Crowdsumers Take Over:

Towards Valid Crowdsourcing of Consumer Research

JOSEPH K. GOODMAN

GABRIELE PAOLACCI *

* Joseph K. Goodman (goodman.425@osu.edu) is assistant professor of marketing, Fisher College of Business, The Ohio State University, 2100 Neil Avenue, 532 Fisher Hall, Columbus, OH 43210. Gabriele Paolacci (gpaolacci@rsm.nl) is assistant professor of marketing, Rotterdam School of Management, Erasmus University, PO Box 1738, 3000 DR Rotterdam, The Netherlands. We thank Jesse Chandler and Joel Huber for their comments, and Heath Cleveland, Gus Passov, Natalie Faust, Katherine Jaruzelski, and Gizem Yalçin for their assistance on the manuscript. Authors contributed equally to this work. Authorship is alphabetical.

Crowdsumers Take Over: Towards Valid Crowdsourcing of Consumer Research

CONTRIBUTION STATEMENT

Crowdsourcing websites such as Amazon Mechanical Turk now represent over 40% of the data published in the *Journal of Consumer Research.* While some research about crowdsourcing has been published recently, the field is widely dispersed across disciplines and there is not a single article/report/editorial that addresses the advantages and potential problems with MTurk to consumer research. In this paper, we address this gap with a thorough and deep review of all the crowdsourcing literature to address the needs of not only consumer researchers, but also editors, associate editors, and reviewers. In addition to providing a very brief overview of MTurk to those who may be unfamiliar with the service, we did three things: (1) Discuss the advantages of crowdsourcing, with explicit reference to how MTurk not only makes consumer science more efficient, but also has the potential to improve the field in a new way; (2) Assess the evidence for problems (both real and perceived) with MTurk, including non-representativeness, nonnaiveté, and the possibility that convenience dictates research programs in consumer research; (3) Present specific guidelines to improve the quality of crowdsourcing in consume research. In sum, we believe the paper makes an important contribution and provides a critical resource to the field of consumer research.

ABSTRACT

Data collection in consumer research has progressively moved away from traditional samples (e.g., university undergraduates) and towards Internet samples. In the last complete volume of the *Journal of Consumer Research* (June 2015-April 2016), 43% of behavioral studies were conducted on the crowdsourcing website Amazon Mechanical Turk (MTurk). The possibility to crowdsource empirical investigations has great efficiency benefits for both individual researchers and the field, but it also poses new challenges and questions for how research should be designed, conducted, analyzed, and evaluated. We assess the evidence on the reliability of crowdsourced populations and the conditions under which crowdsourcing is a valid strategy for data collection. Based on this evidence, we propose specific guidelines for researchers to conduct high quality research via crowdsourcing. We hope this tutorial will strengthen the community's scrutiny on data collection practices and move the field towards better and more valid crowdsourcing of consumer research.

Academic consumer research strongly depends on the availability of study participants. Theories of consumer behavior typically require tests with human participants, and researchers often choose their samples based on convenience. For decades, this translated into an overwhelming reliance on undergraduate students, and the *Journal of Consumer Research* (*JCR*) hosted several debates on the external validity of these investigations (Calder, Phillips, and Tybout 1981, 1982, 1983; Ferber 1977; Lynch 1982, 1983; McGrath and Brinberg 1983; Peterson 2001; Wells 1993). In recent years, however, consumer researchers have increasingly turned to the Internet to recruit study participants and collect data, and in particular to *crowdsourcing.* On websites such as Mechanical Turk (MTurk) and Prolific, researchers act as "employers" and hire and compensate "workers" to participate in computerized tasks (e.g., surveys, choice tasks, and/or simulated shopping environments). Crowdsourcing brings a new meaning to convenience sampling: These platforms grant unprecedented efficiencies, providing researchers with participants who can be accessed at any point in time and are more demographically diverse and more inexpensive to reach than traditional research participants.

Consistent with trends in other social sciences, consumer research is now routinely, if not by default, conducted using online marketplaces. Currently, the most prevalent crowdsourcing destination is MTurk, a marketplace that was launched in 2005 by Amazon. Over 15,000 published papers referenced MTurk in past 10 years (Chandler and Shapiro 2016). We analyzed the last four complete volumes (volumes 39-42, published between June 2012 and April 2016) of the *Journal of Consumer Research* (*JCR*). Across these 24 issues, *JCR* published 1350 surveys and experiments, 27% of which were conducted on MTurk. Most strikingly, the prevalence of MTurk studies steadily increased from 9% of total studies in issue 39 to a whopping 43% in

issue 42. Despite its relative youth, MTurk may well be the most represented participant pool in the history of consumer research.

With its convenience, crowdsourcing has also brought its share of skepticism and controversy. Echoing early concerns with online research (cf. Kraut et al. 2004), some researchers worry about the impossibility to scrutinize participants' behavior—participants might be multitasking or be interrupted during the study. Further, crowdsourced participants self-select into studies and can quit at any time, and samples may vary between studies and between conditions within a study as a result of arbitrary choices in the design and sampling process (Casey et al. 2016, Zhou and Fischbach 2016). Thus, MTurk workers may not provide reliable data nor be particularly representative of real-world consumers—a similar criticism levied for decades on the use of college students in consumer research (Calder et al. 1981, 1982, 1983; Ferber 1977; Lynch 1982, 1983; McGrath and Brinberg 1983; Peterson 2001; Petty and Cacioppo 1996; Wells 1993). Perhaps more concerning to critics is the validity of data obtained from participants who have accumulated experience with social science studies (Pham 2013). At a more philosophical level, there is concern that the efficiency of online samples might increase the attractiveness of research programs and paradigms that can be conducted with the crowdsourcing method (e.g., scenarios), at potential detriment to research that is important but difficult to crowdsource (Pham 2013). These and other concerns are legitimate. While the possibility to crowdsource empirical investigations has great efficiency benefits for both individual researchers and the field, it also poses new challenges for how research should be designed, conducted, and analyzed by researchers, and evaluated by editors, reviewers, and readers.

In this tutorial, we assess the evidence on the reliability and validity of crowdsourced populations and the conditions under which crowdsourcing is a valid strategy for data collection, with the goal to establish *valid crowdsourcing* as a data collection strategy in consumer research. First, after a brief overview of MTurk, we discuss how crowdsourcing provides advantages to individual researchers and the consumer research field. We argue that MTurk and similar websites can improve not only the convenience, but also the quality of consumer research—if used appropriately. Second, drawing from methodological research on crowdsourcing, we address the substance and the magnitude of concerns associated with the use of crowdsourced samples. Third, we offer specific guidelines for consumer researchers to maximize the advantages of crowdsourcing while attenuating the methodological concerns raised in the literature.

Though we will predominantly focus on MTurk—the most popular crowdsourcing destination among behavioral researchers—our findings by and large apply to any crowdsourcing solution that is or will become available in the foreseeable future. As opportunities will flourish to harness the advantages of online samples, we hope our tutorial will trigger new contributions on the methodology of data collection in consumer research.

BACKGROUND: CONSUMER RESEARCH WITH THE CROWD

The use of the Internet as an instrument of data collection in the social sciences dates back to the late nineties, and has been the object of debate ever since (Gosling et al. 2004; Kraut et al. 2004; Skitka and Sargis 2006; see Gosling and Mason 2015 for a recent review). Ten years later, crowdsourcing marketplaces made it much easier for researchers to conduct online

investigations, resulting in an exploding number of online studies. Despite not being originally targeted to academic scholars, MTurk attracted researchers because it provides a constant critical mass of individuals available to complete research studies, as well as an infrastructure that facilitates recruiting and compensating participants.

How does an MTurk study take place? MTurk is a website where *requesters* recruit and compensate a desired number of *workers* to complete tasks, such as identifying information in pictures, transcribing audio files, or completing surveys. Tasks typically last minutes rather than hours or days, and payments range from a few cents to a few dollars, depending on the effort and time required. Requesters post tasks and determine the subpopulation of workers who are qualified to complete them, based on information provided by MTurk (e.g., ratio of approved/submitted tasks, country of residence) or previously collected by the requester (e.g., age, gender). Workers are free to choose and complete any available task for which they are eligible. After a worker completes a task, the requester decides whether to approve the submission (and compensate the worker) or reject it (e.g., because the worker did not comply with the request). Thus, researchers use MTurk by operating it as requesters, and they recruit and compensate workers for participating in online surveys that are hosted on external websites (e.g., Qualtrics). Researchers typically provide participants with a unique alphanumerical code at the end of the study that is entered into MTurk to verify completion.

Who are the workers and why do they complete MTurk tasks? The composition of MTurk workers has fluctuated over time, along with Amazon's openness to international workers. The majority of workers reside in the US, and most researchers restrict participation to US residents in order to increase homogeneity. In the largest recent demographic survey of MTurk (nearly 10,000 workers), Casey and colleagues (2016) found results consistent with

previous similar investigations (e.g., Berinsky, Huber, and Lenz 2012; Paolacci, Chandler, and Ipeirotis 2010; Ross et al. 2010). The average age of US MTurk workers is about 33.5, and males and females are approximately equally represented. Participants are moderately liberal, and more than 80% are White. About 60% of workers are in stable relationship, and 35% are married. About 10% identify as lesbian, gay, or bisexual—slightly higher than the 7% in the general population that identify as LGB for those 18-35, and much higher than the 3.5% national average (Jones and Cox 2015).

