1

INTRODUCTION

Leib Litman and Jonathan Robinson

Do not copy, post, or distribute

A SCIENTIFIC REVOLUTION IN THE MAKING

Research in the social and behavioral sciences is undergoing a profound transformation that is nothing short of a revolution. This revolution consists of an explosion in online participant recruitment practices as well as a proliferation of resources for creating methodologically diverse studies and disseminating them online. Online research makes it possible to study human behavior in exciting and novel ways, at scales not possible in more traditional research settings. The applications of online technologies are limited only by the imagination of the researchers who use them.

At the heart of the web-based research revolution is Amazon's Mechanical Turk (MTurk)[1]. The MTurk platform gives social and behavioral scientists real-time access to thousands of participants from all over the United States and other countries. In traditional laboratory settings, collecting data from 500 participants can take months. On Mechanical Turk, a researcher can launch a study requiring 500 participants, go to lunch, and come back to find a complete dataset that is ready to be downloaded and analyzed.

Mechanical Turk is, however, more than just a way to quickly collect survey responses. MTurk workers write essays for open-ended qualitative research (Schnur, Dillon, Goldsmith, & Montgomery, 2018), grant access to personal data such as from their Fitbits (Brinton, Keating, Ortiz, Evenson, & Furberg, 2017) and Twitter accounts (Braithwaite, Giraud-Carrier, West, Barnes, & Hanson, 2016), allow researchers to study infants by video (Tran, Cabral, Patel, & Cusack, 2017), and engage with other participants in interactive games and group-based social experiments

[1] Note: Amazon Mechanical Turk is a registered trademark of Amazon.com, Inc.

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

(Arechar, Gächter, & Molleman, 2017). MTurk workers also participate in longitudinal studies, including studies that require intensive, daily tracking (Boynton & Richman, 2014). The limits of what is methodologically possible on Mechanical Turk have by no means been fully explored.

The popularity of Mechanical Turk among social and behavioral scientists began skyrocketing in 2011 following the publication of a seminal paper by Buhrmester, Kwang, and Gosling (2011), which showed that high-quality data can be collected on Mechanical Turk quickly and inexpensively. Since then, thousands of published studies have used Mechanical Turk to recruit participants. But interest in online research among scientists began long before 2011. To understand what led to the meteoric rise in the scientific community's adoption of Mechanical Turk, it is helpful to trace the history of online research to its very beginnings. In this chapter, we describe the range of tools that were available to researchers prior to Mechanical Turk and highlight how the limitations of those tools made Mechanical Turk the right platform at the right time for social and behavioral science.

A BRIEF HISTORY OF ONLINE RESEARCH IN THE SOCIAL AND BEHAVIORAL SCIENCES: FROM HTML 2.0 TO MECHANICAL TURK

Research in the social and behavioral sciences is currently shifting from traditional lab-based practices to the web at an ever-increasing pace. But enthusiasm for the opportunities that online research has to offer is nothing new. As early as 2000, researchers were talking about the potential of online research, writing that "the web presents researchers with an unprecedented opportunity to conduct experiments with participants from all over the world rather than with the usual student samples from their local universities. It thus has the potential to serve as an alternative or supplemental source of subjects and research environments for traditional psychological investigations" (Birnbaum & Reips, 2000). Indeed, even before the existence of the World Wide Web, scientists envisioned the vast potential that a global network of interconnected individuals would offer social and behavioral science (Kiesler & Sproull, 1986).

To understand why social scientists were interested in using the internet for research, it helps to consider the nature of human behavioral research. The social and behavioral sciences consist of a wide range of disciplines whose goals are to understand human cognition, behavior, personality, social interactions, health, and lifespan development. These disciplines include psychology, sociology, linguistics, marketing, business, economics, public health, behavioral medicine, and many others. Scientific

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

progress across all of these disciplines depends on having access to human subjects who are willing to participate in research studies. However, gaining access to human participants is often challenging in traditional research settings. A lack of participants limits research opportunities and the speed with which research projects can be completed. For these reasons, scientists have always sought novel venues that offer access to a diverse population of research participants, provided that such venues do not compromise the quality and validity of collected data.

It is not surprising, then, that when the internet started being widely used in the early 1990s, interest in leveraging the web for participant recruitment in scientific studies developed almost immediately. Researchers quickly realized that the web offers access to countless people around the world who may be interested in participating in research studies.

Although the World Wide Web was introduced to the world in 1990, up until 1995 internet browsers were implemented using what is now referred to as HTML 1.0. At that time, web pages were static and did not allow users to interact with the page. Starting in 1995, "fill-out forms" provided the substrate on which online survey software was constructed. Fill-out forms allowed users to enter information on a web page and send that information over the internet. With the introduction of forms, it became possible for the first time to conduct rudimentary online research studies, such as online surveys.

The subsequent introduction of Java and JavaScript provided researchers with additional flexibility to control stimuli over the internet. With JavaScript, researchers were able to randomly assign subjects to different experimental conditions, control the order of stimulus presentation using conditional logic, and create studies that were more sophisticated than simple surveys. With these tools in hand, it did not take long for researchers to start using the web to conduct research studies. As early as 1995, the web was being used for classroom experiments and educational demonstrations in psychoacoustics (Welch & Krantz, 1996). The first published paper using online research participants appeared in *Behavior Research Methods* in 1997 (Smith & Leigh, 1997). Laying out the specifics of this study will highlight the advantages and the limitations of online research at the time, and show how online research has evolved since then.

What the First Online Study Demonstrated About the Benefits of Web-Based Participant Recruitment

The first paper published using online research participants was positioned at the intersection between health psychology and social psychology and addressed the topic of eating disorders. In the 1990s, the rate of eating disorders was on the rise, especially among women. Sociocultural theories suggested this was due, in

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

part, to the influence of the media, which portrayed an ideal of beauty and thinness that is unattainable for most people. The messages portrayed by the media are important because women and men who internalize the ideal of thinness portrayed in magazines and movies are at a higher risk of developing eating disorders. To gain insight into this phenomenon, Smith and Leigh (1997) sought to better understand how the ideal female form is perceived and, in particular, how the perception of female beauty may differ across age groups. The challenge they faced, however, was recruiting participants to study psychological phenomena across the lifespan.

Limitations of the Undergraduate Subject Pool

Although thin ideal internalization had been studied extensively in prior research, the majority of those studies were conducted on college samples using what are referred to as university subject pools. These subject pools typically consist of undergraduate students who participate in research studies either to fulfill a course requirement or for extra credit. Although such undergraduate subject pools provide university faculty with easy access to research participants, they also have many drawbacks that have been discussed extensively in the scientific literature (Gosling & Johnson, 2010). Laboratory-based research with subject pool participants involves scheduling appointments and requires research staff to conduct one-on-one sessions with students. This process is labor intensive and is often time consuming. Another drawback is availability. Commonly, few people are available in the summer and during vacations. Perhaps the most important limitations are methodological. Studies using undergraduate participants are limited by the demographics of the undergraduate population at the institution where the research is taking place (Gosling, Sandy, John, & Potter, 2010; Henrich, Heine, & Norenzayan, 2010). This is especially true for finding older participants, as the average age of undergraduate subject pool participants is approximately 23 years (Gosling, Vazire, Srivastava, & John, 2004).

