Conducting Clinical Research Using Crowdsourced Convenience Samples

Jesse Chandler[1,2] and Danielle Shapiro[3]

[1]Mathematica Policy Research, [2]Institute for Social Research, [3]Department of Physical Medicine and Rehabilitation, University of Michigan, Ann Arbor, Michigan 48103; email: jjchandl@umich.edu

Annu. Rev. Clin. Psychol. 2016. 12:53–81

First published online as a Review in Advance on January 11, 2016

The *Annual Review of Clinical Psychology* is online at clinpsy.annualreviews.org

This article's doi: 10.1146/annurev-clinpsy-021815-093623

Keywords

Amazon Mechanical Turk, Internet research methods

Abstract

Crowdsourcing has had a dramatic impact on the speed and scale at which scientific research can be conducted. Clinical scientists have particularly benefited from readily available research study participants and streamlined recruiting and payment systems afforded by Amazon Mechanical Turk (MTurk), a popular labor market for crowdsourcing workers. MTurk has been used in this capacity for more than five years. The popularity and novelty of the platform have spurred numerous methodological investigations, making it the most studied nonprobability sample available to researchers. This article summarizes what is known about MTurk sample composition and data quality with an emphasis on findings relevant to clinical psychological research. It then addresses methodological issues with using MTurk—many of which are common to other nonprobability samples but unfamiliar to clinical science researchers—and suggests concrete steps to avoid these issues or minimize their impact.

Contents

INTRODUCTION

Crowdsourcing: the distribution of tasks to large groups of individuals via a flexible, open call

Requester: a person who pays workers to complete a task on MTurk

Human intelligence task (HIT): a unit of work that an MTurk worker completes for a requester

Crowdsourcing is the distribution of tasks to large groups of individuals via a flexible, open call. Crowdsourcing has created numerous opportunities to advance science through the efficient allocation of labor to generate, collect, clean, and transform data (for overviews, see Lintott & Reed 2013, Ranard et al. 2014). Interest in crowdsourcing has led to the development of online labor markets, such as Amazon Mechanical Turk (MTurk), that are designed to match people (requesters) requesting the completion of small tasks [referred to here as human intelligence tasks (HITs)] with people willing to do them (workers). The greatest impact of crowdsourcing on social science and clinical research has been the use of these labor markets as a means of recruiting convenience samples.

MTurk is currently the dominant crowdsourcing market used by academic researchers, although a number of other platforms share similar functionality. MTurk is named after a nineteenth-century hoax automaton, the Mechanical Turk, that was purportedly able to play chess. In actuality, the Mechanical Turk contained a human being that directed the Mechanical

Turk's movements. MTurk is intended to provide workers that occupy a similar role for information technology companies, providing "artificial artificial intelligence" that can complete tasks that are difficult to handle through machine computation alone. Examples of commercial applications of crowdsourcing include identifying duplicate products for Amazon.com, determining people's employers based on free text responses for LinkedIn, and conducting near-real-time analysis of sentiment for Twitter.

MTurk has been used widely by academics: A Google Scholar search suggests that approximately 15,000 papers containing the phrase "Mechanical Turk" were published between 2006 and 2014, including hundreds of papers published in top-ranked social science journals using data collected from MTurk (**Figure 1**). Early academic adopters of MTurk worked in computer science–related fields and mostly used workers to produce data, such as corpora of sentences or classifications of pictures used to train or evaluate machine learning software. Early on, computer scientists also explored the potential to conduct research using human subjects (e.g., usability studies; Kittur et al. 2008). From there, MTurk diffused to other disciplines within the umbrella of cognitive science (e.g., linguistics, judgment and decision making, and cognitive psychology) and

Worker: a person who is paid to complete a task by a requester on MTurk

Convenience sample: participants recruited on the basis of ease of access rather than a sampling strategy (e.g., MTurk workers, college students, patients at a clinic)

