Review

Crowdsourcing in Cognitive and Systems Neuroscience

The Neuroscientist
1–13
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10738584211017018
journals.sagepub.com/home/nro
SAGE

Brian P. Johnson[1], Eran Dayan[2*], Nitzan Censor[3*], and Leonardo G. Cohen[1*]

Abstract

Behavioral research in cognitive and human systems neuroscience has been largely carried out in-person in laboratory settings. Underpowering and lack of reproducibility due to small sample sizes have weakened conclusions of these investigations. In other disciplines, such as neuroeconomics and social sciences, crowdsourcing has been extensively utilized as a data collection tool, and a means to increase sample sizes. Recent methodological advances allow scientists, for the first time, to test online more complex cognitive, perceptual, and motor tasks. Here we review the nascent literature on the use of online crowdsourcing in cognitive and human systems neuroscience. These investigations take advantage of the ability to reliably track the activity of a participant's computer keyboard, mouse, and eye gaze in the context of large-scale studies online that involve diverse research participant pools. Crowdsourcing allows for testing the generalizability of behavioral hypotheses in real-life environments that are less accessible to lab-designed investigations. Crowdsourcing is further useful when in-laboratory studies are limited, for example during the current COVID-19 pandemic. We also discuss current limitations of crowdsourcing research, and suggest pathways to address them. We conclude that online crowdsourcing is likely to widen the scope and strengthen conclusions of cognitive and human systems neuroscience investigations.

Keywords
crowdsourcing, cognitive neuroscience, motor control, motor learning, behavioral neuroscience, systems neuroscience

Introduction

Research in systems and cognitive neuroscience has grown exponentially in the past decade. While systems neuroscience is focused on the structure and function of neural circuits and systems, cognitive neuroscience is centered on the biological processes that underlie cognition. The methods utilized in both disciplines often overlap (i.e., behavioral measurements). Sample sizes used in both subdisciplines are often underpowered due to the time, cost, and invasiveness involved with the chosen method(s), as well as the availability of populations of interest. For example, the median sample size of psychology studies has been found to vary between 40 and 120 (Marszalek and others 2011). While these numbers may suffice to test some hypotheses, they are often underpowered (Button and others 2013; Open Science Collaboration 2015).

The Problem: Reproducibility, External Validity, Power

Recent reports underlined the problem of the lack of reproducibility in these fields (Harris 2017; Munafò 2017; Open Science Collaboration 2015). The use of small sample sizes has weakened the external validity (i.e., generalizability of findings) and conclusions drawn from previous investigations (Button and others 2013). Studies using expensive or immobile equipment (e.g., neuroimaging systems, robotic exoskeletons, transcranial magnetic stimulation systems) require that participants

[1]Human Cortical Physiology and Neurorehabilitation Section, National Institute of Neurological Disorders and Stroke, Bethesda, MD, USA
[2]Department of Radiology and Biomedical Research Imaging Center, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[3]School of Psychological Sciences and Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel

*The second, third, and fourth authors contributed equally to this work.

Corresponding Author:
Brian P. Johnson, Human Cortical Physiology and Neurorehabilitation Section, National Institute of Neurological Disorders and Stroke, Building 10, Room 7D50, 9000 Rockville Pike, Bethesda, MD 20892, USA.
Email: brian.johnson2@nih.gov

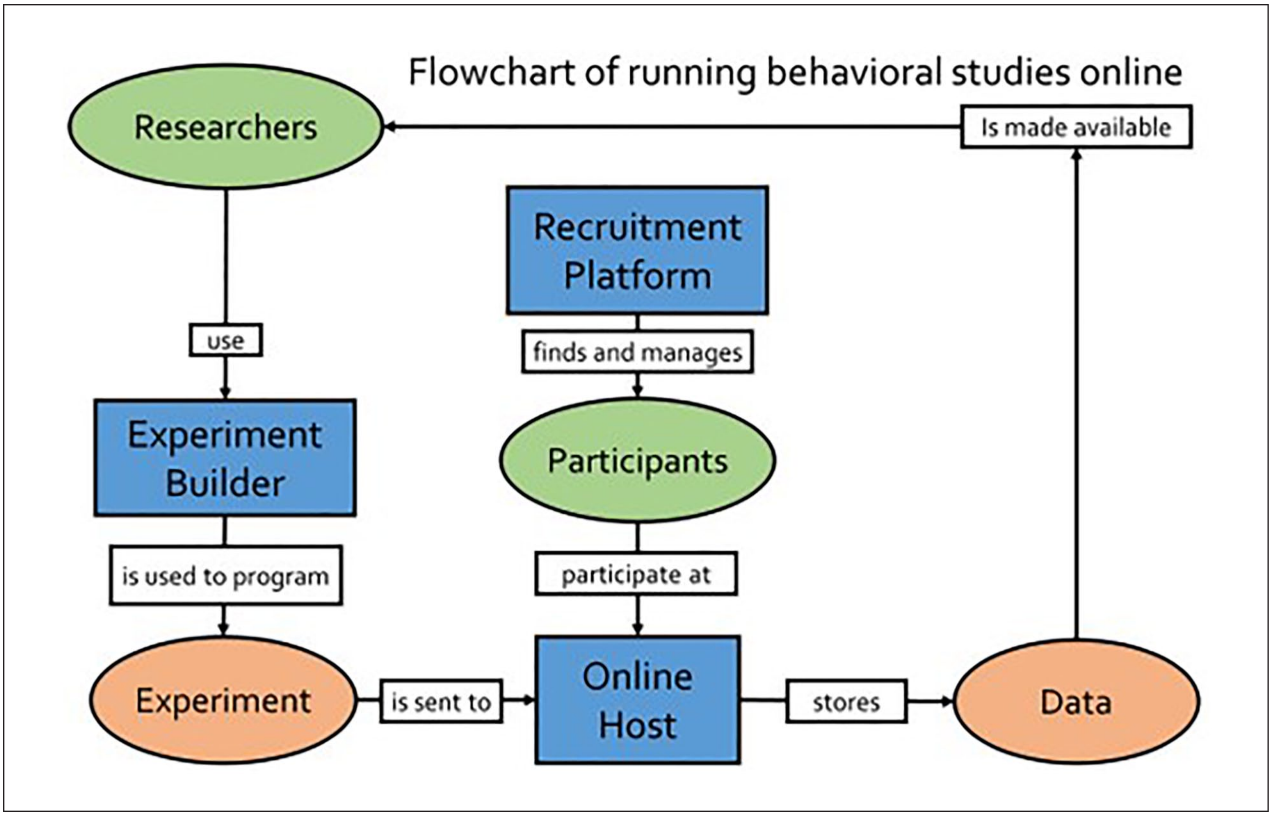


Figure 1. Flowchart of running behavioral studies online as per Sauter and others (2020). After researchers design an experiment, it is set up in the chosen online host website. From there, the experiment is available either on the host website directly or a web link is provided to a third-party website where the online experiment can be accessed by the online participants. The online host recruits participants and facilitates access to the experiment for the participants. The researchers can later access the study data from the online host server and/or third-party web server.

travel to a central location or laboratory to undergo testing, thus limiting the number and diversity of participants (Button and others 2013) (i.e., young college students; "The University Student as a Model Organism," 2010). But even most behavioral investigations that did not require such equipment so far have also been implemented inside laboratories, including small underpowered samples and restricted demographics.

