quality control, payment and the bonus payment deliveries, and analyzing and storing the data. Services that would support the full spectrum of online study processes such as EMPIRICA (Almaatouq et al., 2021), Dallinger,[3] or PsyNet (Harrison & Jacoby, 2020) are all under development at the moment. While some of these might eventually become the convenient way of carrying out online studies, the extra advantage they offer is the capacity to support methods that require participants to interact with each other at a large scale. Studies involving iterated reproduction of rhythms (Jacoby & McDermott, 2017) or iterative altering of the properties of auditory stimuli (Harrison et al., 2020) are good examples of this approach that require integration over the distinct parts of the processes.

In the next section, we will review the pros and cons of the three approaches to online data collection compared to more traditional methodologies: online studies, online studies with recruitment services, and gamified studies.

Review of Approaches

Online Testing in Auditory Research – Pros and Cons

Transparency

Online research fits well with reproducibility and open data initiatives although it is not inherently different from standard lab experiments. From the perspective of recruitment, platforms such as Prolific.co allow eligibility criteria to be applied automatically according to the registration information provided by participants. As a result, subsequent studies/researchers can recruit from a highly similar participant pool according to identical criteria. From the perspective of implementation, online tasks can be shared in their entirety (for example, via the "Library" space in Gorilla or sharing the experiment in Prolific.co or preferably publicly by allowing anyone to duplicate the study settings), including instructions, stimuli and response screens. This allows different teams of researchers to replicate studies with ease, or for data collection to be distributed between collaborators. Studies which are carried out online avoid lab-specific biases in terms of equipment, set-up, experimenter or delivery of instructions; and the ability to deliver the same web-hosted tasks both in the lab and in individual participants' homes allows for the robustness of findings to changes in environment to be assessed.

By sharing recruitment procedures and tasks, the research pipeline for online studies can be clearly defined and automated, and study pre-registration and open data release are also common strategies with this type of research, especially if Dallinger or similar tools are utilized in deploying the study. A growing number of crowdsourcing studies are pre-registered in OSF, although pre-registrations are still overall relatively rare in auditory cognition.

Democratization of Research

Online studies allow scholars with fewer resources – those without dedicated physical infrastructure (e.g., audio labs) or technical support services – to conduct studies by only paying the running costs of the experiment. Online studies also circumvent physical limitations beyond the global pandemic, such as smaller cities and towns having smaller participant pools to recruit from, especially outside of university term time. The limitations of the size of the accessible participant pool may also impact piloting and estimating effect sizes adversely in these settings. Using recruitment services helps to avoid these issues and keeps the pace of research reasonable despite physical bottlenecks created by facilities or regions.

Accessing Specialist Samples

Online research may offer the possibility of accessing specialist samples easily (Smith et al., 2015; Wilkerson et al., 2014). The speakers of specific languages (Turner et al., 2012), regional accents (Njie et al., 2021) or participants with amusia or tinnitus are examples of such special groups of interest to auditory research.

From this brief overview, it is clear that online studies do indeed provide benefits to the researcher on a number of levels. Additionally, the utility of online studies has come to the fore in allowing researchers to pursue experimental procedures where public health concerns or local restrictions would otherwise have made this impossible. It seems plausible that, even allowing for the rising trend demonstrated in Figure 1, we will see a further increase in the number of online studies that also employ recruitment services. However, online data collection also presents several shortcomings, which are discussed next.

Sound Delivery

In online studies, the researcher relies on the hardware and the software (mainly the browser) the participant already has access to on their own device. Participants can be asked to complete headphone checks such as those developed by Woods et al. (2017), which utilizes phase-information to create differences in dynamics that are easy to discern with headphones but not with external speakers such as those in laptops, or Milne et al.2021), which is based on the Huggins pitch test; code and stimuli for both tests are freely available online. Such tests allow researchers to screen out those that do not use headphones. At the time of writing, twelve peer-reviewed studies that use Woods et al.'s (2017) headphone check are available and for which the pass-fail rates are reported. The mean and median failure rates are 16.5% and 17.2% respectively (range: 0–40%). One study (Guang et al., 2020) utilizes the test by Milne et al. (2021), and found a failure rate of 49%. There is also a new headphone check available relying on beating interference to verify the participant's hardware capacity to present stimuli dichotically (Pankovski, 2021).

Even with such headphone checks, the quality of the delivery of the stimuli as well as the quietness of the environment is not under the experimenter's direct control. This puts serious limitations on more psycho-acoustically demanding studies (aiming to establish detection thresholds, just noticeable differences [JNDs], etc.). Moreover, there are ethical issues related to the potential exposure of participants to unexpectedly or unintentionally loud sounds in online studies. Due to this, we recommend that studies should always

start by presenting some kind of representative "calibration sound," with participants instructed to start with their volume turned down and then adjust it to a comfortable level. Experimental stimuli should then never be louder than the calibration sound, and we stress the importance of adjusting the sounds carefully (starting quietly and turning the volume up, rather than vice versa). However, despite such calibration checks and clearly worded instructions, the experimenter cannot fully ensure that the participant's equipment and setup would not lead to uncomfortable listening experiences for the participant.

Aside from those specialist auditory studies that are best carried out in well sound-proofed settings, there can be some merit to the argument that if a phenomenon can be captured in the diverse settings such as those offered by participants' typical headphones and audio devices, the phenomenon is probably a robust one. However, imperfect playback devices may also bias the results in specific ways instead of simply increasing noise. For example, if the effect of interest relies on low frequency information in the auditory signal, this is typically poorly represented by the average headphones (Olive et al., 2018). Even though typical listeners are unable to detect the quality differences between budget and high-end headphones (O'Brien & Schmidt, 2020), results from online studies in such cases may differ from results obtained in more controlled settings. In cases where the sound quality may become an issue for adequately measuring effects of interest, we would recommend at least running validation studies in the lab.

Timing Accuracy

Online experiment platforms such as PsyToolkit, Gorilla, and PsychoPy that run in most current browsers all demonstrate good capacity to bring timing experiments, such as reaction time measures in response to visual and auditory stimuli, to an acceptable level of precision; they usually demonstrate inter-trial variability of 5–10 ms, as compared with lab-based software, which can achieve inter-trial variability under 1 ms (Bridges et al., 2020). For a large-scale comparison of these qualities for lab-based and web-based software across multiple operating systems and browsers, see Bridges et al. (2020) and Anwyl-Irvine, Dalmaijer et al. (2021). In a nutshell, the two evaluation studies demonstrate that web-based solutions provide adequate timing for most cases unless the absolute response times are needed between the individuals. This measure is negatively impacted by different participants providing their responses on different operating systems and browsers. This, however, is not normally required when the comparisons are made within the same participant as is often the case in experimental research. An important caveat to the timing accuracy of the online data collection is that significantly better response timing accuracy is obtained using external response devices (i.e., high-performance button box) than by using the standard keyboard (Bridges et al., 2020). For this reason, the poorer accuracy overall reported by Anwyl-Irvine, Dalmaijer et al. (2021) is closer to the reality of online research as the participants will not have high-performance button boxes installed on their USB ports.