The fact that payments on MTurk tend to be meager in absolute value has historically led many to believe that the MTurk workforce was uneducated and unemployed. Contrary to these speculations, MTurk workers are rather educated and diverse in terms of occupations. Casey and colleagues (2016) found that about half workers are employed full-time in a different job, and less than 10% report being unemployed. About 90% of workers have some university experience, and at least half workers have a college degree. The modal income is between $30,000 and $50,000.

Because MTurk workers are not disproportionately unemployed or uneducated, some find it surprising that they agree to work for nominal amounts of money (Pham 2013). However, many short tasks each paying a small amount can result in substantive earnings for the actual time spent working. Moreover, MTurk is not a perfect substitute to other jobs, as some workers participate in studies during breaks from other work activities (Chandler, Mueller, and Paolacci 2014). Finally, many workers participate in MTurk for reasons additional to earning money, and are often both intrinsically and extrinsically motivated (Chandler and Kapelner 2013; Horton, Rand, and Zeckhauser 2011; Paolacci et al. 2010). This may contribute to wages that are often

below market rates. As we will elaborate later, however, it does not suggest that researchers should compensate less than a fair wage.

THE ADVANTAGES OF CROWDSOURCING

Crowdsourcing has become a dominant data collection technique because it offers several advantages for survey and experimental research. In this section, we discuss the characteristics of crowdsourcing that make it an attractive strategy for data collection. Importantly, crowdsourcing is not only about making it easier, faster, and cheaper to conduct computer-based studies, but it also has the potential to improve how consumer research develops as a field.

Reduced Costs

Crowdsourcing makes many studies cheaper to conduct on many dimensions—from lower participant payments to lower administrative costs. Though norms differ between institutions, participants in a physical lab are usually compensated no less than $5 to cover 30 minutes spent in the lab and the fixed costs of commuting. Crowdsourcing removes commuting costs, and allows compensating participants for the precise time they will spend in the study. As a result, controlling for pay rate, short studies are bound to be less expensive to conduct with crowdsourced samples than in the physical lab. For example, with less than $200 a researcher can conduct a 5-minute study with 200 participants that are compensated more than the US federal minimum wage ($7.25 per hour). Compared to traditional lab studies, MTurk also removes the costs of marketing and recruiting participants, coordinating study times, paying

assistants to administer the lab study, and processing personal financial information for tax purposes.

The reduced cost of crowdsourcing research has obvious budget advantages for researchers and institutions, but perhaps most importantly it can translate into scientific opportunities. First, the convenience of crowdsourcing allows conducting a *larger number of exploratory studies*. The possibility to collect larger amounts of data can help researchers discovering theoretically relevant patterns and refining research hypothesis, accelerating the scientific process.

Second, less expensive data collection allows researchers to conduct *more informative confirmatory studies*. Recent research has warned behavioral researchers of the perils of underpowered investigations, and urged them to use larger samples to increase the validity of hypothesis testing (Simmons, Nelson, and Simonsohn 2011). MTurk allows obtaining sample sizes (and statistical power, all else being equal) that would be prohibitively expensive or even impossible to obtain with traditional participant pools that are available at smaller universities with limited resources. It also provides the necessary power to test non-linear relationships by allowing researchers to manipulate more than two levels of an independent variables (Goldstein 2016). Crowdsourcing can thus improve the information value of consumer research by allowing larger samples, even when participants are paid at the same rate as participants in a physical lab. The same logic applies to replication studies, which may require particularly large samples (Simonsohn 2015).

Finally, whereas crowdsourcing makes data collection easier for many consumer researchers, most strikingly *it makes research possible* for many others. Many academics have scarce access to physical laboratories where they can conduct their empirical investigations.

Platforms such as MTurk open up opportunities for them to conduct valid investigations, democratizing the production of academic results. As advocated by Jonathan Haidt and others (e.g., Duarte et al. 2015), diversity in researchers' viewpoints benefits the social sciences, and inclusive access to study participants is key for this to happen in consumer research.

Participant Diversity

The average characteristics of the MTurk workforce hide the large diversity of the population. For instance, whereas the average MTurk worker is around 33.5 years old and about half of MTurk workers are below 30 years old, older adults are also well represented (Weinberg, Freese, and McElhattan 2014). Despite the belief that working on MTurk implies being very poor, about a quarter of the workers have a household income greater than $75,000 (Casey et al. 2016). Ninety percent of workers have some college experience, and about half of the workers have a four-year college degree or a postgraduate degree—education levels that are obviously not well represented within student pools. Researchers can exploit this participant diversity because crowdsourcing websites allow researchers to target specific subpopulations. On MTurk, researchers can track any measured characteristic of previously recruited workers, potentially building sophisticated panels of participants. These characteristics can then be used as filters for recruitment, allowing researchers to target and recruit specific samples.

Both participant diversity and the ability to recruit specific participants on MTurk facilitate the use of *theory-driven samples*—samples with specific characteristics that are relevant for the situation under study. Compared to using a student sample, conducting a study with the actual population of interest (e.g., people who own a certain product in a product

disposal hypothetical scenario) can increase the external validity of the study (Ferber 1977; Gneezy and Imas 2016; Rapp and Hill 2015), help identify key moderating variables to advance theory (e.g., involvement; Petty and Cacioppo 1979; Calder, Phillips, and Tybout 1982; Lynch 1982), and allow experimental procedures to be tied to the population's specific experiences (e.g., Taylor, Lichtman, & Wood, 1984). The history of *JCR* reveals that during its first decade (1974-1984), student samples and theory-driven samples were equally common in the field. After 40 years, the ratio between student and theory-driven samples is now almost seven-to-one (Rapp and Hill 2015). Though many reasons can explain the decline of theory-driven samples, an increased preference for convenience and lower cost likely played a significant role (as predicted by Ferber 1977).

Consumer researchers are starting to use MTurk participant diversity to recruit theory-driven samples. For example, Connell, Brucks, and Nielsen (2014) studied the effects of childhood exposure to advertisements on product evaluations as adults, and recruited MTurk participants within age ranges compatible with exposure to their advertising stimuli during childhood. Hamerman and Johar (2013) used MTurk to recruit right-handed participants in order to manipulate illusions of control when using one's right hand versus left hand. Others have recruited participants who were married to study self-control in joint-decisions (Dzhogleva and Lamberton 2014) and emotional connections with special life events (Goodman, Malkoc, and Stephenson 2016). Other examples include studies on participants who believed in God (Fergus and Rowatt 2015), were unemployed (Konstam 2014), or had specific psychopathological symptoms (for a review see Chandler and Shapiro 2016).

In addition to allowing researchers to collect samples that closely represent their target populations, crowdsourcing allows the scientific community to grow less dependent on

idiosyncratic samples (e.g., undergraduates at top American universities). MTurk workers include adults of many ages, professions, education levels, and so on. Prolific, a UK-based crowdsourcing research website, provides participants coming from many countries, and more opportunities will certainly follow to move consumer research beyond participants who some think of as ultimately WEIRD (Western, Educated, Industrialized, Rich, and Democratic; Heinrich, Heine, and Norenzayan 2010). This reliance on white, educated, and in particular college students, substantially depended on the additional costs involved in reaching out to other samples. In sum, by reducing such costs, crowdsourcing can improve the ability of consumer researchers to both qualify and generalize their findings.

Importantly, participant diversity is not inherently positive. For theory testing, the heterogeneity of a sample can add unmeasured background factors that might interact with the treatment (Lynch 1982), increasing noise and the rate of false negatives. However, if researchers identify these potential moderators, then they can leverage this diversity to increase both internal and external validity (Lynch 1982). In sum, the diversity of MTurk participants provides new opportunities for researchers, but researchers should be aware of the perils of sampling from a more heterogeneous population than students.

Flexibility

A virtual laboratory is generally thought of as *less* flexible than a physical one. After all, certain studies simply cannot be conducted online, such as those requiring controlled interactions between participants and physical stimuli. Studies that only require a computer to be executed, however, can largely benefit from the flexibility of crowdsourcing. Whereas offline studies are

constrained by the availability of campus participants, research assistants, and laboratory space, crowdsourced studies are not. The constant availability of a critical mass of participants allows researchers to conduct studies with no delay, and to conclude them at unprecedented speed. From the moment a study is ready to be conducted, it can take hours rather than weeks for data to be collected. Thus, the flexibility of crowdsourcing can accelerate the scientific process. Yet, there are many other ways that crowdsourcing increases researcher flexibility, which we discuss next.