The Solution: Online Crowdsourcing

Strategies to address these problems have included multicenter studies and data sharing consortiums. More recently, crowdsourcing has been used in other fields to acquire large quantities of human data samples. Crowdsourcing in other fields allows collection of large sample sizes in a fraction of the time of in-lab studies (e.g., neuroeconomics; Ashar and others 2017; Genevsky and others 2017; Tong and others 2020). This approach enhanced external validity, improved statistical power, and allowed faster investigation of reproducibility of results from these investigations. But there has been little use of crowdsourcing for cognitive and human systems neuroscience research to date. Here, we introduce the concept of crowdsourcing research, highlight emerging crowdsourcing literature in these fields, and lastly identify possible future research contributions of this novel approach. A more diverse and larger population of participants carrying out the task in their own environment leads to the conclusions of crowdsourced studies to be of greater external validity and real-world significance than traditional in-lab experiments.

Crowdsourcing Increases Sample Size

Crowdsourcing research involves the mass distribution of research-related tasks online for people to complete in exchange for monetary compensation (Fig. 1) (Sauter and others 2020). Websites such as Amazon Mechanical Turk (MTurk) (Buhrmester and others 2011) and Prolific (Palan and Schitter 2018) act as a host to allow for the direct connection of researchers and participants to enable crowdsourcing research to take place. Research-related tasks and experiments can be distributed to either all potential participants who are interested or distributed to a group narrowed by specific demographic factors. Potential participants are notified by the online host of the various available research-related tasks for which they are qualified, and then can choose whether to accept the task or not. Tasks can be conducted within the MTurk or Prolific platforms, or via a linked third-party website within the MTurk or Prolific task description (Anwyl-Irvine and others 2020; Barnhoorn and others 2015; Stoet 2010). These platforms have internal payment systems that can transfer money from the researcher to the participant after a task has been completed. As the pool of unique active users continues to grow, researchers have access to tens of thousands of potential participants to allow for entire tasks or studies to be completed within minutes or hours.

Crowdsourcing Data Analysis

Online crowdsourcing has long been utilized by the social sciences through the use of surveys (Chandler and Shapiro 2016) and has become popular within neuroeconomics to predict large scale behavior based on neuroimaging (Ashar and others 2017; Genevsky and others 2017; Tong and others 2020). For example, Tong and others (2020) utilized functional magnetic resonance imaging (fMRI) of a laboratory sample to forecast the aggregate frequency and duration views of 32 online videos (Fig. 2). The 32 videos were selected from 2,950 video thumbnails rated by participants on MTurk who subjectively rated video thumbnails on clarity, affective arousal, affective valence, and desire to watch the video based on the thumbnail. Thus, crowdsourcing allowed for Tong and others (2020) to use videos in the fMRI experiment, which had various levels of the desired affective ratings. By analyzing fMRI data of participants watching the same internet videos, Tong and others (2020) found that nucleus accumbens activity during video onset was positively correlated with online aggregate view frequency while anterior insula activity during video onset was negatively correlated with online aggregate view duration. But within neuroscience, these paradigms have almost exclusively been used to crowdsource data analysis of data sets (Roskams and Popović 2016). Perhaps most notable is Eyewire (Helmstaedter and others 2013; Marx 2013; Tinati and others 2017), which asks citizen scientists to help analyze brain slice images (Fig. 3). But other examples include visual scoring of electroencephalographs to strengthen machine learning algorithms (Lacourse and others 2020; Warby and others 2014), and providing subjective ratings to stimuli prior to their use in neuroimaging (Freeman and others 2013; Kar and others 2019; Mormann and others 2017; Norman-Haignere and others 2015). That is to say, crowdsourced samples thus far have been most commonly used in neuroscience research to hasten manual data analysis, strengthen machine learning algorithms, or in task development.

Crowdsourcing Strategies

Recent advances now allow for the participant's computer mouse, camera, and microphones to be used to investigate a variety of human behaviors, including motor control and motor learning. Recent crowdsourcing studies have collected data via tracking of the user's computer mouse (Tsay and others 2021; Williams and others 2017) or cameras (Chouinard and others 2019; Madsen and others 2021; Semmelmann and Weigelt 2018). For example, Valliappan and others (2020) demonstrated that smartphone cameras can be used to capture oculomotor movements with similar accuracy to mobile eye trackers 100 times more expensive (Valliappan and others 2020) (Fig. 4). In addition, the increasing use, and decreasing cost, of wearable biotechnologies (Lang and others 2017; Yang and Hsu 2010) and brain-computer interfaces (Peterson and others 2020) also brings forth the exciting possibility of integrating these technologies in cognitive and human systems neuroscience.

Improving Reproducibility

Crowdsourcing allows evaluation of reproducibility of in-lab investigations. For example, Bönstrup and others (2020) (Fig. 5) utilized online crowdsourcing to assess reproducibility and generalizability of a previously obtained in-lab result in daily living situations. The initial in-lab investigation characterized a rapid form of consolidation of motor skill on the scale of seconds (n = 27) (Bönstrup and others 2019). They later reproduced the in-lab behavioral finding in a much larger online sample (n = 389) (Bönstrup and others 2020). Assessing reproducibility of that result through an in-lab experiment would have taken years to complete. Online studies also allow investigators to build assessments of reproducibility into original experimental designs. This can be implemented by simultaneous online testing of the same hypothesis in different groups of individuals in a much faster and cost-effective manner than in a laboratory setting. Validating in-lab studies online then allows for new hypotheses to be explored through online crowdsourcing (Bönstrup and others 2020; Enochson and Culbertson 2015).