Answer Format

Many standard answer formats are supported across the various online testing platforms, such as different types of questionnaire responses, forced-choice responses, free text responses and ranked responses among many others. However, not all types of answer

formats that might be constructed in a lab are feasible in an online experiment. Because of the dependency on participants' home setups, it is not always possible to use unusual interfaces which require complex mouse operations, or calibrated monitors or other external devices. Thus, for the most part, experimental procedures must be limited to using standard mouse operations and keyboard responses. Despite these limitations, recent research has demonstrated the feasibility of production tasks, such as tapping (Anglada-Tort et al., 2021) or singing (Pfordresher & Demorest, 2021), or adjusting sliders to create sounds (Harrison et al., 2020) in studies of auditory cognition. Similarly, even in the absence of support of a specific answer format on online testing platforms, alternative ways of implementing tasks can be found: For example, in the absence of a readily available drag-and-drop interface at the time, we have run a sorting study online, asking our participants to download a PowerPoint slide on which the to-be-sorted stimuli were embedded. Participants were then able to sort the stimuli within PowerPoint, save the sorted slide and upload to a file transfer website from which we were able to retrieve the data (Lavan, Burston et al., 2019; Njie et al., 2021).

Lack of Visual Oversight of Participants

Although many lab experiments do not require visual connection during the experiment between the participants and the experimenter, lack of any visual – or auditory – cues during the experiment can amplify problems that sometimes occur in lab experiments such as participants attempting to engage in social media, text message or calls, or becoming perplexed by the experiment instructions or getting stuck at some point. In our view, issues such as these are less likely to occur when participants attend labs in person and interact with experimenters; or in case of confusion with respect to tasks, they can be resolved with timely interactions. Overall, these issues are the crucial part of the quality control that we articulate in more detail in the section about *quality control*.

Copyright and Other Restrictions

Much of the stimuli used in auditory and cognition research might be copyright free sound files created for the experimental needs, but in cases where existing commercial music or audio excerpts are used, the uses of the copyright material need to comply with regional law governing fair use and digital copies of copyrighted materials.

Online Testing Using Recruitment Platforms – Pros and Cons

Diversity

Ethnocentrism – i.e., focusing too heavily on one particular subset of the human population – is one of the main criticisms leveled at psychology in recent years (Rad et al., 2018). By turning to recruitment platforms, we may be able to avoid some of the aspects of so-called WEIRD samples (White, Educated, Industrialized, Rich, and Democratic) (Casler et al., 2013; Henrich et al., 2010) as the participants in these services have more diverse backgrounds (Sheehan, 2018) than typical participant pools. Specifically, Casler, Bickel and Hackett found that MTurk samples are more diverse than traditional volunteer (recruited via social media) samples in terms of both their socio-economic and ethnic backgrounds (2013), also supported by Chandler et al. (2019). Goodman et al. (2013) found a greater degree of linguistic diversity in MTurk samples than in a typical community sample.

However, this is not to say that the diversity provided by MTurk is still not far away from the US population: most Mturk participants are young and older people are under-represented in the participant pool. MTurk participants tend also to be more liberal, and have higher education qualifications when compared to the US population (Casey et al., 2017; Levay et al., 2016). However, some authors point out that the greater degree of variation in the participant pool can also act as a limitation (Feitosa et al., 2015) in some contexts, such as when certain instruments have not been validated in a particular language.

For music and auditory studies, it is worth highlighting that the criticism of WEIRD is extremely relevant as music is highly culturally dependent and notions and preferences about music do vary considerably even within a country depending on sociodemographic background and education. A particular aspect of diversity is cross-cultural research, which normally requires extraordinary connections and resources. Online studies with recruitment services offer a possibility to tackle some, albeit limited cross-cultural research (Cuccolo et al., 2021). Most of the recruitment services allow researchers to define participant recruitment by location, country, and native language at least. This allows for comparisons that relate to geographical location and language of the participants, although it has to be kept in mind that the popularity of the services is not well spread beyond the countries of their originators, and that the people who work in these services are very much reliant on internet and may have fairly Western standards in many of their values (Pollet & Saxton, 2019).

In our experiments using participants obtained from recruitment services such as Profilic.co, we have observed that samples are more representative in terms of their age, gender, and nationality than the average student population, although they do nonetheless deviate from the society at large. The employment details are stable across many experiments and an overall fairly even gender distribution can be readily achieved through pre-screening. The age distribution usually shows that the bulk of the participants are between 25 and 42, and about 20% of the participants are students, which is an improvement from lab studies but it is still clear that students are over-represented in comparison to national statistics (3.5% in the UK[4]).

Affordability

Online studies may provide a cost-effective way of collecting data in auditory sciences. Provided that the quality assurances can be met, the price of data may be cheaper than in lab studies when factoring in the costs involved (researcher time, research assistants, facilities and participant fees). For instance, we carried out a direct economic comparison of lab and online data utilizing recruitment services with respect to a specific study (Armitage & Eerola, 2020). When including the cost of both participant fees and payment to a research assistant, the cost of the lab data was more than double that of the online data. However, the direct lab costs may often be lower if they include free labor available as part of the operation (research assistants working for course credit etc.) although this cost could be assumed to be included elsewhere (such as training of the assistants and general costs of facilities and services). As an example, a typical online dataset requiring 10-minutes of participant time using Prolific.co and a sample of 40 participants would cost $106.40 ($12 for minimum wage × 1/6 h × 40 participants × 1.34 service fee) excluding any piloting. Importantly, there were no significant differences in the attrition

rates, distributions of the data, or effect sizes between the samples obtained via recruitment services and lab (Armitage & Eerola, 2020). However, it should be noted that Buhrmester et al. (2018) advocates for careful investigation of attrition rates, although this is a factor in all internet-mediated research and not a problem that is unique to studies using recruitment services.

Speed and Statistical Power

Online studies with recruitment services allows for rapid data collection. Studies that would take several weeks to run in a lab can be completed in the course of a day via recruitment services depending on the type of task and specialist expertise needed. Clearly, fast data collection offers direct benefits in terms of time saved. The speed of data collection also provides more nuanced benefits beyond speed per se. As data collection is efficient in terms of time, it allows for research to focus on testing specific hypotheses. Funding allowing, hypotheses can therefore be revised and retested at a rate that was not possible previously, allowing for an incremental but thorough advancement of understanding. Also, online studies with recruitment services allow for studies to be appropriately powered, not over or under-powered, if the power analysis is made in advance and with conservative reading of similar studies (Brysbaert, 2019). We therefore argue that the improvements to research are most pronounced if the advantage conferred by the speed and easiness of the data collection is tempered with careful planning and appropriate assessment of the needs.

Data Quality

Traditional volunteer web studies are often subject to significant data wastage or poor data quality (Hochheimer et al., 2019). However, there is research to indicate that the quality of data collected via recruitment services is comparable to lab data and superior to other web data and subject to less participant attrition (Armitage & Eerola, 2020; Hauser & Schwarz, 2016; Kees et al., 2017), although contrasting views also exist (Chmielewski & Kucker, 2020; Grootswagers, 2020). Indeed, we have found that participant attrition and prevalence of outliers in samples obtained through recruitment services is almost identical to lab data and superior to traditional online data collected via convenience sampling (Armitage & Eerola, 2020). Data quality is of course related to the task, instructions, and the quality controls implemented in the study, which we cover later (see *Quality Control*).