Longitudinal Studies. The flexibility of crowdsourcing can also be leveraged to conduct longitudinal studies, which are logistically more complicated than one-shot surveys with a homogeneous population. While retention rates in longitudinal studies will vary depending on payment, tasks, and intervening time, retention rates on MTurk have been reported to be around 70% between waves conducted days, weeks, and even months apart (e.g., Chandler et al. 2015; Reese and Veilleux 2015). Over longer periods of time, retention drops but longitudinal research remains viable. For instance, Chandler and colleagues (2014) found a 44% response rate after one year. Similarly, other forms of longitudinal studies, such as diary studies have been shown to be feasible (e.g., Boynton and Richman 2014), in part because MTurk provides a way to easily contact, motivate, and compensate participants via bonus payments for completing each part of the study.

Cross-Cultural Research. Crowdsourcing websites also allow researchers to conduct cross-cultural research (e.g., Eriksson and Simpson 2010). While Amazon's acceptance of non-US workers has been fluctuating over time, Prolific provides participant populations from multiple countries, and more opportunities will likely follow. Since crowdsourcing allows researchers to reach different samples and control for how these samples are reached, researchers may make particularly valid inferences as they make cross-cultural comparisons, provided that

language barriers do not impair measurement equivalence across samples (Feitosa, Joseph, and Newman 2015).

Interactions between Participants. Studies that require real-time interaction between participants are also possible with crowdsourcing, and often more conveniently than traditional samples. Since MTurk offers researchers access to thousands of workers at any given time, there is always another person online willing to participate in a two-person game or group interaction. Open-source, web-based solutions are emerging (e.g., oTree; Chen, Schonger, and Wickens 2016) that aid researchers in programming experiments (e.g., providing highly customizable templates of standard interactive paradigms) and crowdsourcing them online (e.g., creating "waiting rooms" for queuing MTurk participants before they are matched with one another, Mason and Suri 2011). Researchers have successfully used crowdsourcing for incentivized experiments involving dozens of participants interacting at the same time, numbers that are logistically difficult to achieve in physical labs (Wang, Suri, and Watts 2012; Suri and Watts 2010) and can open new research possibilities. For instance, Watts and Dodds (2007) lamented the lack of empirical consumer research on large influencing networks, and crowdsourcing might offer a way forward.

Alternative Measures. While self-reports and hypothetical choices are the most commonly collected measures with crowdsourcing, there are additional opportunities. For instance, the ability to award bonuses allows for the use of consequential monetary choices (e.g., Dholakia et al. 2016; Goldstein 2016), as well as incentivized games (e.g., Yang and Urminsky 2015). Researchers have also used MTurk as a setting to conduct field experiments, measuring how work decisions depend on features of the crowdsourced tasks (e.g., Chandler and Kapelner 2013). Importantly, the technology available for Internet research has increased over the last few

years, which may also open new possibilities. Response times can be measured reliably (Crump, McDonnell, and Gureckis 2013), webcams and sophisticated software can serve as eye-trackers (e.g., Cheng et al. 2015) or to capture facial expressions for analysis in emotion software (e.g., Den Uyl and Van Kuilenburg 2005), and researchers are developing methods to collect physiological data online (e.g., heart rate, Muender, Miller, Birk, and Mandryk 2016). Of course, simply because a study *may* be crowdsourced, it does not mean it *should*, be crowdsourced. Some studies are best still conducted in a lab environment, such as studies that require direct supervision and/or special equipment/stimuli, last extended periods of time, or contain questions easily answered via a web search. We will discuss the limitations on the crowdsourcing method of data collection in the next section.

Data Quality

A common concern with Internet research is data quality. Intuitively, the impossibility to directly monitor research participants might lead to participant misbehavior of various types, ultimately resulting in low data quality. However, unlike other online populations (and participants in the lab), crowdsourcing marketplaces have incentives in place that are conducive to high data quality. When an MTurk worker submits a task, a requester can choose to reject such submission and forgo paying the worker, and/or block the worker (i.e., preventing the worker from participating in the requester's future tasks). Therefore, workers are motivated to follow instructions and pay attention to the research study (e.g., carefully consider a stimulus before answering the questions that follow), especially if they are aware that attention checks may follow (cf. Hauser and Schwarz 2015, 2016; Oppenheimer, Meyvis, and Davidenko 2009).

In addition to this short-term monetary incentive for conscientiousness, workers have a long-term incentive to avoid being blocked or even rejected. Researchers typically require participants to have a high approval rate (e.g., 95% or higher) to be eligible to participate in their tasks, implying that more rejections will make less work available to workers. In other words, poor work affects participants' immediate payoffs and future employment opportunities.

Given the incentive structure of MTurk, it is not surprising that crowdsourced data has consistently been found to be of high quality (Paolacci and Chandler 2014). Despite the fact that some MTurk workers have admitted to completing tasks while engaged in other activities (e.g., listening to music, Chandler et al. 2014), studies have consistently found that MTurk workers' attention levels are equal or greater than undergraduate and community samples (Hauser and Schwarz 2016; Paolacci et al. 2010; Ramsey, Thompson, McKenzie, and Rosenbaum 2016). One study (Goodman et al. 2013, Study 2) found lower levels of passing an instructional manipulation check (Oppenheimer et al. 2009), but the effect may have been explained by language proficiency. This is consistent with research documenting lower data quality among Indian workers, the second largest population of MTurk (Litman, Robinson, and Rosenzweig 2015). For US participants, research suggests that high-reputation MTurk workers (i.e., those with above 95% approval ratings) produce high quality data without the need to filter based on attention-check questions (Peer, Vosgerau, and Acquisti 2014). In sum, the evidence suggests crowdsourced participants are at least as attentive as lab participants.

MTurk workers have also been shown to be similar in reliability to student and public samples, providing psychometrically sound responses (Buhrmester, Kwang, and Gosling 2011; Holden, Dennie, and Hicks 2013), and they are just as honest, consistent, and conscientious as traditional samples (e.g., Rand 2012; Shapiro, Chandler, and Mueller 2013). Further, they show

the same decision-making heuristics and biases (e.g., present bias, loss aversion, certainty effect) as student and public samples, all with similar effect sizes (Berinsky et al. 2012; Goodman et al. 2013; Paolacci et al. 2010). Cognitive paradigms also consistently replicate (Crump et al. 2013). In sum, there is no evidence that the efficiency gains of crowdsourcing come at the expense of data quality.

ISSUES WITH CROWDSOURCING IN CONSUMER RESEARCH

While crowdsourcing provides efficiencies in data collection with no evidence of a reduction in data quality, crowdsourced samples have unique characteristics that, when unaccounted for, could threaten research validity. Next we address the methodological issues and concerns associated with crowdsourcing.

Representativeness

Because of their diversity, crowdsourced populations are obviously more demographically representative of the general population than students (Paolacci, Chandler, and Ipeirotis 2010); however, this does not mean that they should be treated as a representative, and researchers should be aware of the idiosyncratic characteristics of crowdsourced populations that might moderate treatment effects when developing theory (Lynch 1982).

MTurk workers are different from the general US population (and traditional student samples) in several ways. Compared to the general population, MTurk workers are slightly younger, better educated, less likely to be married, more likely LGBTQ, have slightly lower

income, and are more likely to live with parents (Buhrmester et al. 2011; Casey et al. 2016; Corrigan, Bink, Fokuo, and Schmidt 2015; Paolacci and Chandler 2014; Shapiro et al. 2013; Weinberg et al. 2014). They also score higher on need for cognition (NFC) and civics questions (Berinsky et al. 2012). Workers have been shown to have small but systematic personality differences that typically align with the characteristics of general Internet users, as one might expect from people that enjoy doing solitary tasks on the Internet. For instance, they are slightly more introverted and show higher levels of social anxiety (Goodman et al. 2013), and express slightly lower self-esteem and greater incidence of depression and emotional regulation (Arditte, Çek, Shaw, and Timpano 2015; Shapiro et al. 2013). Table 1 provides a summary of the differences between MTurk workers and the general population found in the literature.

For probability sampling, these results suggest that researchers should not indiscriminately survey MTurk workers to estimate general levels of a target variable. However, researchers can build panels representative of their target populations. Moreover, techniques such as raking and model-based poststratification can be used to statistically adjust the estimates obtained from non-representative samples (Battaglia, Hoaglin, and Frankel 2013, Park, Gelman, and Bafumi 2004), and may be applied to MTurk samples (Goel, Obeng, and Rothschild 2015).

------ PLACE TABLE 1 ABOUT HERE ------

Self-selection

Self-selection is also a potential issue with crowdsourced data. There are several layers of self-selection that a worker has gone through before becoming a participant in a research study.

Workers self-selected into using the Internet and into using MTurk, which leads to observable differences in sample compositions compared to other samples. Critically, however, there is self-selection at the study level: Participants are free to select the tasks they participate in and the ones they eventually complete.

