A peer-reviewed version of this preprint was published in PeerJ on 23 July 2015.

View the peer-reviewed version (peerj.com/articles/1058), which is the preferred citable publication unless you specifically need to cite this preprint.

Woods AT, Velasco C, Levitan CA, Wan X, Spence C. 2015. Conducting perception research over the internet: a tutorial review. PeerJ 3:e1058 https://doi.org/10.7717/peerj.1058

Conducting perception research over the internet: A tutorial review

Andy T. Woods[1,2], Carlos Velasco[1], Carmel A. Levitan[3], Xiaoang Wan[4], & Charles Spence[1]

1. *Crossmodal Research Laboratory, Department of Experimental Psychology, University of Oxford, United Kingdom*
2. *Xperiment, Surrey, United Kingdom.*
3. *Department of Cognitive Science, Occidental College, Los Angeles, United States of America.*
4. *Tsinghua University, Beijing, China*

DATE: MARCH 2015

WORD COUNT: 14,900 WORDS

SUBMITTED TO: *PeerJ* (MARCH, 2015)

CORRESPONDENCE TO: Dr. Andy Woods, Crossmodal Research Laboratory, Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, UK; E-MAIL: andytwoods@gmail.com, TEL: +44 1865 271307; FAX: +44 1865 310447.

ACKNOWLEDGEMENTS: We are grateful for feedback on an earlier draft of this manuscript by Rochelle LaPlante, Kristy Milland, Marcus Munafò, and Six Silberman.

ABSTRACT

This article provides an overview of the literature on the use of internet-based testing to address questions in perception research. Internet-based testing has several advantages over in-lab research, including the ability to reach a relatively broad set of participants and to quickly and inexpensively collect large amounts of empirical data. In many cases, the quality of online data appears to match that collected in laboratory research. Generally speaking, online participants tend to be more representative of the population at large than laboratory based participants. There are, though, some important caveats, when it comes to collecting data online. It is obviously much more difficult to control the exact parameters of stimulus presentation (such as display characteristics) in online research. There are also some thorny ethical considerations that need to be considered by experimenters. Strengths and weaknesses of the online approach, relative to others, are highlighted, and recommendations made for those researchers who might be thinking about conducting their own studies using this increasingly-popular approach to research in the psychological sciences.

KEYWORDS: INTERNET-BASED TESTING; CITIZEN SCIENCE; MECHANICAL TURK.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

Introduction

Over the last few years, the rapid growth of online research has revolutionized the way in which many experimental psychologists choose to conduct (at least some of) their research. On the one hand, it holds the promise of allowing the researcher to go well beyond the typical constraints of the Western, Educated, Industrialised, Rich, and Democratic (WEIRD, see Henrich et al., 2010) pools of participants who form the basis for the vast majority of psychological research. Internet-based testing also opens-up the possibility of conducting research cross-culturally (e.g., Knoeferle et al., 2015; Woods et al., 2013). Furthermore, the experience of many of those researchers who have started to work / publish in this area is that relatively large numbers of participants (>100) can be collected in a relatively short space of time (e.g., in less than 24 hrs, and often in less than 1 hr) at relatively low cost (1-2 USD / participant / 10 minutes). Generally-speaking, such data collection can be achieved with relatively little effort on the part of the experimenters concerned.

On the downside, however, concerns have been expressed about the lack of control over certain factors, such as the inevitable lack of control over the precise parameters of stimulus presentation (for example, screen resolution / display characteristics), not to mention the lack of experimenter supervision of the participants while taking part in these studies. Another issue of concern is just how often supposedly anonymised data makes its way onto the web, whilst still containing details that can indirectly, and often directly, reveal participant identity.

Nevertheless, despite these various concerns and limitations, there has been a rapid and dramatic growth in the number of studies that have been published using online testing over the last few years (see Figure 1). We would argue that the far larger sample sizes that one typically attracts when engaged in online testing, and the much broader diversity of such samples, can more than make up for many of the lacks of control that one is faced with as an experimenter. Indeed, conducting large-scale studies online, when in combination with laboratory-based experiments offering finer control over the testing situation and stimuli may be an attractive, not to mention economic strategy for a variety - or indeed perhaps the majority - of future psychological research in the area of perception. For the time being, though, such studies are limited to the delivery of visual and auditory stimuli.

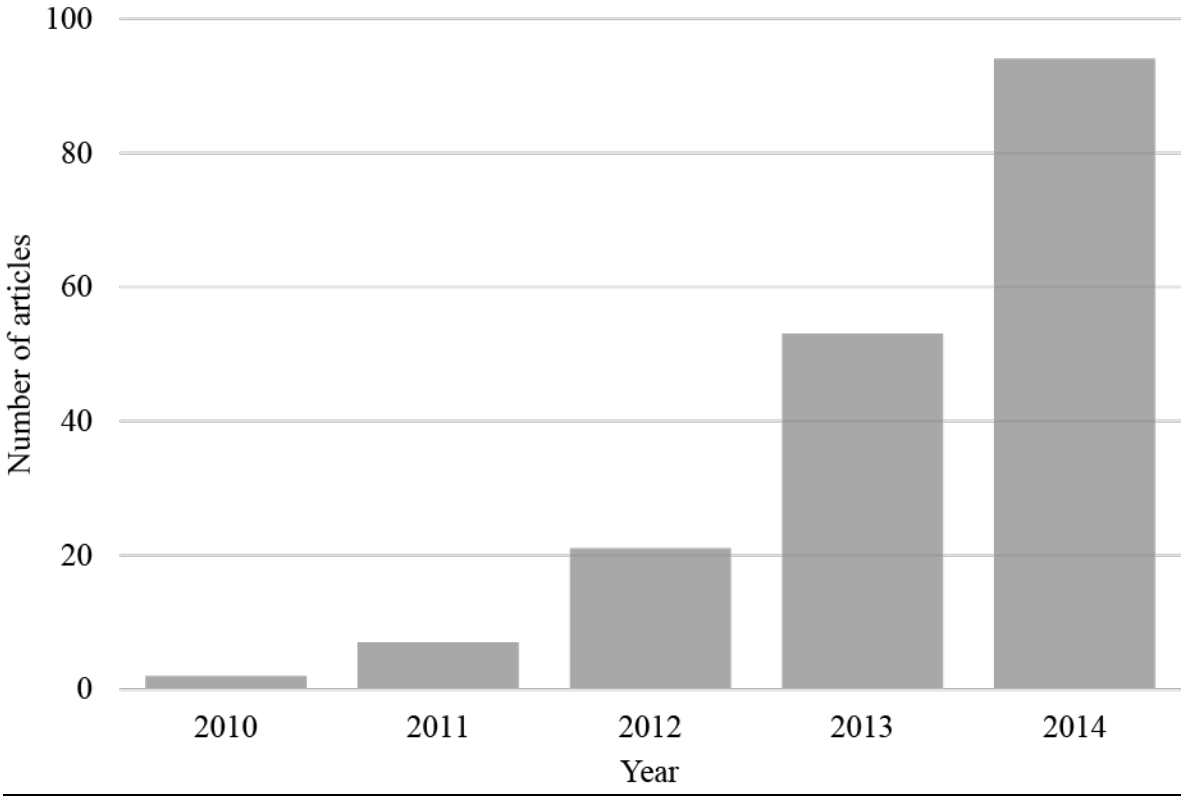


Figure 1: The number of articles found on the Web of Science prior to 2015 with the search term 'Mechanical Turk' within the 'psychology' research area (search conducted on 12-03-2015).

Outline of the present article

In the present article, the challenges and benefits of online research are critically evaluated. First, we highlight how much more representative of the population at large online participants are as compared to their lab-based counterparts, as well as how rapid and economical the collection of data can be online. Following on from this, we explore the various concerns that have been raised with regard to online research, focusing on timing-related issues and how the wide variety of hardware / software that may be used by one's participants can give rise to data problems; the common concern about the lack of supervision of the participants themselves will also be dealt with. Although warranting a paper unto itself, we briefly touch on some of the ethical issues pertaining to online research. We also provide an overview of the main online testing platforms

that are currently available to researchers. Finally, we end by drawing some general conclusions and highlighting what we see as the most promising opportunities for future research.

Benefits of conducting research online

Online research (conducted on a computer with access to the internet), has a number of potential benefits over more traditional laboratory-based studies, which will be evaluated in this section. In particular, we discuss how online research can profit from more representative and diverse samples of participants, as well as the more efficient collection of large amounts of data, and simpler participant payments.

Access to a more representative sample of participants

Online research is less affected by sampling from pools of participants who can be categorized as being WEIRD (Henrich et al., 2010) than traditional laboratory-based research (e.g., Behrend et al., 2011; Berinsky et al., 2012; Chandler et al., 2014; Goodman et al., 2013). So what is known about the characteristics of online recruits? In terms of demographics, the ratio of female to male participants is approximately matched and the average age of the participants is currently estimated to be around 30 years of age (as found in several recent large-sampled online-studies; see Table 1 for a summary). The distribution of ages is typically ex-Gaussian (as often seen with reaction time data; Mason & Suri, 2012). Figure 2 demonstrates this right-sided long-tailed distribution from one of our own recent online studies (Woods et al., in prep). Here it is worth noting that one consequence of the much broader range of ages targeted by online research is that it makes it easier to collect data from older participants.

Table 1: Age and sex characteristics of 4 recent large internet- and phone-based sample studies. Note that 12.5% of Mason and Suri's (2012) participants did not report their gender.

	Recruitment platform	Sample	n	% female	Average age (SD)

Shermer & Levitan (2014)	MTurk	US	2737	40%	29.9 (9.6)
Germine et al. (2012)	TextMyBrain	World	4080 (study 1)	65%	26 (11)
Mason & Suri (2012)	MTurk	World	2896 (5 studies)	55%	32
Buhrmester et al. (2011)	MTurk	World	3006	55%	32.8 (11.5)

In terms of their ethnicity, Berinsky et al. (2012) contrasted US participants recruited through Amazon's Mechanical Turk (MTurk; a popular platform for recruiting participants online) with a sample purported to closely match the US population at large (Mathew, Krosnick, & Arthur, 2010). These researchers found that 83.5% of the Mechanical Turkers (Mturkers) were white (versus 83.0% from the general population), 4.4% were Black (versus 8.9%) and 6.7% were Hispanic (versus 5%; for interested readers, the authors also compared other differences such as marital status, income, housing state and religion and found some between-groups variations). Not restricting themselves to North Americans, Paolacci, Chandler, and Ipeirotis (2010) found that 47% of MTurkers, recruited over a 3-week period in February, 2010, were from the US and 34% from India. As a side note, see Milland (2014b) for a thought-provoking account on some of the hurdles faced by those outside of the US trying to earn a living via MTurk.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

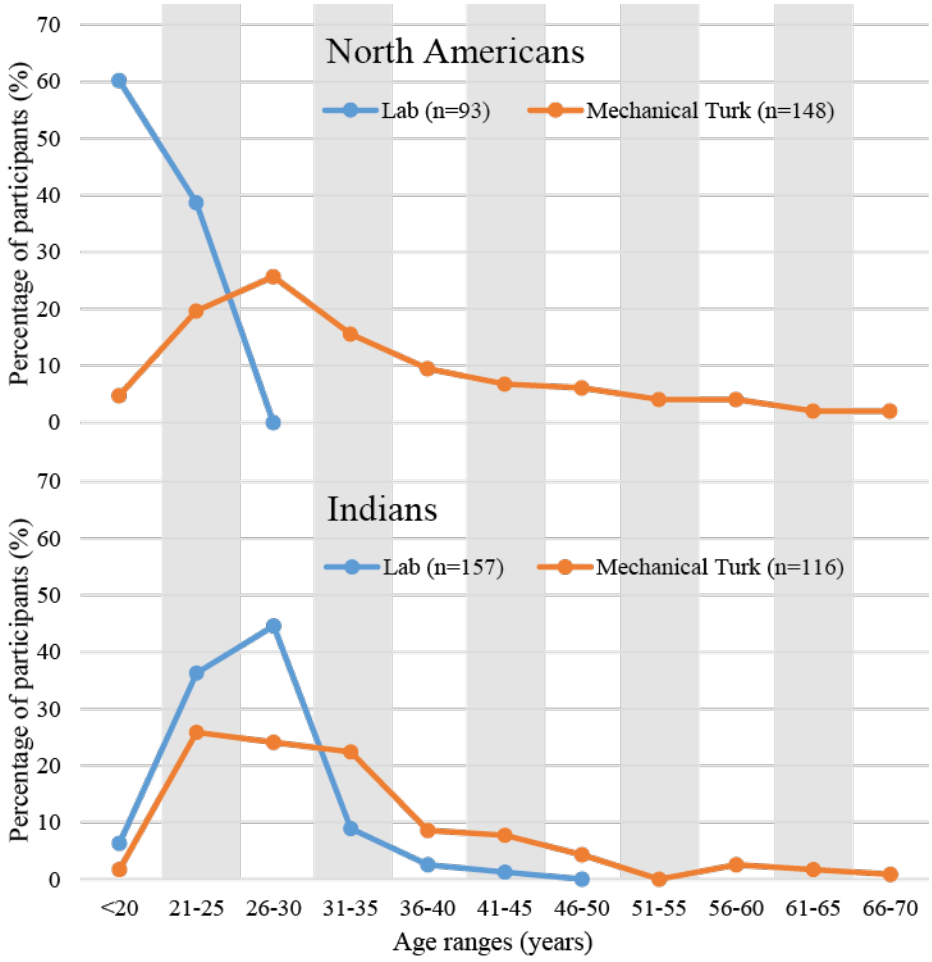


Figure 2: The distribution of ages for US and Indian participants recruited via Mechanical Turk or tested in a lab-based setting (Woods et al., in prep).