Scalability

Online studies with recruitment services allow the possibility of automating the data collection at a high level, which allows easy transportability, replicability and the possibility to alter the method, concepts, stimuli or a measure by a simple option or even run several variant experiments in parallel. This is also possible at some level in traditional lab experiments, but with considerable more effort and customization. The real benefit of the scalability comes from experiments where participants' responses are taken as input for other participants such as in iterated rhythm production tasks (Jacoby & McDermott, 2017) or in Gibbs sampling with humans (Harrison et al., 2020). Scalability usually requires that the researchers utilize an automation service such as Dallinger or Pushkin (Hartshorne

et al., 2019), which allow a high level of abstraction and the automation of all practicalities (recruitment, running the experiment, paying participants, and managing data). Collecting data with recruitment services also brings several possible shortcomings, presented next.

Quality Control

The reliability of online studies utilizing recruitment services has been explored with several well-known psychometric instruments (personality etc.) and these indices (obtained with test-retest evaluations and calculating Cronbach alphas) are typically at the same level as in lab experiments or in surveys (Buhrmester et al., 2018). However, in addition to quality control covered later, there have been reports of a number of instances of fraud and other quality issues, especially in MTurk. Some of these have been attributed to fraudulent respondents who typically use VPS/VPN (virtual private servers and networks) to hide their identity (at least nationality or the specific IP address to allow multiple submission, etc.) from the service (Kennedy et al., 2020). For this reason, it has been suggested that data from respondents connecting to recruitment services via VPS/VPN should be discarded. In a similar vein, Kan and Drummey(2018) found that a significant minority of participants were willing to carry out experiments despite not meeting inclusion criteria. Due to these concerns, Kan and Drummey (2018) have proposed some mitigating measures, such as using the demographic pre-screening offered by the recruitment services where possible rather than by stating inclusion criteria when advertising the study to participants.

In addition to fraudulent participants, there have been reports of bots being present in MTurk (Chmielewski & Kucker, 2020; Kennedy et al., 2020; Moss & Litman, 2018). Depending on their sophistication, bots can be identified by them failing even simple quality control checks and should provide low-quality data on even the simplest task. Similarly, bots could be, for example, identified by anomalous response time behavior and by coming from a few specific geolocations, although much of the prevention has to take place at the recruitment services, since they are able to see the full pattern of data and also enforce policies on how new accounts are verified and approved.[5] Analyses carried out by the recruitment services themselves suggest that the problems were not created by bots, but by a small number of foreign workers related to a few server farms. As such, the quality issues should therefore not be a problem as long as sufficient data quality controls and checks are in place within an experiment.

Longitudinal or Interconnected Studies

Longitudinal studies can be challenging in many of the recruitment services. For instance MTurk does not natively support longitudinal studies or studies where participants need to be pre-screened by criteria that are not included in MTurk's standard demographic profile. There are ways to work around this limitation (Stoycheff, 2016), however, and services such as CloudResearch facilitate implementing various operations with the same participants over time that allow running longitudinal studies and they also provide additional quality control, see Chandler et al. (2019). Follow-up studies are supported in Prolific.co and Gorilla, where it is possible to re invite participants who completed previous parts of studies, or to exclude participants who have completed previous related studies.

Sampling Issues

Although participants from recruitment services might offer a more diverse sample than those typically obtained in most of the lab studies using well-educated undergraduate students, participants from recruitment services tend to be younger and less likely to be fully employed than national averages (Mellis & Bickel, 2020). As such, recruitment services do not really provide a population sample but a convenience sample. This is unlikely to be a large problem for auditory research. If prevalence estimates are needed, or research questions require a representative sample, the recruitment service needs to be used in a specific way to capture the characteristics of a representative sample. Some of these services, Prolific.co and Qualtrics Panel, offer a separate service where the researcher can request a representative or a stratified sample from the population for an additional fee.

Research Practices

An obstacle for research utilizing online data collection with recruitment services may be that peer-reviewers may not yet be familiar with using online testing and recruitment services. This can lead to queries voicing concerns about experimental control, sample characteristics and the overall data quality. Depending on the task, it may, of course, be valid to ask for a lab validation. However, validations have been run already for many routine experimental paradigms and tasks, limiting the usefulness of further lab validations. In our experience, reviewers' concerns can be addressed by having included quality control measures, such as headphones screening and attention checks. Given the increased use of online studies with or without recruitment services, we expect that this type of data collection will soon be considered as valid and standard, provided that appropriate quality controls are in place.

Another issue related to samples is that the ease and the speed of research may also tempt scholars to adopt questionable principles such as p-hacking, HARKing, or other problematic practices that go against the traditional use of statistical thresholds (Wicherts et al., 2016). There are solutions to some of these issues such as performing sequential analyses during the data collection (Lakens, 2017) to avoid p-hacking, or pre-registering the study intentions, outcome measures, sample size, and inclusion/exclusion criteria. The use of preprint servers (e.g., PsyArXiv[6] or arXiv[7]) to post all studies soon after they are concluded will guard against the danger of only reporting those iterations of the study that delivered results under the conventional statistical thresholds. But overall, these problems exist in any empirical research and require commitments to research integrity rather than special measures to control online studies using recruitment services.

Online Testing Using a Gamification Approach – Pros and Cons

Gamified online studies utilize "game-design elements in any non-game system context to increase users' intrinsic and extrinsic motivation, help them process information, help them to better achieve goals, and/or change their behavior." (Treiblmaier et al., 2018, p. 134). Gamification has been occasionally utilized in music and audio-related online studies; there are several online games addressing rhythm (Bellec et al., 2013; Duffy et al., 2018), an online game for collecting music similarity data (Wolff et al., 2015), and an

online game for detecting hooks in music (Burgoyne et al., 2013) as well as ongoing projects about musicality and other topics.[8] The benefits of gamified data collection can be substantial (Honing, 2021). Online games can potentially lead to a very large number of participants without paying them anything. They also may help to spread the word about the topic and increase engagement and the impact of research. However, not all studies can be gamified and the very nature of creating a game suitable for anyone may work against research goals or prevent researchers from imposing necessary controls or from collecting crucial background information. Gamified data collection tends to require very bespoke development and typically requires considerable investment in app development or web technologies, although some of the online testing platforms are now starting to include "game-builders" in their services.[9] Generally gamification studies are difficult to replicate as the public interest will wane after the initial wave of curiosity. All critical issues of implementation that are relevant for online studies in the auditory domain in general are relevant to gamified studies as well, but we see gamification as a special approach that may offer a unique combination of engagement that brings in limited data from a diverse yet large sample of interested people.

Key Commitments When Utilizing Recruitment Services in Online Auditory Research

Conducting research with recruitment services rather than traditional online surveys or lab experiments raises ethical questions that we want to address next. We address issues that may be of concern to individual Institutional Review Boards as well as considering good practice more broadly. Whilst some of the points we raise could be discussed in the context of all empirical research, many are unique to online studies using recruitment services and auditory research.

Ethical Considerations

The first assumption is that all empirical research – including both online studies in general and those studies which use recruitment services – should be subject to the local ethics policies, which usually implement the national research integrity and ethics regulations.