Some tasks will be generally more attractive to complete for workers. Higher pay rates affect the attractiveness of a task (e.g., Buhrmester et al. 2011; Mason and Watts 2010), and may affect the attractiveness of further tasks posted by the same researcher via reputation effects (e.g., Higgins, McGrath, and Moretto 2010). Because tasks are by default sorted by recency, more recently posted tasks are more likely to be selected (Chilton, Horton, Miller, and Azenkot 2010), and there is evidence that paying in multiples of 5 cents increases task attractiveness (Horton and Chilton 2010). The fact that some tasks are more attractive for *every* worker might surprisingly affect sample composition: If a task receives publicity on worker forums that are not representative of the MTurk workforce (e.g., as a task "worth turking for" on Reddit), this may translate into biased samples (e.g., because these forums are more likely to attract prolific MTurk users and are disproportionately populated by males, Chandler, Mueller, and Paolacci 2014).

An advantage of crowdsourcing is that researchers can post tasks at any time; however, the day of the week or the hour in the day in which a study is posted can affect sample composition. Casey et al. (2016) conducted a large MTurk study on intertemporal demographic differences among US residents, and found effects of posting times that go above and beyond attracting people from different time zones. Most interestingly, they found that completing the survey in the night (vs. morning) was associated with higher likelihood of being single, using a smartphone to complete the survey, and being a less prolific MTurk worker. They also found that

"early" participants who complete the first observations in a study tend to be older and male, and report higher levels of emotional stability, conscientiousness, and agreeability.

Self-selection is also based on the characteristics of the crowdsourced study. Certain studies (e.g., those dealing with a certain topic, or perceived as cognitively demanding) might be more attractive to certain people (e.g., with an interest in the topic, or with higher need for cognition). As a result, the starting samples might be biased in theoretically meaningful ways. This problem is exacerbated by *previewing*—when workers inspect the survey before deciding whether to complete it. Moreover, dropping out in the middle of a study is bound to be more common online than in a physical lab, due to the lower material investment in participation and visibility. This affects the final sample in a study, which consists of people who decided to enroll and not quit during the study. Attrition (i.e., low completion rates) is always problematic for external validity, as findings might not generalize to people who (would) decide to quit a study. But attrition is particularly troublesome when it differs systematically by condition in a between subject design. If participants are more likely to quit in one condition (e.g., a condition that first requires a long essay about feeling powerless), and quitting correlates with theoretically relevant characteristics (e.g., low need for cognition, or low self-esteem), then assumptions of random assignment will fail, with serious threats to internal validity (Chandler and Shapiro 2016, Horton et al. 2011; Zhou and Fishbach 2016). In the next section we will discuss strategies to mitigate this problem.

Perhaps the most dangerous threat posed by self-selection concerns the studies of specific subpopulations (e.g., racial minorities or owners of a certain product) that recruit participants based on self-reported eligibility (e.g., "Only participate if you own [product x]"). By crossing data provided by participants across studies, Chandler and Paolacci (2016) found that a

substantial amount of respondents in such studies might in fact be imposters. This is the result of a small, though nonnegligible, amount of workers who misrepresent their relevant characteristics (especially when the payment is high) and the fact that ineligible respondents in a study are a function not only of the prevalence of liars, but also of the rarity of the target subpopulation. In other words, even if the proportion of MTurk workers who will lie to get access to a study is small, researchers who blatantly recruit members of rare populations (e.g., owners of a Gucci hand bag) may still find themselves with a substantial number of ineligible responses that are hard to detect. This threatens the validity of a study, particularly because eligible and ineligible participants may answer in systematically different ways (Siegel, Navarro, and Thomson 2015). We discuss how to properly screen participants in the next section.

Participant Nonnaiveté

The MTurk population is large but not infinite, and researchers are not sampling from all the registered users. Rather, at any point in time there might be a few tens of thousands workers available, and in any quarter the average laboratory may be sampling from a population of less than 10,000 (Fort, Adda, and Cohen 2011, Stewart et al. 2015). Because researchers crowdsource thousands of tasks every day, MTurk workers may have become accustomed to participating in social science studies. Compounding this possibility, researchers are not sampling uniformly across the population. Chandler et al. (2014) found that the 10% most productive workers were responsible for 41% of the observations of a sample of behavioral research studies. Classic paradigms in psychology are widely known, especially among prolific workers (Chandler et al. 2014; Thomson and Oppenheimer 2016). Many worry that participants might become nonnaive

via exchange of information on MTurk worker forums. However, whereas vivid anecdotes exist (e.g., MTurk workers ironizing researchers' overuse of certain experimental procedures), crosstalk about critical content of a novel study may be practically negligible (Chandler et al. 2014) and is discouraged if not prohibited by worker forum managers. In sum, because crowdsourced pools are shared by a huge number of researchers, MTurk samples might often contain many "professional survey-takers," who are experienced with research participation and might be knowledgeable about specific studies.

There is evidence that participant nonnaiveté affects the validity of research instruments relevant to consumer research. For example, performance on the Cognitive Reflection Test (Frederick 2005), a commonly employed measure of people's tendency to resist intuitive responses (e.g., Simonson and Sela 2011), depends on how often people may have seen the test and has become a confounded measure of reflexivity on MTurk (Chandler et al. 2014; Thomson and Oppenheimer 2016). Rand and colleagues (2014) conducted a series of studies to test whether people have an intuitive preference for cooperation in interpersonal dilemmas, and attributed the declining size of the effect over time to workers' increased experience. Similarly, there are suggestions of effects that might not be replicable with experienced research participants (Connors et al. 2016, DeVoe and House 2016). In an investigation of the effects of study-specific nonnaiveté, Chandler and colleagues (2015) found that completing a two-condition experiment a second time resulted in smaller effect sizes, particularly when the time elapsed between participations was small and when participants were assigned to different conditions. Whereas these results together seem to suggest that nonnaiveté might generally reduce the likelihood of observing true effects in the data, more research should be conducted on the effects of general and study-specific participant nonnaiveté.

Can MTurk Dictate Research Programs?

There are also philosophical concerns with crowdsourcing, which ironically stem from its very advantages. Some worry that the lure of MTurk may lead researchers to develop a preference for hypotheses and designs that are "crowdsourceable" instead of hypotheses and designs that are theoretically or substantively interesting (Pham 2013). Studies that are important but more difficult to conduct, by this account, would become less likely to be conducted. We empathize with this concern, though it is not something specific to crowdsourcing. Some studies are inevitably easier to conduct than others, even when conducted in a physical lab. If the state of consumer research had been negatively affected by a disproportionate preference for convenient procedures, this would predate online samples, and similar worries of convenience have been expressed before the advent of crowdsourcing (Baumeister, Vohs, and Funder 2007; Ferber 1977). If anything, the possibility to crowdsource the studies that *can* be crowdsourced (typically those that can be executed via a computer) should free up resources in laboratories (e.g., participants, lab space, time) that can be dedicated to studies that strictly require the physical presence of participants (e.g., experiments that require touching or tasting products, or physical interactions with others). Moreover, the costs or the cumbersomeness of a study should not be treated as indicators of quality or validity. *All else being equal*, a study that is easier and less costly to conduct, especially when publicly funded, should be preferable.

Some also worry that as the opportunity cost of studies decreases, people might become less mindful in their designs and procedures, conducting more studies than they would otherwise do and with less than optimal designs. This is arguably a questionable research practice, and it

certainly is questionable when people persist on conducting studies deliberately to capitalize on the chance that they will obtain "publishable" results. This, however, is a concern with the researcher's integrity and rigor when planning and reporting studies that is independent of the tools employed to conduct such studies, and issues such as selective reporting, file drawer, and preregistration have been recently receiving the attention they deserve (e.g., Moore 2016; Simonsohn, Nelson, and Simmons 2014; van't Veer and Giner-Sorolla 2016; Wagenmakers et al. 2012). On the contrary, there is nothing inherently wrong with serendipitous explorations that are followed by appropriately powered (and ideally preregistered) confirmatory research (Alba 2012; Lynch et al. 2012; Sakaluk 2016; Wagenmakers et al. 2012). Ultimately, it is the researcher's responsibility to leverage the efficiency of crowdsourcing to conduct valid investigations and future research should continue to address these issues.

CROWDSOURCING CONSUMER RESEARCH: GUIDELINES

Given the issues with crowdsourcing consumer research, we next propose several guidelines for researchers to minimize these concerns and maximally enjoy the benefits of MTurk and other crowdsourcing sites.

Minimize the risks of self-selection. As researchers recruit participants in crowdsourcing marketplaces, they should minimize the risks connected to self-selection. Specifically, we encourage researchers to describe tasks generically, making sure that participants' expectations are aligned with the nature of the study without revealing details that would make the study more or less attractive to different participants with different dispositions or characteristics. To

maximize quality, researchers should make full use of quality filters (e.g., on MTurk, recruiting workers with approval ratings superior to 95%, Peer et al. 2014).

Although more sophisticated platforms (e.g., Prolific) allow the selective recruiting of participants with certain characteristics (e.g., demographics), these screeners may not be sufficient for very specific samples (e.g., people with extreme attitudes towards a brand). In these cases, researchers need to collect the relevant information (e.g., attitudes towards a brand) from MTurk workers, and then recruit only the participants who belong to the target subpopulation. To avoid recruiting participants who misrepresent themselves in order to participate, prescreening surveys should always conceal the required characteristic (e.g., asking about attitudes towards a brand without disclosing that only people with extreme attitudes will later have the possibility to participate in a study). Importantly, *any* survey is a screening survey, to the extent that it records information (including mere participation in a study) that might be subsequently used as a recruitment filter. Associating a participant's response with the participant's Mturk WorkerID allows researchers to build their own panel of participants (e.g., Litman, Robinson, and Abberbock 2016, Peer et al. 2012). Though the WorkerID is simply an alphanumerical string, it does have the potential to reveal personally identifying information (Lease et al. 2013); thus, researchers should treat WorkerIDs as confidential information.