Therefore, a key motivation of Smith and Leigh (1997) for conducting the study online was to expand the age range of participants so as to examine whether the perception of female beauty remains constant with age. A second motivation was to examine whether any differences in the results would emerge between the online sample and the subject pool sample. An obvious concern with conducting online research is that researchers lose control over several aspects of the study that may affect participants' experience and attentiveness. As the first research study conducted online, this study sought to examine how the quality of online samples compares to a more well-established approach.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Conducting the First Online Study

The stimuli for the study consisted of black-and-white drawings that varied in the hip-to-bust ratio. Color images could not be accommodated by most monitors at the time and so were not used. Each image was rated by participants on how attractive it appeared. Participants were not recruited from any particular platform, as participant recruitment platforms did not exist. Instead, the researchers hoped that random people surfing the web would come across the study and be intrigued enough to spend 10 minutes or so completing the task.

The data collection was slow. Only around 25 people per day initiated the study. Of those who started, more than 70% dropped out. Nevertheless, after a few months the authors collected data from more than 550 participants. Comparing the web-based and college samples revealed a remarkable similarity in the results among younger participants. Most important, the online study found a systematic difference between the way in which older and younger people perceived the attractiveness of the female form. Younger participants tended to find the thinner images more attractive than did older participants. The spectrum of attractiveness ratings was shifted toward the heavier figures for the older participants, who found the heavier body types more attractive. These findings raised the intriguing possibility that younger individuals are more susceptible to beauty ideals presented by the media, or perhaps that older individuals prefer heavier body types.

This initial study demonstrated two key principles of online data collection. First, the quality of data collected online could be comparable to the data collected in more traditional research settings. Second, the demographic diversity of online samples could open significant opportunities to pursue novel research questions in a way that would be less time consuming and expensive than more traditional methods.

It was of course possible to claim, as many did (see, for example, Gosling et al., 2004), that such clean results may not be obtained for other types of online studies. A major concern was that participating in a study in an unregulated environment, as opposed to a laboratory, would increase distractibility, which would surely invalidate the results of studies that required sustained attention. Furthermore, the initial study did not reveal whether online participants could be relied on to carefully read instructions or long questionnaires, participate in tiring or tedious tasks, or complete reaction-time tasks. But this initial study did show that it was possible to collect high-quality data online, at least in some circumstances.

At the same time, this initial study demonstrated many of the limitations of online research. Data collection was generally slow, and the dropout rate was high. Significant technical limitations inhibited the range of stimuli that could be reliably

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

presented online. Finally, significant expertise was required to program and implement the study. Indeed, within a few years, multiple books would be published describing the programming necessary to create online studies because the initial technological hurdles were steep (Birnbaum, 2004; Fraley, 2004; Hewson, 2003). Despite these challenges, this pioneering study established the feasibility of collecting data online and highlighted a key advantage of the demographic diversity of online samples—thus launching a new era in online data collection methodology.

The Emergence of Large-Scale Online Projects

Starting in 1997, the number and scale of online studies in the social and behavioral sciences rose dramatically (see Krantz & Dalal, 2000). In 1998, Mahzarin Banaji and her colleagues initiated an online data collection project called Project Implicit (Nosek, Banaji, & Greenwald, 2002). This project examined implicit prejudice by allowing online participants to complete a short version of the Implicit Association Test (IAT). Project Implicit dwarfed previous online studies in its scope, collecting an average of more than 1,000 responses per day. Within the project's initial year and a half, more than 600,000 responses were collected. And more than two and a half million responses were collected over a five-year period. An important methodological advancement of Project Implicit was that implicit prejudice was measured by reaction time.

In the most popular version of the IAT task, positive words like *good* and *happy* are sometimes presented with White faces or stereotypically White names, and at other times with Black faces or stereotypically Black names. On other trials, negative words like *bad* and *evil* are shown together with either White or Black faces. Participants are asked to press a key with their left index finger as fast as they can if the word is positive and with their right index finger if the word is negative. Implicit prejudice is revealed when responses to positive words are faster when those words are paired with White faces than if those same words are paired with Black faces. What was methodologically critical here is that in order to detect these minor differences in reaction times subjects had to be highly attentive and carefully follow instructions. Prior to Project Implicit many researchers assumed that a controlled laboratory environment was necessary to detect such reaction-time differences.

Like the Smith and Leigh (1997) study, access to demographically diverse samples allowed Project Implicit to examine implicit prejudice among groups that are difficult to access in traditional research settings. Due to the scale and diversity of the sample, Project Implicit was able to establish that implicit prejudice exists among all age groups, all ethnic groups, all political parties, and all education levels. At the same time, the data revealed that social group membership moderates both implicit and explicit prejudice in important ways. Implicit prejudice toward African Americans

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

was higher among Whites compared to Blacks. The implicit prejudice effect was significantly higher for participants over age 50 than it was for younger people. Thus, a widely diverse group of participants provided unprecedented opportunities to examine group differences as moderators of implicit and explicit prejudice and with a sample size that could not be imagined with traditional lab-based approaches.

At the same time Project Implicit launched, large-scale studies in personality psychology began to be conducted online. Gosling and his colleagues (Gosling et al., 2004) collected a dataset of more than 350,000 participants who filled out the Big Five inventory. Because large extant lab-based datasets of the Big Five were available, it was possible to examine the data quality of online surveys with a high level of fidelity. Gosling and colleagues examined the internal reliabilities and correlation patterns among the five major personality traits in comparison to published datasets. Additionally, they conducted the first systematic examination of the diversity of internet samples relative to samples that were published in the *Journal of Personality and Social Psychology*, one of the most prestigious psychology journals.

In terms of diversity, Gosling et al.'s findings reinforced results previously presented by Nosek et al. (2002) and Smith and Leigh (1997), demonstrating that online samples were considerably more diverse across all demographic variables compared to samples of college students. In addition, they found that internal reliabilities of all personality dimensions were within a few decimal points of published norms, and that convergent and discriminant validity of all of the personality measures were consistent with published data.

Together, these early studies spoke to the concern among researchers that the quality of data collected online and outside of a well-controlled laboratory setting could not be trusted. The data from reaction-time tasks and the internal reliabilities of surveys were similar in quality to more traditional lab-based studies. Combined with the opportunities that diverse online datasets had to offer, by the early 2000s the web was beginning to look more and more appealing as a place to conduct research in the social and behavioral sciences.

THE USE OF ONLINE SAMPLES IN APPLIED BEHAVIORAL RESEARCH

As the popularity of online data collection was increasing within basic social science, web-based data collection methods were also being adopted in applied areas of behavioral science such as market research. Market research helps organizations better understand the needs of customers by collecting data about customers' opinions,

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

their experiences and satisfaction with products, and the effectiveness of advertising messages. For example, Volvo conducted a study of female drivers living in California to help with the development of a new Volvo model (Parasuraman, Grewal, & Krishnan, 2006).

Before the market research industry began using online data collection methods, respondents were typically contacted by mail, by phone, or in face-to-face interviews and focus groups (Evans & Mathur, 2005; Schibrowsky, Peltier, & Nill, 2007). These practices were inefficient, time consuming, and costly. As a result, much like their academic counterparts, market researchers began looking at the web to streamline finding and collecting data from research participants.