It is important to recognize that online participants might also have their own peculiarities. So, for example, Paolacci and Chandler (2014) have recently highlighted how, as a group, MTurkers are typically more computer literate, better educated, and less likely to be employed than the regular population; indeed, the authors argue that it is just such differences in computer literacy that may drive some of the key differences between the online participants and the population at large (Shapiro et al., 2013). Perhaps fitting with this 'geek' stereotype, Mason and Suri (2012) found that MTurkers tend to be less extraverted and less emotionally stable than those recruited from off the street, whilst also being more open to new experiences. The authors also reported that over half of the participants whom they tested reported being on a relatively low wage (≤30,000 USD).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

Goodman et al. (2012) directly tested how participants recruited through Mechanical Turk differed from those recruited on the street in a middle class area, presumably near Washington University in the United States. No discernible difference was found between the groups in terms of their age, sex, or level of education. However, 27.5% of the MTurkers had English as their second language as compared to just 10.5% of those recruited from the street. The prevalence of self-reported clinical conditions such as depression matched that seen in the general population (Shapiro et al., 2013), and 95.5% of MTurkers started some form of college education (Martire & Watkins, 2015).

Thus, despite the above-mentioned variations from the general population, online participants do seem to be more representative of the population at large than those typically recruited for laboratory-based studies, in that a broader age range and a more equal distribution of males and females sign up to take part in studies, who would appear to be equally susceptible to clinical conditions such as depression (which has been shown to impact on perceptual processing, Fitzgerald, 2013), but may be more educated than others in their offline community.

Here, it is also worth asking, whether online participants might not be WEIRD *enough* to successfully complete studies? Could it be, for example, that those with too little research experience respond haphazardly, thus distorting the pattern of results that is obtained? Not so, according to Chandler, Mueller, and Paolacci, (2014) who found that over 132 experiments, 16,408 participants had undertaken an average of 2.24 studies (standard deviation of 3.19), with participants actually likely having taken part in tens or even hundreds of studies. Indeed, Rand et al (2014) explicitly asked 291 MTurkers how many studies they had taken part in, and found that the median worker had participated in approximately 300 scientific studies (20 in the previous week; n.b. some Mturkers actively avoid academic research, Rochelle LaPlante & Kristy Milland, personal communication, March 18, 2015) compared to 15 studies as self-reported by 118 students as part of a participant pool in Harvard (1 in the previous week).

So rather than potentially impacting results due to participant naivety, the results of research conducted online may instead be skewed because of participant overfamiliarity. Indeed, the repercussions of conducting many studies throughout the day has led to a discussion about whether certain MTurkers may not end up becoming rather 'robotic' in their responding (Marder, 2015). It is likely though that the field of perceptual psychology that focuses on the more automatic features of the human brain, would be less affected by such issues as compared to more cognitive fields of

psychology. Another topic of concern is the high dropout rate that is sometimes exhibited by online studies (e.g., Crump et al., 2013). Here the interested reader is directed to Marder's (2015) excellent article on the topic of familiarity.

It is not surprising that a variety of forums and online tools have arisen to help those taking part in online research (especially on Mechanical Turk) and one concern is that online experiments and their presumed goals are a focus of discussion via these tools (note that on some forums such as MTurkGrind.com, reddit.com/r/mturk, and TurkerNation, such comments are quickly reported to moderators and deleted, Rochelle LaPlante & Kristy Milland, personal communication, March 18, 2015). However, according to Chandler, Mueller, and Paolacci (2014), whilst 28% of their 300 participants reported visiting Mechanical Turk orientated forums and blogs, it was the amount a task paid (ranked as most important) and its duration (ranked second most important) that were discussed most often as compared to, for example, a task's purpose (which was ranked sixth).

One can wonder what the impact would be of an experimenter recruiting through their own social media channels or their laboratory websites, as it is likely that people similar to the experimenter are the ones likely to undertake a study so advertised (a phenomenon known as homophily, e.g., Aiello et al., 2012). Indeed, in January 2015, the Pew Research Center reported a range of substantial differences of opinion between the North American public at large and their scientific community, ranging from issues pertaining to eating genetically modified foods (88% scientists in favour, versus 37% of the general public), that humans have indeed evolved over time (98% versus 65%) and that climate change is mostly attributable to human activity (87% versus 50%). In some sense, then, one might want to consider whether we scientists might actually not be the WEIRDest of them all?

In summary, although online participants most certainly have their own peculiarities as compared with the population at large, it is doubtful whether this WEIRDness is any more pronounced than that shown by the undergraduates who take part in our lab-based research. The very fact that classical studies have been successfully replicated in both groups, each with their own peculiarities, is actually reassuring in itself.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

Access to large pools of participants

One of the most important advantages of conducting online research is the speed and ease with which large amounts of data can be collected. In laboratory-based experiments, researchers typically test participants individually or in small groups over a period of several days, weeks, or even months. Unfortunately, this in-person testing of participants can introduce noise, attributable to, for instance, differences in task explanation (though see the article by Mirams, Poliakoff, Brown, & Lloyd, 2013, highly praised on twitter, where the researchers attempted to avoid this issue by, amongst other things, making sure that each participant received their instruction by means of an audio-recording) or even basic demographic differences can influence performance on psychological tasks (e.g., Marx & Goff, 2005; Rumenik, Capasso, & Henrick, 1977). Perhaps most pertinently, research assistants / researchers can provide subtle unintentional cues to the participants regarding how to respond to the task at hand (e.g., see Doyen, Klein, Pichon, & Cleeremans, 2012; Intons-Peterson, 1983; Orne, 1962). As Orne noted a little over half a century ago, there is a social psychological element to any in-lab psychology study. Furthermore, the scheduling of participants takes time, and depending on the specific participant pool, there may be a significant number of participants who do not turn up or else who turn up late to their appointed experimental session. That said, paid for tools such as SonaSystems and Experimetrix nowadays help by automating much of the sign-up process and can also send out reminder emails (https://www.sona-systems.com/, http://www.experimetrix.com/; see also the soon to be released open source LabMan toolbox, https://github.com/TheHandLaboratory/LabMan/). Another drawback of much of the laboratory-based research is that it can be difficult to run multiple participants in parallel, because of experimenter constraints, as well as limits on experimental set-ups / space.

By contrast, with online research, when utilizing the appropriate recruitment platform (the focus of the next section), massive numbers of people can undertake a study at any one time. What is more, the availability of participants is not limited by the vagaries of the academic year, with participation in many university settings being much more prevalent in term time than out of term time (unfortunately compounding this issue, students who receive course credit as opposed to payment for taking part in studies are both less motivated and have been shown to display less sustained attention at the end of the term as compared to the start, Nicholls, Loveless, Thomas, Loetscher, & Churches, 2015). Note that outside of term time there are more participating

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

MTurkers, which in all likelihood correlates with an increased number of student MTurkers looking to earn some money (Rochelle LaPlante, personal communication, March 18, 2015). There can also be severe challenges associated with scaling up one's sample sizes in the laboratory-setting, whereas online, the pool of potential participants would appear to be more than large enough for most questions (Mason & Suri, 2012). Another practical benefit of conducting research online is that the payment of participants can often be automated; that is, the researcher need only make one payment instead of many, and does not need to collect hundreds of individual receipts from participants, minimising their interaction with their financial department.

Recruitment platforms

There are several online resources for the recruitment of participants online, with perhaps the most well-known being Mechanical Turk. Although this platform is primarily aimed at letting those working in industry recruit many individuals to do tasks related to business such as categorising photos or rating website content (see also Innocentive, oDesk, and CloudCrowd; Chandler et al., 2013), the last few years have seen an increasing number of psychological studies starting to use the service (e.g., Crump et al., 2013). In 2014, Mechanical Turk claimed to have half a million individuals registered (Paolacci & Chandler, 2014). However, more recent research suggests that the active potential (US) participants available for a typical study are more likely to number *only* ten thousand (Stewart, Ungemach, Harris, Bartels, & Newell, submitted). Unfortunately, in the summer of 2014, Mechanical Turk stopped allowing new 'requesters' (individuals wanting others to complete a task), and new 'workers' in 2012 (participants), to sign-up who did not have sufficient credentials identifying them as residing in the United States[1], such as US bank accounts and Social Security Numbers (do see http://ai.reddit.com/r/mturk/#ai, for some personal testimonials on the issue). Consequently, many researchers have begun to explore alternatives to Mechanical Turk (or rely on third-party tools such as www.mTurkData.com, to continue having access to Mechanical Turk). One alternative service aimed specifically at academic research is

[1] Although the reasons for this are still unknown, a first wave cull in 2012 had been thought to be due issues of fraud regarding unscrupulous MTurkers gaming the system (Admin, 2013). It is the first author's belief though that the 2014 change occurred as pressure had been put on MTurk to ensure all its workers were tax identifiable.

Prolific Academic (https://prolificacademic.co.uk/), which, as of January 2015, had just over 4000 people signed up to take part in research, with just under 1000 new recruits signing up each month.