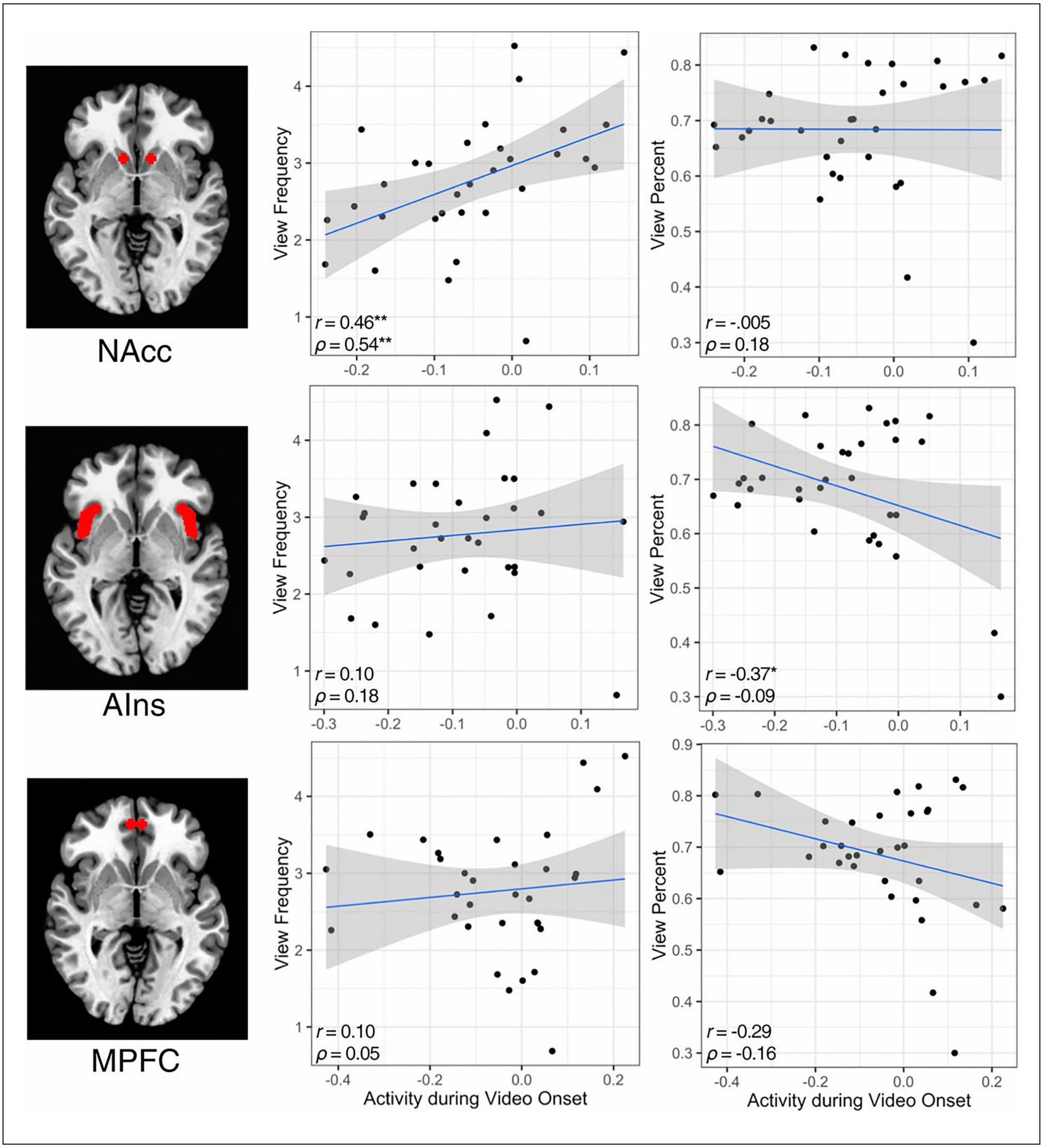
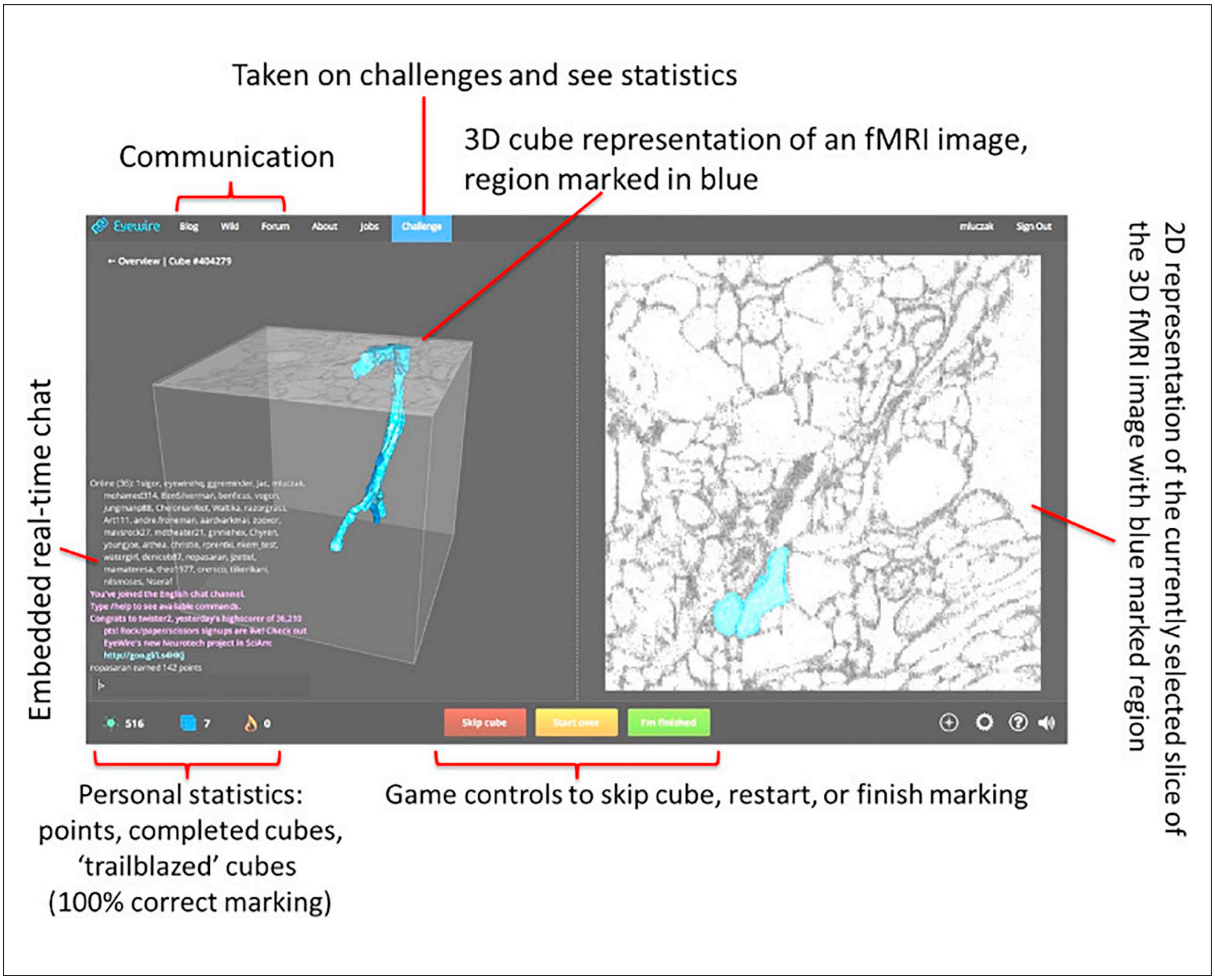


Figure 2. Use of in-lab functional magnetic resonance imaging (fMRI) activity to forecast behavior of a population on the internet. Figure as per Tong and others (2020). Use of fMRI to investigate whether in-lab group neural activity while watching internet videos could forecast aggregate online behavior (i.e., view frequency and duration of views as percentage of video watched) of the same internet videos. (Top) NAcc, (Middle) AIns, and (Bottom) MPFC. Prior to initiation of the fMRI study, a pilot study used online crowdsourcing to select 32 out of 2950 videos to use which had various levels of affective ratings from the online participants (i.e., clarity, affective arousal, affective valence, and desire to watch the video based on the thumbnail). The main findings shown are that NAcc activity during video onset was positively correlated with online aggregate view frequency and AIns activity during video onset was negatively correlated with online aggregate view percentage (i.e., view duration). NAcc, nucleus accumbens; AIns, anterior insula; MPFC, medial prefrontal cortex. $^{*}P < 0.05$; $^{**}P < 0.01$.



Online (36): 1sigor, eyewirehq, ggreminder, jac, mluczak, mohamed314, BenSilverman, benficus, vogon, jungmanp88, ChelonianRiot, Waltka, razorgrass, Art111, andre.froneman, aardvarkmai, zoexor, mavsrock27, mdtheater21, pmnehex, Chyren, youngjoe, athea, chrisba, rprentki, nkem_test, watergirl, denicebb87, nopasaran, jbettel, mamateresa, theo1977, crenico, tikenkani, nlsmoses, Nseraf

You've joined the English chat channel.

Type /help to see available commands.

Congrats to twister2, yesterday's highscorer of 36,210 pts! Rock/paper/scissors signups are live! Check out EyeWire's new Neurotech project in SciAm: http://goo.gl/Ls4HKJ

nopasaran earned 142 points

Taken on challenges and see statistics

Communication

3D cube representation of an fMRI image, region marked in blue

Embedded real-time chat

2D representation of the currently selected slice of the 3D fMRI image with blue marked region

Personal statistics: points, completed cubes, 'trailblazed' cubes (100% correct marking)

Game controls to skip cube, restart, or finish marking

Figure 3. User interface of a crowdsourcing data analysis task called Eyewire. Figure as per Tinati and others (2017). Eyewire is a citizen science project where people online help to analyze neuroimaging data. The work of the citizen scientists is combined with machine learning classifiers to help with analyses of the connectome. The citizen scientists are asked to repetitively follow a single retinal neuron through slices of serial block-face electron microscopy (SBEM) images (right) to create a three-dimensional representation (left). Gamification elements (e.g., real-time chat, points, challenges, leaderboard) are integrated into Eyewire to encourage engagement and motivation.