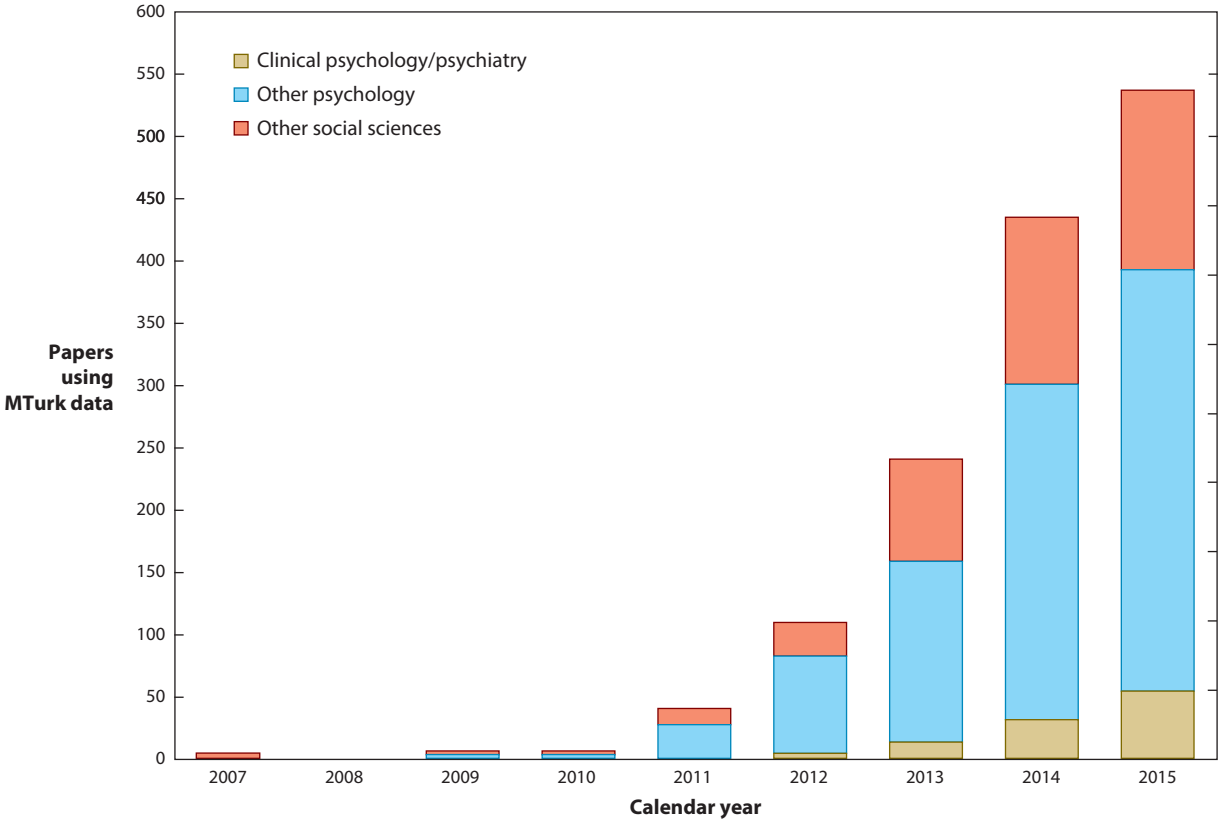

Figure 1

The number of papers using Amazon Mechanical Turk (MTurk) data published in social science journals with an impact factor greater than 2.5. Journals are categorized and assigned impact factors according to the Thomson Reuters Journal Citation Reports. Discipline categories were aggregated to Clinical Psychology or Psychiatry ($N = 76$), Other Psychology ($N = 82$), and Other Social Science ($N = 158$). Journals belonging to more than one category were assigned to the most specific applicable category within this coding scheme.

then to closely related disciplines such as social psychology, personality psychology, and consumer behavior. More recently, researchers interested in political science, clinical science, and sociology have embraced MTurk as a source of high-quality convenience samples. The first clinically relevant papers focused on topics of shared interest with decision making (i.e., gambling; Weatherly & Cookman 2014) and with personality psychology (i.e., narcissism and psychopathy; Jones & Paulhus 2011). More recently, clinical psychologists have used MTurk to study a broader range of psychological symptoms and interventions.

Advances in Internet technology have made it easier to reach large and diverse samples of research participants at low cost (for a detailed discussion, see Gosling & Mason 2015), allowing researchers with limited access to community or clinical populations to conduct research. Even well-funded researchers have benefited from these tools by using data from online convenience samples to make data-driven decisions about which ideas to prioritize within more expensive modes of data collection. The use of MTurk to recruit research participants is a special case of this more general trend.

MTurk provides a number of features that are attractive to researchers: A shared data security and payment infrastructure is provided by Amazon, which lowers fixed overhead costs, reduces the need for technical expertise or administrative support, and avoids payment hassles that may disincentivize respondent participation (Mason & Suri 2012). A rudimentary reputation system assigns unique identifiers to each worker and tracks worker performance, which makes it difficult for the same individual to submit multiple responses to a particular research study and allows requesters to avoid workers who have a history of providing poor-quality responses.

One of the most attractive features of MTurk is the size of the sample it offers. It has attracted enough users to create a market that is sufficiently liquid to quickly fulfill the intermittent needs of individual researchers (for a discussion, see Chandler et al. 2013). Consequently, MTurk provides better-quality data in less time than do other extant means of recruiting convenience samples (e.g., advertisements on Facebook; for a comprehensive comparison of online convenience sample recruitment strategies, see Shao et al. 2015). The cost of collecting data is also an attractive feature of MTurk: Although commercially available online research panels provide additional services related to sample selection, including efforts to make samples more representative, they typically cost several times more than samples collected from MTurk (Berinsky et al. 2012, Mullinix et al. 2014, Weinberg et al. 2014).