The use of recruitment platforms – mainly MTurk – has received negative press in recent years both within and outside the academic community (see, for instance, Semuels, 2018). The concerns raised have suggested shortcomings in the behavior of individual researchers and in the governance of the platforms. The focus of these concerns is frequently financial, with very low rates of remuneration being reported as common and quite possibly the norm (Hitlin, 2016). Despite the median hourly wage of $10.20 reported in Table 2, Hara et al. (2018) have calculated that when nonpayment and returned tasks are taken into account, the median hourly wage on MTurk is in the range of $1.77 – $2.11. Nonpayment refers to the proportion of participants in the service that report being not paid at least once for their work. These numbers are surprisingly large: 60% for Mechanical Turk, 11% for Appen, and 29% for Prolific.co according to http://faircrowd.work. It is also worth pointing out that services such as faircrowd.work are able to pool together experiences from hundreds of participants in these services. In addition

to fair payment issues, there are additional concerns such as anonymity, misrepresentation of task duration or complexity by researchers, or unfair rejection of work (Salehi et al., 2015).

As outlined in the section about *recruitment services*, Prolific.co was launched as a recruitment platform specifically for academic research. Alongside the benefits to researchers that we have presented already, Prolific.co has inbuilt safeguards to ensure fair treatment of participants. Payment is at a recommended minimum rate of $10.26 (£7.50) per hour with a hard minimum of $6.84 (£5) per hour. If researchers reject (i.e., do not pay) a participant, then they must provide the participant with a reason, and there is clear guidance to researchers as to when a participant's submission can and can not be rejected, with suggested alternatives such as allowing the participant to redo the task or offering partial payment. There is also guidance for participants on how to appeal rejection decisions. Prolific' supporting documentation for researchers provides reminders for researchers about these policies which aim to ensure fairness.

Reporting Commitments

As for studies run in the laboratory, researchers should commit to reporting all technical solutions and decisions for studies using online testing and recruitment in a format that enables the replication of these studies (e.g., by specifying the recruitment filter(s), headphone check(s) and its pass-rate, stimulus preparation, visual materials, inclusion/exclusion criteria). It is also important to report the date range of the data collection as the construct of interest may change over time, or we may learn that the recruitment service's participant pool was compromised with "bots," fraudulent participants or a surge of newcomers. There is also the possibility to share the experiment fully via the online platform and also share the recruitment service details within the recruitment tool, thus allowing others to capitalize on the exact same tasks, protocols, stimuli, instruction and sampling criteria. Nothing prevents researchers from releasing the stimuli, design, and the de-individualized data[10] in an Open Access repository[11] and also include the experiment scripts (and analysis scripts, for that matter). Prolific.co promotes transparency in reporting practices, advocating for use of mechanisms such as pre-registration or registered reports. Gorilla hosts Open Materials to share the protocols, tasks, and questionnaires.[12] Likewise Buhrmester et al. (2018) suggests that researchers report in detail on the use of the recruitment platform, for instance, any restrictions on participants' experience, attrition rates, rates of payment, and so on. We suggest that these details would be well-suited to be formal reporting under online experiment protocols, similar to guidelines for reporting experimental protocols in life sciences (Giraldo et al., 2018) or bio- and nanosciences (Faria et al., 2018).

Financial Commitments

Participants from recruitment services often have a dual role as both 'workers' and "participants," and a proportion of participants fall below the Federal Poverty Line (Ipsen et al., 2021), yet several participants report hourly remuneration less than US Federal minimum wage (Hitlin, 2016). Thus, it is the researchers' responsibility to ensure

that rates of remuneration are fair, and do not simply reflect the minimum remuneration possible. To align with lab-based studies, we encourage researchers to commit to pay the respondents *at least the minimum wage*. In case of estimation of the duration of the task being overall too short in comparison to the actual time spent on the task, it is recommended – and is relatively easy in most recruitment systems – to increase the participant fees – or pay bonuses – to reflect the actual time spent on the task. An interesting dilemma is the currency and wage differences between countries. Is the compensation tied to the country of the participant or the researcher? It might be safe to err on the side of higher pay, but that will also set up pressures for participants from low income countries to take part with VPS/VPN posing as coming from another country to make significantly higher earnings.

Fair Use of Recruitment Services

In response to the financial and ethical issues raised about the use of recruitment services by academic researchers, several universities have enacted policies on the use of recruitment platforms in research, and the platforms themselves provide (often non-binding) guidelines on fair and professional treatment of participants.

Transparency

Ensure that tasks are transparent in terms of ownership: There is a clear ownership that can be traced back to the PIs and ethics approvals, much the same as there would be for in-lab experiments or more traditional internet-mediated research. For participants, it should be very clear and transparent what the remuneration is for a task, the amount of time it will take, timescale for payment, and whom to contact if questions arise.

Valuing Participants

Ensure participants are aware of the value of their contribution to the research. Offer a debrief at the closing stages, a lay summary, suggestions for further reading, and thank the participants along with the payment.

Professional Standards

Recruitment services handle the payments in different ways, but one of the main issues for participants is that they are paid fairly and promptly. Some of the services auto-approve participant payments and partial payments (due to failing attention checks or otherwise not completing) are organized in various ways. Also make sure that all responses to participant queries are prompt and professional and possibly utilize similar standards even if handled by separate researchers and research assistants.

Rejection Policy

The policy of rejecting participant contributions should be clear. Rejecting participants' work may have implications in some recruitment services for both payment for the present task and the participant's ability to access other tasks. There are cases where rejection is likely such as when the participant fails to pass attention or headphone checks. In these circumstances participants can be reimbursed only partially (e.g., if

they contributed a few minutes of their time before failing the headphone check) or not at all (e.g., if they fail attention checks over a certain tolerance such as 20% of the attention checks). If it becomes known that a participant has started the task, but not completed it (e.g., from timestamps or direct contact from the participant), it is possible to pay a pro-rata equivalent for the time spent on the task. For data that are not usable, it may be possible to give participants the opportunity to redo the task. It should be noted that some platforms such as Prolific.co reserve the right to overturn rejection decisions.

Quality Control

Over the past 5 years, the authors have utilized Gorilla, Qualtrics and PsyToolkit for online studies and when the need has arisen, turned mostly to Prolific.co or occasionally to MTurk as the recruitment service. Most of the studies we have carried out online have been relatively straightforward data collection exercises without complex elements such as follow-ups or pre-screening, although these have been implemented in some cases. Based on these experiences and following the ongoing scholarly discussion about online studies and use of recruitment services, we want to highlight the topic of quality control, which was not explored in detail during the earlier discussion of pros and cons.

The topic of quality control has received considerable attention with regard to online studies, since the assumption is that the participants working remotely will have more distractions, a wider range of backgrounds and life situations, less uniform expectations of what to do in the studies, or even fraudulent motivations to participate in studies, all of which could lead to unwanted variability in the responses. However, Rodd (2019) suggests that many of the quality control checks designed for online studies should actually also be implemented in lab studies. We agree that quality control should be an inherent part of any data collection, not just limited to online studies. One such operation is to design the experiments in labs and online contexts to be within-subject designs, which mitigates the differences between the different set ups (equipment, volume, etc.) as well as some of the individual differences.