Crowdsourcing in Exploratory and Hypothesis-Driven Investigations

Beyond reproducibility, crowdsourcing provides advantages to conducting both exploratory and hypothesis-driven research. In exploratory studies, increasingly more systematic investigations can be conducted into phenomena of interest. For example, different study groups can be tested serially or in parallel to carry out a parametric investigation of the effect of varying a task parameter on a given behavior. In hypothesis-driven studies, the ability to recruit large samples will allow for testing of hypotheses with small expected effect sizes, thus requiring a large sample size not commonly feasible during in-lab investigations. The ease of participation in online experiments can allow for longitudinal studies to be conducted with greater frequency of timepoints and longer overall duration. However, it should be noted that longitudinal online experiments pose challenges that single sessions do not. The most important one is the likelihood of higher dropout rates and poorer compliance after the initial session. The exponentially larger

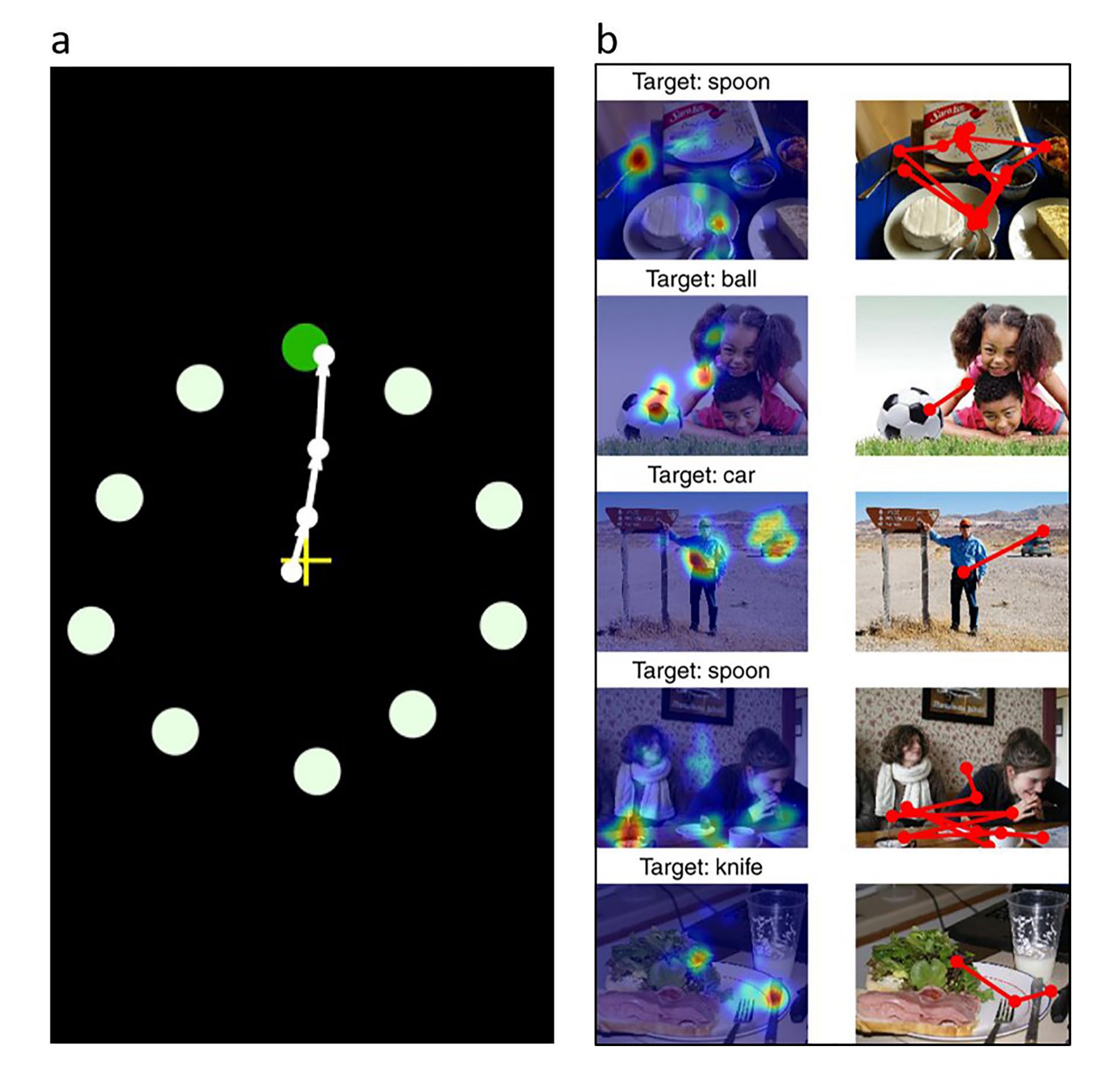


Figure 4. Use of smartphone cameras to track visual gaze scanpath. Figure adapted from Valliappan and others (2020). Valliappan and others (2020) utilized the built-in cameras on smartphones to track gaze scanpaths of participants during visual search tasks. (a) Participants were asked to focus their gaze on a target with high contrast to the other targets. The image shown is an example from a single participant displaying scanpath while searching for the specified target. (b) Participants were asked to focus their gaze on a specified target in each image. Images on the left are examples from a single participant which indicate fixation heatmaps during visual search for a target object. Images on the right are examples from a single participant which indicate visual scanpath while searching for a target object. The authors found that the accuracy of their method used for eye tracking via smartphones is similar to that of mobile eye trackers 100 times more expensive.

recruitment numbers relative to lab investigations are likely to uncover novel mechanisms of behavior. There have indeed been recent unique uses of crowdsourcing platforms in the fields of linguistics (Enochson and Culbertson 2015), and of visual (Kim and others 2019; Panichello and others 2019) and auditory (McWalter and McDermott 2019; Mehr and others 2018) perception.

The New Problem: Internal Validity, Causality

The use of crowdsourcing carries additional challenges like inherently decreased internal validity (i.e., confidence in the scientific methods used to determine causality). Factors contributing to the variability in results from