The novelty and rapid spread of MTurk have drawn considerable attention and have led to research efforts aiming to evaluate the benefits, trade-offs, and methodological concerns of using this platform. Many of these are shared with Internet convenience samples in general (e.g., Gosling & Mason 2015, Reips 2002) and web panels in particular (e.g., Downes-Le Guin et al. 2006) but are new to researchers for whom MTurk has provided a convenient point of entry into collecting data online. The popularity of MTurk and the resulting scrutiny it has invited make it one of the most well understood convenience samples available to researchers and have contributed to renewed interest in online research methodology. This article discusses who MTurk workers are, reviews the kinds of research studies most suitable for the platform, and identifies potential methodological limitations and best practices when using this sample.

DEMOGRAPHIC CHARACTERISTICS OF THE MECHANICAL TURK POPULATION

Amazon claims that MTurk has more than 500,000 registered users. The number of active users is unknown but is likely close to 15,000 individual US workers at any given time (Stewart et al. 2015; see also Fort et al. 2011), with perhaps a similar number of Indian users (see Indian Workers

INDIAN WORKERS

The large population of Indian workers on MTurk presents an opportunity for researchers interested in cultural psychology (e.g., Nouri & Traaum 2014, Raihani & Bshary 2012) but also presents unique challenges. Indian workers are highly educated (Ipeirotis 2010, Khanna et al. 2010) and as honest as US workers (Suri et al. 2011), yet data are consistently of lower quality (Kazai et al. 2012, Khanna et al. 2010, Shaw et al. 2011). In particular, Indian workers struggle with instructional manipulation checks and reverse-coded items (Litman et al. 2015), suggesting that language difficulties are a key issue.

Quality differences underscore the importance of restricting participation to US samples when international workers are not necessary, as increased error undermines the psychometric properties of instruments (Feitosa et al. 2015) and raises serious questions about the interpretability of cultural differences observed using English language materials. Unlike US worker data, the quality of Indian worker data is influenced by financial compensation (Litman et al. 2015), but it is not known if this occurs because increased payment motivates workers or because it attracts more skilled workers (Gupta et al. 2014).

sidebar). Although people from any country can join MTurk, studies examining the nationality of workers have found that the vast majority of workers reside in the United States or India (Paolacci et al. 2010).

Most researchers restrict their sample to US workers, and therefore US worker demographic characteristics are the best understood and of greatest practical value. The US worker population is diverse but not representative of the population as a whole. In particular, MTurk workers tend to be younger and better educated (Paolacci & Chandler 2014; for large-sample validations, see Casey et al. 2015, Greenblatt 2014, Huff & Tingley 2014, Palmer et al. 2015). European- and Asian-Americans are overrepresented, and Hispanics of all races and African Americans are underrepresented (Paolacci & Chandler 2014; for large sample validations, see Casey et al. 2015, Greenblatt 2013, 2014, Huff & Tingley 2014). These differences parallel differences between Internet users and the US population as a whole and mirror the demographic characteristics of other online nonprobability samples (Hillygus et al. 2014). Workers are also less religious and more liberal than the population as a whole, which may reflect differences in age and education (Berinsky et al. 2012, Mullinix et al. 2014).

As is generally true of people who take paid surveys (Hillygus et al. 2014), workers tend to report lower personal incomes, are more likely to be unemployed or underemployed (Corrigan et al. 2015, Shapiro et al. 2013), and are less likely to own their residence (Berinsky et al. 2012) than Americans in general. Interestingly, measures of household income are closer to national averages (Greenblatt 2013, Huff & Tingley 2014) and workers are less likely to live alone (Greenblatt 2013, 2014). One possible explanation for this discrepancy is that a large proportion of MTurk users are under- and unemployed millennials living with their parents. A recent survey of approximately 10,000 MTurk workers provides some support for this conjecture: Nearly 20% of MTurk workers live in a household in which the oldest member is more than 20 years older than they are (Casey et al. 2015).

Reflecting their younger age and cohort, MTurk workers are less likely to be married than the population as a whole (Berinsky et al. 2012, Shapiro et al. 2013), but they are only slightly less likely to have biological children and might be more likely to have stepchildren than the population as a whole (Shapiro et al. 2013), which suggests ample opportunity to conduct research on family dynamics. Researchers have also consistently observed that workers are more likely (7–9%) to report identifying as LGBTQ (lesbian, gay, bisexual, transgender, queer) (Corrigan et al.

2015, Reidy et al. 2014, Shapiro et al. 2013), again in part because the sample is younger than the population as a whole (Gates 2014).

Probability sample: a sample consisting of individuals selected at random from a defined population

In direct comparison studies, MTurk samples are found to be more representative than college student samples, community samples recruited from college towns (Berinsky et al. 2012), or other online sources (Casler et al. 2013). However, MTurk samples are less representative than web-based probability samples (i.e., individuals recruited through random digit dialing and invited to participate in an online panel) in terms of gender, race, income, and marital status (Berinsky et al. 2012, Mullinix et al. 2014, Weinberg et al. 2014). On some dimensions, such as age and home ownership, MTurk worker samples are biased in opposite directions relative to online probability samples (Weinberg et al. 2014), telephone samples (Simons & Chabris, 2012), and face-to-face interviews (Berinsky et al. 2012), which suggests that MTurk is good at reaching populations that are typically underrepresented through traditional recruitment techniques (see Blumberg & Luke 2007).