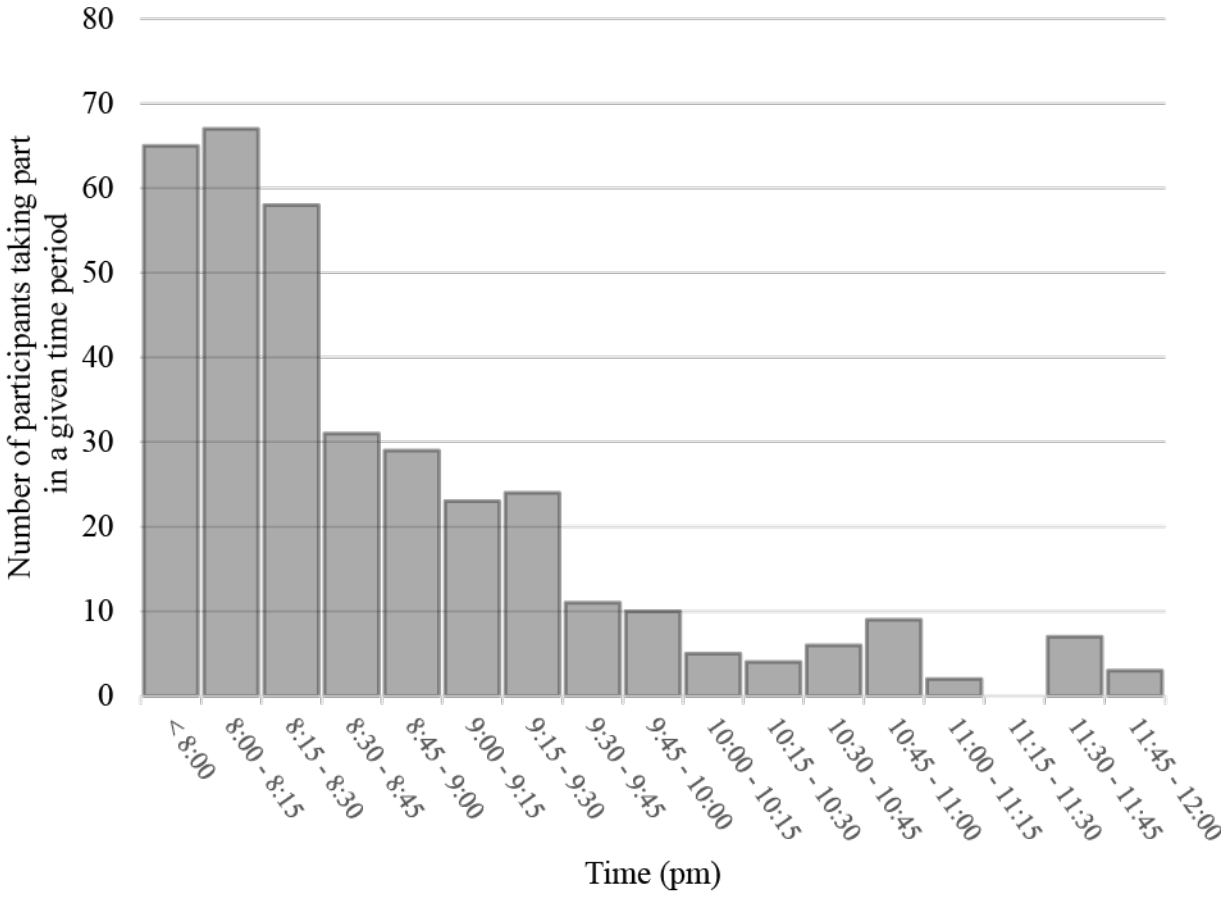
Besides providing a ready source of participants, recruitment platforms also let researchers recruit from specific sub-populations. With Mechanical Turk, for example, researchers can specify whether they wish to recruit participants just from the US, or from several countries, or from anywhere in the world (permitting that there are Mturkers from those countries). Going one step further, Prolific Academic lets researchers specify a range of criteria for recruitment, such as native language, age, sex, and even ethnicity.

Unfortunately, one limitation with the existing platforms is that there is little variability in terms of the country from which one can recruit. For example, in 2010, 47% of MTurkers were North American and 34% from India (Paolacci, et al., 2010). It is no surprise that given the aforementioned new sign-up policy for MTurk, the percentage of Americans taking part in recent years has become much larger (87%, Lakkaraju, 2015). Prolific Academic, on the other hand, has no such demographic blockade, with, as of 7/11/2014, participants predominately coming from the US, UK, Ireland, and India (42%, 33%, 4%, and 2% respectively; https://prolificacademic.co.uk/about/pool; note though, that individuals without an academic email address could only sign-up as from a few months ago). It would seem likely that the list of represented countries will grow as the platform continues to expand.

It is important to note that large swathes of potential participants from around the world still remain untapped! How would one go about recruiting participants from China, for instance, or from Colombia? Whilst sites such as TestMyBrain.org demonstrate that it is possible to recruit large numbers of participants from English-speaking countries (76%) via social networking sites and search engines (e.g., n=4080, in Germine et al.'s, 2012, first study), it is much harder to directly recruit only from specific countries. One option here is to create your own research panel and recruit people via the local / social media (e.g., Wilson, Gosling, & Graham, 2012). A whole range of commercial software solutions exists for such a purpose; unfortunately, we are not aware of any open source alternatives (instead, we have developed our own, https://github.com/ContributeToScience/participant-booking-app).

Speed of data collection

As online research is typically conducted in parallel, a large number of participants can be recruited and take part in a study in a short space of time. With Mechanical Turk, for example, 100s of participants can sign up to take part within just 15 minutes of publicly releasing a study, as shown from one of our recent studies (see Figure 3; Pechey, Attwood, Munafo, Scott-Samuel, Woods & Marteau, in prep). The obvious benefit is the ability to rapidly explore a given scientific issue, as demonstrated by the 'what colour is that dress' viral, and Tarja Peromaa's admiral effort of collecting data from 884 participants within a few days of 'that dress' going public (Peromaa, 2015).



PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

Figure 3: The rate of experiment completion over a four-hour period (n=360; collected February, 2015, from 8pm onward, Eastern Standard Time; Pechey et al., in prep). The first author's suspicion is that 'long tail' sign-ups typically observed in MTurk are the result of some participants signing up and then quitting a study, and the resultant 'time-out' delay before a new person can take the unfinished slot.

Another benefit of the rapid collection of data is that it allows the researcher to explore the impact of day-based events, such as Valentine's Day, Christmas, or Ramadan. Indeed, as 100s of participants can be tested in less than an hour, this opens up the opportunity of testing individuals on even finer-grained timescales. Needless to say, global time differences come into their own with such rapid sign-up.

Of course, one important caveat with the rapid large-scale sign-up of participants is that if there happens to be a flaw in one's study, the experimenter could potentially (and rightly) receive hundreds of angry emails, each of which often require individual attention (failure to do so in the past has led to some researchers being blacklisted en-masse by MTurkers; Kristy Milland, personal communication, March 18, 2015). Not only should the experiment work flawlessly with no ambiguity in terms of the instructions given (as there are no experimenters present to clarify the situation), the server-hardware running the study also needs to be up to scratch. From our own experience with the Xperiment research platform (http://www.xperiment.mobi), our first studies were run from a basic Amazon Cloud server (t1.micro; see https://aws.amazon.com/ec2/previous-generation/), assuming that it could meet any demands that could be thrown at it. However, we had not expected the sheer volume of requests to the server for one particularly demanding video-streaming study, which caused the server to crash. We now run our studies from a more substantial server (m1.small), and are in the process of providing our participants with a live public messaging tool to contact the experimenter regarding any vagaries of experimental design. Needless to say, it is particularly important with online research to pilot one's study on several systems (perhaps with your colleagues) before releasing it to the online community. Indeed, if the study is to be done on MTurk, be advised that the platform provides a 'testing ground' on which you can do your own experiments, and ensure that MTurk and your software are properly communicating (https://requester.mturk.com/developer/sandbox). We also suggest gradually increasing the

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

required sample size of one's early studies (perhaps testing 10 participants with a study, then 50, then…) to ensure that the equipment can deal with demand.

Economical

In general, collecting data from participants online provides an economical means of conducting research, with, for example, the experimenter not having to cover the fees sometimes associated with participants travelling to and from the lab. Those who have looked at whether the payment amount influences how many people are willing to take part (i.e., how many sign-up) and / or how seriously they take the experiment (discussed later in the section entitled 'Random responding') have, perhaps somewhat surprisingly, shown little relation between reward and effort; only the rate of recruitment seems to be influenced by payment size (Paolacci & Chandler, 2014; Mason & Watts, 2009; though do see Ho, Slivkins, Suri, & Vaughan, 2015, who show that bonuses can sometimes improve task performance). A different picture emerges however, when you ask MTurkers themselves what motivates people to take part in low wage research. In an excellent blog article, Spamgirl, aka Kristy Milland (2014a), highlights how low wage tasks will tend to be avoided by all, except by, for example, those using 'bots' to help automate their answers, and those in a desperate situation – which very likely impacts on data quality (see also Silberman, Milland, LaPlante, Ross, & Irani, 2015).

There most certainly is responsibility on the side of the experimenter to ensure fair payment[2] for work done, there being no minimum wage on, for example, Mechanical Turk (see the guidelines written by MTurkers themselves for conducting research on this platform, http://guidelines.wearedynamo.org/). A sensible approach to payment may be to establish what the participant would fairly earn for an hour-long study and then scale the payment according to task duration. For example, Ipeirotis (2010) reported paying participants $1 for a 12.5 minute study ($4.80 / hour) and Berinsky et al. (2012) $6 / hour. An online discussion by MTurkers themselves suggests that 10 cents / minute as a *minimum* going rate (Iddemonofelru, 2014); do note that this

[2] Tangentially, the lead author of this paper was approached after a recent talk given on the topic of this manuscript and was queried as to whether it was fair to pay online participants more than others do in their online studies, as this would presumably drive up the cost for all. Although a valid point, it is the lead author's view that such a payment ethos is not fair on the participants themselves.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

is below the current minimum wage in the US, which is $7.25 per hour, and 10 cents / minute in actual fact is especially unfair for many of those trying to earn a living on MTurk (Rochelle LaPlante, personal communication, March 18, 2015). A fairer rate that we have decided to adopt in our own research is 15 cents / minute (as used by the third party service www.mTurkData.com). A keen eye will spot that the wages reported above seem to increase year-by-year, which may be down to the increasing proportions of North American MTurkers compared to other nationalities. Do be advised though, that those researchers using Mechanical Turk who are seen as offering too small a financial incentive to participants for taking part in their study are often, and rightly so, the focus of negative discussion on social media. Indeed, researchers should be aware that tools have been developed that let the MTurkers evaluate the people providing their tasks (e.g., https://turkopticon.ucsd.edu/), one of the parameters being 'fairness' in terms of pay (the others being 'fast' again in terms of pay, 'fair' in terms of disputes, and 'communication' in terms of ease of reaching the scientist; see Figure 4 for the first author's TurkOpticon Profile).

Figure 4: an example TurkOpticon requester profile (74 MTurkers having provided feedback on the requester).

Of course, money is by no means the only motivating factor for those individuals who take part in online research. Germine et al.'s (2012) successful replication of three classical studies online (more on this study later) were based on data collected at http://www.TestMyBrain.org, where, in exchange for partaking in the study, the participants were told how they performed in comparison to the 'average' participant. To put into perspective how popular this kind of approach can be, between 2009 and 2011, half a million individuals took part in a study on TestMyBrain. Indeed, the rising popularity of such 'citizen science' projects would, we argue here, also offer an incredible opportunity for other areas of science (see also https://www.zooniverse.org/ and https://implicit.harvard.edu/implicit/).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

Cross-cultural research

The ability to write one experiment and run it with many participants from different cultures is appealing (although language translation can be effortful). For example, identifying the extent to which certain percepts are culturally-mediated has been useful for understanding a range of perceptual phenomena, including colour vision (e.g., Berlin, 1991; Kay & Regier, 2003), music perception (e.g., Cross, 2001), and, from some of us, crossmodal correspondences / expectations (e.g., Levitan et al., 2014; Wan et al, 2014a, 2014b, 2015).

Focusing on one example in more detail, Eriksson and Simpson (2010) were able to explicitly test both US and Indian participants in their study on emotional reactions to playing the lottery, as they collected their data from Mechanical Turk where the Workers are mostly from these two countries. Via an online questionnaire, they asked their participants about whether they were willing to enter a specific lottery, as well as how they would feel about losing or winning it. Their results revealed that the female participants were less willing to enter the lottery than the male participants, though the Indian participants were generally more willing to enter the lottery than the North Americans. What is more, their results also revealed that both male and female participants who were willing to enter the lottery gave lower ratings on how bad they would feel about losing than their counterparts who were not willing to enter the lottery. These findings allowed the researchers, at least partially, to attribute the gender difference in risky behaviour to the different emotional reactions to losing. Importantly, the researchers were able to observe the same result patterns (of gender difference, and of the linkage between willingness-to-risk and anticipated emotional reactions to losing) in two samples from different countries, and also documented the cross-country difference in risky behaviour.