As the data collection environment is not under the experimenters' control in an online study, it is important to provide quality checks that ascertain whether the participant is paying attention to the task at hand and understands the instructions properly. Here we divide these checks into generic attention, technical, consistency, expertise, and honesty checks:

Generic Attention Checks

Checks such as *Instructional Manipulation Checks* (IMCs) can be used. In this the respondent is shown the following text: "You should not answer this question if you read it; it is to check your attention: (1) Strongly Disagree; (2) Disagree; (3) Don't Disagree/Don't Agree; (4) Agree; (5) Strongly Agree." Past research has demonstrated that 16% to 18% of respondents fail IMCs, although this rate is not higher than in lab studies (Paas et al., 2018). Variant attention checks can be tasks that resemble captchas (Completely Automated Public Turing test to tell Computers and Humans Apart) where participants are asked to pick a color, word, or an image.

Domain Specific Attention Checks

Attention checks that rely on auditory information can be used. Such a task can ask participants to "type the two digits you hear in a speech excerpt into the box below," e.g., (Sauter et al., 2020) or be presented as a variant of a captcha relying on timbre, pitch height or another auditory property of interest.

Technical Checks

Technical control tasks typically relate to audio quality such as checks for headphones and general ability to discriminate volume or pitch differences (Milne et al., 2021; Pankovski, 2021; Woods et al., 2017). For production studies such as capturing tapping or singing, there are usually initial checks to test the recording and timing capacities of the computer (Anglada-Tort et al., 2021). One can also use a "honeypot" check which targets only bots by implementing two forms on top of each other, where a human participant only sees one, but any automated script will see and offer responses to both (Downs et al., 2010).

Consistency Checks

Build-in duplication in the form of repeated test items is another way of measuring attention and allows the researcher to analyze the possible inconsistencies in the repeated items.

Expertise Checks

Expertise checks relate to self-disclosed expertise, which is again not unique to online studies as it applies to all studies, but the accessibility of these experiments and the potential payments received from these may encourage prospective participants to mislead the experimenter about their background or expertise. To ensure that participants meet the expertise criteria, it is recommended to test the specific expertise rather than rely on self-reported expertise. To give an example, for expertise about a specific musical genre (for instance, Hindustani classical music), one can devise short, timed statements that require the specific expertise to answer correctly. Or preferably, the questions can be in the form of audio examples ("Is this sound example in North Hindustani (a) Dhrupad, (b) Khyal, (c) Ghazal or (d) Thumri style? (please choose one)").

Honesty Checks

Various types of honesty checks can be implemented after the main task: Participants can be asked whether they truly fit the recruitment criteria, while being assured that their answer will in no way affect their payment. It is known across multiple studies that a noteworthy proportion (3–28%) of participants are dishonest about their qualifications (MacInnis et al., 2020).

The quality control in lab studies is normally implemented post-experiment, with unreliable responses/participants discarded based on pre-defined criteria (such as inter-subject reliability, response speed, or another task-dependent criterion). It would be possible to implement the control protocols within the online experiment and eliminate the inattentive or noisy respondents during the experiment, but in our experience and according to the principles outlined in the payment policies above, we have deemed it safer to assess the quality of the responses and participants after the experiment.

Conclusions

Overall, online studies and studies with recruitment services can be set up relatively easily compared to traditional lab studies, and can often be carried out more quickly. However, this comes with a trade-off in terms of implementation (control of environmental conditions, attention, and audio setup) and recruitment. It is possible to mitigate the implementation issues to some extent, for instance, by use of headphone checks or attention checks embedded within the experiment, but online studies will always be carried out on equipment of variable quality in situations that are varied across the participants. Online studies that draw the participants from recruitment services seem to be less subject to the degradation of quality between lab and online data compared to volunteer web samples recruited for instance via social media. Indeed, we have not found a significant difference in the quality of the data between lab studies and online studies using recruitment services, either in terms of participant attrition or in the distribution of the data itself (which has been the case with online volunteer samples).

Online studies may not be inherently more transparent than traditional lab studies, but many of the open science principles such as pre-registration, sample size determination, and replication are at least somewhat easier to implement than in traditional studies. The transparency is also promoted by some recruitment services (Prolific.co[13]). They also allow the eligibility criteria to be applied automatically and in principle facilitate the recruitment of a highly similar set of participants in subsequent studies. It is also possible to share entire experiments within the online experiment system, making at least direct replications straightforward. Finally, online studies can also avoid specific biases that labs may have in terms of facilities, equipment, experimenter, or instructions.

Overall, online studies with or without recruitment services do not remove the necessity for lab studies in auditory research, but they do allow for good quality data to be collected outside of a lab. Recruitment services can be seen to offer several advantages over convenience samples. As well as offering opportunities for data collection when access to labs is restricted, for instance, during the Covid-19 pandemic, online studies offer benefits in their own right. As scholars in music cognition and other auditory disciplines grow more accustomed to the benefits and challenges of online studies, they are bound to become more frequent in coming years. We hope that our reflections and summaries above are helpful to researchers embarking on their online studies and promote good practice in terms of research transparency, quality control and ethics.

Notes

1. A pool of recent studies was defined by searching Google Scholar with the search terms "headphone check Prolific.co" and "headphone check MTurk", as headphone checks have become an indispensable part of online auditory studies. These keyword combinations generated a corpus of around 100 studies from which we have drawn the majority of our examples.
2. https://pavlovia.org/.
3. https://github.com/Dallinger/Dallinger.
4. https://www.hesa.ac.uk/news/17-01-2019/sb252-higher-education-student-statistics/numbers.

5. Prolific.co's anti-bot measures, see https://blog.prolific.co/bots-and-data-quality-on-crowdsourcing- platforms/ and https://gorilla.sc/online-experiments-and-bots-what-can-be-done/ and analysis of these events at CloudResearch https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it/.
6. https://psyarxiv.com.
7. https://arxiv.org.
8. e.g., https://www.themusiclab.org/.
9. See https://gorilla.sc/product/gorilla-game-builder/.
10. Data from recruitment services is not fully anonymous since the participant recruitment service IDs and IP addresses are often logged into the data.
11. E.g., https://osf.io, https://github.com, or https://dataverse.org.
12. https://app.gorilla.sc/open-materials.
13. https://researcher-help.prolific.co/hc/en-gb/categories/360000850653-Prolific-s-Best-Practice-Guide.

Acknowledgments

We thank Peter M. C. Harrison for insightful comments about gamification approach and for detailing many of the issues of scalability that come with the use of systems such as Dallinger and Pushkin.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

ORCID

Tuomas Eerola http://orcid.org/0000-0002-2896-929X
James Armitage http://orcid.org/0000-0001-9802-7479
Nadine Lavan http://orcid.org/0000-0001-7569-0817
Sarah Knight http://orcid.org/0000-0002-5013-9364

References

Aljanaki, A., Yang, Y.-H., Soleymani, M., & Papadelis, C. (2017). Developing a benchmark for emotional analysis of music. *PloS One*, *12*(3), e0173392. https://doi.org/10.1371/journal.pone.0173392

Almaatouq, A., Becker, J., Houghton, J. P., Paton, N., Watts, D. J., & Whiting, M. E. (2021). Empirica: A virtual lab for high-throughput macro-level experiments. *Behavior Research Methods*, *53*(5), 2158–2171. https://doi.org/10.3758/s13428-020-01535-9

Anglada-Tort, M., Harrison, P. M. C., & Jacoby, N. (2021). REPP: A robust cross-platform solution for online sensorimotor synchronization experiments. *bioRxiv*.