MTurk workers seem to differ from other commonly used samples on a number of psychological dimensions, either for reasons directly related to their self-selection into MTurk or because of correlated demographic differences. Numerous findings suggest that workers are above average in cognitive aptitude: They score higher than the general population on measures in a range of areas including civics knowledge (Berinsky et al. 2012), financial literacy (Krische 2014), science knowledge (Cooper & Farid 2014), and computer literacy (Behrend et al. 2011). Perhaps more convincingly, workers report higher SAT scores and score higher on sample SAT items with less evidence of cheating than do college students (Cavanagh 2014). Workers also tend to score highly on learning goal orientation (Behrend et al. 2011) and need for cognition—an individual difference reflecting enjoyment of and motivation to engage in difficult cognitive tasks (Berinsky et al. 2012).

In terms of the prevalence of clinical symptoms, the extant literature is a bit more nuanced. Initial estimates of the proportion of MTurk workers taking psychotropic medication range from 10% to 12%, and initial estimates of lifetime incidence of diagnosed mental illness range from 20% to 30% (Rojas & Widiger 2014, Shapiro et al. 2013, Wiens & Walker 2015). Data about the prevalence of depression and anxiety are inconsistent, with some researchers finding that levels of depression and anxiety reported by MTurk workers are in line with those observed in other community samples (Shapiro et al. 2013), others reporting lower rates (Veilleux et al. 2015), and still others reporting elevated rates (Arditte et al. 2015). The proportion of individuals who score above clinically significant cutoff scores for obsessive-compulsive symptoms (Arditte et al. 2015, Fergus & Bardeen 2014) and attention-deficit/hyperactivity disorder (ADHD) (Wymbs & Dawson 2015) appears similar across MTurk and other community samples.

A number of studies provide converging evidence that MTurk workers differ in specific and clinically relevant ways, displaying a cluster of probably interrelated differences in social anxiety, emotion regulation, and autism spectrum disorder (ASD) features. Workers consistently fall about one standard deviation above college students on measures of social anxiety (Arditte et al. 2015, Shapiro et al. 2013; in addition, compare Fergus 2014b to Fergus et al. 2012 and Nichols & Webster 2015, study 3 to study 1) and report more difficulty functioning in social, school, and work situations (Gootzeit 2014). These findings are consistent with personality research that suggests that MTurk workers are more introverted than college or community samples (Behrend et al. 2011, Goodman et al. 2013, Kosara & Ziemkiewicz 2010) and report lower self-esteem (Goodman et al. 2013).

Some evidence indicates that workers also experience more generalized difficulties with emotional regulation. MTurk workers report experiencing more anxiety or negative affect in stressful situations (Johnson & Borden 2012, Veilleux et al. 2014) and a lower tolerance for physical discomfort and psychological distress than do college students (Arditte et al. 2015, Gootzeit 2014;

but see also Veilleux et al. 2014). These findings are consistent with evidence from personality research that workers are somewhat more neurotic and less agreeable than are college or community samples (Goodman et al. 2013, Kosara & Ziemkiewicz 2010).

In addition, some evidence suggests that MTurk workers are more likely to possess traits associated with autism spectrum disorders (ASDs). In one study ($N = 823$), researchers found that 1.8% of MTurk users reported having been diagnosed with an ASD—almost double the rate reported in a community sample (Mitchell & Locke 2015). Although it is possible that some of this difference is attributable to self-report measurement error, the total proportion of individuals who reported living in a household containing at least one person with an ASD was roughly the same as the community sample, suggesting that there wasn't a general tendency to inflate reports of a diagnosis within this particular study. Similarly, Autism Spectrum Quotient (AQ) scores observed within the MTurk population are about one-third of a standard deviation above those previously observed in a large college student sample (Eriksson 2013 and Palmer et al. 2015 compared to Ruzich et al. 2015). AQ scores are correlated with both decreased extraversion and increased neuroticism (Austin 2005), providing convergent validity with the population differences in personality cited above.

Turning to alcohol and substance abuse, Veilleux and colleagues (2014) found that MTurk workers binge drink less frequently than do college students. Although MTurk workers were quite likely to screen positive on the CAGE-AID—a clinical assessment of problematic substance use—they generally reported only light to moderate consumption of alcohol and recreational drugs (Shapiro et al. 2013). Approximately 10% of MTurk workers report regularly smoking marijuana (Shapiro et al. 2013), and approximately 22% to 25% of workers define themselves as tobacco smokers (Johnson et al. 2015, Reese & Veilleux 2016), which is slightly higher than the US population as a whole.