Complimenting in-lab research

As will become apparent in the sections below, it is unlikely that online research will subsume everything that is done in the lab anytime soon. We believe, though, that online research can certainly provide an especially helpful tool with which to complement laboratory-based research. For example, if research is exploratory in nature, conducting it online first may help the researcher scope out hypotheses, and prune out those alternatives that have little support. Subsequent lab based research can then be run on the most promising subset of hypotheses. For example, our own

research on crossmodal associations between basic tastes and the visual characteristics of stimuli, started out 'online', where we explored basic associations between these elements. After having found a link between round shapes and the word sweet, we then moved into the lab to test with real sweet tastants in order to tease out the underlying mechanisms (see Velasco et al., 2015a; Velasco, Woods, Liu, & Spence, in press; see also Velasco et al., 2014, 2015b, for another example of complementary online and offline research).

As online participants are less WEIRD than those in lab-based studies, following-up a lab-based study with one conducted online may help strengthen the generalizability of one's initial findings or, by means of a much larger sample, offer more conclusive proof of one's findings (e.g., Knoeferle et al., 2015, Woods et al., 2012).

Comparing online with in-lab

Whilst questionnaire-based research readily lends itself to the online environment, a common belief is that reaction-time (RT) studies, or those requiring fine temporal / spatial stimulus control cannot readily be conducted online. Perhaps surprisingly then, to date, the majority of the comparative non-questionnaire based studies that have been conducted, running essentially the same study online and in the laboratory, have provided essentially consistent results. Indeed, this is perhaps especially surprising, given the current replication crisis sweeping through the field of psychology (Pashler & Wagenmakers, 2012); although do consider that this preference to publish findings based on low-power statistically significant effects as opposed to insignificant effects may in actual fact be why the majority of online findings mirror those in lab. Of course, an alternative scenario is that when faced with significant lab findings that don't replicate online, only the lab findings eventually get published – the insignificant findings, as sadly is often the case, getting relegated to the file drawer (Simonsohn, Nelson, & Simmons, 2014; Spellman, 2012; refreshingly, the journal Psychological Science now require authors to declare, amongst other things, that "all independent variables or manipulations, whether successful or failed, have been reported in the Method section(s)", Eich, 2014, p4).

Explicitly testing whether traditional lab based studies would work online, Germine et al. (2012) successfully replicated five tasks that were thought to be particularly susceptible to issues such as lapses in attention by participants and satisficing ('cheating', see Oppenheimer et al., 2009).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

Examples of such tasks were the Cambridge Face Memory Test, where faces are shown for three seconds to be remembered later (Duchaine & Nakayama, 2006) and the Forward Digit Span task, which is concerned with the number of digits that can be recalled after being shown serially, one after the other each for one second (Wechsler, 2008). Germine et al. (2012, p. 847) concluded that *'...web samples need not involve a trade-off between participant numbers and data quality.'*

Similarly, Crump et al. (2013) replicated eight relatively well-established laboratory based tasks online, which the authors categorised as being either RT-based (such as the Erikson Flanker task, Erikson, 1995), focused on memory (e.g., concept learning, Shepard, Hovland, & Kenkins, 1961), or requiring the stimuli to be presented for only a short period of time. The only task that was not completely replicated, a masked priming task (Eimer & Schlaghecken, 1998), was in this latter category, where visual leftward or rightward pointing arrows were (it was assumed; more on this later) presented for 16, 32, 48, 64, 80, 96 ms and the participant's task was to indicate the direction in which the arrows pointed. In contrast to the original lab based study by Eimer and Schlaghecken, the authors did not replicate the expected effects for stimuli of durations of 16-64 ms and concluded that short duration stimuli cannot be reliably shown when conducting internet-based research.

In the same vein, the Many Labs study (Klein et al., 2014) directly compared 13 effects across 36 samples and settings, including multiple online samples. The online samples came from universities, from Mechanical Turk, and from a different online platform (Project Implicit) that did not pay participants. Across all of these samples, very little difference in effect size was seen between online and in-lab data.

The majority of the replication attempts by Germine et al. (2012), Crump et al. (2012) and Klein et al. (2014) were successful. It would seem that only a subset of studies, specifically those requiring short stimulus presentation, are not so well suited to online research. Indeed, as mentioned by Crump et al., as technology improves, it is likely that even this category of task may be achieved satisfactorily online. Who knows, there may come a time when laboratory and online research are on a par in terms of data quality — indeed, given a disagreement between such studies, one could argue that the effects from more ecologically valid scenario, the person being tested at home in online research, should be treated preferentially as they would more likely also occur in the population at large. We will turn our attention to the issue of temporal precision later on, and demonstrate that in some circumstances, such precision can actually be achieved today.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

A popular argument is that, even if online research were more prone to error than traditional laboratory-based research, simply by increasing the number of participants in one's study, the researcher can offset such issues. In Simcox and Fiez's (2014, Experiment 2), 100 MTurkers took part in a successful replication of a classic Erickson Flanker task (Nieuwenhuis, Stins, Posthuma, Polderman, Boomsma, & de Geus, 2006). In order to assess how many participants would need to be tested in order to achieve an effect of a similar power to that observed in laboratory settings, the authors systematically varied the number of participants contributing to identical analyses (10,000 random re-samples per analysis). Reassuringly, a comparable number of online participants and in-lab participants were required for the replicated effect to be observed. The authors also noted that by increasing the sample size from 12 to 21, the chance of a Type 2 error (wrongly concluding that no effect is present) dropped from 18% to 1% – in their study, this could be achieved by recruiting additional participants for a total of $6.30. Or expressed less sensationally, by collecting 75% more participants for an additional cost of 75% of the original total participant fees.

So, although tasks requiring that visual stimuli be presented for especially short durations are seemingly less suited for online research at present, in a few years, as proposed by Crump et al. (2012), this position will likely change, thus making such research valuable to the research community. Indeed, offsetting the reduced power of such experiments online with more participants may help us bridge the gap between now and then.

Potential concerns with online research

It is important to acknowledge that there are a number of potential concerns with online research. Below we try to answer some of the most common concerns that we have encountered in our own research. Many of them, it has to be said, were raised by the inquisitive, sceptical, and downright incredulous reviewers of our own papers.

1) Timing

Getting a stimulus to appear on screen at the exact millisecond-specific time, and for the right duration, is indeed very hard to achieve, even for lab-based software (see Garaizar et al., 2014);

with online studies, the issue mostly boils down to the fact that the browser does not know when the monitor refreshes (although see Gibhub, 2014) and so cannot synchronize stimulus presentation with a given screen refresh. A consequence is that if a visual image is set to appear/disappear between refreshes, it will only do so on the next refresh. Indeed, if a stimulus is to appear and disappear within a period of time smaller than a refresh interval, it may not appear at all, or could appear for (often much) longer than desired, and not at the right time. This is probably why Crump et al. (2013) were unable to replicate the Flanker task for short duration stimuli.

We tested this appearance issue in a simulation where we varied the duration of visual stimulus, starting at a random time during the refresh cycle (10,000 virtual presentations per stimulus duration). Figure 5 shows the likelihood of short duration stimuli being shown at all, or being shown for the wrong duration, or starting / stopping at the wrong time (https://github.com/andytwoods/refreshSimulation; available to run / tweak online here http://jsfiddle.net/andytwoods/0f56hmaf/). As most people use LCD monitors which typically either refresh 60 (78.1% of monitors) or 59 times a second (21.9% of monitors), we know that the majority of screens refresh every 16.67 ms or 16.95 ms (Witzel et al., 2013). As shown in Figure 5, thus, by having none of your stimuli shown for less than 16.95 ms, the stimulus should appear on screen for about the correct duration and (>90% of the time). Specifying your stimulus durations as multiples of 16.95ms will also lead to more accurately presented longer-duration stimuli. Indeed, one may wonder why the majority of research software packages do not allow experimenters to specify their stimuli in terms of refresh intervals (as only done by DMDX, to the best of our knowledge).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

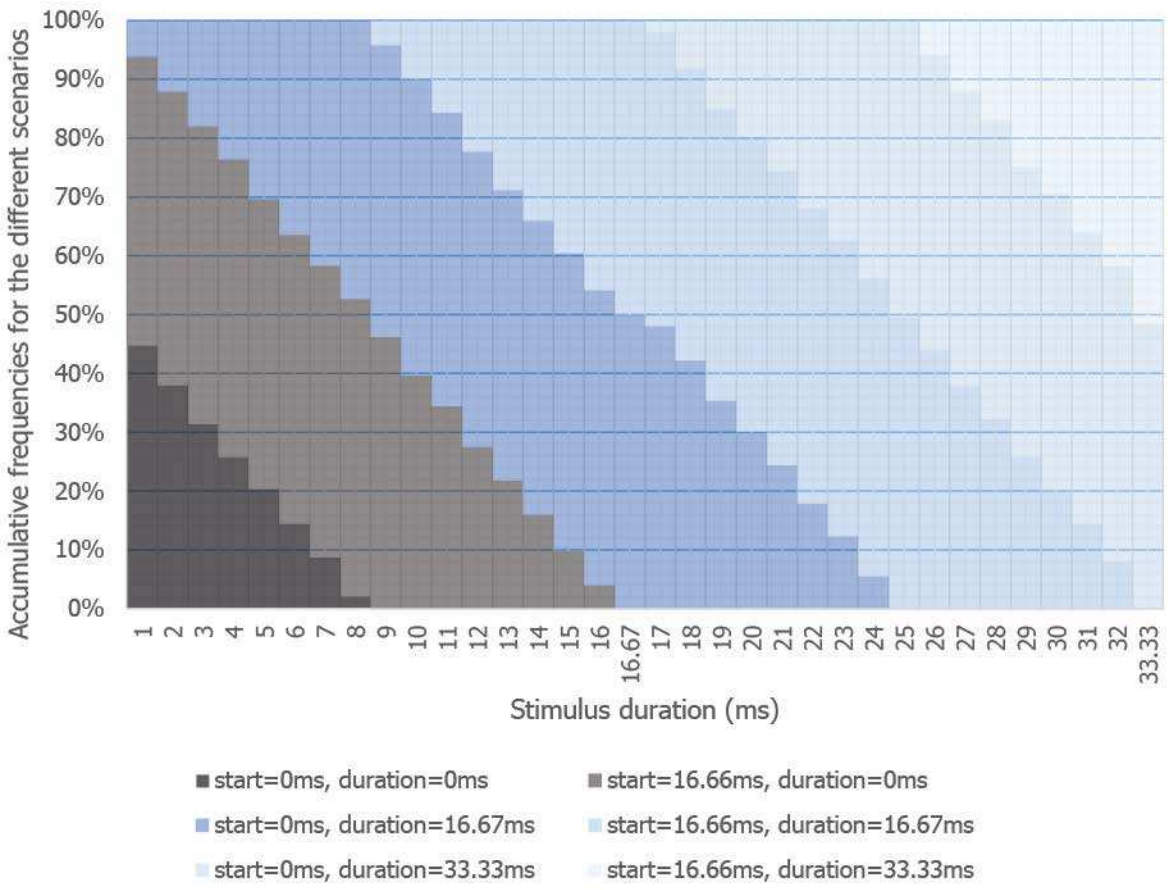


Figure 5: Likelihood of stimuli of different presentation durations appearing on screen, or doing so with the wrong start time, end time, and/or duration (screen refresh of 16.67).

A consequence of not knowing when the screen refreshes, and thus not knowing when a stimulus will appear on the participant's screen, is that, it is hard to know from when exactly RTs should be measured. Another issue is that RTs unfortunately vary quite considerably depending on the brand of keyboard used in a study, which is most certainly a big issue with online research. Plant and Turner (2009) found, for example, that the mean delay between button press and reported time was between 18.30 ms to 33.73 ms for 4 different PC keyboards (standard deviations ranged between .83 ms and 3.68 ms). With Macintosh computers, Neath et al (2011) found that keyboards added a delay between 19.69 ms and 39.56 ms (standard deviations were between 2.67 ms and 2.72 ms). In a laboratory setting, this is not such an issue where typically participants are tested using the same experimental apparatus and thus same keyboard (e.g. 5 ms response delays with a

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

random variation of -2.7 ms to +2.7 ms: 22.11, 18.07, 17.59, 20.9, 22.3; mean=20.19, stdev=2.23). However, when lots of different keyboards are used, a whole variety of different latencies act to introduce troublesome variation into your data (e.g., 5 random response delays of 20ms to 40ms, with the same random variation added: 19.45, 37.8, 37.57, 22.7, 31.23; mean=29.75, stdev=8.43).