Market research differs from basic research in the social and behavioral sciences in a number of critical ways. First, because market research is integrated with commerce, it is a better-funded industry. In 2015, market research generated $21 billion in revenue (Rivera, 2015). In the early days of online research, these considerable financial resources were used to develop a robust infrastructure for recruiting online participants for this industry. A second way market research differs from academic research is that market researchers are often interested in specific market segments. In the Volvo study, for example, the specific market segment was women living in California who drive regularly. In other studies the market segment might be people who shop at Walmart or people who take cruise ship vacations at least once per year. Although general population studies are also common in market research, without the ability to break the market into specific segments, much of market research would not be feasible. Therefore, in addition to the need to reach a large and diverse online participant population, which market research shared in common with research in the social and behavioral sciences, there was a need to develop the tools to sample specific groups.

To meet the tremendous demand of market research companies, dozens of sample suppliers such as Survey Sampling International (now known as Dynata) and SurveyMonkey emerged to provide participants and robust web-based solutions. Using these platforms, companies conducting market research could request a sample with a thousand participants that are matched to the U.S. Census in terms of age, gender, and ethnicity, and who have specific shopping profiles. Companies that specialize in recruiting such participants for market research studies are known as panel providers or online access panels (Baker et al., 2010; Comley & Beaumont, 2011).

Panel providers often use two approaches to reach online participants. First, there are opt-in panels, which are made up of people who have agreed to be contacted for online research studies. Once people express an interest in being part of an opt-in panel, they are profiled by being asked to provide demographic information and to answer questions about their background, interests, and preferences. Panel companies use this information

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

to make individuals of interest available for specific market research studies. The second commonly used approach is called river sampling. In river samples, respondents are recruited as they surf the web (Baker et al., 2010). Individuals might see an advertisement to participate in a survey when they are playing a video game or shopping. In both opt-in panels and river samples, participants are given incentives to take surveys, for example, by being given the opportunity to earn cash or rewards points, and via sweepstakes, as we discuss in more detail in Chapter 10.

In addition to creating large online panels with tens of millions of participants, panel providers developed a robust infrastructure for research. Specifically, panel providers developed the technology to recruit and compensate participants, to target participants within specific geographic locations, and to protect the integrity of the data collection process.

At about the same time as the creation of online panels, there was a parallel explosion in the tools available for conducting online market research. Survey platforms such as Qualtrics and SurveyMonkey were created to make it easy to design and implement methodologically sophisticated research studies. In addition, numerous platforms emerged with specialized tools to accommodate different types of stimuli. At G2 Analytics, for example, researchers can employ technology that allows them to conduct studies where respondents watch videos such as political speeches or advertisements and respond any time they see or hear something they like or dislike. Such tools allow for real-time, in-depth assessment of video content.

By 2005, dozens of market research panel companies were providing access to millions of participants around the world. Driven by the need to reach specific market segments, hundreds of companies began specializing in recruiting specific populations. For example, some companies focused on recruiting Latino participants and others focused on recruiting participants in Asia. Eventually, companies started specializing in recruiting participants from social networks like Facebook, and through mobile technology (Poynter, Williams, & York, 2014). After 2005, many companies began to collaborate with one another in an attempt to meet the increasing demand for sampling hard-to-reach groups. Companies realized quickly that, alone, they could not meet the demand of many market research studies, especially if those studies targeted rare and difficult-to-access groups. Collectively, however, panel providers could find nearly any group of participants and, as a result, many companies either merged or developed partnerships to increase their overall reach. Currently, panel providers typically work closely with multiple industry partners. If any one sample provider cannot meet the demands of a study, it reaches out to its partners to increase the likelihood of meeting the required quota.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

While the market research industry developed on its own for a time, beginning in 2010 there existed two communities that increasingly sought access to online research participants: academic researchers in the social and behavioral sciences and applied researchers in market research and related fields. At that time, online academic research was in its infancy—enough online research had been conducted to demonstrate the vast potential of web-based recruitment, but virtually no infrastructure existed to support this field. Academic researchers interested in conducting online studies had no specific platform they could use to recruit online participants. Instead, researchers had to post studies on their own websites or various online forums in hopes of attracting enough interested participants. Significant expertise was required to ensure that participants were not completing the study multiple times and that the dataset was not corrupted by unscrupulous respondents. Beyond the ability to conduct a simple one-time study, no infrastructure existed for longitudinal follow-up or for recruiting participants within specific population segments. In short, social and behavioral scientists interested in conducting online studies had virtually no resources at their disposal, had to have significant programming expertise, and had to improvise ways of recruiting participants for each study.

At the same time, the infrastructure for conducting web-based market research was at a significantly more advanced stage of development. Sophisticated technological solutions were in place to manage the recruitment process. A market researcher wanting to conduct an online study had numerous online sample suppliers to choose from to recruit general population samples, or samples based on specific targeting criteria. These platforms had built-in protections against fraud, and sophisticated technology was available to conduct methodologically complex studies.

At first glance, online market research providers would seem to be a perfect recruiting tool that academics should have turned to for meeting their participant recruitment needs. However, panel companies had little penetration into the academic research space. The primary reason for this was cost. Many academic studies are not well funded and could not afford access to online research panels. For this reason, even though significant resources were available for online recruitment of participants through online panels, little research in social and behavioral sciences was conducted using these panels.

In summary, multiple barriers prevented research in the social and behavioral sciences from being conducted online on a large scale. This set the stage for the emergence of Amazon Mechanical Turk.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

AMAZON MECHANICAL TURK

Amazon's Mechanical Turk was created in 2005. With Mechanical Turk, Amazon sought to create a platform on which people could solve problems that did not have automated computer-based solutions. The term *Mechanical Turk* is based on an 18th-century chess machine that started to travel the world in 1770 and exhilarated crowds by outplaying many of the top chess players of that time. Among the more illustrious opponents of the Mechanical Turk were Benjamin Franklin and Napoleon Bonaparte, both of whom were beaten by the machine, and the world's then top chess player, Philidor, who beat the Turk though not without significant effort.

Unbeknownst to anyone at the time, inside the machine sat a chess master who used a magnet to move the pieces. The human player was concealed by an ingenious set of moving contraptions that may go down in history as the greatest magic act of all time. In almost 100 years of exhibitions around Europe, no one figured out the trick. Only after a fire tragically destroyed the machine in 1854 did its owner reveal the secret. In hindsight it is now obvious that in the 1800s no machine could have been sophisticated enough to play chess, much less to pose a challenge to great players like Philidor. Indeed, it was not until Deep Blue beat Gary Kasparov in 1997 that chess programs were able to pose a serious threat to the world's top players.

The Mechanical Turk chess machine demonstrated a principle that is at the heart of Amazon's MTurk platform: Certain tasks are better done by people than by machines. Although the ability of machines to conduct complex tasks is dramatically more advanced today than it was in 1800, the principle that people are able to do many things that computers can't is as true now as it was back then. For this reason, tasks that are conducted on Mechanical Turk are called Human Intelligence Tasks, or HITs for short. This name highlights that the problems being worked on through Mechanical Turk require the kind of human intelligence that machines do not yet possess. Often, using people is simply more efficient even when the task can be accomplished by machines. Increasingly, human input via Mechanical Turk is being used to train neural networks and is used for tasks in which humans and machines solve problems synergistically.