In sum, MTurk is not representative, but it is more diverse than samples typically used in clinical research (e.g., students and community samples). The diverse sample demographics make it well suited for recruiting specific populations, including people who identify as LGBTQ, are unemployed or underemployed, are married, or are parents. The number of users available through MTurk also makes it feasible to identify and recruit sufficiently large samples of people with relatively common psychological conditions, but MTurk may be especially useful for researchers interested in social anxiety or ASDs because of their somewhat greater prevalence within this population. Conversely, it is likely that some populations are underrepresented, such as individuals with intellectual disabilities or severe psychopathology (e.g., psychosis or severe depression) that makes using a computer difficult. Perhaps more importantly, the characteristics of the MTurk worker population are transparent and increasingly well understood, making it possible to articulate potential limitations of recruiting from this sample in a way that is not feasible with other nonprobability samples.

WORKER MOTIVATION

Studies consistently show that money, fun, and learning new skills (in that order) are the primary motivating forces for completing MTurk HITs (Behrend et al. 2011, Horton et al. 2011, Litman et al. 2015, Paolacci et al. 2010; but see also Buhrmester et al. 2011). That financial incentives are motivating is surprising given the low pay workers are willing to accept relative to the mainstream workforce. One possibility is that workers are in financial need and lack viable alternative income sources (Brawley & Pury 2016), which is consistent with the high number of workers who are unemployed and underemployed (Shapiro et al. 2013). Additionally (or alternatively), the autonomous nature of MTurk offers advantages that may partially offset low pay: Workers select what

they want to do and when and where they want to do it. Tasks can fill gaps of time between, or even during, regular jobs or other small segments of time throughout the day that otherwise have little economic value. They can also be completed in an environment of the worker's choosing (e.g., at home in pajamas or while performing another task that does not require undivided attention). Finally, MTurk may be particularly appealing for workers for whom traditional workplaces are inaccessible or undesirable due to a disability, a mental health condition, personal preference (e.g., not wanting to interact with people), or other barrier (e.g., lack of transportation).

Other motives influence workers' decisions to complete tasks on MTurk (for a qualitative analysis, see Martin et al. 2014). In fact, compensation was listed as the primary motivation for completing tasks by fewer MTurk workers (45%) than by college participants completing a task for college credit (78%) (Behrend et al. 2011). The mixture of motives reported by MTurk workers has led some to suggest that MTurk should be regarded as paid leisure rather than a replacement for work (Jiang & Wagner 2014), a perspective that would view time spent clicking through surveys on MTurk as a substitute for time spent clicking through online games. Reflecting their high need for cognition and learning motivation, workers may find tasks such as translating, completing surveys, and tagging photos to be reasonable alternatives to other less engaging activities that people typically use to kill time (e.g., watching television). The opportunity to learn new skills is a tertiary but still important motivating factor (Behrend et al. 2011, Horton et al. 2011, Paolacci et al. 2010) that is probably of particular relevance to tasks related to translation and transcription rather than participation in research studies.

VALIDITY OF WORKER DATA

MTurk workers are virtually anonymous and complete HITs in an unsupervised environment, with clear incentives to accomplish tasks as quickly as possible. Thus, there has been understandable concern that workers may not respond seriously or truthfully. Initial concerns centered on whether workers were sufficiently attentive (e.g., Paolacci et al. 2010), a worry reinforced by reports that workers often complete surveys in environments that are far from ideal, such as while watching TV or while in the room with another person (Chandler et al. 2014, Clifford & Jerit 2014). However, there is little evidence that these distractions have a negative effect on data quality.

Scale reliability obtained from MTurk samples is consistently identical to or even superior to that of other samples (Behrend et al. 2011, Buhrmester et al. 2011, Jahnke et al. 2015, Johnson & Borden 2012). There are not many comparisons of responses to individual-scale items across populations. The Big Five inventory (a measure of personality) is the only multifaceted scale that has been examined carefully across MTurk and other populations. The available evidence demonstrates that equivalent five-factor solutions are obtained across MTurk, undergraduate, and community samples (Feitosa et al. 2015), with most items functioning more-or-less equivalently across samples (Behrend et al. 2011). Other studies comparing measures of narcissism across MTurk workers and a population of psychiatric outpatients (Miller et al. 2013) and measures of body image across college and MTurk samples (Tylka & Wood-Barcalow 2015) have reached similar conclusions.

Data provided by workers also have high concurrent and convergent validity. In initial research, unemployed MTurk workers reported more negative affect than those who were employed, women reported more anxiety and were more likely to report clinically significant levels of depression, and men reported more alcohol and drug consumption, consistent with findings from large and representative samples (Shapiro et al. 2013). Arditte and colleagues (2015) did not replicate the elevated levels of depression and anxiety among women reported by Shapiro and colleagues (2013), but they did observe that women reported elevated levels of social anxiety and symptoms related to eating

disorders. Among a large sample of MTurk workers, Wymbs & Dawson (2015) found that men were more likely to be diagnosed with ADHD as children but less so as adults, and that people with ADHD diagnoses had lower educational attainment, paralleling observations in other samples.