All is not lost however. Systematically exploring how well RTs could actually be measured online, Reimers and Stewart (2014) recently tested RTs on 5 different computers, 3 web-browsers, and 2 types of web-technology (Adobe Flash, HTML 5) using a Black Box Toolkit (http://www.blackboxtoolkit.com/; a piece of hardware that can be used to accurately measure response times and generate button presses). The authors used the device to detect screen flashes generated by the testing software by means of a photodiode, and to generate button presses at precise times by completing the circuit of a button of a hacked keyboard. Although there was some variability across machines, and although RTs were generally overestimated by 30ms (comparable to the delays reported above, although standard deviations were typically 10ms), the authors concluded that the noise introduced by such technical issues would only minimally reduce the power of online studies (the authors also suggested that the within-participant design is particularly suited to online research given this variability). Once again bolstering the support for conducting valid RT research online, Schubert, Murteira, Collins, and Lopes (2013) found comparable RT measurement variability when comparing their own online Flash-based research software ScriptingRT (mean 92.80 ms, standard deviation 4.21) with laboratory-based software using the photodiode technique mentioned above (millisecond means for DMDX, E-prime, Inquisit and Superlab, were respectively 68.24, 70.96, 70.05, 98.18; standard deviations 3.18, 3.30, 3.20 and 4.17; the authors must be commended for their 'citizen science' low-cost Arduino-based timing solution, which Thomas Schubert fleshes out on his blog https://reactiontimes.wordpress.com/). These findings were broadly mirrored by de Leeuw and Motz (in press) who compared accuracy for recording RTs in a visual search task that run either via Matlab's Psychophysics ToolBox or in a webbrowser via JavaScript. Whilst RTs for the latter were about 25ms longer than the former, reassuringly there were no real differences in data variability over platforms. Simcox and Fiez (2014) found that browser timing accuracy was only compromised when unusually large amounts of system resources were in use. The authors measured timing by externally measuring screen flashes with a photodiode that were placed 1000ms apart in time, and concluded that browser based timing is in most scenarios as accurate as lab-based software. In summary, then, it would seem

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

then that the variability introduced by participants using different computers / monitors / web-browsers, is negligible in comparison to the variability introduced by the participants themselves (Brand & Bradley, 2012; Reimers & Stewart, 2014), although, one has to wonder whether using the participant's smartphone-camera to detect screen refresh / stimulus presentation parameters (from say a few pixels devoted to this purpose in the top of the participants' screen) and appropriately feeding this knowledge back to the testing software may help with accuracy. Some modern day cameras certainly are able to capture video at high enough frame rates (e.g., 120Hz, Haston, 2014).

One way to get around the browser-related limitations of not knowing when the screen refreshes is to ask participants to download experimental software to run outside of the browser (unfortunately MTurk does not permit the downloading of external software). One problem here though is that the experimenter cannot really ask their participants to undertake the fine calibrations normally required to set up experimental lab-based software (e.g., timeDX, http://psy1.psych.arizona.edu/~jforster/dmdx/help/timedxhtimedxhelp.htm), so more superficial means of calibration must be automatically undertaken. Seeing if their own compromise solution were sufficient for the downloadable webDMDX, Witzel et al (2013) tested whether the results of classical time critical studies differed across lab based DMDX (Forster & Forster, 2003) and webDMDX and found good consistency across software platforms. Curiously, however, the results of a stimulus that had been set up to appear for 50ms in the lab-based software matched those for a 67ms duration stimulus in the web based software. The authors found that the lab-based stimulus was just over 3 refreshes in length (16.67ms * 3 = 50.01ms) and so was actually shown for an additional interval, for 66.68 ms, as was 67ms stimulus (n.b., DMDX rounds to the nearest refresh interval), which was easily corrected. Thus, if your participants trust your software, and your participant panel permits it, it may be advisable to use software like webDMDX for those experiments requiring fine temporal control of stimuli.

2) Variability in hardware

Perhaps the most obvious issue with online research, as alluded to above, is the sheer variety of hardware and software used by participants. Although it can be argued that online research is more ecologically valid because of this varied hardware compared to lab-based studies that all run on

the same device, hardware variability, nevertheless, poses some unique challenges for the experimenter; especially when considering that the web browser can only determine a few device parameters such as screen resolution and operating system (but see Lin et al., 2012). For example, the resolutions of monitors differ massively over participants; we found in 2013 an average resolution of 1422 x 867 pixels over 100 participants' monitors, with large respective standard deviations of 243 and 136 pixels (Woods et al 2013). As there is no way to assess the physical size of monitors via a web browser, standardising the size of one's stimuli over participants is extremely difficult. As a work around, Bechlivanidis & Lagnado (submitted) had their participants hold up a CD, a credit card, or a 1 US dollar bill to their screen, and then adjust a shape on the screen to match the size of the object (see also Yung, Cardoso-Leite, Dale, Bavelier & Green, 2015). The authors also asked their participants whether they were an arm's distance away from their monitor to get an idea of their distance from the monitor (see also Krantz, who suggests a real world rule of thumb—by holding your thumb an arm's distance from the monitor, perpendicular elements directly beneath the thumb are approximately 1 or 2 visual degrees, 2001). Another approach is to find your participant's blind spot—by asking the participant to focus on a shape whilst another shape horizontally moves relative to it, and indicate when the moving shape disappears from view—and then resize experimental images appropriately. Sadly though, we cannot anchor our online participants' heads in place to prevent fidgeting, although, as suggested by a helpful audience member in a recent talk by the first author, monitoring the participant via a webcam and resizing stimuli appropriately may be one future strategy to help cope with this.

Another issue is that the many dials and buttons that adorn the modern-day computer often make it impossible to quantify properties such as volume, brightness and colour. There are ways to counter this issue. For example, the participant could be asked to adjust the volume until an audio stimulus is just audible, or indicate when elements in a visual image are most contrasting (To, Woods, Goldstein, & Peli, 2013). Yung et al. (2015) did the latter by presenting on screen a band of grey bars and asking their participants to adjust the brightness of the bar (in their software) until all of the bars were visible. We have also started to include an audio password (or AudibleCaptcha) in our experiments that can only be answered when the volume is set appropriately (Knöferle, Woods, Käppler, & Spence, 2015). The daring may even consider using staircases to establish a variety of thresholds for audio stimuli. Although it is impossible really to control for background

noise levels, by using webcam microphones, it may be possible to quantify background noise levels and re-run noisy trials or add noise levels as a covariate in subsequent data analyses.

Perhaps one of the hardest challenges is colour. Although one approach to combat this is to use colour words instead of the colours themselves (e.g., Piqueras-Fiszman, Velasco, & Spence, 2012; Velasco et al., 2014), this solution would only be suitable for a small number of studies (those that only use colour categories). An initially promising solution would be to run studies on identical devices such as the same generation iPad device.

Unfortunately, however, even purportedly identical screens viewed in identical environmental conditions vary in terms of colour and brightness (Cambridge Research Systems, personal communication, February 17-18, 2015). Others have suggested using psychophysics to identify issues with the current monitor and then dynamically adjusting the presented images appropriately. Hats off to To, Woods, Goldstein, and Peli (2013), who presented participants with a variety of coloured and hashed line patches in different shades and asked their participants to adjust their properties so, for example, two such patches would match in terms of their brightness. The authors found that participants performed to a similar ability to a photometer (.5% sensitivity difference). A potential future solution could be to ask participants to use the camera on their mobile devices to video both their computer screen being used for a study, and a common, colourful, household object, (e.g., a bottle of CocaCola™; cf. the size solution of Bechlivanidis & Lagnado, submitted). Software on the mobile device could then potentially liaise with the research software to calibrate screen colour to the reference object. Thus, although presenting the same colour to participants irrespective of device is probably not achievable with current technologies, there are some nice 'work-arounds' that may help somewhat offset any variability in one's data due to inconsistent colour (as can also be done by collecting data from many more participants).

Auditory stimuli and the variability in the hardware they are generated by pose similar problems. For example, Plant and Turner (2009) found that computer speaker systems introduced a delay before audio presentation, that ranged anywhere from 3.31 ms all the way up to 37 ms (respective standard deviations of 0.02 and 1.31ms), with the duration of the sound varying by 1-2 ms across systems. Previous work has also found that auditory information is sometimes treated differently depending on whether participants wear headphones or hear sounds through speakers (Di Luca, Machulla, & Ernst, 2009; though see also Spence, 2007). One option is that the researcher may

wish to include questions pertaining to the participants' audio hardware. Needless to say, tasks that require the fine temporal control of auditory *and* visual stimuli, such as needed in the visual flash illusion (Shams, Kamitani & Shamojo, 2002) and McGurk effect (McGurk & MacDonald, 1976), perhaps would best be undertaken in the laboratory. Although do consider that if such an illusion / effect were reliable enough, a staircase procedure could be used to identify the delay required for auditory and visual elements to be temporally synchronous, which could then be used to calibrate subsequent auditory-visual testing on that computer.

Briefly summarising, the variability in hardware used by participants in online studies pose unique problems that with the current level of technology are hard to directly address. Several workarounds exist for each issue however, and in the end of the day, collecting more data (as always) is a healthy way to offset some of these issues.

3) Unique participants?

How can you be sure that the same subject is not taking part in the experiment multiple times? Participants recruited through Mechanical Turk or Prolific Academic must have an online profile that theoretically prevents them from taking part in the same study more than once. Although potentially an individual can have multiple accounts, it is harder to do these days with increasingly tight security-conscious sign-up criteria. Indeed, if the participant wishes to get paid, they must provide unique bank account and Social Security Number details (for MTurk), each of which requires a plethora of further identification checks.

The research software itself can also provide some checks for uniqueness, for example, by storing a unique ID in each participant's web browser cache or Flash cache, thus making it easier to identify repeat participants. Although it is sometimes possible to identify potential repeaters by means of their (identical) IP address, Berinsky et al. (2012) noted that the 7 out of 551 participants in their Mechanical Turk study who had identical IP addresses, could well have done the study on the same computer, or same shared internet connection; indeed, this day and age, the participants could even have done the study through the same Virtual Private Network and be in quite different geographic locations from those determined via IP address (or indeed through self-report).

A related concern arises when an experimenter conducts multiple different online experiments using the same platform. Preventing previous participants from participating in future experiments

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

is difficult using Mechanical Turk (but see http://mechanicalturk.typepad.com/blog/2014/07/new-qualification-comparators-add-greater-flexibility-to-qualifications-.html), so typically the experimenter ends up having to manually, tediously, exclude repeats after participation. Bear in mind here that relying on participants to not undertake a task if they have done a similar one in the past is unfair given the sheer number of past studies each likely will have undertaken. Perhaps a much more impactful issue is when participants become overly familiar with popular experimental methods / questionnaires that are used by different researchers. Highlighting this issue, Chandler, Mueller, and Paolacci (2014) found that out of 16,409 participants in over 132 studies, there were only 7,498 unique workers with the most active 1% completed 11% of hits (see also Stewart, Ungemach, Harris, Bartels, & Newell, submitted; Berinsky, Huber, & Lenz, 2012).

Although these issues most certainly are a concern for researchers focusing on the study of perception, it is likely that repeat participants would be far more problematic for more cognitive-focused areas of psychology. It may simply be the case for the psychologist interested in perception to ask participants how often they have undertaken similar tests in the past and use this data a covariate in their subsequent statistical analysis.