Workers and Requesters

The MTurk platform has two types of actors: workers and requesters. Typical tasks that workers do on Mechanical Turk are transcriptions of audio files, identification of objects in visual images, and categorization tasks. HITs may ask workers to review a website's usability and functionality, or to download an app and provide feedback about the user experience. Within a few years, Mechanical Turk reported 500,000

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

registered workers all over the globe on its platform. Thousands of HITs were available for these workers to choose from at any one time.

Mechanical Turk has some aspects in common with other platforms that recruit people online to perform research studies for monetary compensation, such as those run by market research panel providers. There are also fundamental differences. At its core Mechanical Turk is a platform that connects people who are willing to do work for monetary rewards with people who need that work done. The two key differences between Mechanical Turk and the market research platforms that preceded it are that monetary compensation on Mechanical Turk is set by market forces on the requester-worker exchange and that Mechanical Turk hosts a much wider range of tasks than traditional market research platforms.

Mechanical Turk as an Exchange

As an exchange, Mechanical Turk is a marketplace where wages are driven by market forces. Requesters set prices for how much they want to pay workers for any given HIT, and workers are free to choose whether to work on those HITs or not. Mechanical Turk charges the requester a standard transaction fee for each HIT when that requester agrees that the HIT was completed properly. When Mechanical Turk was launched in 2005, the transaction fee was 10%, but that fee was increased in 2015 to 20% or 40% depending on the task.

The MTurk exchange operates in stark contrast to other platforms where monetary compensation for participants is set not by the requester but by the platform itself. On Mechanical Turk it is possible to conduct a study where participants are paid $0.10 or even less for a HIT (Buhrmester et al., 2011; Litman, Robinson, & Rosenzweig, 2015). Although it is not common practice to pay participants so little, it is in principle possible to conduct a study that costs orders of magnitude less than it would on other platforms. For example, on SurveyMonkey, participants who complete a five-minute task are typically given a reimbursement of $0.50 donated to their charity of choice. SurveyMonkey typically charges the researcher $5 for each participant, thus adding a $4.50 transaction fee. A study with 500 participants could cost $2,500 on SurveyMonkey. On Mechanical Turk, by contrast, a five-minute study can be conducted paying each participant $0.10, which means that with 500 participants, the study can be completed for $50 (or 1/50th the cost). Even when participants are paid $1 for a five-minute study—a rate of $12 per hour—the price for the study would total $600 including MTurk fees. Interestingly, when conducted on Mechanical Turk, this study is not only cheaper for the researcher but also provides a higher level of compensation to the participant. (See Chapter 11 for data on compensation rates and a discussion

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

of ethics concerning pay.) The difference in price between Mechanical Turk and its alternatives is dramatic.

Reputation Mechanism and Data Quality

Because Mechanical Turk was created to provide solutions for technically challenging tasks, it did not originally focus on profiling its workers. After all, it doesn't matter whether workers categorize images in Los Angeles or in New York City, or whether they are 20 or 50 years old. For such projects all that matters is that the images are categorized correctly. Thus, rather than creating an infrastructure for selectively targeting specific groups, Mechanical Turk focused on mechanisms for ensuring data quality. Key among these mechanisms was a reputation system that allowed requesters to select workers based on their experience and past performance. Unlike other online panel providers, MTurk's reputation system was unique and eventually became so successful that workers became hyperattentive to task demands for fear of damaging their reputation.

The Right Platform at the Right Time for Academic Research

Mechanical Turk emerged as a platform on which many thousands of workers were available to complete tasks at a high level of quality, but where the demographic characteristics and geographic locations of those workers could not be controlled (this functionality would be added in 2016). For this reason, Mechanical Turk initially had limited use for most market researchers. For researchers in the social and behavioral sciences, however, Mechanical Turk provided the solutions they were seeking. The limited ability to selectively target participants was not a major barrier because most studies in the social and behavioral sciences were being conducted using university subject pools anyway, and those pools generally lacked the ability to target specific populations. For social and behavioral scientists, general population studies were the norm.

Because the price of transactions was significantly lower on Mechanical Turk than on other platforms, accessing online participants was now affordable in the academic community. The built-in reputation mechanism helped maintain data quality, as did the built-in mechanisms for preventing duplicate workers. And, because a payment mechanism was built into the system, participants could be paid seamlessly. Mechanical Turk eliminated the need for academic researchers to develop the technological expertise to host their own websites for data collection and the time needed to search internet forums in hopes of finding enough people to participate in their studies. With Mechanical Turk, social and behavioral researchers finally had access to a platform that could be used to recruit online participants easily, quickly, and affordably.

But numerous questions needed to be answered before the platform would be adopted by the research community. The most important question was whether data collected

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

on Mechanical Turk would meet the rigorous standards of quality required for scientific research. Many rightly wondered, is it realistic to expect high data quality from studies conducted on a platform where workers usually make less than a dollar to participate? In addition, in the early days of MTurk it was unclear what tasks researchers could realistically expect workers to perform—it is one thing to expect workers to stay attentive for 5 minutes, but it's another to expect them to carefully attend to task demands for 30 minutes or more. Was research on Mechanical Turk limited to studies that ask only survey questions, or could more sophisticated studies like reaction-time tasks also be conducted?

These questions were addressed in multiple early studies (e.g., Berinsky, Huber, & Lenz, 2012; Buhrmester et al., 2011; Goodman, Cryder, & Cheema, 2013; Horton, Rand, & Zeckhauser, 2011; Paolacci, Chandler, & Ipeirotis, 2010). In particular, Gosling and colleagues had been conducting research in personality psychology for over a decade using both online and subject pool participants (e.g., Gosling, Rentfrow, & Swann, 2003; Robins, Tracy, Trzesniewski, Potter, & Gosling, 2001). They thus had large datasets against which the performance of MTurk workers could be compared. Their deep experience with online research in the decade preceding the launch of Mechanical Turk perfectly positioned them to explore how data quality on Mechanical Turk stacked up against other online forums and more traditional laboratory settings.

The First Explorations of Data Quality on Mechanical Turk

One of the very first questions researchers addressed when assessing the utility of Mechanical Turk for academic research was whether participants who were paid little money could be relied on to provide high-quality data for survey tasks. Buhrmester et al. (2011) used the Big Five inventory to examine whether MTurk workers attentively fill out surveys. Focusing on the unique capability of Mechanical Turk to set any price for a study, they examined how different levels of monetary compensation affected data quality. Four monetary compensation conditions were used: $0.02, $0.05, $0.10, and $0.50. Participants in all four conditions were asked to fill out the Big Five personality inventory as well as multiple other surveys. Additionally, the number of survey instruments was changed across conditions to vary the length of the survey. In some conditions the study included many surveys and on average took 30 minutes to complete. Other conditions had fewer survey instruments and took just five minutes to complete. The goal of the study was to gain insight into the relationships among survey length, monetary compensation, and data quality.