Other studies have directly compared the magnitude of concurrent and convergent validity statistics across populations. For example, MTurk workers and college students show equivalent relationships between state and trait empathy measures (Johnson & Borden 2012) and attachment and self-esteem (Wickham et al. 2015). Likewise, Veilleux and colleagues (2015) observed similar but stronger relationships between lay theories of emotional self-regulation and depression, binge eating, and anxiety in MTurk workers relative to a college population.

Paralleling evidence of the strong psychometric properties of worker self-report data, available evidence suggests that experiments produce equivalent effect size estimates within MTurk and other samples. Initial demonstrations of the equivalence of effect sizes across MTurk worker and college student samples used individual experiments and small sample sizes (Berinsky et al. 2012, Horton et al. 2011, Paolacci et al. 2010). These have been largely corroborated by high-powered batteries of experiments administered to both MTurk and non-MTurk samples (Klein et al. 2014, Mullinix et al. 2014), with the caveat that phenomena that are moderated by demographic characteristics upon which MTurk differs from representative samples show corresponding differences in the magnitude of observed effects (Mullinix et al. 2014).

Some researchers have directly investigated whether workers are attentive and honest. Worker attentiveness was among the first issues addressed by researchers investigating the validity of MTurk data (Paolacci et al. 2010). In this study and subsequent research, workers have consistently passed "catch" trials and other attention checks at an equal or higher rate than participants in other samples, although recent pass levels are inflated by extensive worker practice at answering these items (Hauser & Schwarz 2015a). Taking a somewhat different approach, within self-report scales the number of inconsistent responses to synonymous items (indicative of random responding) and the maximum string length of identical numerical responses (indicative of straightlining) do not differ across MTurk and college undergraduate samples (Behrend et al. 2011).

Questions of psychometrics and the replicability of experimental effects on MTurk are often ultimately about the veracity of worker responses. Worker responses are usually consistent, suggesting that they are likely true. Demographic details provided at different time-points are also usually identical, with more than 95% of respondents reporting age, race, gender, or location consistently across two data collection points (Mason & Suri 2012, Rand 2012, Shapiro et al. 2013). As a point of comparison, participant gender is collected twice in the General Social Survey (a high-quality probability sample), with agreement rates of about 96% (Black et al. 2000). Similarly, reports of user location corresponded to logged IP addresses in 97% of cases (Rand 2012).

Test-retest reliability for psychological instruments is also generally high on MTurk. Test-retest reliabilities for individual differences in personality characteristics (Big Five Inventory) averaged around $r = 0.85$ when measured three weeks apart and were consistently higher than relationships noted in previous studies of other samples (Buhrmester et al. 2011). Subsequent research has also found high test-retest reliability for clinical instruments. To illustrate, measures of depression are highly correlated one week apart (Beck Depression Inventory: $r = 0.87$, Shapiro et al. 2013; Patient Health Questionnaire-9: $r = 0.78$, Carr 2014), as are more general measures of psychological well-being (Acceptance and Action Questionnaire-II: $r = 0.83$, Brief Experiential Avoidance Questionnaire: $r = 0.85$, Carr 2014). Schleider & Weisz (2015) examined several variables related to family functioning across three months and found that parent assessments of parent and child psychological functioning remained consistent across this time period (r's > 0.82). They did, however, observe weaker relationships across time points between reported family functioning ($r = 0.66$) and parenting stress ($r = 0.36$).

CLINICAL SCIENCE APPLICATIONS OF MECHANICAL TURK

Some researchers have used MTurk to recruit convenience samples of typically functioning adults to examine public attitudes about clinically relevant issues, such as mental health stigma (Corrigan et al. 2015) and attitudes toward therapy (Arch et al. 2015). Several research groups have studied MTurk workers' perceptions of mental health professionals (Lebowitz et al. 2015) and of people diagnosed with specific psychiatric and behavioral disorders, including ASDs (Mitchell & Locke 2015), depression (Burke et al. 2014), and pedophilia (Jahnke et al. 2015).

MTurk workers have occasionally been used as pilot subjects (Green et al. 2014) or as a control group for studies on clinical populations (for an overview, see Azzam & Jacobson 2013), under the assumption that workers, though not representative, are "representative enough" on the variables of interest. MTurk has also gained interest as a recruitment tool for researchers interested in specific and potentially clinically relevant behaviors such as tobacco use (Cougle et al. 2014, Johnson et al. 2015), alcohol and drug use (Boynton & Richman 2014, DeWall et al. 2014, Kim 2014, Tobin et al. 2014), and gambling (Weatherly & Cookman 2014).

As awareness of MTurk has grown, clinical scientists have started to use MTurk to examine a variety of psychopathological symptoms within the general population, including compulsive buying (Rose & Segrist 2012), hoarding (Raines et al. 2015), obsessive-compulsive disorder (Fergus & Bardeen 2014), generalized anxiety disorder (Lebowitz et al. 2014), depression (Winer et al. 2014, Yang et al. 2014), and hypomania (Devlin et al. 2015). Personality psychologists have studied the so-called dark triad of narcissism, psychopathy, and Machiavellianism extensively using MTurk samples (e.g., Davenport et al. 2014, Jones & Olderbak 2014, Jones & Paulhus 2011). Other researchers have used MTurk to conduct experimental studies investigating the processes underlying clinically relevant symptoms such as the effects of priming of certain religious beliefs on scrupulosity (Fergus & Rowatt 2015) and the relationship between disgust sensitivity and borderline personality disorder (Standish et al. 2014).