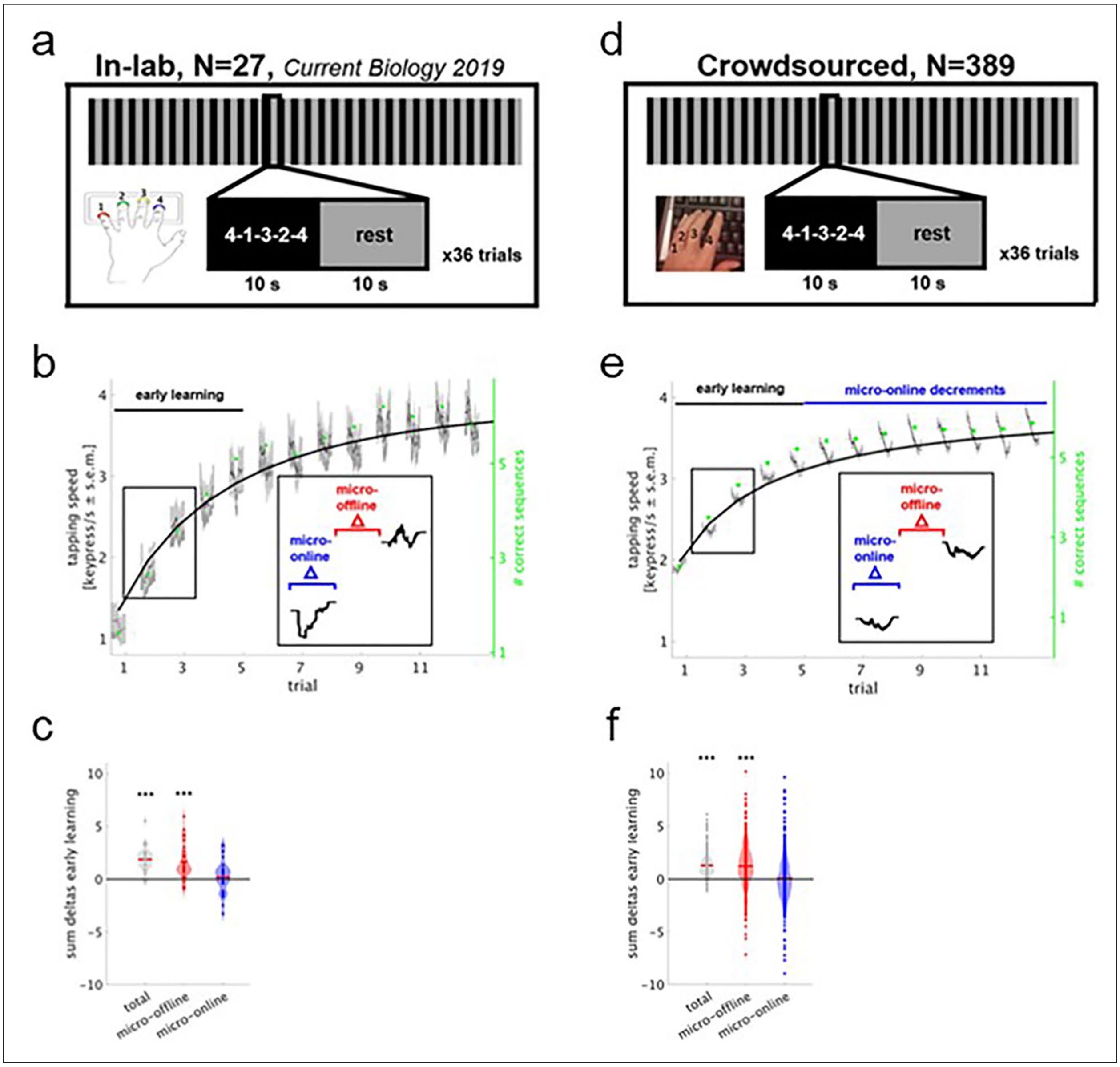


Figure 5. Crowdsourced replication of in-lab motor learning study results. Figure adapted from Bönstrup and others (2020). (a-c) In-lab study. (d-f) Crowdsourced study. (a, d) Motor sequence learning task used in both studies. (b, e) Learning curves reporting tapping speed as mean inter-tap interval within correct sequences (correct keypresses/s) for the in-lab (b) and crowdsourced (e) studies. Both studies characterized micro-online changes (change in tapping speed between first and last correct sequence during a practice period) and micro-offline changes (change in tapping speed between last correct sequence during a practice period and the first correct sequence during the next practice period) in motor skill learning. (c, f) Sum of changes in performance (mean = red line) during early learning (first five trials) for each participant. Note the reproducibility of the learning curves (b and e) and core finding that all early learning was accounted for by micro-offline gains during wakeful rest intervals (c and f). ***$P < 0.001$.

online experiments relative to results from in-lab studies may include differences in participants' setups, environments, hardware and software, as well as engagement (Anwyl-Irvine and others 2020; Bridges and others 2020; Clifford and Jerit 2014; Simcox and Fiez 2014; Garaizar and others 2014; Paolacci 2010; Plant and Turner 2009; Pronk and others 2020; Yung and others 2015). These factors are highlighted below, as well as some of the unique ethical issues that are raised by online crowdsourcing studies.

Crowdsourcing Demographics

While the worldwide pool of registered MTurk users is in the hundreds of thousands (Difallah and others 2018; Robinson and others 2019), the majority of all tasks are carried out by a small percentage of active users (Berinsky and others 2012; Chandler and others 2014; Stewart and others 2015). For example, Chandler and others (2014) report that 41% of all completed task submissions were accounted for by 10% of users on MTurk. Furthermore, a 2010 study reported that almost all MTurk users were located in the United States or India (Paolacci 2010), biasing external validity (i.e., confidence in the generalization of the study conclusions beyond the specific context of the study) in international studies. The MTurk user population has previously not been found to be representative of the U.S. population (Berinsky and others 2012; Corrigan-Gibbs and others 2015; Huff and Tingley 2015; Paolacci and Chandler 2014; Shapiro and others 2013). However, Berinsky and others (2012) noted that MTurk participants were more representative of the U.S. population in general than commonly used college student samples in university based in-lab investigations. As a more recently developed platform, there is less published on the demographics of Prolific. Prolific has been found to have more active users than MTurk (Robinson and others 2019). Fortunately, screening filters are available in both platforms to attempt to recruit demographically representative or specific populations (Kim and others 2018; Yang and others 2014).

Online participants must also contend with the temptation to finish a study as quickly as possible in order to move onto the next study and maximize monetary compensation over time. For some participants, this could be accomplished through not complying with instructions or speeding up in a way that compromises accuracy. This financial incentive also brings the risk of decreased motivation from participants, or even fraudulent responses (Chandler and Paolacci 2017). To contend with this, online crowdsourcing paradigms have allowed for researchers to rate the quality of performance from a given participant, with the ability to also ban the participant from any future studies or even deny payment, though the ethical concern with the latter must be noted.

Because there is a relatively limited number of active users who complete many online studies, there is the possibility that these active users may re-encounter the same or similar commonly used experimental tasks requiring naïve participants, eventually becoming familiar with them. The effect of non-naivete on performing cognitive tasks in online environments has been mixed and seems to depend on the task (Chandler and others 2015; Zwaan and others 2018). Although not explicitly investigated yet, non-naivete is of particular concern for motor learning studies, which often utilize a handful of tasks such as the motor sequence task (Karni and others 1998) and serial reaction time task (Nissen and Bullemer 1987), and can result in long-term retention of the learned motor skill. Participants could then likely demonstrate uncharacteristically fast performance improvements via re-learning (or savings) (Krakauer and others 2005), subsequently skewing study results (Chandler and others 2014; Chandler and others 2015). In addition, individuals could even display transfer or generalization of skill from study tasks that they previously completed to novel study tasks (Seidler 2004). MTurk and Prolific have developed tools to address, at least partially, these potential confounds. These platforms allow for researchers to track the user identification number of previous participants so that research groups can exclude individuals from participating in future studies using the same or similar tasks. However, as the use of online platforms grows and multiple research groups start to study the same tasks, recruiting completely naïve participants will only become more difficult.

Variability of Environment, Setup, Hardware, and Software in Crowdsourcing

Another factor to consider with online crowdsourcing is that participants complete studies in a variety of environments. It has been found that MTurk users tend to be present in distracting environments, where other people may be present or have other visual and auditory stimuli co-occurring (Clifford and Jerit 2014; Simcox and Fiez 2014). In addition, the environment and physical setup of participants will vary with regard to room lighting, physical posture, distance, and gaze angle from the device (Yung and others 2015), among others. Last, researchers must also contend with the inherent variability brought by the many hardware and software options that participants may use. Examples include the dimensions of the devices and hardware being used, as well as response times of software and hardware (Anwyl-Irvine and others 2020; Bridges and others 2020; Garaizar and others 2014; Plant and Turner 2009; Pronk and others 2020). Similarly, internet speed and reliability will be inherently variable across participants. Hypothesis generation should consider that all these factors will influence end-point measures.