Intuitively, one might expect the data quality from lower paying HITs to be worse than HITs that pay more. One might also expect data quality to be worse in longer HITs than shorter HITs, especially when the longer HITs do not pay well.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Surprisingly, however, data quality was consistently high across all conditions and remained the same whether participants were paid $0.02 or $0.50 and whether the study took 5 minutes or 30 minutes to complete (Buhrmester et al., 2011). The alpha reliability coefficients for every one of the Big Five personality dimensions was within a few decimal points of the reliabilities that had been reported both in previous online studies and in studies using undergraduate samples. Thus, this initial study showed that the data quality of surveys on Mechanical Turk was as good as the data quality using more traditional methods, including undergraduate and online samples, and was independent of task length and payment.

At the same time that Buhrmester et al. (2011) were conducting their research, the present authors were conducting similar studies to examine the quality of data collected on Mechanical Turk, some also using the Big Five (Litman et al., 2015). We found exactly the same pattern of results but under an extended range of payment conditions. We also used multiple data-quality screens within these studies, such as the squared discrepancy procedure, which examines the consistency with which workers respond to forward and reversed questions, and multiple attention manipulation checks (see Chapter 5). Our data-quality results were consistent with Buhrmester et al.'s findings and showed that whether participants were paid $0.02 or $1 had no impact on data quality.

Initially, these results seemed counterintuitive. How could paying $0.02 for a 30-minute task result in the same level of data quality as paying $1 for a 5-minute task? The answer to this question lies in Mechanical Turk's reputation mechanism. As both Buhrmester et al.'s (2011) and our data showed, the differences between the various payment and length conditions are revealed in the speed with which participants accepted and completed HITs. The longer and lower paying tasks took many days to recruit a sufficient number of participants. The shorter and better paying tasks, by contrast, were completed within minutes. Most workers were simply not willing to take HITs that did not pay well. However, because workers were careful to protect their reputation, they did not want to click through a task quickly just to receive a small monetary reward—and risk being rejected and having their reputation suffer. As a result, the majority of workers simply refused to work on those tasks, but those who did accept the task worked on it diligently. Chapter 7 provides an in-depth examination of the effect of pay rate on sample composition.

The key findings from these early studies revealed that Mechanical Turk's reputation mechanism is extremely successful at maintaining high levels of data quality on the platform, even when pay is relatively low. That is not to say monetary compensation does not play any role in data quality on Mechanical Turk. HITs that require high levels of creativity and effort are likely to be affected by pay rate. For example, as a later study conducted by researchers at the NYU Law School revealed, tasks that

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

require high levels of creativity are significantly affected by monetary compensation in the form of bonuses (Buccafusco, Burns, Fromer, & Sprigman, 2014). Further, monetary compensation affects retention rates in longitudinal studies (see Chapter 10). Monetary compensation thus has important implications for ethics (see Chapter 11), sample composition (see Chapter 7), and data quality for demanding tasks such as open-ended responses. But in terms of workers meeting basic task demands, the Buhrmester et al. (2011) and Litman et al. (2015) data revealed that high-quality data can be attained quickly and at low cost for many types of tasks.

Collecting Reaction-Time Data on Mechanical Turk

As interest in Mechanical Turk grew, researchers began exploring the extent to which Mechanical Turk could be used as a valid source of data in a variety of areas within the social and behavioral sciences. During the initial stages of this vetting era, researchers asked many of the same questions about Mechanical Turk that were originally asked about web-based research (see Gosling et al., 2004). One central question was whether reaction-time data could be collected reliably over the internet. The collection of reaction-time data presents several challenges that are of specific interest to cognitive scientists. First, reaction-time data are collected at the millisecond (ms) level, and small differences between groups and/or conditions often lead to important findings. Second, there is considerable heterogeneity in the computers used by online participants, including differences in processor speeds, internet connectivity, and monitor refresh rates. Finally, there is little or no control over the experimental environment in which such studies are conducted. Due to instrument heterogeneity and a lack of experimenter control, there were real concerns about whether Mechanical Turk could accommodate studies that rely on reaction-time data recorded with millisecond precision. Because the collection of reaction time affects numerous research areas in the social and behavioral sciences, this question seemed particularly important to address.

Shortly after academic researchers began running studies on Mechanical Turk, one group of researchers examined the efficacy of running reaction-time studies by attempting to replicate several classic cognitive psychology experiments (Crump, McDonnell, & Gureckis, 2013). Across multiple studies, they demonstrated that stimuli can be presented with at least 80 ms presentation speed and that effect sizes of at least 20 ms can be reliably detected. This level of precision meets the requirements of the majority of reaction-time studies. These effects are much better than may have been expected given the technical demands and lack of experimenter control over the study settings on Mechanical Turk. Later studies showed that significantly faster presentation speeds and even finer levels of reaction-time resolution are also possible when using more specialized software (Barnhoorn, Haasnoot, Bocanegra, & van Steenbergen, 2015). The Crump et al. (2013) results showed that data quality for studies that rely on extremely fine levels

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

of reaction-time measurement can be reliably attained on MTurk, leading to increasingly widespread adoption of Mechanical Turk for studies requiring reaction-time measures (for an overview, see Stewart, Chandler, & Paolacci, 2017).

Expanding the Range of What Is Possible on Mechanical Turk

Initially, research looking at data quality on Mechanical Turk showed that it held considerable promise for collecting high-quality data with studies that used both survey instruments and reaction-time measures as outcomes. There were, however, formidable challenges to running more complex social and behavioral research studies on Mechanical Turk. Even the most rudimentary requirements of behavioral research studies were difficult to implement. For example, researchers typically did not want participants from a first study to participate in subsequent studies, since the workers were now familiar with the study's protocols. But on Mechanical Turk it was extremely difficult to prevent workers from participating in a study based on their previously taken HITs. After all, as far as Mechanical Turk was concerned, if a worker was experienced at categorizing images, why wouldn't someone want that individual to do it again? Thus there was no easy mechanism put in place to flexibly manage participant recruitment. On the flip side, it was also difficult to conduct longitudinal studies. Recruiting participants from previous studies on Mechanical Turk was complex and required the management of what are called qualifications (see Chapters 3 and 4).

Various solutions were developed by the research community to solve these problems. Eventually, specialized third-party platforms such as psiTurk (Gureckis et al., 2016), TurkServer (Mao et al., 2012), and TurkPrime (now CloudResearch; Litman, Robinson, & Abberbock, 2017) were developed to help researchers navigate various intricacies of Mechanical Turk. With the robust capabilities of Mechanical Turk now augmented by third-party solutions, the use of Mechanical Turk among scientists exploded with unprecedented speed.

The rate of adoption of Mechanical Turk was dramatic. Within a few years following the publication of Buhrmester et al.'s (2011) seminal paper in the journal *Perspectives on Psychological Science*, the paper was cited more than 5,000 times. As of 2015, 40% of all research papers published in the *Journal of Personality and Social Psychology* had at least one study with MTurk workers (Zhou & Fishbach, 2016). Researchers quickly began to explore how Mechanical Turk could accommodate research in each of their specific subfields of the social and behavioral sciences. Within a short time, researchers from more than 30 different academic disciplines were routinely using Mechanical Turk for their research studies (Bohannon, 2016).