Avoid attrition. To minimize non-selective and selective attrition, researchers should increase participants' initial investment in the study. Specifically, we suggest researchers require participants to formally enroll in a study (i.e., "accept the HIT" in MTurk jargon) before accessing the study. Requiring enrollment prevents *previewing* of a study and raises the time costs required by participants to return the task to MTurk (Litman et al. 2016, Peer et al. 2012). This strategy, combined with study descriptions on MTurk that are generally vague, also ensures

that a study's content does not affect a worker's choice as to whether to participate or not. Similarly, increasing the effort demanded *before* the experimental manipulation (and increasing the payment accordingly) will decrease the attractiveness of quitting the study mid-way (Horton, Rand, and Zeckhauser 2011).

Nonetheless, these strategies that can be used to prevent attrition may not remove it entirely. For this reason, researchers should measure attrition ex post, and particularly whether attrition was different between experimental conditions. On Qualtrics, surveys should be "closed" before downloading the data, to make sure that partial responses are recorded and any imbalance in the number (and characteristics) of participants between conditions is detected and reported (see Zhou and Fishbach 2016). Tools such as *TurkPrime* report *bounce rates* (the percentage of participants that previewed the HIT's description but did not accept the HIT) and *completion rates* (the percentage of participants that accepted the HIT and completed it) by default.

Manage the pool. The efficiency of the research experience on MTurk makes it easy to underestimate the importance of adequately performing some administrative actions connected to the research. In physical laboratories, participants receive prompt responses by researchers and/or lab managers to their inquiries, and are rarely expelled from the pool. When conducting studies online, we propose that researchers maintain the same behavior: respond promptly to worker questions and issues and be cautious when blocking any MTurk worker, which can put the worker's account at risk of removal. To get a perspective on MTurk workers, we suggest requesters monitor the forums used by workers (e.g., Mturkgrind, Turkernation, mTurklist, reddit.com/r/mturk, and Turkopticon), which can contain information about requesters'

reputations and ultimately inform researchers about the paradigms that might be less promising to conduct on MTurk.

Manage nonnaive participants. Previous exposure to studies or stimuli may sometimes affect the validity or the power of research paradigms. In such cases, researchers should ensure that they do not recruit participants who participated in related studies. This is a prebuilt filter on certain platforms (e.g., Prolific), and on MTurk it can be done by manually assigning qualifications or relying on external solutions such as *TurkPrime* (Litman et al. 2016). When WorkerIDs are linked to study responses, researchers can also exclude previous participants ex post from the data analysis (or test the effects of their inclusion).

Researchers should also avoid using "classic" paradigms (e.g., trolley problem, Asian disease, common manipulations of power, etc.) or attention checks (e.g., Oppenheimer et al. 2009) that participants are very likely to have encountered in the past, and strive to use novel variations (Chandler et al. 2014). Worker experience (i.e., the number of studies completed in the past) can also be used as a covariate in data analysis. Prolific allows exporting study metadata that include the total number of studies taken in the past by each participant. On MTurk, the self-reported number of academic studies completed in the past (DeVoe and House 2016, Rand et al. 2014), or the number of studies that the participant completed for the same researcher (which is visible in the downloadable "Results" file associated with the study) might be useful proxies of a participant's experience with research studies, though we are not aware of systematic investigations of their explanatory power.

Pay a fair wage. Unlike Prolific, MTurk does not impose any minimum wage, and wages are set at the requesters' discretion. The question of whether a minimum wage policy should be enforced is notoriously empirical in addition to philosophical, as this might benefit participating

workers but also reduce the work and the total payments available on the marketplace. However, there are ethical considerations to paying a fair wage, and whereas most evidence suggests that pay rates do not affect data quality in a typical consumer study (Burhmester et al. 2011; Goodman et al. 2013; Mason and Watts 2009; but see Fort et al. 2011, Litman et al. 2015), adequate payments are in the researchers' best interest too. A researcher's reputation among the participant population strongly depends on whether the researcher consistently pays fair wages. As illustrated on *Turkopticon*, a website that collects workers' reviews about requesters, paying overly low wages lowers requesters' reputation, and may decrease the attractiveness of subsequent tasks. Though we are not aware of any systematic research examining reputation effects, low reputations may plausibly affect data quality and certainly impact the credibility of the scientific community as a whole. Thus, we urge researchers to pay a fair wage and consider their reputation and the reputation of the field before they set a wage.

Diversify Samples. As the previous discussion suggests, every participant population—from student samples to MTurk—has its own idiosyncrasies. When methodologically feasible, we propose that researchers test theories across different samples to identify theoretically important moderating background factors. When results converge, this is evidence of the robustness of a result and suggests that either sample might be usable in future studies of a phenomenon. On the contrary, failures to replicate a finding using a different sample imply that important theoretical or methodological moderators might be at work. Identifying these moderators strengthens both internal and external validity (Lynch 1982). In other words, consumer researchers should not think of alternative samples (e.g., MTurk and students) as pure substitutes, but as complements to one another in theory development.

Behave ethically. Amazon provides ethically-informed Terms of Service for requesters that include critical aspects, such as not collecting identifiable information, but these terms are ultimately underspecified and not tied to the specificities of researchers. However, a group of academics and MTurk workers developed and continually update *Guidelines for Academic Requesters* (http://wiki.wearedynamo.org/index.php/Guidelines_for_Academic_Requesters) which suggest researchers (a) clearly identify themselves, (b) provide reasonable time estimates for the required work, (c) approve work as soon as possible, (d) maintain worker privacy, (e) do not block workers to avoid duplicate participants, (f) maintain a responsive line of communication, and (g) pay fairly. We urge consumer researchers to comply with these guidelines. In addition, the use of deception is more problematic compared to physical labs. Crowdsourced participants can belong to the pool for several years, and the collective problems associated with contamination (e.g., paradigms becoming less credible) can be stickier. A shared participant pool is ultimately a public good that researchers should contribute to with ethical behaviors (see Gleibs 2016 for a more thorough discussion of the ethical aspects on MTurk experimentation).

Report. Consumer research relied for decades on a relatively homogeneous participant population (e.g., students), and this might have attenuated the attention that journals dedicated to details of sampling in data collection. Given the dynamic diversity of crowdsourced samples, however, it is more important that researchers report details of their sample. In light of selection issues described in the previous section, we highly encourage researchers to include the following details of their participants (in method sections or appendix): compensation (including pay rate), country of residence, approval cutoffs (e.g., > 95%), whether and how nonnaïveté was dealt with, and basic demographics (e.g., gender, age). Especially for between subject

experiments that might impose different burdens for participants across conditions, attrition/completion rates (as a function of condition) should also be reported.

CONCLUSION

Researchers across disciplines have turned to web-based opportunities to conduct empirical research, and consumer scientists are no exception. Crowdsourcing websites like MTurk make survey and experimental investigations more efficient. When used virtuously, crowdsourcing can also contribute to improve consumer science by allowing more numerous and informative studies and increasing participant and researcher diversity. However, online research and crowdsourcing in particular is not immune to risks, which the community should not neglect. "Crowdsumers," like other participant pools for consumer research, are a public good that we should manage with the greatest care. Our guidelines allow researchers to minimize the short- and long-term threats to validity, and contribute to move the field towards valid crowdsourcing of consumer research.