Strategies to Improve Data Quality in Online Studies

Fortunately, there are methods that the researcher can implement to decrease the frequency of poor data quality caused by misunderstanding of instructions, decreased attentiveness, decreased motivation, or fraudulent responses. Examples include quizzing participants on task instructions (Crump and others 2013), increasing meaningfulness of the task by informing participants of the scientific and real-world importance of the study (Chandler and Kapelner 2013; Goncalves and others

Box 1. The Importance of Pilot Testing in Online Crowdsourcing Research.

Online crowdsourcing allows for hundreds or thousands of participants to complete a single study within a single day. Given the speed with which data can be collected, it is of the utmost importance that the experiment be fully piloted and troubleshot before being launched online. While it is possible to cancel a study after being launched on MTurk, potential participants who have already seen the link to the study on MTurk prior to the cancellation being initiated can still complete the study. That means, for example, if 700 participants are requested through the MTurk platform and the study is launched but subsequently cancelled seconds or minutes later, there is the potential for hundreds of participants to still complete the study. This makes piloting of the experiment interface, data acquisition, and data quality extremely important before launching the experiment online. But while these factors can typically be piloted and troubleshooted in-house, it is more difficult to determine what instructions will be most effective in optimizing participant understanding of the study and subsequent data quality.

Given that the demographics of individuals in online crowdsourcing studies are more heterogenous than the undergraduate students who are commonly recruited for studies, crowdsourced samples may be less familiar with the goals and methods of some studies. In addition, many workers on MTurk and other platforms participate in multiple studies per day, sometimes in distracting environments, both of which can lead to attentional strains. There is also the temptation to complete as many studies online as possible within a given timeframe to maximize monetary reward, which may lead to skipping through instructions. This is why it is important for researchers to make study instructions as concise as possible, with important points emphasized via underlining, bolding, or highlighting. Video and/or audio instructions could be utilized as well. There is also the possibility of incorporating a set temporal interval for each instructional prompt, or a minimum temporal interval before participants can voluntarily advance to the next prompt. Other strategies to enhance participant understanding include additional prompts repeating key information, questions probing participant understanding of instructions, allowing practice trials (which can include automated corrective feedback), and monetary bonuses for high quality data.

But these various methods also bring forth the possibility of fundamentally changing the study experience for the participant (e.g., potential for a monetary bonus for high data quality), confounding the interpretation and analysis of data collected via multiple instructional methods. It is therefore important to pilot what instructional method(s) produce acceptable data quality from small pilot samples before subsequently launching an experiment seeking a large sample size with one standard instructional experience for all participants.

2015), asking participants to commit to providing high-quality data (Elmalech and Grosz 2017), as well as requiring individuals to list, or choose from multiple choice, demographic screening variables rather than simply asking whether or not they meet the listed criteria (Chandler and Paolacci 2017). Increasing participant compensation has been found to increase motivation and attentiveness (Aker and others 2012; d'Eon and others 2019; Durward and others 2020) (but see Litman and others 2015 regarding data quality) while also increasing the risk for fraudulent responses (Chandler and Paolacci 2017). Investigators may consider incorporating additional measures in their task design that can capture data quality (i.e., measurements of reaction times or eye movements during task performance that can inform about subjects' attention) and later control for such factors during analysis. All of these factors highlight the importance of pilot testing in online crowdsourcing research (see Box 1).

In-Lab and Online Hybrid Studies

The use of in-lab/online crowdsourced hybrid studies can contribute to provide complementary information on a given hypothesis. For example, while crowdsourcing can characterize a reproducible behavioral phenomenon, in-lab investigations can identify the neural substrates underlying that behavior through neuroimaging or the causality of the involvement of those brain regions through brain stimulation or pharmacological interventions. The reverse is also possible. Researchers may consider first validating known in-laboratory findings with crowdsourced samples before adding research questions that can possibly be answered through crowdsourcing (Barnhoorn and others 2015; Bönstrup and others 2020; Crump and others 2013; Enochson and Culbertson 2015; Hilbig 2016; Slote and Strand 2016; Zwaan and Pecher 2012).

Future Directions

Expansion of Crowdsourcing Research

Online crowdsourcing research will likely see a sharp rise in utilization in cognitive and human systems neuroscience research in the near future. Internet access and usage continues to increase worldwide, with over 4.6 billion currently estimated internet users. In addition, the so-called "gig economy" has gained immense popularity worldwide, with workers increasingly turning to part-time work (e.g., ride sharing, apartment rentals, grocery delivery) for additional income. Participating in online studies is in line with the combination of these trends, and rather than necessitating the acquisition of a mode of transportation or living quarters, users require nothing more than an internet-compatible device to earn money by participating in research. As such, it should be expected that the number of users of crowdsourcing research platforms will follow.

Another catalyst of increasing popularity of crowdsourcing research is the COVID-19 pandemic and resulting economic downturns. As a result, many people are unemployed or underemployed, and stay at home orders make other gig economy jobs difficult to perform as well. Additionally, many research institutes and universities have halted or severely limited ongoing studies during the COVID-19 pandemic, including cognitive and human systems neuroscience. It should then be expected that online crowdsourcing platforms will see a rise in the number of users looking to participate in studies, as well as researchers turning to crowdsourcing as an option to continue research.

Conclusions

Laboratory behavioral research in cognitive and human systems neuroscience has long been limited by underpowering, limited demographics, reproducibility, and external validity. Online crowdsourcing research allows for researchers to quickly recruit larger, more representative sample sizes than traditional laboratory experiments. While crowdsourcing has been used in past neuroscience research to aid in data analysis, task development, and strengthen machine learning algorithms, other recent behavioral studies of cognitive and human systems neuroscience have utilized participants' computer (e.g., mouse, camera, keyboard) to conduct hypothesis-driven and exploratory research in a fraction of the time of in-lab studies. However, limitations of online crowdsourcing research include limited internal validity and increased variability of data. While online crowdsourcing research has become more popular in cognitive and human systems neuroscience over recent years, it is likely that the limiting of in-lab research during the COVID-19 pandemic will result in an increase in crowdsourcing research.

Acknowledgments

We are thankful to Michele Richman (Clinical Monitoring Research Program Directorate, Frederick National Laboratory for Cancer Research) for providing helpful feedback regarding this manuscript.

Author Contributions

Conceptualization and methodology: B.P.J., E.D., N.C., and L.G.C. Writing: B.P.J., E.D., N.C., and L.G.C.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported (in part) by the Intramural Research Program of the National Institutes of Health (NIH), National Institute of Neurological Disorders and Stroke (NINDS).