4) Random responding

A common concern with online research is that those taking part in a paid study do not do so with the same care and diligence as those in a lab-based study. In fact, however, the research that has been conducted in this area shows that lab-based studies are not necessarily the gold standard we often presume. In one such study by Oppenheimer, Meyvis, and Davidenko (2009), immediately after completing two classic judgement and decision-making studies, participants were presented with a catch-trial where they were explicitly told to click a small circle at the bottom of the screen, as opposed to one of 9 response buttons making up a line scale that was shown in the centre of the screen. Not only did a disquieting 46% of the participants fail the task, but only by excluding these individuals were both the classic tasks were successfully replicated. Thus one cannot assume that participants in lab are attending as carefully as one might hope. As an example from our own experiences, one author received a text message from such a participant who was 'mid study', saying they would be late for his later experiment! Reassuringly though, perhaps again highlighting that perceptual psychology is more robust to such issues than other areas of our

discipline, we used the Oppenheimer et al. (2009) attention check for an online face emotion task and found that only 1% of MTurkers failed the task (Dalili, 2015; for an in depth discussion see Hauser & Schwarz, 2015). We return to this shortly.

Perhaps one key issue scientists have with online research is the absence of the experimenter who can be quizzed to clear up uncertainties, or make sure the participant follows instructions. Painting a bleak picture, Chandler, Mueller, and Paolacci (2014) asked 300 MTurkers what they were doing whilst completing a study, and found that 18% of responders were watching TV, 14% listening to music and 6% were communicating with others online (the interested reader is directed to a video where a MTurker discusses this issue in reference to looking after her baby whilst participating in research, Marder, 2015). Several strategies, besides the catch trial mentioned earlier (Oppenheimer et al., 2009), have been developed to deal with the consequences of such distraction and potential disinterest (Downs, Holbrook, Sheng, & Cranor, 2010, Crump, McDonnel, & Gureckis, 2013; Germine et al., 2012), perhaps the most simple being to quiz the participants as to the nature of the task before proceeding to the study. Crump et al. found that this approach led to a closer replication of a classic rule-based classification learning study (Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994), compared to an earlier study where there was no such intervention.

Indicating that this is not such an issue, when Hauser and Schwarz (2015) directly set about comparing the performance of lab-based and internet recruited participants on the Oppenheimer, Meyvis, and Davidenko catch trial (2009), and found the latter group much less likely to fail at the task. Hauser and Schwarz first found that lab-based participants failed an astounding 61% of the time – even more than the original study – whilst online participants recruited on MTurk only failed 5% of the time. This broad pattern of results was replicated for a novel version of the catch trial in Experiment 2. To test whether MTurkers were just very vigilant for such catch trials (as they may have had similar ones in the past; see the 'overfamiliarity' discussion above) or whether, indeed, MTurkers paid more attention, in a third study both groups were tested on a soda-pricing task (adapted from Oppenheimer et al., 2009) that has been shown to be sensitive to levels of attention. Supporting the latter account, online participants scored much better in a test sensitive to levels of attention compared to their lab-based counterparts.

In summary, whilst the lack of experimenter supervision for participants recruited online most certainly is worrying, it is important to bear in mind that lab-based research does not necessarily

ensure attentive participants either. The very fact that a lot of past research has been replicated would indicate that the different issues with online and in lab research may be similarly impactful on our results.

5) Ethics

While it is relatively clear where the responsibility for ethics lies in a study conducted within a given department, online research is often an unknown area for both the researcher and the local ethics committee. The British Psychology Society recently weighed in on this topic (British Psychological Society, 2006, 2013; see also the American Psychological Association's overview on this, Kraut et al., 2002; Ross 2014), highlighting the key issue that it is the physical absence of the experimenter during the study, preventing, for example, the experimenter from stopping the study early if the participant starts showing any sign of distress. Presumably though, the online participant would feel less obligation to actually finish a study they were uncomfortable with, compared to if it were lab-based study.

There are several other issues as well. Besides issues of fair payment (highlighted earlier), online anonymity is also a key issue. For example, with a bit of deduction, it is often possible to extrapolate the identity of an individual from their pattern of responses (El Emam & Arbuckle, 2013; King, 2011; see also some such high-profile examples from the Netflix challenge, Narayanan & Shmatikov, 2008, and social networks, Narayanan & Shmatikov, 2009). Highlighting this, MTurker Worker IDs are made available to research software when people take part in an MTurk study. We asked 100 MTurkers to enter their Worker ID into Google and tell us "Did your Google search of your Worker ID find links to 'raw data' (answers you gave) from previous studies you have taken part in?" and "Did your Google search results contain information that revealed your name or IP address?" A staggering 47 Mturkers reported finding such past data in their search results, whilst 5 MTurkers reported finding their name / IP-address. Further exploration is warranted to check just what information past researchers are making publicly available online *alongside* potentially identity revealing MTurker Worker IDs, as this clearly goes against ethical guidelines. Several MTurkers also emailed telling us that their past Amazon store reviews of books appeared in their search results—with a bit of investigation it transpired that Amazon Ids and MTurker Worker IDs are one and the same! (see Lease et al. 2013, who discuss

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

this and other issues in detail). In light of the above, we would urge researchers to carefully select the information that is stored alongside collected data, and to remove Worker IDs before sharing data online. If Worker ID data must be stored online (e.g., to be shared by the members of a specific lab), that data should be adequately encrypted, and not left as 'plain text' as was seen often in the just mentioned survey.

The recent drive to opensource datasets coupled with the ethical requirement of allowing participants to withdraw their data after data collection (up to a certain point in time, anyway, such as the conclusion of the analysis) unfortunately muddies the waters regarding anonymity. One strategy for this we have adopted is to ask participants to provide a password that can be used if they wish their data removed by a later date; although given the large number of passwords one must remember these days, it is not clear if this will prove effective.

Conducting research online

In 2004, the lead author devoted several months creating a one-off website to run a study on crossmodal face perception using videos (Stapleton, Woods, Casey, & Newell, in prep). Things have progressed far since then! There are now a variety of software platforms aimed at collecting questionnaire based data online, with a smaller number of packages now aimed specifically at conducting online behavioural research. Some of the latter, alongside their strengths and weakness as reported by their main developers, have been listed in Table 2.

One way the packages differ is in terms of whether they are opensource or commercial in nature. Whilst licensed software is often thought to be easier to use and have a better support network than opensource software, a downside is that if a bug is found, you are at the mercy of a limited number of developers to fix the problem, instead of being able to immediately explore code yourself or ask for help from the lively opensource community. Conventional wisdom would also suggest that commercial software would be easy and more versatile than opensource 'freely contributed to' software but the reality is that this is often not the case (e.g., The Gimp, http://www.gimp.org/, is an opensource feature rich alternative to Adobe Photoshop). Moreover, commercial software is typically configured with the needs of large-scale corporate users in mind, whereas the opensource community may be more receptive to suggestions that benefit academic users.

If you have no programming experience, deciding on a testing package that does not require the coding may be a quicker option for getting a study on the web, if your task only requires typical experimental features (such as buttons, scales, the ability to show pictures, etc). Some packages such as Qualtrics, for example, let you create studies by dragging and dropping components on screen. An intermediate option offering more flexibility is to use software that relies of scripts to run an experiment (e.g. ScriptingRT, WebDMDX, Xperiment).

Whether or not the research software is Adobe Flash based or not is another consideration. Although Flash has purported to have been 'dying' for a number of years now, it is in fact present on most modern computers; for example, it is installed automatically within the Chrome web browser which has 61.9% market share (Browser Statistics, 2015), and can be automatically installed in other popular browsers such as Firefox (23.4% market share). Flash is also making a comeback by re-inventing itself as a cross-platform tool capable of making apps for both Android and IOS; indeed, it won the 2015 Consumer Electronics Show best mobile application development platform. As we found out recently though with the lead author's package called Xperiment, reliance on the proprietary closed-source Adobe Flash environment meant that when bugs in closed source code did arise, we were entirely dependent upon Adobe engineers to fix issues. At the start of 2014, Adobe updated their software and thus 'broke' an experiment we were running, leading to a loss of 31.3% of participant data (see the bug here https://productforums.google.com/forum/m/#!topic/chrome/ifL98iTMhPs). This may well have been due to 'teething issues' due to Google Chrome releasing its own version of Flash around that time called 'pepperFlash'. In light of this though, the lead author is considering porting over the Xperiment package to the opensource cross-platform Haxe toolkit, which allows software to natively run in the browser (without Flash), as well on several platforms such as IOS and Android.

Table 2: Popular online research platforms, their main features, strengths and weaknesses, as reported by their developers (survey conducted through Google Forms, on 13-3-2015, which is not listed in the below table on account of being mostly questionnaire-focused and thus 'neutral territory' for responders).

		JsPsych	Inquisit	LimeSurvey	ScriptingRT	Qualtrics	Unipark	WebDMDX	Xperiment
Opensource		yes	no	yes	yes	no	no	no	yes (in beta)
Yearly Fee for one researcher (USD)			1495			?[1]	138.12		
Publish directly to crowd sourcing sites[2]		no[3]	with addons [4]	no	no	MTurk	no	no[5]	MTurk, ProlificAc
Questionnaire vs Research focus (Q vs R)									
Coding required for	Software setup	yes	no	no	yes	no	no	no	yes
	Creating a study	yes	script based	no	script based	no	no	script based	script based
Possible trial orderings	Random	yes	yes	yes	yes	yes	yes	yes	yes
	Counterbalanced	yes	yes	no	no	yes	yes	yes	yes
	Blocked	yes	yes	yes	no	yes	yes	yes	yes

1. Many academic institutions have licenses with Qualtrics already. Individual academic pricing was not disclosed to us and could not be found via search engines. Note also that some features (e.g., more advanced randomization) may require a more expensive package.

2. Although all platforms let the researcher provide a URL where the participant can undertake a study, some crowd-sourcing sites need to communicate directly with the testing software in order to know, for example, if the participant should be paid.

3. "None directly; but it can be used to publish on any platform that allows for custom JavaScript and HTML content"

4. See http://www.millisecond.com/support/docs/v4/html/howto/interopsurveys.htm

5. "It uses an HTML POST command so pretty much anything, depends how skilled you are. We provide a site running a general purpose script to gather data and email it to experimenters should people not be in a position to setup a site to gather the data."

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

On the future of online perception research

Smart devices will come into their own in the coming years (e.g. Brown et al. 2014; Dufau et al., 2011; Millar, 2013). Making them seem particularly well suited for online perceptual psychology are their plethora of sensors (light levels, global position system, proximity) and actuators (vibration, flashing light), as well as their range of peripherals such as smart watches (letting you measure for example, heart rate, cadence and even sleep quality), other wearables such as motion tracking (e.g., http://www.xensr.com/) and even intelligent cushions that measure seated posture quality (http://darma.co/). Of course, these new technologies may well be affected by the same issues we highlighted before. For example, in our own tentative steps down the road of smart phone research, we have found large differences in terms of vibration levels that different smartphones can produce, which is, presumably due to the devices using a variety of vibration motors.

Not only are smart devices rich in sensors and actuators, they can add a new dimension to research by being able to contact participants at any point during the day using Push notifications, or to link with software services to provide even richer sources of information for your investigation. If a study were concerned say, with vibration detection in noisy environments, the device could be made to vibrate only when the background noise level was at a desired level. Alternatively, GPS could be used if your paradigm required participants only be tested in certain geographical locations. We predict such 'mashups' of technologies (e.g. Paredes-Valverde, Alor-Hernández, Rodríguez-González, Valencia-García, & Jiménez-Domingo, 2015) will really be a game changer for perceptual psychology (for some predictions on future ways we will interact with our devices, see Nijholt, 2014).

Unfortunately, the current state of affairs mirrors that for online research in 2005 where one-off experiment apps must be made, typically for either IOS or Android devices. An early example of such an app, reported in 2010, by Killingsworth and Gilbert, had participants' iPhones, randomly throughout the day ask their users a series of questions to do with their current levels of mind wandering and happiness. The authors curiously found that mind-wandering was negatively associated with happiness (although more recent findings suggests that this affect depends upon the mind wandering being negative itself in terms of emotion, Poerio, Totterdell, & Miles, 2013).

Conducting research on gaming devices such as the Xbox One and Playstation 4 is surprisingly not that far away. Transpilers, or source-to-source compilers (Source-to-source compiler, n.d.)

allow developers to write code once and port that code to different platforms and programming languages. A transpiler that can *currently* port code to such gaming devices is the commercially available Unity 3D package (http://unity3d.com/; see also Adobe Air and the opensource Haxe platform; as of yet though, neither package can port to gaming devices; https://www.adobe.com/uk/products/air.html, http://haxe.org/).