The scientific literature is replete with research questions explored using the MTurk participant pool. Table 1.1 lists some examples of research questions that have been

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

TABLE 1.1 ● EXAMPLES OF RESEARCH QUESTIONS ADDRESSED USING MECHANICAL TURK

Research Question	Citation
Does cognitive reappraisal ability buffer against the indirect effects of perceived stress reactivity on Type 2 diabetes?	Sagui and Levens (2016)
How can individuals be mobilized to action against voter ID laws?	Valentino and Neuner (2017)
Is there support for a general factor of well-being?	Longo, Coyne, Joseph, and Gustavsson (2016)
How might men who are depressed express themselves?	Nadeau, Balsan, and Rochlen (2016)
Is there a connection between eating pathology and risk for engaging in suicidal behavior?	Skinner, Rojas, and Veilleux (2017)
What factors predispose some individuals toward holding beliefs in God?	Wlodarski and Pearce (2016)
How does charismatic leadership affect cooperation?	Grabo and van Vugt (2016)
Are people more authentic after having vivid thoughts about their death?	Seto et al. (2016)
How does psychopathy relate to individuals' ability to discriminate trustworthy and untrustworthy faces and genuine expressions?	Sacco et al. (2016)
Does a feminine appearance for women pursuing STEM erroneously signal that they are not well suited for science?	Banchefsky et al. (2016)
Is belief in God linked to social and emotional cognition?	Jack et al. (2016)
Does pathogen-avoidance motivation lead to health protective behaviors?	Gruijters et al. (2016)
Do parenting styles affect parental feeding practices?	Kiefner-Burmeister et al. (2016)
Are there sociocultural differences in people's beliefs in the utility of preventive medicine?	Dye et al. (2016)
What factors influence customers' likelihood to join a program in the casino industry?	Quigno and Zhang (2016)
How can privacy concerns help predict mobile commerce activity?	Eastin et al. (2016)

(Continued)

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

TABLE 1.1 ● *(Continued)*

Research Question	Citation
What are the characteristics of an ideal dairy farm?	Cardoso et al. (2016)
Does smiling when giving people service impact evaluations of service providers?	Andrzejewski and Mooney (2016)
Does an ad background with a warm color make people's judgments toward a company more positive?	Choi et al. (2016)
How do consumer preferences influence moral judgments of corporate misconduct?	Lewis et al. (2016)
How does the way people organize semantic information change with age?	Unger et al. (2016)
How can automatic methods for assessing credibility and relevance of social media posts be improved?	Figueira et al. (2016)
Do people update their representations when making judgments from memory, or do they maintain their representations based on the initial encoding?	Sharif and Oppenheimer (2016)
Does thinking about a limited future enhance the positivity of subsequently recalled information?	Barber et al. (2016)
How do the instructions jurors receive affect the way they consider confessional evidence?	O'Donnell and Safer (2017)
What are the characteristics of an ideal mentor/ advisor?	Bailey et al. (2016)

addressed using MTurk workers as research participants. As this table illustrates, the studies conducted on Mechanical Turk span multiple disciplines and topics.

CONCLUSION

Prior to Mechanical Turk, online research remained a niche field, existing largely outside of mainstream science. Despite interest from a growing number of researchers (Birnbaum, 2004; Fraley, 2004; Gosling & Johnson, 2010; Hewson, 2003), several factors kept online research from entering the mainstream. First, significant skepticism remained about the level of data quality of online samples (see Gosling et al.,

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

2010). The inability to control what participants were doing during the study raised researchers' suspicions. Perhaps most important, at the time, significant technical skills were required to conduct even the simplest study online, putting online research out of reach of most researchers who did not have the technical expertise to carry out such studies.

The emergence of Mechanical Turk has democratized online research. With Mechanical Turk, no longer does online research require significant programming expertise and high levels of research funding. Online research in general, and Mechanical Turk in particular, has moved on from being a niche field. Soon, almost every researcher will be dealing with Mechanical Turk in one form or another, either for data collection for their own studies or in reading and evaluating the research of others.

While research on Mechanical Turk is becoming increasingly more sophisticated and complex, new research tools and methodologies are increasingly becoming available. Researchers are now able to collect data via live audio (Gašić, Jurčíček, Thomson, Yu, & Young, 2011; Gašić et al., 2013; McGraw, 2013) and video chat (Miller, Mandryk, Birk, Depping, & Patel, 2017). It is becoming increasingly possible to collect physiological data during live video interviews (Muender, Miller, Birk, & Mandryk, 2016) and by accessing databases from wearable devices (Brinton et al., 2017). Clinical research is also becoming increasingly more commonplace (see Chandler & Shapiro, 2016). For example, workers grant access to their Twitter accounts, which can be used to gain clinical insights into critical areas such as suicidality and depression (Braithwaite et al., 2016). Additionally, since Mechanical Turk was originally intended as a marketplace for work, researchers are increasingly learning how to use MTurk workers as their research assistants for tasks like stimulus development and validation (see Chapter 2, where we use MTurk workers to provide feedback about parts of this book).

With exponential increases in the number of tools available for conducting research on Mechanical Turk, the amount of data available about the MTurk worker population, and the quality of the available data, it is becoming increasingly difficult for researchers to keep track of all the new advances and resources.

Getting started on Mechanical Turk requires familiarity with the basic mechanisms and concepts that make Mechanical Turk unique. MTurk workers function within a unique online culture that many researchers are not familiar with. Third-party platforms are becoming more sophisticated and provide researchers with increasingly powerful new tools to enhance Mechanical Turk's functionality. The demographic composition of MTurk workers, including the question of how many workers are available, is also generally poorly understood. Important questions about ethical concerns surrounding research on Mechanical Turk commonly arise. Finally, as

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

researchers increasingly begin to look to the web for their data collection needs, interest in using other platforms is also beginning to increase. Although Mechanical Turk has substantial advantages over other platforms, alternative sources of data collection such as market research platforms have much to offer social and behavioral scientists. However, the relative advantages and disadvantages of Mechanical Turk versus market research platforms are often not clear.

This book aims to address these and many other issues. In the upcoming chapters we provide an overview of basic concepts that are unique to Mechanical Turk, including an overview of the MTurk culture and ecosystem. The first part of the book (Chapters 1–5) offers an introduction. Its aim is to provide an overview of stimulus development platforms, show how to set up a study on Mechanical Turk, discuss best practices for study setup, and provide a conceptual introduction to the MTurk application programming interface (API) and third-party systems such as CloudResearch. The second part of the book addresses advanced topics, including data quality on Mechanical Turk and other platforms, the demographic composition of Mechanical Turk, the activity levels of workers, best sampling practices, the representativeness of studies conducted on Mechanical Turk, best practices for conducting longitudinal research, and the ethics of conducting research on Mechanical Turk.

Finally, this book is not limited to Mechanical Turk. We aim to provide a comprehensive and up-to-date overview of the complex and quickly evolving ecosystems available for online participant recruitment. Mechanical Turk is one of many approaches for recruiting participants online. Each approach has its advantages and limitations, which are discussed in Chapter 10. We provide an overview of market research platforms, data quality and demographic differences between Mechanical Turk and online panels, and advantages and disadvantages of using Mechanical Turk relative to other platforms. Overall, we see Mechanical Turk and online panels as being complementary, with each well suited for specific research questions. This book aims to introduce readers to the many options available to help researchers complete their projects successfully using online research participants.