ORCID iD

Brian P. Johnson https://orcid.org/0000-0003-0555-8946

References

Aker A, El-Haj M, Albakour M-D, Kruschwitz U. 2012. Assessing crowdsourcing quality through objective tasks. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association. p. 1456–61. http://www.lrec-conf.org/proceedings/lrec2012/pdf/583_Paper.pdf

Anwyl-Irvine A, Dalmaijer E, Hodges N, Evershed J. 2020. Online timing accuracy and precision: a comparison of platforms, browsers, and participant's devices. https://doi.org/10.31234/osf.io/jfeca

Anwyl-Irvine A, Massonnié J, Flitton A, Kirkham N, Evershed JK. 2020. Gorilla in our midst: an online behavioral experiment builder. Behav Res Methods 52(1):388–407. doi:10.3758/s13428-019-01237-x

Ashar YK, Andrews-Hanna JR, Dimidjian S, Wager TD. 2017. Empathic care and distress: predictive brain markers and dissociable brain systems. Neuron 94(6):1263–1273.e4. doi:10.1016/j.neuron.2017.05.014

Barnhoorn JS, Haasnoot E, Bocanegra BR, van Steenbergen H. 2015. QRTEngine: an easy solution for running online reaction time experiments using Qualtrics. Behav Res Methods 47(4):918–29. doi:10.3758/s13428-014-0530-7

Berinsky AJ, Huber GA, Lenz GS. 2012. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. Polit Anal 20(3):351–68. doi:10.1093/pan/mpr057

Bönstrup M, Iturrate I, Hebart MN, Censor N, Cohen LG. 2020. Mechanisms of offline motor learning at a microscale of seconds in large-scale crowdsourced data. NPJ Sci Learn 5(1):7. doi:10.1038/s41539-020-0066-9

Bönstrup M, Iturrate I, Thompson R, Cruciani G, Censor N, Cohen LG. 2019. A rapid form of offline consolidation in skill learning. Curr Biol 29(8):1346–1351.e4. doi:10.1016/j.cub.2019.02.049

Bridges D, Pitiot A, MacAskill MR, Peirce JW. 2020. The timing mega-study: comparing a range of experiment generators, both lab-based and online. PeerJ 8:e9414. doi:10.7717/peerj.9414

Buhrmester M, Kwang T, Gosling SD. 2011. Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? Perspect Psychol Sci 6(1):3–5. doi:10.1177/1745691610393980

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, and others. 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14(5):365–76. doi:10.1038/nrn3475

Chandler D, Kapelner A. 2013. Breaking monotony with meaning: motivation in crowdsourcing markets. J Econ Behavior Organ 90:123–33. doi:10.1016/j.jebo.2013.03.003

Chandler J, Mueller P, Paolacci G. 2014. Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. Behav Res Methods 46(1):112–30. doi:10.3758/s13428-013-0365-7

Chandler J, Paolacci G. 2017. Lie for a dime: when most prescreening responses are honest but most study participants are impostors. Soc Psychol Pers Sci 8(5):500–8. doi:10.1177/1948550617698203

Chandler J, Paolacci G, Peer E, Mueller P, Ratliff KA. 2015. Using nonnaive participants can reduce effect sizes. Psychol Sci 26(7):1131–9. doi:10.1177/0956797615585115

Chandler J, Shapiro D. 2016. Conducting clinical research using crowdsourced convenience samples. Annu Rev Clin Psychol 12(1):53–81. https://doi.org/10.1146/annurev-clinpsy-021815-093623

Chouinard B, Scott K, Cusack R. 2019. Using automatic face analysis to score infant behaviour from video collected online. Infant Behav Dev 54:1–12. doi:10.1016/j.infbeh.2018.11.004

Clifford S, Jerit J. 2014. Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies. J Exp Polit Sci 1(2):120–31. doi:10.1017/xps.2014.5

Corrigan-Gibbs H, Gupta N, Northcutt C, Cutrell E, Thies W. 2015. Deterring cheating in online environments. ACM Trans Comput Hum Interact 22(6):28.1–28.23. doi:10.1145/2810239

Crump MJC, McDonnell JV, Gureckis TM. 2013. Evaluating Amazon's Mechanical Turk as a Tool for experimental behavioral research. PLoS One 8(3):e57410. doi:10.1371/journal.pone.0057410

d'Eon G, Goh J, Larson K, Law E. 2019. Paying crowd workers for collaborative work. Proc ACM Hum Comput Interact 3(CSCW):125. doi:10.1145/3359227

Difallah D, Filatova E, Ipeirotis P. 2018. Demographics and dynamics of Mechanical Turk workers. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. ACM. p. 135–43. doi:10.1145/3159652.3159661

Durward D, Blohm I, Leimeister JM. 2020. The nature of crowd work and its effects on individuals' work perception. J Manage Inform Syst 37(1):66–95. doi:10.1080/07421222.2019.1705506

Elmalech A, Grosz BJ. 2017. "But you promised": Methods to improve crowd engagement in non-ground truth tasks. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. https://dash.harvard.edu/handle/1/34787806

Enochson K, Culbertson J. 2015. Collecting psycholinguistic response time data using Amazon Mechanical Turk. PLoS One 10(3):e0116946. doi:10.1371/journal.pone.0116946

Freeman J, Ziemba CM, Heeger DJ, Simoncelli EP, Movshon JA. 2013. A functional and perceptual signature of the second visual area in primates. Nat Neurosci 16(7):974–81. doi:10.1038/nn.3402

Garaizar P, Vadillo MA, López-de-Ipiña D, Matute H. 2014. Measuring software timing errors in the presentation of visual stimuli in cognitive neuroscience experiments. PLoS One 9(1):e85108. doi:10.1371/journal.pone.0085108

Genevsky A, Yoon C, Knutson B. 2017. When brain beats behavior: neuroforecasting crowdfunding outcomes. J Neurosci 37(36):8625–34. doi:10.1523/JNEUROSCI.1633-16.2017

Goncalves J, Hosio S, Rogstadius J, Karapanos E, Kostakos V. 2015. Motivating participation and improving quality of contribution in ubiquitous crowdsourcing. Comput Networks 90:34–48. doi:10.1016/j.comnet.2015.07.002

Harris R. 2017. Rigor mortis: how sloppy science creates worthless cures, crushes hope, and wastes billions. Basic Books.

Helmstaedter M, Briggman KL, Turaga SC, Jain V, Seung HS, Denk W. 2013. Connectomic reconstruction of the inner plexiform layer in the mouse retina. Nature 500(7461):168–74. doi:10.1038/nature12346

Hilbig BE. 2016. Reaction time effects in lab- versus web-based research: experimental evidence. Behav Res Methods 48(4):1718–24. doi:10.3758/s13428-015-0678-9

Huff C, Tingley D. 2015. "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. Res Polit 2(3). doi:10.1177/2053168015604648

Kar K, Kubilius J, Schmidt K, Issa EB, DiCarlo JJ. 2019. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. Nat Neurosci 22(6):974–83. doi:10.1038/s41593-019-0392-5

Karni A, Meyer G, Rey-Hipolito C, Jezzard P, Adams MM, Turner R, and others. 1998. The acquisition of skilled motor performance: Fast and slow experience-driven changes in primary motor cortex. Proc Natl Acad Sci U S A 95(3):861–8. doi:10.1073/pnas.95.3.861

Kim J, Cao X, Meczkowski E. 2018. Does stigmatization motivate people to quit smoking? Examining the effect of stigmatizing anti-smoking campaigns on cessation intention. Health Commun 33(6):681–9. doi:10.1080/10410236.2017.1299275

Kim JS, Elli GV, Bedny M. 2019. Knowledge of animal appearance among sighted and blind adults. Proc Natl Acad Sci U S A 116(23):11213–22. doi:10.1073/pnas.1900952116

Krakauer JW, Ghez C, Ghilardi MF. 2005. Adaptation to visuomotor transformations: consolidation, interference, and forgetting. J Neurosci 25(2):473–8. doi:10.1523/JNEUROSCI.4218-04.2005

Lacourse K, Yetton B, Mednick S, Warby SC. 2020. Massive online data annotation, crowdsourcing to generate high quality sleep spindle annotations from EEG data. Sci Data 7(1):190. doi:10.1038/s41597-020-0533-4

Lang CE, Waddell KJ, Klaesner JW, Bland MD. 2017. A method for quantifying upper limb performance in daily life using accelerometers. J Vis Exp 122:55673. doi:10.3791/55673

Litman L, Robinson J, Rosenzweig C. 2015. The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. Behav Res Methods 47(2):519–28. doi:10.3758/s13428-014-0483-x

Madsen J, Júlio SU, Gucik PJ, Steinberg R, Parra LC. 2021. Synchronized eye movements predict test scores in online video education. Proc Natl Acad Sci U S A 118(5):e2016980118. doi:10.1073/pnas.2016980118