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 , 1407–1425. doi: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

Armitage, J., & Eerola, T. (2020). Reaction time data in music cognition: A comparison of pilot data sets from lab, crowdsourced and convenience web samples. *Frontiers in Psychology*, *10* 2883 . https://doi.org/10.3389/fpsyg.2019.02883

Armitage, J., Lahdelma, I., & Eerola, T. (2021). Automatic responses to musical intervals: Contrasts in acoustic roughness predict affective priming in western listeners. *The Journal of the Acoustical Society of America*, *150*(1), 551–560. https://doi.org/10.1121/10.0005623

Athanasopoulos, G., Eerola, T., Lahdelma, I., & Kaliakatsos-Papakostas, M. (2021). Harmonic organisation conveys both universal and culture-specific cues for emotional expression in music. *PloS One*, *16*(1), e0244964. https://doi.org/10.1371/journal.pone.0244964

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*(3), 800–813. https://doi.org/10.3758/s13428-011-0081-0

Bellec, G., Elowsson, A., Friberg, A., Wolff, D., & Weyde, T. (2013). A social network integrated game experiment to relate tapping to speed perception and explore rhythm reproduction. *Proceedings of the Sound and Music Computing Conference*, 30 July - 3 August Stockholm, Sweden Bresin, R (Berlin, Germany: Logos Verlag Berlin) (pp. 19–26).

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. https://doi.org/10.1093/pan/mpr057

Bhatti, S. S., Gao, X., & Chen, G. (2020). General framework, opportunities and challenges for crowdsourcing techniques: A comprehensive survey. *Journal of Systems and Software*, *167*, 110611. https://doi.org/10.1016/j.jss.2020.110611

Boas, T. C., Christenson, D. P., & Glick, D. M. (2020). Recruiting large online samples in the United States and India: Facebook, Mechanical Turk, and qualtrics. *Political Science Research and Methods*, *8*(2), 232–250. https://doi.org/10.1017/psrm.2018.28

Bradshaw, A., & McGettigan, C. (2021). *Convergence in voice fundamental frequency during synchronous speech*. PloS one 16 10 e0258747 doi:https://doi.org/10.1371/journal.pone.0258747 .

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

Brown, M. A. Z., Violet, A., & Hedayati, A. N. D. (2018). What accounts for individual differences in susceptibility to the McGurk effect? *PloS One*, *13*(11), 1–20. https://doi.org/10.1371/journal.pone.0207160

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

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, *13*(2), 149–154. https://doi.org/10.1177/1745691617706516

Burgoyne, J. A., Bountouridis, D., Balen, J. V., & Honing, H. (2013). Hooked: A game for discovering what makes music catchy. *Proceedings of the 14th Society of Music Information Retrieval Conference (ISMIR)* November 4-8 (New York, US: ISMIR). 245–250 Curitiba, Brazil.

Casey, L. S., Chandler, J., Levine, A. S., Proctor, A., & Strolovitch, D. Z. (2017). Intertemporal differences among MTurk workers: Time-based sample variations and implications for online data collection. *SAGE Open*, *7*(2), 2158244017712774. https://doi.org/10.1177/2158244017712774

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

Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, *51*(5), 2022–2038. https://doi.org/10.3758/s13428-019-01273-7

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, *11*(4), 464–473. https://doi.org/10.1177/1948550619875149

Cuccolo, K., Irgens, M. S., Zlokovich, M. S., Grahe, J., & Edlund, J. E. (2021). What crowdsourcing can offer to cross-cultural psychological science. *Cross-Cultural Research*, *55*(1), 3–28. https://doi.org/10.1177/1069397120950628

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). Are your participants gaming the system? Screening Mechanical Turk workers. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, April 10-15 (New York, US: ACM) Atlanta, USA. (pp. 2399–2402).

Duffy, S., Pearce, M., & Yasin, I. (2018). What makes rhythms hard to perform? An investigation using Steve Reich's clapping music. *PloS One*, *13*(10), e0205847. https://doi.org/10.1371/journal.pone.0205847

Escudero, P., Smit, E. A., & Angwin, A. (2021). *Investigating orthographic versus auditory cross-situational word learning with online and lab-based research*. PsyArXiv. https://doi.org/10.31234/osf.io/tpn5e

Faria, M., Björnmalm, M., Thurecht, K. J., Kent, S. J., Parton, R. G., Kavallaris, M., Johnston, A. P. R., Gooding, J. J., Corrie, S. R., Boyd, B. J., Thordarson, P., Whittaker, A. K., Stevens, M. M., Prestidge, C. A., Porter, C. J. H., Parak, W. J., Davis, T. P., Crampin, E. J., & Caruso, F. (2018). Minimum information reporting in bio–nano experimental literature. *Nature Nanotechnology*, *13*(9), 777–785. https://doi.org/10.1038/s41565-018-0246-4

Feitosa, J., Joseph, D. L., & Newman, D. A. (2015). Crowdsourcing and personality measurement equivalence: A warning about countries whose primary language is not English. *Personality and Individual Differences*, *75*, 47–52. https://doi.org/10.1016/j.paid.2014.11.017

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). *LabVanced: A unified JavaScript framework for online studies*. International Conference on Computational Social Science July 10-13 (Cologne, Germany: IC2S) Cologne, Germany.

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

Giraldo, O., Garcia, A., & Corcho, O. (2018). A guideline for reporting experimental protocols in life sciences. *PeerJ*, *6*, e4795. https://doi.org/10.7717/peerj.4795

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224. https://doi.org/10.1002/bdm.1753

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

Guang, C., Lefkowitz, E., Dillman-Hasso, N., Brown, V., & Strand, J. (2020). Recall of speech is impaired by subsequent masking noise: A replication of Rabbitt (1968) experiment 2. *Auditory Perception & Cognition*, *3*(3), 158–167. https://doi.org/10.1080/25742442.2021.1896908

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., & Bigham, J. P. (2018). A data-driven analysis of workers' earnings on Amazon Mechanical Turk. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* April 21-26 (New York, US: ACM) Montreal, Canada (pp. 1–14 doi:https://doi.org/10.1145/3173574.3174023).

Harrison, P. M. C., & Jacoby, N. (2020). *PsyNet: The online human behavior lab of the future* Accessed16 8 2021. https://www.aesthetics.mpg.de

Harrison, P. M. C., Marjieh, R., Adolfi, F., Rijn, P. V., Anglada-Tort, M., Tchernichovski, O., Larrouy-Maestri, P., & Jacoby, N. (2020). Gibbs sampling with people Advances in Neural Information Processing Systems 33 Larochelle, H, Ranzato, M, Hadsell, R, Balcan, F, and Lin, H (New York, US: Curran Associates, Inc.)10659–10671 https://proceedings.neurips.cc/paper/2020/file/7880d7226e872b776d8b9f23975e2a3d-Paper.pdf .