REFERENCES

Andrzejewski, S. A., & Mooney, E. C. (2016). Service with a smile: Does the type of smile matter? *Journal of Retailing and Consumer Services*, *29*, 135–141. doi:10.1016/j.jretconser.2015.11.010

Arechar, A. A., Gächter, S., & Molleman, L. (2017). Conducting interactive experiments online. *Experimental Economics*, 1–33.

Bailey, S. F., Voyles, E. C., Finkelstein, L., & Matarazzo, K. (2016). Who is your ideal mentor? An exploratory study of mentor prototypes. *Career Development International*, *21*(2), 160–175. doi:10.1108/CDI-08-2014-0116

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., & Kennedy, C. (2010). AAPOR report on online panels. *Public Opinion Quarterly*, *74*(4), 711–781.

Banchefsky, S., Westfall, J., Park, B., & Judd, C. M. (2016). But you don't look like a scientist! Women scientists with feminine appearance are deemed less likely to be scientists. *Sex Roles*, *75*(3–4), 95–109. doi:10.1007/s11199-016-0586-1

Barber, S. J., Opitz, P. C., Martins, B., Sakaki, M., & Mather, M. (2016). Thinking about a limited future enhances the positivity of younger and older adults' recall: Support for socioemotional selectivity theory. *Memory & Cognition*, *44*(6), 869–882. doi:10.3758/s13421-016-0612-0

Barnhoorn, J. S., Haasnoot, E., Bocanegra, B. R., & van Steenbergen, H. (2015). QRTEngine: An easy solution for running online reaction time experiments using Qualtrics. *Behavior Research Methods*, *47*(4), 918–929. doi:10.3758/s13428-014-0530-7

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

Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology*, *55*(1), 803–832. doi:10.1146/annurev.psych.55.090902.141601

Bohannon, J. (2016). Psychology. Mechanical Turk upends social sciences. *Science*, *352*(6291), 1263–1264. doi:10.1126/science.352.6291.1263

Boynton, M. H., & Richman, L. S. (2014). An online daily diary study of alcohol use using Amazon's Mechanical Turk. *Drug and Alcohol Review*, *33*(4), 456–461. doi:10.1111/dar.12163

Braithwaite, S. R., Giraud-Carrier, C., West, J., Barnes, M. D., & Hanson, C. L. (2016). Validating machine learning algorithms for Twitter data against established measures of suicidality. *JMIR Mental Health*, *3*(2), e21. doi:10.2196/mental.4822

Brinton, J. E., Keating, M. D., Ortiz, A. M., Evenson, K. R., & Furberg, R. D. (2017). Establishing linkages between distributed survey responses and consumer wearable device datasets: A pilot protocol. *JMIR Research Protocols*, *6*(4), e66. doi:10.2196/resprot.6513

Buccafusco, C., Burns, Z. C., Fromer, J. C., & Sprigman, C. J. (2014). Experimental tests of intellectual property laws' creativity thresholds. *Texas Law Review*, *93*, 1921–1980.

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

Cardoso, C. S., Hötzel, M. J., Weary, D. M., Robbins, J. A., & von Keyserlingk, M. A. (2016). Imagining the ideal dairy farm. *Journal of Dairy Science*, *99*(2), 1663–1671. doi:10.3168/jds.2015-9925

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology*, *12*, 53–81.

Choi, J., Chang, Y. K., Lee, K., & Chang, J. D. (2016). Effect of perceived warmth on positive judgment. *Journal of Consumer Marketing*, *33*(4), 235–244. doi:10.1108/JCM-02-2015-1309

Comley, P., & Beaumont, J. (2011). Online market research: Methods, benefits and issues—Part 1. *Journal of Direct, Data and Digital Marketing Practice*, *12*(4), 315–327. doi:10.1057/dddmp.2011.8

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE*, *8*(3), e57410. doi:10.1371/journal.pone.0057410

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Dye, T., Li, D., Demment, M., Groth, S., Fernandez, D., Dozier, A., & Chang, J. (2016). Sociocultural variation in attitudes toward use of genetic information and participation in genetic research by race in the United States: Implications for precision medicine. *Journal of the American Medical Informatics Association*, *23*(4), 782–786. doi:10.1093/jamia/ocv214

Eastin, M. S., Brinson, N. H., Doorey, A., & Wilcox, G. (2016). Living in a big data world: Predicting mobile commerce activity through privacy concerns. *Computers in Human Behavior*, *58*, 214–220. doi:10.1016/j.chb.2015.12.050

Evans, J. R., & Mathur, A. (2005). The value of online surveys. *Internet Research*, *15*(2), 195–219. doi:10.1108/10662240510590360

Figueira, A., Sandim, M., & Fortuna, P. (2016). March). An approach to relevancy detection: Contributions to the automatic detection of relevance in social networks. *WorldCIST*, *1*, 89–99.

Fraley, R. C. (2004). *How to conduct behavioral research over the Internet: A beginner's guide to HTML and CGI/Perl*. New York: Guilford Press.

Gašić, M., Breslin, C., Henderson, M., Kim, D., Szummer, M., Thomson, B., & Young, S. (2013). On-line policy optimisation of Bayesian spoken dialogue systems via human interaction. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8367–8371). New York: IEEE.

Gašić, M., Jurčíček, F., Thomson, B., Yu, K., & Young, S. (2011). On-line policy optimisation of spoken dialogue systems via live interaction with human subjects. In *2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 312–317). New York: IEEE.

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

Gosling, S. D., & Johnson, J. A. (2010). *Advanced methods for conducting online behavioral research*. Washington, DC: American Psychological Association.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. doi:10.1016/S0092-6566(03)00046-1

Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (2010). Wired but not WEIRD: The promise of the Internet in reaching more diverse samples. *Behavioral and Brain Sciences*, *33*(2–3), 94–95. doi:10.1017/S0140525X10000300

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

Grabo, A., & van Vugt, M. (2016). Charismatic leadership and the evolution of cooperation. *Evolution and Human Behavior*, *37*(5), 399–406. doi:10.1016/j.evolhumbehav.2016.03.005

Gruijters, S. L. K., Tybur, J. M., Ruiter, R. A. C., & Massar, K. (2016). Sex, germs, and health: Pathogen-avoidance motives and health-protective behaviour. *Psychology & Health*, *31*(8), 959–975. doi:10.1080/08870446.2016.1161194

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., . . .Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842. doi:10.3758/s13428-015-0642-8

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29–29. doi:10.1038/466029a

Hewson, C. (2003). *Internet research methods: A practical guide for the social and behavioural sciences*. Newbury Park, CA: Sage.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: conducting experiments in a real labor market. *Experimental Economics*, *14*(3), 399–425. doi:10.1007/s10683-011-9273-9

Jack, A. I., Friedman, J. P., Boyatzis, R. E., & Taylor, S. N. (2016). Why do you believe in God? Relationships between religious belief, analytic thinking, mentalizing and moral concern. *PLoS ONE*, *11*(3), e0149989. doi:10.1371/journal.pone.0149989

Kiefner-Burmeister, A., Hoffmann, D., Zbur, S., & Musher-Eizenman, D. (2016). Implementation of parental feeding practices: Does parenting style matter? *Public Health Nutrition*, *19*(13), 2410–2414. doi:10.1017/S1368980016000446

Kiesler, S., & Sproull, L. S. (1986). Response effects in the electronic survey. *Public Opinion Quarterly*, *50*(3), 402–413.