MTurk has also been used to recruit participants with specific characteristics. Yang and colleagues (2014) prescreened workers and recruited individuals reporting depressive symptoms, and Usinger (2014) recruited participants who reported mild to moderate anxiety. Others have used MTurk to recruit cigarette smokers (Kim 2014), overweight people (Pearl et al. 2014), Catholics and Protestants (Fergus & Rowatt 2015), the long-term unemployed (Konstam et al. 2015), and immigrants (Bernal 2014). Of particular note, researchers have found a high representation of fathers, a notoriously difficult group to recruit both online and in person (Parent et al. 2015, Schleider & Weisz 2015).

MTurk seems particularly useful for studying certain topics. Workers report greater comfort disclosing psychologically relevant information online than through an in-person interview (Shapiro et al. 2013), facilitating the study of clinical symptoms of a sensitive nature such as self-reported pedophiliac interests (Wurtele et al. 2014), intimate partner violence (Jones & Olderbak 2014, Reidy et al. 2014), and self-injury (Andover 2014, Victor & Klonsky 2014). As noted previously, MTurk workers are highly computer literate, making the platform particularly useful for studying phenomena related to computer use, such as the association between clinically relevant variables and social media use (Davenport et al. 2014) and the properties of Internet-related disorders like cyberchondria (Fergus 2014a, Norr et al. 2015). The greater frequency with which social anxiety (Allan et al. 2015) and ASD features (Eriksson 2013, Palmer et al. 2015) appear among MTurk users may facilitate the study of these issues.

The large number of available participants can even make it possible to obtain small but adequate samples of rare or hard-to-access populations that may be of particular interest to some clinical psychologists. Arch (2014) recruited pregnant women for a study on preferences for anxiety

treatment. Papa and colleagues (2014) examined bereavement among people who had recently experienced the death of a loved one, become divorced, or lost a job. A study of cancer survivors was able to recruit 166 participants, though with a significant time investment (Carr 2014). Lynn (2014) was able to recruit veterans of Operation Iraqi Freedom. Although, as is discussed later, efforts to recruit specific populations—particularly those that are rare—can face data quality challenges if people mistakenly or deceptively misidentify themselves, the large number of workers on MTurk represents an opportunity to access populations that might otherwise be impossible to reach.

Mechanical Turk Affords a Variety of Research Methods

Although most clinical psychological research focuses on single-shot administration of self-report scales, or simple experiments, advances in computing technology have enabled the use of more elaborate research methods. It is now possible to present stimuli and measure reaction time with millisecond precision using Flash (Simcox & Fiez 2014), Java script (de Leeuw 2014), or even existing survey platforms such as Qualtrics (Barnhoorn et al. 2014), opening up a wide range of implicit and behavioral measures for use online. Although MTurk workers complete tasks in an unsupervised environment, experiments that use these techniques to measure small differences in response latencies, including the Implicit Association Task (Klein et al. 2014) and Stroop Task (Crump et al. 2013), or that rely on brief presentation of stimuli, such as attentional blink, flanker, and subliminal priming paradigms (Crump et al. 2013), have all been successfully implemented using MTurk samples. More recently, researchers have even demonstrated proof of concept of the use of webcams to conduct remote eye-tracking studies (Lebreton et al. 2015).

Workers on MTurk are assigned unique identifiers that make it possible to identify their responses, recontact them, and regulate their participation across multiple surveys. Consequently, waves of data can be collected from workers over a period of days, allowing for diary studies and other forms of experience sampling (Boynton & Richman 2014, Lanaj et al. 2014, Usinger 2014). Longitudinal research studies can also be conducted over longer time frames. When workers are contacted weeks or even months apart, attrition rates are typically approximately 30% (Reese & Veilleux 2016, Schleider & Weisz 2015, Shapiro et al. 2013, Wiens & Walker 2015) and increase to about 55% after one year (Chandler et al. 2014, Daly & Nataraajan 2015). Based on estimates that about one-quarter of the pool quits MTurk and is replaced by new workers every three months (Stewart et al. 2015), these attrition rates probably represent the ceiling of what can be obtained on this platform.

At any given time, many workers are completing HITs on MTurk, making it possible to set up and run remote interactive group tasks (e.g., Hawkins 2014, Mason & Watts 2012, Suri & Watts 2011). Coordinating groups to arrive in a physical lab is a hassle, making MTurk an attractive alternative to traditional subject pools for running experiments on how dyads or groups of people interact. Because interactive tasks are familiar to MTurk workers, it is also possible to create believable experiments that involve only pseudointeraction with others (Rand et al. 2014). Although clinical scientists may not typically conduct basic scientific research on group dynamics (for a recent exception, see ten Brinke 2015), there is clearly an exciting potential to use MTurk to understand how clinically relevant traits predict behavior in interpersonal contexts, such as cooperative and competitive games.