Harrison, P. M. C., & Müllensiefen, D. (2018). Development and validation of the computerised adaptive beat alignment test (CA-BAT). *Scientific Reports*, *8*(1), 1–19. https://doi.org/10.1038/s41598-018-30318-8

Hartshorne, J. K., Leeuw, J. R., De, 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

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

Heffner, C. C., Newman, R. S., & Idsardi, W. J. (2017). Support for context effects on segmentation and segments depends on the context. *Attention, Perception & Psychophysics*, *79*(3), 964–988. https://doi.org/10.3758/s13414-016-1274-5

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and Brain Sciences*, *33*(2–3), 111–135. https://doi.org/10.1017/S0140525X10000725

Hitlin, P. (2016). *Research in the crowdsourcing age: A case study*. Pew Research Center. https://www.pewresearch.org/internet/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/

Hochheimer, C. J., Sabo, R. T., Perera, R. A., Mukhopadhyay, N., & Krist, A. H. (2019). Identifying attrition phases in survey data: Applicability and assessment study. *Journal of Medical Internet Research*, *21*(8), e12811. https://doi.org/10.2196/12811

Honing, H. (2021). Lured into listening: Engaging games as an alternative to reward-based crowdsourcing in music research. *Zeitschrift für Psychologie*, *229*(4). https://doi.org/10.1027/2151-2604/a000474

Howe, P. D. L., & Lee, S. B. W. (2021). Attribute amnesia in the auditory domain. *Perception*, 03010066211022175 doi:https://doi.org/10.1177/03010066211022175.

Ipsen, C., Kurth, N., & Hall, J. (2021). Evaluating MTurk as a recruitment tool for rural people with disabilities. *Disability and Health Journal*, *14*(1), 100991. https://doi.org/10.1016/j.dhjo.2020.100991

Jacoby, N., & McDermott, J. H. (2017). Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Current Biology*, *27*(3), 359–370. https://doi.org/10.1016/j.cub.2016.12.031

Jakubowski, K., Belfi, A. M., & Eerola, T. (2021). Phenomenological differences in music-and television-evoked autobiographical memories. *Music Perception: An Interdisciplinary Journal*, *38*(5), 435–455. https://doi.org/10.1525/mp.2021.38.5.435

James, E., Gaskell, M. G., Pearce, R., Korell, C., Dean, C., & Henderson, L. (2020). *The role of prior lexical knowledge in children's and adults' word learning from stories*. PsyArXiv. https://doi.org/10.31234/osf.io/vm5ad

Kan, I. P., & Drummey, A. B. (2018). Do imposters threaten data quality? An examination of worker misrepresentation and downstream consequences in Amazon's Mechanical Turk workforce. *Computers in Human Behavior*, *83*, 243–253. https://doi.org/10.1016/j.chb.2018.02.005

Kanber, E., Lavan, N., & McGettigan, C. (2021). Highly accurate and robust identity perception from personally familiar voices. *Journal of Experimental Psychology. General*. https://doi.org/10.1037/xge0001112

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of Advertising*, *46*(1), 141–155. https://doi.org/10.1080/00913367.2016.1269304

Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, *98*(3), 630–644. https://doi.org/10.1016/j.neuron.2018.03.044

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. G. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, *8*(4), 614–629. https://doi.org/10.1017/psrm.2020.6

Knight, S., Lavan, N., Torre, I., & McGettigan, C. (2021). The influence of perceived vocal traits on trusting behaviours in an economic game. *Quarterly Journal of Experimental Psychology*, *74*(10), 1747–1754. https://doi.org/10.1177/17470218211010144

Lahdelma, I., Armitage, J., & Eerola, T. (2020). Affective priming with musical chords is influenced by pitch numerosity. *Musicae Scientiae*, 102986492091112. https://doi.org/10.1177/1029864920911127

Lahdelma, I., Athanasopoulos, G., & Eerola, T. (2021). Sweetness is in the ear of the beholder: Chord preference across United Kingdom and Pakistani listeners. *Annals of the New York Academy of Sciences*, *1502*(1), 72–84. https://doi.org/10.1111/nyas.14655

Lahdelma, I., & Eerola, T. (2020). Cultural familiarity and musical expertise impact the pleasantness of consonance/dissonance but not its perceived tension. *Scientific Reports*, *10*(1), 8693. https://doi.org/10.1038/s41598-020-65615-8

Lakens, D. (2017). *Performing high-powered studies efficiently with sequential analyses*. PsyArXiv. https://doi.org/10.1002/ejsp.2023

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. https://doi.org/10.1002/ejsp.2023

Lange, K., Kühn, S., Filevich, E., & Margulies, D. (2015). "Just Another Tool for Online Studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PloS One*, *10*(6), e0130834. https://doi.org/10.1371/journal.pone.0130834

Lavan, N., Burston, L. F., & Garrido, L. (2019). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology*, *110*(3), 576–593. https://doi.org/10.1111/bjop.12348

Lavan, N., Burston, L. F., Ladwa, P., Merriman, S. E., Knight, S., & McGettigan, C. (2019). Breaking voice identity perception: Expressive voices are more confusable for listeners. *Quarterly Journal of Experimental Psychology*, *72*(9), 2240–2248. https://doi.org/10.1177/1747021819836890

Lavan, N., Knight, S., Hazan, V., & McGettigan, C. (2019). The effects of high variability training on voice identity learning. *Cognition*, *193*, 104026. https://doi.org/10.1016/j.cognition.2019.104026

Lavan, N., Knight, S., & McGettigan, C. (2019). Listeners form average-based representations of individual voice identities. *Nature Communications*, *10*(1), 2404. https://doi.org/10.1038/s41467-019-10295-w

Lavan, N., Kreitewolf, J., Obleser, J., & McGettigan, C. (2021). Familiarity and task context shape the use of acoustic information in voice identity perception. *Cognition*, *215*, 104780. https://doi.org/10.1016/j.cognition.2021.104780

Lavan, N., Merriman, S. E., Ladwa, P., Burston, L. F. K., Knight, S., & McGettigan, C. (2020). 'Please sort these voice recordings into 2 identities': Effects of task instructions on performance in voice sorting studies. *British Journal of Psychology*, *111*(3), 556–569. https://doi.org/10.1111/bjop.12416

Lavan, N., Mileva, M., Burton, A. M., Young, A. W., & McGettigan, C. (2021). Trait evaluations of faces and voices: Comparing within-and between-person variability. *Journal of Experimental Psychology. General*. https://doi.org/10.1037/xge0001019

Lavan, N., Mileva, M., & McGettigan, C. (2021). How does familiarity with a voice affect trait judgements? *British Journal of Psychology*, *112*(1), 282–300. https://doi.org/10.1111/bjop.12454

Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of Mechanical Turk samples. *Sage Open*, *6*(1), 2158244016636433. https://doi.org/10.1177/2158244016636433

MacInnis, C. C., Boss, H. C., & Bourdage, J. S. (2020). More evidence of participant misrepresentation on MTurk and investigating who misrepresents. *Personality and Individual Differences*, *152*, 109603. https://doi.org/10.1016/j.paid.2019.109603

Mellis, A. M., & Bickel, W. K. (2020). Mechanical Turk data collection in addiction research: Utility, concerns and best practices. *Addiction*, *115*(10), 1960–1968. https://doi.org/10.1111/add.15032

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 , 1551–1562 doi:https://doi.org/10.3758/s13428-020-01514-0.

Moss, A., & Litman, L. (2018). After the bot scare: Understanding what's been happening with data collection on MTurk and how to stop it. Accessed17 8 2021 https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it/ ,, .