Krantz, J. H., & Dalal, R. (2000). Validity of web-based psychological research. In *Psychological experiments on the Internet* (pp. 35–60). New York: Academic Press.

Lewis, R., Anderson, R., & Pounders, K. (2016). Morality shifting: How consumer preferences influence moral judgments of corporate misconduct. *Journal of Promotion Management*, *22*(1), 1–15. doi:10.1080/10496491.2015.1107014

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442. doi:10.3758/s13428-016-0727-z

Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behavior Research Methods*, *47*(2), 519–528. doi:10.3758/s13428-014-0483-x

Longo, Y., Coyne, I., Joseph, S., & Gustavsson, P. (2016). Support for a general factor of well-being. *Personality and Individual Differences*, *100*, 68–72. doi:10.1016/j.paid.2016.03.082

Mao, A., Chen, Y., Gajos, K. Z., Parkes, D., Procaccia, A. D., & Zhang, H. (2012). Turkserver: Enabling synchronous and longitudinal online experiments. *Proceedings of HCOMP*, *12*.

Mao, A., Chen, Y., Gajos, K. Z., Parkes, D. C., Zhang, H., & Procaccia, A. D. (2012). Turkserver: Enabling synchronous and longitudinal online experiments. *In Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence.*

McGraw, I. (2013). Collecting speech from crowds. In M. Eskenazi, G. A. Levow, H. Meng, G. Parent, & D. Suendermann (Eds.), *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment* (pp. 38–71). Hoboken, NJ: John Wiley & Sons.

Miller, M. K., Mandryk, R. L., Birk, M. V., Depping, A. E., & Patel, T. (2017). Through the looking glass: The effects of feedback on self-awareness and conversational behaviour during video chat. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 5271–5283). New York, NY: Association for Computing Machinery.

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Muender, T., Miller, M. K., Birk, M. V., & Mandryk, R. L. (2016). Extracting heart rate from videos of online participants. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 4562–4567). New York, NY: Association for Computing Machinery.

Nadeau, M. M., Balsan, M. J., & Rochlen, A. B. (2016). Men's depression: Endorsed experiences and expressions. *Psychology of Men & Masculinity*, *17*(4), 328–335. doi:10.1037/men0000027

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101–115. doi:10.1037/1089-2699.6.1.101

O'Donnell, C. M., & Safer, M. A. (2017). Jury instructions and mock-juror sensitivity to confession evidence in a simulated criminal case. *Psychology, Crime & Law*, 1–21.

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

Parasuraman, A., Grewal, D., & Krishnan, R. (2006). *Marketing research*. Boston, MA: Cengage Learning.

Poynter, R., Williams, N., & York, S. (2014). *The handbook of mobile market research: Tools and techniques for market researchers*. West Sussex, United Kingdom: John Wiley & Sons.

Quigno, J., & Zhang, L. (2016). Casino customers' intention to join a loyalty rewards program: The effect of number of tiers and gender. *Cornell Hospitality Quarterly*, *57*(2), 226–230.

Reips, J. M. U. D., & Musch, J. (2000). A brief history of web experimenting. In M. H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 61–87). New York, NY: Academic Press.

Rivera, E. (2015). Market research in the US. *IBISWorld Industry Report*, *54191*.

Robins, R. W., Tracy, J. L., Trzesniewski, K., Potter, J., & Gosling, S. D. (2001). Personality correlates of self-esteem. *Journal of Research in Personality*, *35*(4), 463–482. doi:10.1006/jrpe.2001.2324

Sacco, D. F., Merold, S. J., Lui, J. H. L., Lustgraaf, C. J. N., & Barry, C. T. (2016). Social and emotional intelligence moderate the relationship between psychopathy traits and social perception. *Personality and Individual Differences*, *95*, 95–104. doi:10.1016/j.paid.2016.02.031

Sagui, S. J., & Levens, S. M. (2016). Cognitive reappraisal ability buffers against the indirect effects of perceived stress reactivity on type 2 diabetes. *Health Psychology*, *35*(10), 1154–1158. doi:10.1037/hea0000359

Schibrowsky, J. A., Peltier, J. W., & Nill, A. (2007). The state of Internet marketing research: A review of the literature and future research directions. *European Journal of Marketing*, *41*(7/8), 722–733.

Schnur, J. B., Dillon, M. J., Goldsmith, R. E., & Montgomery, G. H. (2018). Cancer treatment experiences among survivors of childhood sexual abuse: A qualitative investigation of triggers and reactions to cumulative trauma. *Palliative & Supportive Care*, *16*(6), 767–776.

Seto, E., Hicks, J. A., Vess, M., & Geraci, L. (2016). The association between vivid thoughts of death and authenticity. *Motivation and Emotion*, *40*(4), 520–540. doi:10.1007/s11031-016-9556-8

Sharif, M. A., & Oppenheimer, D. M. (2016). The effect of relative encoding on memory-based judgments. *Psychological Science*, *27*(8), 1136–1145. doi:10.1177/0956797616651973

Skinner, K. D., Rojas, S. M., & Veilleux, J. C. (2017). Connecting eating pathology with risk for engaging in suicidal behavior: The mediating role of experiential avoidance. *Suicide and Life-Threatening Behavior*, *47*(1), 3–13. doi:10.1111/sltb.12249

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.

Smith, M. A., & Leigh, B. (1997). Virtual subjects: Using the Internet as an alternative source of subjects and research environment. *Behavior Research Methods*, *29*(4), 496–505.

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*, *21*(10), 736–748. doi:10.1016/j.tics.2017.06.007

Tran, M., Cabral, L., Patel, R., & Cusack, R. (2017). Online recruitment and testing of infants with Mechanical Turk. *Journal of Experimental Child Psychology*, *156*, 168–178. doi:10.1016/j.jecp.2016.12.003

Unger, L., Fisher, A. V., Nugent, R., Ventura, S. L., & MacLellan, C. J. (2016). Developmental changes in semantic knowledge organization. *Journal of Experimental Child Psychology*, *146*, 202–222. doi:10.1016/j.jecp.2016.01.005

Valentino, N. A., & Neuner, F. G. (2017). Why the sky didn't fall: Mobilizing anger in reaction to voter ID laws. *Political Psychology*, *38*(2), 331–350. doi:10.1111/pops.12332

Welch, N., & Krantz, J. H. (1996). The world-wide web as a medium for psychoacoustical demonstrations and experiments: Experience and results. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 192–196.

Wlodarski, R., & Pearce, E. (2016). The God allusion. *Human Nature*, *27*(2), 160–172. doi:10.1007/s12110-016-9256-9

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493–504. doi:10.1037/pspa0000056

Do not copy, post, or distribute

Copyright ©2021 by SAGE Publications, Inc.
This work may not be reproduced or distributed in any form or by any means without express written permission of the publisher.