TABLE 1

SIMILARITIES/DIFFERENCES OF MTURK WORKERS TO TRADITIONAL SAMPLES

Difference/ Similarity	Details	· References
Representative	· More representative of US population than college samples or other online sources · But not representative of general population · Less representative than web-based probability samples (i.e., random digit dialing) · Younger MTurkers tend to resemble young people in general more than older MTurkers resemble older people in general	· Berinsky et al. 2012 · Bohannon 2011 · Casler et al. 2013 · Huff and Tingley 2015 · Simons and Chabris 2012
Demographics	*Compared to general population, workers are*: · Younger (means range from 29-35) · Better educated · Less likely to be married · Lower income · Less home ownership, · More likely LGBTQ · More likely to live with parents · More likely to be unemployed or underemployed · More professionally diverse (though not representative of general pop.); over-representative of tech-related fields · Less Hispanic and African American representation · Similar to other samples in terms of urban vs. rural and region of residence (slightly more Northeast) · More liberal · Less religious · Slightly less likely to have biological children; more likely to have stepchildren	· Berinsky et al. 2012 · Buhrmester et al. 2011 · Casey et al. 2016 · Corrigan et al. 2015 · Huff and Tingley 2014 · Keith and Harms 2016 · Paolacci and Chandler 2014 · Shapiro et al. 2013 · Weinberg et al. 2014
Personality	· Introversion: Scored higher (vs. college and community samples) · Neuroticism: scored higher (vs. college samples and general pop.) · Agreeableness: scored lower · Self-esteem: scored lower (vs. college samples and general pop.) · Empathy: Score higher on trait measures of empathy and state measures of involvement in a story · Psychological distress and physical discomfort: Some evidence suggesting lower tolerance	· Arditte et al. 2015 · Goodman et al. 2013 · Johnson and Borden 2012 · Holden et al. 2013 · Shapiro et al. 2013 · Veilleux et al. 2014
Psychopathology	· Depression and anxiety: mixed results · Social anxiety: Scored higher (vs. college and community samples) · ADHD and OCD: no difference from general population	· Arditte et al. 2015 · Eriksson 2013 · Palmer et al. 2015 · Ruzich et al. 2015 · Shapiro et al. 2013

	· Autism spectrum disorders: may be more likely to possess traits of ASDs and had higher Autism Spectrum Quotient scores than community sample	· Veilleux et al. 2014 · Wymbs and Dawson 2015
Substance Use	· Binge drink less frequently than college students · Smoke tobacco and marijuana slightly more than US average	· Johnson et al. 2015 · Reese and Veilleux 2016 · Shapiro et al. 2013 · Veilleux et al. 2014
Attention & Involvement	For the most part, show the same levels of attention or even higher levels of attention depending on the task. Pay more attention to community samples and some college student samples. Workers have reported higher state involvement in a story presented in a study. May depend more on task, native language, and length of task (no evidence that compensation increases attention of US participants).	· Behrend et al. 2011 · Goodman et al. 2013 · Hauser and Schwarz 2016 · Johnson and Borden 2012 · Ramsey et al. 2016
Cheating & Honesty	Cheat Less: Answered fewer fake items correctly than college sample. 97% of reported location information validated with IP addresses.	· Cavanagh 2014 · Rand 2012
Disclosure	Greater comfort disclosing sensitive information than in-person interviews.	· Shapiro et al. 2013
SAT	Higher SAT: Workers self-reported SAT scores in 75th percentile.	· Cavanagh 2014
Knowledge	· Greater civics knowledge (vs. average Americans) · Greater scientific knowledge · Greater computer/internet knowledge	· Behrend et al. 2011 · Berinsky et al. 2012 · Cooper and Farid 2014
Need for Cognition (NFC) & Learning Goal Orientation	· Greater NFC: Scored higher on NFC (vs. average Americans) · Score highly on learning goal orientation	· Behrend et al. 2011 · Berinsky et al. 2012
Cognitive Reflection Test (CRT)	Mixed. Scored 1.21 on average, which is better than the community sample (.96), but worse than a student sample from a top tier private school (1.69)	· Goodman et al. 2013
Psychometrics	Several studies have found no differences or superior psychometric properties to MTurk data (vs. college and community samples)	· Behrend et al. 2011 · Buhrmester et al. 2011 · Feitosa et al. 2015 · Jahnke et al. 2015 · Johnson and Borden 2012
Validity of Data	· Workers often complete surveys in less-than-ideal environments, but little evidence of negative effect on data · Scale reliability identical or superior to other samples · Produce similar effect size estimates in standard tasks · High test-retest reliability · Score higher on malingering (11%), may reflect outdated measures · Repeated participation may lead to practice effects · Payments do not appear to affect data quality (except perhaps for Indian workers), even at low compensation rates	· Behrend et al. 2011 · Buhrmester et al. 2011 · Chandler and Shapiro 2016 · Chandler et al. 2014 · Clifford and Jerit 2014 · Holden et al. 2013 · Jahnke et al. 2015 · Johnson and Borden 2012 · Litman et al. 2015 · Shapiro et al. 2013

REFERENCES

Alba, Joseph W. (2012), "In Defense of Bumbling," *Journal of Consumer Research*, 38 (6), 981-7.

Arditte, Kimberly A., Demet Çek, Ashley M. Shaw, and Kiara R. Timpano (2015), "The Importance of Assessing Clinical Phenomena in Mechanical Turk Research*," Psychological Assessment,* 28 (6), 684-91.

Battaglia, Michael P., David C. Hoaglin, and Martin R. Frankel (2013), "Practical Considerations in Raking Survey Data." *Survey Practice*, 2 (5).

Baumeister, Roy F., Kathleen D. Vohs, and David C. Funder (2007), "Psychology as the Science of Self-reports and Finger Movements: Whatever Happened to Actual Behavior?." *Perspectives on Psychological Science*, 2 (4), 396-403.

Behrend, Tara S., David J. Sharek, Adam W. Meade, and Eric N. Wiebe (2011), "The Viability of Crowdsourcing for Survey Research," *Behavioral Research Methods*, 43 (3), 800-13.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz (2012), "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk," *Political Analysis*, 20 (3), 351-68.

Bohannon, John (2011), "Social Science for Pennies," *Science*, 334 (21), 307.

Boynton, Marcella. H., and Laura Smart Richman (2014), "An Online Daily Diary Study of Alcohol Use Using Amazon's Mechanical Turk" *Drug and Alcohol Review*, 33 (4), 456-61.

Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling (2011), "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?" *Perspectives on*

Psychological Science: A Journal of the Association for Psychological Science, 6 (1), 3-5.

Calder, Bobby J., Lynn W. Phillips, and Alice M. Tybout (1981), "Designing Research for Application," *Journal of Consumer Research*, 8 (September), 197–207.

_____ (1982), "The Concept of External Validity," *Journal of Consumer Research*, 9 (December), 240–244.

_____ (1983), "Beyond External Validity," *Journal of Consumer Research*, 10 (June), 112–114.

Casey, Logan, Jesse Chandler, Adam Levine, Andrew Proctor, and Dara Strolovitch (2016), "Demographic Characteristics of a Large Sample of US Workers," *Working Paper.*

Casler, Krista, Lydia Bickel, and Elizabeth Hackett (2013), "Separate but Equal? A Comparison of Participants and Data Gathered via Amazon's MTurk, Social Media, and Face-to-Face Behavioral Testing," *Computers in Human Behavior*, 29 (6), 2156-60.

Cavanagh, Thomas M. (2014), "Cheating on Online Assessment Tests: Prevalence and Impact on Validity," PhD Thesis, Colorado State University.

Chandler, Dana and Adam Kapelner (2013), "Breaking Monotony with Meaning: Motivation in Crowdsourcing Markets," *Journal of Economic Behavior & Organization*, 90, 90123-33.

Chandler, Jesse, Pam Mueller, and Gabriele Paolacci (2014), "Nonnaïveté Among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers," *Behavior Research Methods*, 46 (1), 112-30.

_____, Gabriele Paolacci, Pam Mueller, Eyal Peer and Kate A. Ratliff (2015), "Using Nonnaïve Participants Can Reduce Effect Sizes," *Psychological Science*, 26 (7), 1131-9.

_____ and Danielle Shapiro (2016), "Conducting Clinical Research Using Crowdsourced Convenience Samples," *Annual Review of Clinical Psychology*, 12, 53-81.

Chen, Daniel L., Martin Schonger, and Chris Wickens (2016), "oTree—An Open-source Platform for Laboratory, Online, and Field Experiments," *Journal of Behavioral and Experimental Finance*, 9, 88-97.

Cheng, Shiwei, Zhiqiang Sun, Xiaojuan Ma, Jodi L. Forlizzi, Scott E. Hudson, and Anind Dey (2015), "Social Eye Tracking: Gaze Recall with Online Crowds," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, New York: ACM, 454-463.

Chilton, Lydia B., John J. Horton, Robert C. Miller, and Shiri Azenkot (2010), "Task Search in a Human Computation Market," *Proceedings of the ACM SIGKDD Workshop on Human Computation*, New York: ACM, 1-9.

Clifford, Scott and Jennifer Jerit (2014), "Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory and Online Studies," *Journal of Experimental Political Science*, 1 (2), 120-31.

Connell, Paul M., Merrie Brucks, and Jesper H. Nielsen (2014), "How childhood advertising exposure can create biased product evaluations that persist into adulthood," *Journal of Consumer Research*, 41 (1), 119-34.

Connors, Scott, Mansur Khamitov, Sarah Moroz, Lorne Campbell, and Claire Henderson (2016), "Time, Money, and Happiness: Does Putting a Price on Time Affect Our Ability to Smell the Roses?." *Journal of Experimental Social Psychology*, 67, 60-4.

Cooper, Emily A. and Hany Farid (2014), "Does the Sun Revolve Around the Earth? A Comparison Between the General Public and Online Survey Respondents in Basic Scientific Knowledge, *Public Understanding of Science*, 25 (2), 146-53.

Corrigan, Patrick W., Andrea B. Bink, J. Konadu Fokuo, and Annie Schmidt (2015), "The Public Stigma of Mental Illness Means a Difference Between You and Me," *Psychiatry Res.* 226 (1), 186–91

Crump, Matthew JC, John V. McDonnell, and Todd M. Gureckis (2013), "Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research," *PloS One*, 8 (3).