'Big data' is most certainly part of the future for psychological research, where hundreds of thousands of participants contribute data as opposite to tens or hundreds as seen in typical lab-based studies. To attract these numbers, researchers, for example, gamify their paradigm to make it fun for people to take part, offer feedback about how people have done after task completion (e.g., testMyBrain that we mentioned earlier providing score feedback). An alternative strategy is to piggyback existing sources of data, as Stafford and Dewar (2014) nicely demonstrate with their n=854,064 study exploring skill learning whilst people played an online game called Axon, that was developed for the Wellcome Trust (http://axon.wellcomeapps.com/).

In China, around 10:34 pm on Thursday 19, 2015 (The Chinese New Year), apparently 810 million smartphones were shaken at their TVs every minute. Over the course of a 4-hour long show (China Central Television Spring Festival gala), the shake count totalled an incredible 11 billion! What had happened was that weChat (a Chinese instant messaging service) in collaborating with a plethora of retail companies had offered the public the possibility of winning 'red envelopes' containing small amounts of money (for example, 2 Chinese Yuan, or about 0.32 USD), by just shaking their phones. One can only wonder what could be achieved if an intrepid researcher managed somehow to piggyback this, a la Stafford and Dewar (2014). Careful care would needed to ensure such a study was ethnically sound, however (c.f. the recent Facebook emotion manipulation study, as discussed by Ross, 2014).

Conclusions

Over the last 5 years or so, we have found the internet to be a very effective means of conducting online research to address a number of perceptual research questions. It offers a number of potential benefits over in-lab testing and is particularly useful for quickly collecting a large amount of data across a relatively wide range of participants. On the flip-side, there are a number of potential limitations that also need to be borne in mind. In terms of ethics, it seems that online

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

participants are not as anonymous as they should be, which needs addressing. It is also tricky to account for differences in hardware over machines and there still remain some issues related to the fine control of timing. Over the coming years though, it is likely that such issues will become less of a problem as technology develops, new solutions arise, and clearer ethical guidelines become available to researchers. In the meantime, a simple approach to deal with some of these issues though is, as always, to collect more data, which fortunately is easy, economical and fast in online research. Taken together, we believe that online testing will continue to become an ever-more popular approach to testing perception based research questions in the years to come.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

REFERENCES

Admin. (2013, January 17). The reasons why Amazon Mechanical Turk no longer accepts international Turkers. Retrieved March 7, 2015, from http://turkrequesters.blogspot.co.uk/2013/01/the-reasons-why-amazon-mechanical-turk.html

Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., & Menczer, F. (2012). Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)*, **6(2)**, 9.

Bechlivanidis, C. & Lagnado, D.A. (submitted). Time Reordered: Causal perception Guides the Interpretation of Temporal Order. *Cognition*.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, **43**, 800-813.

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**, 351-368.

Berlin, B. (1991). *Basic color terms: Their universality and evolution*. Oxford, UK: University of California Press.

Brand, A., & Bradley, M. T. (2012). Assessing the effects of technical variance on the statistical outcomes of web experiments measuring response times. *Social Science Computer Review*, **30**, 350-357.

British Psychological Society (2006). *Report of the Working Party on Conducting Research on the Internet*: *Guidelines for ethical practice in psychological research online.* REP62/06.2007.

British Psychological Society (2013). *Ethics Guidelines for Internet-mediated Research*. INF206/1.2013. Available from:www.bps.org.uk/publications/policy-andguidelines/research-guidelines-policydocuments/research-guidelines-poli

Brown, H. R., Zeidman, P., Smittenaar, P., Adams, R. A., McNab, F., et al. (2014). Crowdsourcing for cognitive science – The utility of smartphones. *PLoS ONE*, **9**, e100662.

Browser Statistics. (2015, January 1). Retrieved February 7, 2015, from http://www.w3schools.com/browsers/browsers_stats.asp

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

1 Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of
2 inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, **6**, 3-5.

3 Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk
4 workers: Consequences and solutions for behavioural researchers. *Behavioural Research Methods,*
5 **46**, 112-130.

6 Chandler, J., Paolacci, G., & Mueller, P. (2013). Risks and rewards of crowdsourcing marketplaces.
7 In P. Michelucci (Ed.), *Handbook of human computation* (pp. 377-392). New York, NY: Springer.

8 Cross, I. (2001). Music, cognition, culture, and evolution. *Annals of the New York Academy of*
9 *Sciences*, **930**, 28-42.

10 Crump, J. C., McDonnel, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as
11 a tool for experimental behavioural research. *PLoS ONE*, **8**, e57410.

12 Dalili, M. (2015, March 20). TARG Blog » Research doesn't just happen in the lab anymore:
13 Mechanical Turk, Prolific Academic, and online testing. Retrieved March 20, 2015, from
14 http://targ.blogs.ilrt.org/2015/03/20/research-doesnt-just-happen-in-the-lab-anymore-
15 mechanical-turk-prolific-academic-and-online-testing/

16 De Leeuw, J.R. & Motz, B.A. (in press). Psychophysics in a web browser? Comparing response times
17 collect with JavaScript and Psychophyiscs Toolbox in a visual search task. *Behavior Research*
18 *Methods.*

19 Di Luca, M., Machulla, T. K., & Ernst, M. O. (2009). Recalibration of multisensory simultaneity:
20 Cross-modal transfer coincides with a change in perceptual latency. *Journal of Vision*, **9**, 7.

21 Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). Are your participants gaming the
22 system? Screening Mechanical Turk workers. In *Proceedings of the SIGCHI Conference on*
23 *Human Factors in Computing Systems* (pp. 2399-2402). New York, NY. ACM.

24 Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: it's all in the mind,
25 but whose mind?. *PloS ONE*, **7**, e29081.

26 Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically
27 intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic
28 participants. *Neuropsychologia*, **44**, 576-585.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

1 Dufau, S., Dun˜abeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F.-X., Balota, D.
2 A., Brysbaert, M., Carreiras, M., Ferrand, L., Ktori, M., Perea, M., Rastle, K., Sasburg, O., Yap,
3 M. J., Ziegler, J. C., & Grainger, J. (2011). Smart phone, smart science: How the use of
4 smartphones can revolutionize research in cognitive science. *PLoS ONE*, **6**, e24974.

5 Eimer, M., & Schlaghecken, F. (1998). Effects of masked stimuli on motor activation: Behavioral and
6 electrophysiological evidence. *Journal of Experimental Psychology: Human Perception and*
7 *Performance*, **24**, 1737-1747

8 El Emam, K., & Arbuckle, L. (2013). *Anonymizing health data: Case studies and methods to get you*
9 *started*. Sebastopol, CA: O'Reilly Media.

10 Eriksen, C. W. (1995). The flankers task and response competition: A useful tool for investigating a
11 variety of cognitive problems. *Visual Cognition*, **2**, 101-118.

12 Eriksson, K., & Simpson, B. (2010). Emotional reactions to losing explain gender differences in
13 entering a risky lottery. *Judgment and Decision Making*, **5**, 159-163.

14 Fitzgerald, P. J. (2013). Gray colored glasses: Is major depression partially a sensory perceptual
15 disorder? *Journal of Affective Disorders*, **151**, 418-422.

16 Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond
17 accuracy. *Behavior Research Methods, Instruments, & Computers*, **35**, 116-124.

18 Garaizar, P., Vadillo, M. A., López-de-Ipiña, D., & Matute, H. (2014). Measuring software timing
19 errors in the presentation of visual stimuli in cognitive neuroscience experiments. *PLoS ONE*, **9**,
20 e85108.

21 Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012).
22 Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual
23 experiments. *Psychonomic Bulletin & Review*, **19**, 847-857.

24 GitHub issue (2014, October 6). Use requestAnimationFrame when possible to pseudo-sync with
25 display refresh #75. Retrieved from https://github.com/jodeleeuw/jsPsych/issues/75/

26 Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths
27 and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, **26**, 213-
28 224.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

1 Haston, S. (2014, April 22). Ultimate slow mo smartphone camera comparison, with fire hand!
2 Retrieved March 12, 2015, from http://www.mobilegeeks.com/slow-motion-tested-4-top-
3 smartphones-fire-hand/

4 Hauser, D. J., & Schwarz, N. (2015). Attentive Turkers: MTurk participants perform better on online
5 attention checks than do subject pool participants. *Behavior Research Methods*, 1-8. DOI:
6 10.3758/s13428-015-0578-z

7 Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and*
8 *Brain Sciences*, **33**, 61-135.

9 Ho, C. J., Slivkins, A., Suri, S., & Vaughan, J. W. (2015, May X). Incentivizing High Quality
10 Crowdwork. In *Proceedings of the International World Wide Web*. Florence, Italy: IW3C2.

11 Holmes, N. P., & Spence, C. (2005). Multisensory integration: Space, time, and superadditivity.
12 *Current Biology*, **15**, R762-R764.

13 Iddemonofelru (2014, May10). Anyone wanna raise awareness of fair pay? [Web blog] Retrieved
14 from
15 https://www.reddit.com/r/mturk/comments/257cco/anyone_wanna_raise_awareness_of_fair_pay
16 /

17 Intons-Peterson, M. J. (1983). Imagery paradigms: How vulnerable are they to experimenters'
18 expectations? *Journal of Experimental Psychology: Human Perception and Performance*, **9**, 394-
19 412.

20 Ipeirotis, P. G. (2010). Analyzing the Amazon Mechanical Turk marketplace. *XRDS: Crossroads, The*
21 *ACM Magazine for Students*, **17(2)**, 16-21.

22 January 2015 Market Share (2015, January 30). Retrieved February 7, 2015, from
23 http://www.w3counter.com/globalstats.php?year=2015&month=1

24 Kay, P., & Regier, T. (2003). Resolving the question of color naming universals. *Proceedings of the*
25 *National Academy of Sciences*, **100**, 9085-9089.

26 Killingsworth, M. A., & Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science*, **330**,
27 932-932.

28 King, G. (2011). Ensuring the data-rich future of the social sciences. *Science*, **331**, 719-721.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

1 Klein, R. A., Ratliff, K. A., Vianello, M., Adams, Jr, R. B., Bahník, Š., Bernstein, M. J., et al. &
2 Woodzicka, J. A. (2014). Investigating variation in replicability. *Social Psychology*, **45**, 142-152.

3 Knöferle, K. M., Woods, A., Käppler, F., & Spence, C. (2015). That sounds sweet: Using crossmodal
4 correspondences to communicate gustatory attributes. *Psychology & Marketing*, **32**, 107-120.

5 Krantz, J. H. (2001). Stimulus delivery on the Web: What can be presented when calibration isn't
6 possible? In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of internet science* (pp. 113-130).
7 Lengerich, GE: Pabst Science.

8 Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J. & Couper, M. (2002). Psychological
9 Research Online: Opportunities and Challenges. *Psychological Research*, **412**, 268-7694.

10 Lakkaraju, K,
11 A Preliminary Study of Daily Sample Composition on Amazon Mechanical Turk (March 11,
12 2015). Available at
13 SSRN: http://ssrn.com/abstract=2560840 or http://dx.doi.org/10.2139/ssrn.2560840

14 Lease, M., Hullman, J., Bigham, J. P., Bernstein, M. S., Kim, J., Lasecki, W., Bakhshi, S., Mitra, T.,
15 & Miller, R. C. (2013). Mechanical turk is not anonymous. Available at SSRN 2228728.

16 de Leeuw, J. R. (2014). jsPsych: A JavaScript library for creating behavioral experiments in a Web
17 browser. *Behavior Research Methods*, **39**, 365-370.

18 Levitan, C. A., Ren, J., Woods, A. T., Boesveldt, S., Chan, J. S., McKenzie, K. J., Dodson, M., Levin,
19 J.A., Leong, C. & van den Bosch, J. J. (2014). Cross-cultural color-odor associations. *PLoS ONE*,
20 **9**, e101651.