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, *2*(2), 109–138. https://doi.org/10.1017/XPS.2015.19

Nacke, L. E., & Deterding, C. S. (2017). The maturing of gamification research. *Computers in Human Behavior*, *71*, 450–454. https://doi.org/10.1016/j.chb.2016.11.062

Njie, Lavan, N., & McGettigan, C. (2021). *Talker and accent familiarity yield advantages for voice identity perception: A voice sorting study*. PsyArXiv. https://doi.org/10.31234/osf.io/b6ftg

O'Brien, A. M., & Schmidt, J. L. (2020). Typical listeners are unable to detect sound quality differences between luxury and value headphones. *Cognition, Brain, Behavior*, *24*(1), 57–74 doi:10.24193/cbb.2020.24.04.

Olive, S., Khonsaripour, O., & Welti, T. (2018). *A survey and analysis of consumer and professional headphones based on their objective and subjective performances*. Audio Engineering Society. http://www.aes.org/e-lib/browse.cfm?elib=19774

Paas, L. J., Dolnicar, S., & Karlsson, L. (2018). Instructional manipulation checks: A longitudinal analysis with implications for MTurk. *International Journal of Research in Marketing*, *35*(2), 258–269. https://doi.org/10.1016/j.ijresmar.2018.01.003

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

Pankovski, T. (2021). Screening for dichotic acoustic context and headphones in online crowd-sourced hearing studies. *Canadian Acoustics*, *49*(2). https://jcaa.caa-aca.ca/index.php/jcaa/article/view/3403

Payne, B., Lavan, N., Knight, S., & McGettigan, C. (2021). Perceptual prioritization of self-associated voices. *British Journal of Psychology*, *112*(3), 585–610. https://doi.org/10.1111/bjop.12479

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

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Pfordresher, P. Q., & Demorest, S. M. (2021). The prevalence and correlates of accurate singing. *Journal of Research in Music Education*, *69*(1), 5–23. https://doi.org/10.1177/0022429420951630

Pfordresher, P. Q., & Demorest, S. M. (2020). Construction and validation of the Seattle Singing Accuracy Protocol (SSAP): An automated online measure of singing accuracy Russo, F. A., Ilari, B., and Cohen, A. J. In *The Routledge Companion to Interdisciplinary Studies in Singing* (pp. 322–333). New York, US: Routledge.

Pollet, T. V., & Saxton, T. K. (2019). How diverse are the samples used in the journals 'evolution & human behavior' and 'evolutionary psychology?' *Evolutionary Psychological Science*, *5*(3), 357–368. https://doi.org/10.1007/s40806-019-00192-2

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

Rezlescu, C., Danaila, I., Miron, A., & Amariei, C. (2020 More time for science: Using Testable to create and share behavioral experiments faster, recruit better participants, and engage students in hands-on research Parkin, B. L.). . *Real-World Applications in Cognitive Neuroscience*, *253* (New York, US: Elsevier), 243–262 doi:https://doi.org/10.1016/bs.pbr.2020.06.005.

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

Salehi, N., Irani, L. C., Bernstein, M. S., Alkhatib, A., Ogbe, E., & Milland, K. (2015). We are dynamo: Overcoming stalling and friction in collective action for crowd workers. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* April 18-23 (New York, US: ACM) Seoul, Republic of Korea (pp. 1621–1630 doi:http://dx.doi.org/10.1145/2702123.2702508).

Sauter, M., Draschkow, D., & Mack, W. (2020). Building, hosting and recruiting: A brief introduction to running behavioral experiments online. *Brain Sciences*, *10*(4), 251. https://doi.org/10.3390/brainsci10040251

Schmidtke, D., Gagné, C. L., Kuperman, V., Spalding, T. L., & Tucker, B. V. (2018). Conceptual relations compete during auditory and visual compound word recognition. *Language, Cognition and Neuroscience*, *33*(7), 923–942. https://doi.org/10.1080/23273798.2018.1437192

Semuels, A. (2018). *The internet is enabling a new kind of poorly paid hell*. The Atlantic. https://www.theatlantic.com/business/archive/2018/01/amazon-mechanical-turk/551192/

Sheehan, K. B. (2018). Crowdsourcing research: Data collection with Amazon's Mechanical Turk. *Communication Monographs*, *85*(1), 140–156. https://doi.org/10.1080/03637751.2017.1342043

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*(1), 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

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

Smith, N. A., Sabat, I. E., Martinez, L. R., Weaver, K., & Xu, S. (2015). A convenient solution: Using MTurk to sample from hard-to-reach populations. *Industrial and Organizational Psychology*, *8*(2), 220. https://doi.org/10.1017/iop.2015.29

Speck, J. A., Schmidt, E. M., Morton, B. G., & Kim, Y. E. (2011). A comparative study of collaborative vs. Traditional musical mood annotation. *Proceedings of the 11th Society of Music Information Retrieval Conference (ISMIR)* October 24-28 Miami, US (Vol. *104* (New York, US: ISMIR), pp. 549–554 https://ismir2011.ismir.net/papers/PS4-13.pdf).

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., Bartels, D. M., Newell, B. R., Paolacci, G., and Chandler, J. 2015 The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers Judgment and Decision making. 10(5): 479–491

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

Stoycheff, E. (2016). Please participate in part 2: Maximizing response rates in longitudinal MTurk designs. *Methodological Innovations*, *9*, 2059799116672879. https://doi.org/10.1177/2059799116672879

Tierney, A., Patel, A., Jasmin, K., & Breen, M. (2021). Individual differences in perception of the speech-to-song illusion are linked to musical aptitude but not musical experience. *PsyArXiv*.

Treiblmaier, H., Putz, L.-M., & Lowry, P. B. (2018). Setting a definition, context, and theory-based research agenda for the gamification of non-gaming applications. *Association for Information Systems Transactions on Human-Computer Interaction (THCI)*, *10*(3), 129–163 doi:10.17705/1thci.00107.

Turner, A. M., Kirchhoff, K., & Capurro, D. (2012). Using crowdsourcing technology for testing multilingual public health promotion materials. *Journal of Medical Internet Research*, *14*(3), e79. https://doi.org/10.2196/jmir.2063

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wilkerson, J. M., Iantaffi, A., Grey, J. A., Bockting, W. O., & Rosser, B. S. (2014). Recommendations for internet-based qualitative health research with hard-to-reach populations. *Qualitative Health Research*, *24*(4), 561–574. https://doi.org/10.1177/1049732314524635

Wolff, D., MacFarlane, A., & Weyde, T. (2015). Comparative music similarity modelling using transfer learning across user groups. *Proceedings of the 15th Society of Music Information Retrieval Conference (ISMIR)* 26-30 October (New York, US: ISMIR) Malaga, Spain (pp. 24–30 https://archives.ismir.net/ismir2015/paper/000151.pdf).

Woods, K. J., & McDermott, J. H. (2018). Schema learning for the cocktail party problem. *Proceedings of the National Academy of Sciences*, *115*(14), E3313–E3322. https://doi.org/10.1073/pnas.1801614115

Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception & Psychophysics*, *79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361–2

Yoho, S. E., Borrie, S. A., Barrett, T. S., & Whittaker, D. B. (2019). Are there sex effects for speech intelligibility in American English? Examining the influence of talker, listener, and methodology. *Attention, Perception & Psychophysics*, *81*(2), 558–570. https://doi.org/10.3758/s13414-018-1635-3

Zack, E. S., Kennedy, J., & Long, J. S. (2019). Can nonprobability samples be used for social science research? A cautionary tale. *Survey Research Methods*, *13*(2), 215–227. https://doi.org/10.18148/srm/2019.v13i2.7262