Den Uyl, M. J., and H. Van Kuilenburg (2005), "The FaceReader: Online Facial Expression Recognition," in *Proceedings of Measuring Behavior*. 30, 589-90.

DeVoe, Sanford E., and Julian House (2016), "Replications with MTurkers who are naïve versus experienced with academic studies: A comment on Connors, Khamitov, Moroz, Campbell, and Henderson," *Journal of Experimental Social Psychology*, 67, 65-7.

Dholakia, Utpal, Leona Tam, Sunyee Yoon, and Nancy Wong (2016), "The Ant and the Grasshopper: Understanding Personal Saving Orientation of Consumers," *Journal of Consumer Research*, 43, 134-155.

Dzhogleva, Hristina, and Cait Poynor Lamberton (2016), "Should Birds of a Feather Flock Together? Understanding Self-control Decisions in Dyads," *Journal of Consumer Research*, 41 (August), 361-80.

Duarte, José L, Jarret T. Crawford, Charlotta Stern, Jonathan Haidt, Lee Jussim, and Philip E. Tetlock. (2015), "Political Diversity Will Improve Social Psychological Science" *Behavioral and Brain Sciences, 38*, e130.

Eriksson, Kimmo (2013), “Autism-Spectrum Traits Predict Humor Styles in the General Population,” *Humor,* 26 (3), 461–75.

_____, and Brent Simpson (2010), “Emotional Reactions to Losing Explain Gender Differences in Entering a Risky Lottery,” *Judgment and Decision Making*, *5*(3), 159.

Feitosa, Jennifer, Dana L. Joseph, and Daniel A. Newman (2015), “Crowdsourcing and Personality Measurement Equivalence: A Warning about Countries Whose Primary Language is not English,” *Personality and Individual Differences,* 75 (March), 47-52.

Ferber, Robert (1977), "Research by Convenience," *Journal of Consumer Research*, 4 (June), 57-58.

Fergus, Thomas A., and Wade C. Rowatt (2015), “Uncertainty, God, and Scrupulosity: Uncertainty Salience and Priming God Concepts Interact to Cause Greater Fears of Sin,” *Journal of Behavior Therapy and Experimental Psychiatry*, 46, 93-98.

Fort, Karën, Gilles Adda, and K. Bretonnel Cohen (2011), “Amazon Mechanical Turk: Gold Mine or Coal Mine?,” *Computational Linguistics*, 37(2), 413-20.

Frederick, Shane (2005), “Cognitive Reflection and Decision Making,” *Journal of Economic Perspectives,* 19 (4), 25-42.

Gleibs, Ilka H. (2016), “Are all “Research Fields” Equal? Rethinking Practice for the Use of Data from Crowdsourcing Market Places,” *Behavior Research Methods,* doi:10.3758/s13428-016-0789-y

Gneezy, Uri and Alex Imas (2016), “Lab in the Field: Measuring Preferences in the Wild,” in *Handbook of Field Experiments*, ed. Abhijit Banerjee and Esther Duflo, Elsevier.

Goel, Sharad, Adam Obeng, and David Rothschild (2015), "Non-Representative Surveys: Fast, Cheap, and Mostly Accurate," working paper, retrieved from http://researchdmr.com/FastCheapAccurate.

Goldstein, Daniel (2016), Presidential Address, Society for Judgment and Decision Making Annual Conference, Boston, MA.

Goodman, Joseph K., Cynthia Cryder, and Amar Cheema (2013), "Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples," *Journal of Behavioral Decision Making*, 26 (3), 213-24.

_____, Selin A. Malkoc, and Brittney Stephenson (2016), "Celebrate or Commemorate? A Material Purchase Advantage when Honoring Special Life Events," *Journal of the Assocation for Consumer Research*, 1 (4), 497-508.

Gosling, Samuel D., Simine Vazire, Sanjay Srivastava, and Oliver P. John (2004), "Should We Trust Web-based Studies? A Comparative Analysis of Six Preconceptions about Internet Questionnaires," *American Psychologist*, 59 (2), 93.

Gosling, Samuel D., and Winter Mason (2015), "Internet Research in Psychology," *Psychology*, 66.

Hauser, David J. and Norbert Schwarz (2015), "It's a Trap! Instructional Manipulation Checks Prompt Systematic Thinking on "Tricky" Tasks," *Sage Open*, (April-June), 1-6.

_____ (2016), "Attentive Turkers: MTurk Participants Perform Better on Online Attention Checks than Subject Pool Participants," *Behavior Research Methods,* 48 (1), 400-407.

Hamerman, Eric J. and Gita V. Johar (2013), "Conditioned Superstition: Desire for Control and Consumer Brand Preferences," *Journal of Consumer Research,* 40 (3), 428-443.

Heinrich, Joseph, Steven J. Heine, and Ara Norenzayan (2010), "The Weirdest People in the World?" *Behavioral and Brain Sciences,* 33 (2-3), 61-135.

Higgins, Chiara, Elizabeth McGrath, and Lailla Moretto (2010), "MTurk Crowdsourcing: A Viable Method for Rapid Discovery of Arabic Nicknames?," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk,* Association for Computational Linguistics, 89-92.

Holden, Christopher J., Trevor Dennie, and Adam D. Hicks (2013), "Assessing the Reliability of the M5-120 on Amazon's Mechanical Turk," *Computers in Human Behavior*, 29 (4), 1749-54.

Horton, John J., David G. Rand, and Richard J. Zeckhauser (2011), "The Online Laboratory: Conducting Experiments in a Real Labor Market," *Experimental Economics,* 14 (3), 399-425.

_____ and Lydia B. Chilton (2010), "The Labor Economics of Paid Crowdsourcing," in *11th Association for Computing Machinery Conference on Electronic Commerce*, Cambridge, MA.

Huff, Connor and Dustin Tingley (2015), "Who Are These People?" Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents," *Research & Politics,* 2 (3), 1-12.

Irani, Lilly and Six Silberman (2016), *TurkOpticon,* Univesrity of California, San Diego, https://turkopticon.ucsd.edu/.

Jahnke Sara, Roland Imhoff, and Juergen Hoyer (2015), "Stigmatization of People with Pedophilia: Two Comparative Surveys," *Archives of Sexual Behavior*, 44 (1), 21-34.

Johnson, Dan R. and Lauren A. Borden (2012), "Participants at Your Fingertips Using Amazon's Mechanical Turk to Increase Student-Faculty Collaborative Research," *Teaching of Psychology*, 39 (4), 245–5.

Johnson, Patrick S., Evan S. Herrmann, and Matthew W. Johnson (2015), "Opportunity Costs of Reward Delays and the Discounting of Hypothetical Money and Cigarettes," *Journal of the Experimental Analysis of Behavior*, 103 (1), 87-107.

Jones, Robert P. and Daniel Cox (2015), "How Race and Religitoin Shape Millenial Attitudes on Sexuality and Reproductive Health," *Public Religion Research Institute*, http://www.prri.org/wp-content/uploads/2015/03/PRRI-Millennials-Web-FINAL.pdf .

Keith, Melissa G. and Peter D. Harms (2016), "Is Mechanical Turk the Answer to Our Sampling Woes?" *Industrial and Organizational Psychology*, 9 (1), 162-7.

Kraut, Robert, Judith Olson, Mahzarin Banaji, Amy Bruckman, Jeffrey Cohen, and Mick Couper (2004), "Psychological Research Online: Report of Board of Scientific Affairs' Advisory Group on the Conduct of Research on the Internet," *American Psychologist*, 59 (2), 105.

Konstam, Varda, Sara Tomek, Selda Celen-Demirtas, and Kay Sweeney (2015), "Volunteering and Reemployment Status in Unemployed Emerging Adults a Time-worthy Investment?," *Journal of Career Assessment*, 23 (1), 152-165.

Levay, Kevin E., Jeeremy Freese, and James N. Druckman (2016), "The Demographic and Political Composition of Mechanical Turk Samples," *Sage Open*, 1-17.

Litman, Leib, Jonathan Robinson, and Cheskie Rosenzweig (2015), "The Relationship Between Motivation, Monetary Compensation, and Data Quality Among US- and India-based Workers on Mechanical Turk," *Behavior Research Methods*, 47 (2), 519-28.

_____, _____ , and Tzvi Abberbock (2016), "TurkPrime.com: A Versatile Crowdsourcing Data Acquisition Platform for the Behavioral Sciences," *Behavior Research Methods*, 1-10.

Lynch, John G., Jr. (1982), "On the External Validity of Experiments in Consumer Research," *Journal of Consumer Research*, 9 (December), 225–39.

_____ (1983), "The Role of External Validity in Theoretical Research," *Journal of Consumer Research*, 10 (June), 109–11.

_____ (1999), "Theory and External Validity," *Journal of the Academy of Marketing Science*, 27 (Summer), 367–76.

_____, Joseph W. Alba, Aradhna Krishna, Vicki Morwitz, and Zeynep Gurhan-Canli (2012), "Knowledge creation in consumer research: Multiple routes, multiple criteria." *Journal of Consumer Psychology* 22, 473-85.