Marszalek JM, Barber C, Kohlhart J, Holmes CB. 2011. Sample size in psychological research over the past 30 years. Percept Motor Skills 112(2):331–48. doi:10.2466/03.11.PMS.112.2.331-348

Marx V. 2013. Neuroscience waves to the crowd. Nat Methods, 10(11):1069–74. doi:10.1038/nmeth.2695

McWalter R, McDermott JH. 2019. Illusory sound texture reveals multi-second statistical completion in auditory scene analysis. Nat Commun 10(1):5096. doi:10.1038/s41467-019-12893-0

Mehr SA, Singh M, York H, Glowacki L, Krasnow MM. 2018. Form and function in human song. Curr Biol 28(3):356–368.e5. doi:10.1016/j.cub.2017.12.042

Mormann F, Kornblith S, Cerf M, Ison MJ, Kraskov A, Tran M, and others. 2017. Scene-selective coding by single neurons in the human parahippocampal cortex. Proc Natl Acad Sci U S A 114(5):1153–8. https://doi.org/10.1073/pnas.1608159113

Munafò M. 2017. Metascience: reproducibility blues. Nature 543(7647):619–20. doi:10.1038/543619a

Nissen MJ, Bullemer P. 1987. Attentional requirements of learning: evidence from performance measures. Cogn Psychol 19(1):1–32. doi:10.1016/0010-0285(87)90002-8

Norman-Haignere S, Kanwisher NG, McDermott JH. 2015. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron 88(6):1281–96. doi:10.1016/j.neuron.2015.11.035

, and Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. Science 349(6251):aac4716. doi:10.1126/science.aac4716

Palan S, Schitter C. 2018. Prolific.ac—a subject pool for online experiments. J Behav Exp Finance 17:22–7. doi:10.1016/j.jbef.2017.12.004

Panichello MF, DePasquale B, Pillow JW, Buschman TJ. 2019. Error-correcting dynamics in visual working memory. Nat Commun 10(1):3366. doi:10.1038/s41467-019-11298-3

Paolacci G. 2010. Running experiments on Amazon Mechanical Turk. Judgment Decis Making: 5(5):9.

Paolacci G, Chandler J. 2014. Inside the Turk: understanding Mechanical Turk as a participant pool. Curr Dir Psychol Sci 23(3):184–8. doi:10.1177/0963721414531598

Peterson V, Galván C, Hernández H, Spies R. 2020. A feasibility study of a complete low-cost consumer-grade brain-computer interface system. Heliyon 6(3):e03425. doi:10.1016/j.heliyon.2020.e03425

Plant RR, Turner G. 2009. Millisecond precision psychological research in a world of commodity computers: new hardware, new problems? Behav Res Methods 41(3):598–614. doi:10.3758/BRM.41.3.598

Pronk T, Wiers RW, Molenkamp B, Murre J. 2020. Mental chronometry in the pocket? Timing accuracy of web applications on touchscreen and keyboard devices. Behav Res Methods 52(3):1371–82. doi:10.3758/s13428-019-01321-2

Robinson J, Rosenzweig C, Moss AJ, Litman L. 2019. Tapped out or barely tapped? Recommendations for how to harness the vast and largely unused potential of the Mechanical Turk participant pool. PLoS One 14(12):e0226394. doi:10.1371/journal.pone.0226394

Roskams J, Popović Z. 2016. Power to the people: addressing Big Data challenges in neuroscience by creating a new cadre of citizen neuroscientists. Neuron 92(3):658–64. doi:10.1016/j.neuron.2016.10.045

Sauter M, Draschkow D, Mack W. 2020. Building, hosting and recruiting: a brief introduction to running behavioral experiments online. Brain Sci 10(4):251. doi:10.3390/brainsci10040251

Seidler RD. 2004. Multiple motor learning experiences enhance motor adaptability. J Cogn Neurosci 16(1):65–73. doi:10.1162/089892904322755566

Semmelmann K, Weigelt S. 2018. Online webcam-based eye tracking in cognitive science: a first look. Behav Res Methods 50(2):451–65. doi:10.3758/s13428-017-0913-7

Shapiro DN, Chandler J, Mueller PA. 2013. Using Mechanical Turk to study clinical populations. Clin Psychol Sci 1(2):213–20. doi:10.1177/2167702612469015

Simcox T, Fiez JA. 2014. Collecting response times using Amazon Mechanical Turk and Adobe Flash. Behav Res Methods 46(1):95–111. doi:10.3758/s13428-013-0345-y

Slote J, Strand JF. 2016. Conducting spoken word recognition research online: validation and a new timing method. Behav Res Methods 48(2):553–66. doi:10.3758/s13428-015-0599-7

Stewart N, Ungemach C, Harris AJL, Bartels DM, Newell BR, Paolacci G, and others. 2015. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. Judgment Decis Making 10(5):13.

Stoet G. 2010. PsyToolkit: a software package for programming psychological experiments using Linux. Behav Res Methods 42(4):1096–104. doi:10.3758/BRM.42.4.1096

, and The university student as a model organism. 2010. Nat Neurosci 13(5):521. doi:/10.1038/nn0510-521

Tinati R, Luczak-Roesch M, Simperl E, Hall W. 2017. An investigation of player motivations in Eyewire, a gamified citizen science project. Comput Hum Behav 73:527–40. doi:10.1016/j.chb.2016.12.074

Tong LC, Acikalin MY, Genevsky A, Shiv B, Knutson B. 2020. Brain activity forecasts video engagement in an internet attention market. Proc Natl Acad Sci U S A 117(12):6936–41. doi:10.1073/pnas.1905178117

Tsay JS, Lee AS, Ivry RB, Avraham G. 2021. Moving outside the lab: the viability of conducting sensorimotor learning studies online. BioRxiv 2021.01.30.181370. doi:10.1101/2021.01.30.181370

Valliappan N, Dai N, Steinberg E, He J, Rogers K, Ramachandran V, and others. 2020. Accelerating eye movement research via accurate and affordable smartphone eye tracking. Nat Commun 11(1):4553. doi:10.1038/s41467-020-18360-5

Warby SC, Wendt SL, Welinder P, Munk EGS, Carrillo O, Sorensen HBD, and others. 2014. Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. Nat Methods 11(4):385–92. doi:10.1038/nmeth.2855

Williams P, Jenkins J, Valacich J, Byrd M. 2017. Measuring actual behaviors in HCI research—a call to action and an example. AIS Trans Hum Comput Interact 9(4):339–52. doi:10.17705/1thci.00101

Yang C-C, Hsu Y-L. 2010. A review of accelerometry-based wearable motion detectors for physical activity monitoring. Sensors (Basel) 10(8):7772–88. doi:10.3390/s100807772

Yang K, Friedman-Wheeler DG, Pronin E. 2014. Thought acceleration boosts positive mood among individuals with minimal to moderate depressive symptoms. Cogn Ther Res 38(3):261–9. doi:10.1007/s10608-014-9597-9

Yung A, Cardoso-Leite P, Dale G, Bavelier D, Green CS. 2015. Methods to test visual attention online. J Vis Exp 96:52470. doi:10.3791/52470

Zwaan RA, Pecher D. 2012. Revisiting mental simulation in language comprehension: six replication attempts. PLoS One 7(12):e51382. doi:10.1371/journal.pone.0051382

Zwaan RA, Pecher D, Paolacci G, Bouwmeester S, Verkoeijen P, Dijkstra K, and others. 2018. Participant nonnaiveté and the reproducibility of cognitive psychology. Psychon Bull Rev 25(5):1968–72. doi:10.3758/s13423-017-1348-y