21 Lin, K., Chu, D., Mickens, J., Zhuang, L., Zhao, F., & Qiu, J. (2012, June). Gibraltar: Exposing
22 hardware devices to web pages using AJAX. In *Proceedings of the 3rd USENIX conference on*
23 *Web Application Development* (pp. 7-7). Boston, MA: USENIX Association.

24 Marder, J. (2015, February 11). *The Internet's hidden science factory*. Retrieved February 12, 2015,
25 from http://www.pbs.org/newshour/updates/inside-amazons-hidden-science-factory/

26 Martire, K. A., & Watkins, I. (2015). Perception problems of the verbal scale: A reanalysis and
27 application of a membership function approach. Science & Justice. Available online 28 January
28 2015, ISSN 1355-0306, http://dx.doi.org/10.1016/j.scijus.2015.01.002.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

1 Marx, D. M., & Goff, P. A. (2005). Clearing the air: The effect of experimenter race on target's test
2 performance and subjective experience. *British Journal of Social Psychology*, **44**, 645-657.

3 Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk.
4 *Behavior Research Methods*, **44**, 1-23.

5 Mason, W., & Watts, D. J. (2010). Financial incentives and the performance of crowds. *ACM SigKDD*
6 *Explorations Newsletter*, **11**, 100-108.

7 Matthew, D., Krosnick J, A., & Arthur, L. (2010). *Methodology report and user's guide for the 2008-*
8 *2009 ANES Panel Study*. Palo Alto, CA and Ann Arbor, MI: Stanford University and the
9 University of Michigan.

10 McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing speech. *Nature*, **264**, 746-748.

11 Milland, K. (2014a, January 1).The myth of low cost, high quality on Amazon's Mechanical Turk.
12 Retrieved March 12, 2015, from http://turkernation.com/showthread.php?21352-The-Myth-of-
13 Low-Cost-High-Quality-on-Amazon-s-Mechanical-Turk

14 Milland, K. (2014b). Indian Turkers: Shining a spotlight on the invisible masses. Unpublished
15 manuscript.

16 Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, **7**,
17 221-237.

18 Mirams, L., Poliakoff, E., Brown, R. J., & Lloyd, D. M. (2013). Brief body-scan meditation practice
19 improves somatosensory perceptual decision making. *Consciousness and Cognition*, **22**, 348-359

20 Narayanan, A., & Shmatikov, V. (2008, May). Robust de-anonymization of large sparse datasets. In
21 *IEEE Symposium on Security and Privacy, 2008* (pp. 111-125). Oakland, CA: IEEE.

22 Narayanan, A., & Shmatikov, V. (2009, May). De-anonymizing social networks. In *Security and*
23 *Privacy, 2009 30th IEEE Symposium* (pp. 173-187). Oakland, CA: IEEE.

24 Neath, I., Earle, A., Hallett, D., & Surprenant, A. M. (2011). Response time accuracy in Apple
25 Macintosh computers. *Behavior Research Methods*, **43**, 353-362.

26 Nicholls, M. E., Loveless, K. M., Thomas, N. A., Loetscher, T., & Churches, O. (2014). Some
27 participants may be better than others: Sustained attention and motivation are higher early in
28 semester. *The Quarterly Journal of Experimental Psychology*, **68**, 1-19.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

1 Nieuwenhuis, S., Stins, J. F., Posthuma, D., Polderman, T. J., Boomsma, D. I., & de Geus, E. J. (2006).
2 Accounting for sequential trial effects in the flanker task: Conflict adaptation or associative
3 priming? *Memory & Cognition*, **34**, 1260-1272.

4 Nijholt, A. (2014). Breaking fresh ground in human–media interaction research. *Frontiers in ICT*, **1**,
5 4.

6 Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing
7 modes of rule-based classification learning: A replication and extension of Shepard, Hovland, and
8 Jenkins (1961). *Memory & Cognition*, **22**, 352-369.

9 Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks:
10 Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, **45**,
11 867-872.

12 Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular
13 reference to demand characteristics and their implications. *American Psychologist*, **17**, 776-783.

14 Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical
15 Turk. *Judgment and Decision Making*, **5**, 411-419.

16 Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant
17 pool. *Current Directions in Psychological Science*, **23**, 184-188.

18 Paredes-Valverde, M. A., Alor-Hernández, G., Rodríguez-González, A., Valencia-García, R., &
19 Jiménez-Domingo, E. (2015). A systematic review of tools, languages, and methodologies for
20 mashup development. *Software: Practice and Experience*, **45**, 365-397.

21 Pechey R, Attwood A, Munafò M, Scott-Samuel NE, Woods A, & Marteau TM (in prep). Wine glass
22 size and shape impact on judgements of volume.

23 Peromaa, T. (2015, March 7). Colors of "The Dress" (original & red-green versions): Percepts of 884
24 readers. Retrieved March 7, 2015, from
25 http://tperomaa.vapaavuoro.uusisuomi.fi/kulttuuri/189048-colors-of-"the-dress"-original-red-
26 green-versions-percepts-of-884-readers

27 Pew Research Center (2015, Jan). *Public and scientists' views on science and society.* Retrieved from
28 http://www.pewinternet.org/files/2015/01/PI_ScienceandSociety_Report_012915.pdf

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

1 Piqueras-Fiszman, B., Velasco, C., & Spence, C. (2012). Exploring implicit and explicit crossmodal
2 colour–flavour correspondences in product packaging. *Food Quality and Preference*, **25**, 148-155.

3 Plant, R. R., & Turner, G. (2009). Millisecond precision psychological research in a world of
4 commodity computers: New hardware, new problems? *Behavior Research Methods*, **41**, 598-614.

5 Poerio, G. L., Totterdell, P., & Miles, E. (2013). Mind-wandering and negative mood: Does one thing
6 really lead to another? *Consciousness and Cognition*, **22**, 1412-1421.

7 Reimers, S., & Stewart, N. (2014). Presentation and response timing accuracy in Adobe Flash and
8 HTML5/JavaScript Web experiments. *Behavior Research Methods*, doi:10.3758/s13428-014-
9 0471-1.

10 Ross, M. W. (2014). Do research ethics need updating for the digital age? *Monitor on Psychology*, **45**,
11 64.

12 Rumenik, D. K., Capasso, D. R., & Hendrick, C. (1977). Experimenter sex effects in behavioral
13 research. *Psychological Bulletin*, **84**, 852.

14 Schubert, T. W., Murteira, C., Collins, E. C., & Lopes, D. (2013). ScriptingRT: A software library for
15 collecting response latencies in online studies of cognition. *PLoS ONE*, **8**, e67769.

16 Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain*
17 *Research*, **14**, 147-152.

18 Shapiro, K. L., Raymond, J. E., & Arnell, K. M. (1997). The attentional blink. *Trends in Cognitive*
19 *Sciences*, **1**, 291-296.

20 Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications.
21 *Psychological Monographs: General and Applied*, **75(13)**, 1-42.

22 Shermer, D. Z., & Levitan, C. A. (2014). Red hot: The crossmodal effect of color intensity on
23 piquancy. *Multisensory Research,* **27***,* 207-223.

24 Silberman, S., Milland, K., LaPlante, R., Ross, J., & Irani, L. (2015, March 16). Stop citing Ross et
25 al. 2010. Retrieved March 22, 2015, from https://medium.com/@silberman/stop-citing-ross-et-al-
26 2010-who-are-the-crowdworkers-b3b9b1e8d300

27 Simcox, T., & Fiez, J. A. (2014). Collecting response times using Amazon Mechanical Turk and
28 Adobe Flash. *Behavior Research Methods*, **46**, 95-111.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

1 Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of*
2 *Experimental Psychology: General*, **143**, 534-547.

3 Source-to-source compiler. (n.d.). *Wikipedia*. Retrieved February 13, 2015, from
4 https://en.wikipedia.org/wiki/Source-to-source_compiler Spence, C. (2007). Audiovisual
5 multisensory
6 integration. *Acoustical Science & Technology*, **28**, 61–70.

7 Spellman, B. A. (2012). Introduction to the special section data, data, everywhere... especially in my
8 file drawer. *Perspectives on Psychological Science*, **7**, 58-59.

9 Stafford, T., & Dewar, M. (2014). Tracing the trajectory of skill learning with a very large sample of
10 online game players. *Psychological Science*, **25**, 511-518.

11 Stapleton, J., Woods, A. T., Casey, S., & Newell, F. N. (in prep). The contribution of facial and vocal
12 cues on the perceived attractiveness of others: A role for visual 'capture'.

13 Stewart, N., Ungemach, C., Harris, A. J. L., Bartels, D. M., & Newell, B. R. (submitted). The size of
14 the active Amazon Mechanical Turk population. *Judgment and Decision Making*.

15 To, L., Woods, R. L., Goldstein, R. B., & Peli, E. (2013). Psychophysical contrast calibration. *Vision*
16 *Research*, **90**, 15-24.

17 Velasco, C., Wan, X., Salgado-Montejo, A., Woods, A., Onate, G., Mi, B., & Spence, C. (2014). The
18 context of colour-flavour associations in crisps packaging: A cross-cultural study comparing
19 Chinese, Colombian, and British consumers. *Food Quality and Prefere*nce, **38**, 49-57.

20 Velasco, C., Woods, A. T., Deroy, O., & Spence, C. (2015a). Hedonic mediation of the crossmodal
21 correspondence between taste and shape. *Food Quality and Preference*, **41**, 151-158.

22 Velasco, C., Wan, X., Knoeferle, K., Zhou, X., Salgado-Montejo, A., & Spence, C. (2015b). Searching
23 for flavor labels in food products: The influence of color-flavor congruence and association
24 strength. *Frontiers in Psychology*, **6**:301.

25 Velasco, C., Woods, A., Liu, J., & Spence, C. (in press). Assessing the role of taste intensity and
26 hedonics in taste/shape correspondences. *Multisensory Research*.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015

1 Wan, X., Velasco, C., Michel, C., Mu, B., Woods, A. T., & Spence, C. (2014a). Does the shape of the
2 glass influence the crossmodal association between colour and flavour? A cross-cultural
3 comparison. *Flavour*, **3**, 3.

4 Wan, X., Woods, A. T., Seoul, K.-H., Butcher, N., & Spence, C. (2015). When the shape of the glass
5 influences the flavour associated with a coloured beverage: Evidence from consumers in three
6 countries. *Food Quality & Preference*, **39**, 109-116.

7 Wan, X., Woods, A. T., van den Bosch, J., McKenzie, K. J., Velasco, C., & Spence, C. (2014b). Cross-
8 cultural differences in crossmodal correspondences between tastes and visual features. *Frontiers*
9 *in Psychology: Cognition*, **5**:1365.

10 Wechsler, D. (2008). *Wechsler adult intelligence scale–Fourth Edition (WAIS–IV)*. San Antonio, TX:
11 NCS Pearson.

12 Wilson, R. E., Gosling, S. D., & Graham, L. T. (2012). A review of Facebook research in the social
13 sciences. *Perspectives on Psychological Science*, **7**, 203-220.

14 Witzel, J., Cornelius, S., Witzel, N., Forster, K. I., & Forster, J. C. (2013). Testing the viability of
15 webDMDX for masked priming experiments. *The Mental Lexicon*, **8**, 421-449.

16 Woods, A. T., Butcher, N., & Spence, C. (in prep). Revisiting fast lemons and sour boulders: Testing
17 crossmodal correspondences using an internet-based testing methodology over four cultures.

18 Woods, A. T., Spence, C., Butcher, N., & Deroy, O. (2013). Fast lemons and sour boulders: Testing
19 the semantic hypothesis of crossmodal correspondences using an internet-based testing
20 methodology. *i-Perception*, **4**, 365-369.

21 Yung, A., Cardoso-Leite, P., Dale, G., Bavelier, D., & Green, C. S. (2015). Methods to test visual
22 attention online. *Journal of Visualized Experiments*, **96**, e52470.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.921v1 | CC-BY 4.0 Open Access | rec: 24 Mar 2015, publ: 24 Mar 2015