Mason, Winter and Siddharth Suri (2011), "Conducting Behavioral Research on Amazon's Mechanical Turk," *Behavior Research Methods*, 44 (1), 1-23.

_____ and Duncan J. Watts (2010), "Financial Incentives and the Performance of Crowds," in *Proceedings of ACM SIGKDD Workshop on Human Computation, pp 77–85.*

Muender, Thomas, Matthew K. Miller, Max V. Birk, and Regan L. Mandryk (2016), "Extracting Heart Rate from Videos of Online Participants," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'2016).*

Oppenheimer, Daniel M., Tom Meyvis, and Nicolas Davidenko (2009), "Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power," *Journal of Experimental Social Psychology*, 45 (4), 867-72.

Palmer, Colin J. Bryan Paton, Peter G. Enticott, and Jakob Hohwy (2015), "Subtypes' in the Presentation of Autistic Traits in the General Adult Population," *Journal of Autism and Developmental Disorders,* 45 (5), 1291–301.

Paolacci, Gabriele and Jesse Chandler (2014), "Inside the Turk: Understanding Mechanical Turk as a Participant Pool," *Current Directions in Psychological Science*, 23 (3), 184-8.

_____ and Panagiotis G. Ipeirotis (2010), "Running Experiments on Amazon Mechanical Turk," *Judgment and Decision Making,* 5 (5), 411-9.

Park, David K., Andrew Gelman, and Joseph Bafumi (2004), "Bayesian Multilevel Estimation with Poststratification: State-level Estimates from National Polls," *Political Analysis*, 12 (4), 375-385.

Peer, Eyal, Gabriele Paolacci, Jesse Chandler, and Pam Mueller (2012), "Selectively Recruiting Participants from Amazon Mechanical Turk Using Qualtrics," *SSRN*, 2100631.

_____, Joachim Vosgerau, and Alessandro Acquisti (2014), "Reputation as a Sufficient Condition for Data Quality on Amazon Mechanical Turk," *Behavioral Research Methods,* 46 (4), 1023-31.

Peterson, Robert A. (2001), "On the Use of College Students in Social Science Research: Insights from a Second-Order Meta-analysis," *Journal of Consumer Research*, 28 (December), 450-61.

Petty, Richard C. and John T. Cacioppo (1979), "Issue Involvement Can Increase or Decrease Persuasion by Enhancing Message-Relevant Cognitive Responses," *Journal of Personality and Social Psychology*, 37, 1915-1926.

_____ (1996), "Addressing Disturbing and Disturbed Consumer Behavior: Is It Necessary to Change the Way We Conduct Behavioral Science?" *Journal of Marketing Research,* 33 (February), 1–8.

Pham, Michel Tuan (2013), "The Seven Sins of Consumer Psychology," *Journal of Consumer Psychology,* 23 (4), 411-23.

Ramsey, Sarah R., Kristen L. Thompson, Melissa McKenzie, Alan Rosenbaum (2016), "Psychological Research in the Internet Age: The Quality of Web-Based Data," *Computers in Human Behavior*, 58, 354-60.

Rand, David G. (2012), "The Promise of Mechanical Turk: How Online Labor Markets Can Help Theorists Run Behavioral Experiments," *Journal of Theoretical Biology,* 299, 172-9.

Rand, David G., Alexander Peysakhovich, Gordon T. Kraft-Todd, George E. Newman, Owen Wurzbacher, Martin A. Nowak, and Joshua D. Greene (2014), "Social Heuristics Shape Intuitive Cooperation," *Nature Communications,* 5 (3677).

Rapp, Justine M., and Ronald Paul Hill (2015), "Lordy, Lordy, Look Who's 40! The *Journal of Consumer Research* Reaches a Milestone," *Journal of Consumer Research*, *42*(1), 19-29.

Reese, Elizabeth D. and Jennifer C. Veilleux (2016), "Relationships Between Craving Beliefs and Abstinence Self-Efficacy Are Mediated by Smoking Motives and Moderated by Nicotine Dependence," *Nicotine & Tobacco Research*, 18 (1), 48-55.

Ross, Joel, Lilly Irani, M. Silberman, Andrew Zaldivar, and Bill Tomlinson (2010), "Who are the Crowdworkers?: Shifting Demographics in Mechanical Turk," in *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, New York: ACM, 2863-72.

Ruzich, Emily, Carrie Allison, Paula Smith, Peter Watson, Bonnie Auyeung, Howard Ring and Simon Baron-Cohen (2015), "Measuring Autistic Traits in the General Population: A Systematic Review of the Autism-Spectrum Quotient (AQ) in a Nonclinical Population Sample of 6,900 Typical Adult Males and Females," *Molecular Autism*, 6 (2), 1-12.

Sakaluk, John K. (2106), "Exploring Small, Confirming Big: An Alternative System to the New Statistics for Advancing Cumulative and Replicable Psychological Research." *Journal of Experimental Social Psychology*, 66, 47-54.

Shapiro, Danielle N., Jesse Chandler J, and Pam Mueller (2013), "Using Mechanical Turk to Study Clinical Populations," *Clinical Psychological Science*, 1, 213–20

Siegel, Jason T., Mario A. Navarro, and Andrew L. Thomson (2015), "The Impact of Overtly Listing Eligibility Requirements on MTurk: An Investigation Involving Organ Donation, Recruitment Scripts, and Feelings of Elevation," *Social Science & Medicine*, 142, 256-60.

Simmons, Joseph, Leif D. Nelson, and Uri Simonsohn (2011), "False-positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, 22 (11), 1359-66.

Simonson, Itamar, and Aner Sela (2011), "On the Heritability of Consumer Decision Making: An Exploratory Approach for Studying Genetic Effects on Judgment and Choice," *Journal of Consumer Research*, 37 (6), 951-66.

Simonsohn, Uri, (2015), "Small Telescopes: Detectability and the Evaluation of Replication Results," *Psychological Science*, 26, 559-569.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons (2014), "P-Curve and Effect Size Correcting for Publication Bias Using Only Significant Results," *Perspectives on Psychological Science*, 9 (6), 666-81.

Linda. J. and Edward G. Sargis (2006), "The Internet as Psychological Laboratory," *Annual Review of Psychology*, 57, 529-55.

Stewart, Neil, Christoph Ungemach, Adam J. L. X Harris, Daniel M. Bartels, Ben R. Newell, Gabriele Paolacci, and Jesse Chandler (2015), "The Average Laboratory Samples a Population of 7,300 Amazon Mechanical Turk Workers," *Judgment and Decision Making*, 10 (5), 479-91.

Suri, Siddharth, and Duncan J. Watts (2011), "Cooperation and Contagion in Web-based, Networked Public Goods Experiments," *PLoS One, 6*(3), e16836.

Taylor, Shelley E., Rosemary R. Lichtman, and Joanne V. Wood (1984), "Attributions, Beliefs About Control, and Adjustment to Breast Cancer." *Journal of Personality and Social Psychology*, 46 (3), 489-502.

Thomson, Keela S. and Daniel M. Oppenheimer (2016), "Investigating an Alternate Form of the Cognitive Reflection Test," *Judgment and Decision Making*, 11 (1), 99-113.

Van 't Veer, Anna Elizabeth and Roger Giner-Sorolla, (2016), "Pre-Registration in Social Psychology—A Discussion and Suggested Template," *Journal of Experimental social Psychoogy*, 67, 2-12.

Veilleux, Jennifer C., Kayla D. Skinner, Elizabeth D. Reese, and Jennifer A. Shaver (2014), "Negative Affect Intensity Influences Drinking to Cope Through Facets of Emotion Dysregulation," *Personality and Individual Differences,* 56 (March), 96–101.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas, and Rogier A. Kievit (2012), "An Agenda for Purely Confirmatory Research," *Perspectives on Psychological Science,* 7 (6), 632-8.

Wang, Jing, Siddharth Suri, and Duncan J. Watts (2012), "Cooperation and Assortativity with Dynamic Partner Updating," *Proceedings of the National Academy of Sciences*, 109 (36), 14363-8.

Watts, Duncan J., and Peter Sheridan Dodds (2007), "Influentials, Networks, and Public Opinion Formation," *Journal of Consumer Research*, *34*(4), 441-58.

Weinberg, Jill, Jeremy Freese, and David McElhattan (2014), "Comparing Data Characteristics and Results of an Online Factorial Survey Between a Population-Based and a Crowdsource-Recruited Sample," *Sociological Science*, 1 (August), 292–310

Wells, William (1993), "Discovery-Oriented Consumer Research," *Journal of Consumer Research*, 19 (March), 489–504.

Wymbs, Brian T. and Anne E. Dawson (2015), "Screening Amazon's Mechanical Turk for Adults with ADHD," *Journal of Attention Disorders*, 1-10.

Yang, Adelle X. and Oleg Urminsky (2015), "The Foresight Effect: Local Optimism Motivates Consistency and Local Pessimism Motivates Variety," *Journal of Consumer Research*, 42 (3), 361-77.

Zhou, Haotian and Ayelet Fishbach (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.