Perhaps of greatest interest to clinical researchers, several researchers have demonstrated that MTurk can have a role in the development of online psychological interventions (see Andersson 2016). Through MTurk, psychological interventions can be pilot tested for usability on healthy participants (Howard & MacCalla 2014) or even implemented and assessed using individuals with clinically significant symptoms. Usinger (2014) conducted a randomized controlled trial of

mindfulness meditation as a treatment for anxiety. O'Connell and colleagues (2016) asked people to engage in other-focused acts of kindness and observed improved interpersonal relationships relative to controls. In two especially innovative examples, MTurk workers have even demonstrated their potential as practitioners, blending the role of MTurk workers as research participants and as a source of crowdsourced labor. Morris & Picard (2014) had MTurk workers provide empathic support and cognitive reappraisals to other users and found that they were generally effective at this task. Along the same lines, Marietta and colleagues (2014) used MTurk workers as participants in an online antibullying intervention.

Human computation: the use of a human or human-machine hybrid system to solve computational problems that are difficult for machines alone

Mechanical Turk as a Human Computation Tool

MTurk can also be used by researchers for its intended purpose as a human computation tool to support transcription of written (Lang & Rio-Ross 2011) or spoken (Marge et al. 2010) language, content coding, generating experimental stimuli (Sina et al. 2014), and other tasks typically undertaken by research assistants (for an overview, see Chandler et al. 2013). In a clinically relevant example, Vlahovic and colleagues (2014) had MTurk workers code forum posts of breast cancer survivors for support-seeking and support-giving behavior. Although workers are not experts, for statistical reasons their aggregated beliefs often meet or exceed the accuracy of smaller groups with more knowledge. Illustrating this potential, 9 workers evaluated speech audio files and produced judgments equivalent to those of 3 trained speech pathologists (McAllister Byun et al. 2014). Using CrowdFlower (a service similar to MTurk), Benoit and colleagues demonstrated that 15 workers can produce judgments of political content equivalent to judgments of 5 political science PhD students and faculty (Benoit et al. 2015). Equally important, Benoit and colleagues nicely illustrated the speed and scale of content coding that is possible using crowdsourcing: In one iteration, the crowd was able to content code 22,000 sentences in under five hours for $360.

METHODOLOGICAL CHALLENGES OF MECHANICAL TURK

Mechanical Turk Is a Nonprobability Sample of the Population

Studies using MTurk are a special case of convenience sampling and, as such, many of the methodological limitations known to apply to other convenience samples—like those recruited from clinics and from college student subject pools—also apply to MTurk. In particular, MTurk is not representative of any particular population. A lack of representativeness is usually (but not always) a minor concern to psychologists, who are more interested in associations between variables than in point estimates for the population at large. Because they are more interested in modeling relationships than in describing a population, psychologists typically respond to potential differences across subpopulations by recruiting them in sufficient numbers to test potential differences rather than by recruiting them in proportions that mirror the population as a whole (Groves 2004). However, because MTurk workers are more diverse than the samples psychologists typically use, the sample may be assumed to include a specific subpopulation or mistaken as representative, tempting researchers to draw inappropriate inferences, such as estimates of population prevalence (although this can be done in some situations provided appropriate sample stratification and weighting procedures are used; Greenblatt 2013, 2014).

As is true of all convenience samples, if moderator variables are correlated with the probability of joining MTurk, then relationships between predictor and criterion variables can be inflated or attenuated, whereas main effects observed in experimental designs may more closely resemble an interaction between the experimental treatment and an individual difference variable (e.g.,

Mullinix et al. 2014). For example, the criteria by which liberals and conservatives evaluate other people may differ, and MTurk workers tend to be politically liberal. Thus, it may be reasonable to worry about the generalizability of findings related to mental illness stigma to the population at large.

Representativeness of MTurk workers can also differ across subgroups. For example, young people on MTurk tend to resemble young people in general more than older people on MTurk tend to resemble older people in general (Huff & Tingley 2014), perhaps because using MTurk deviates more from the range of typical behavior for older cohorts than for younger cohorts. Thus, it may be reasonable to worry that some findings related to age (e.g., reduced levels of clinical symptoms among older adults; Arditte et al. 2015) reflect something about the kind of older adults likely to use MTurk rather than something about aging itself (but see Bui et al. 2015). That said, theory development is spurred by researchers who disagree with the default presumed generalizability of a finding and can produce evidence that illustrates the moderating influence of other variables, and these concerns should not automatically preclude the use of MTurk samples.

A somewhat more concerning issue is that the degree of representativeness of a sample is knowable only for variables that are measured. It is always possible that a convenience sample can differ from the population in general in some critical yet unmeasured way. One small advantage that MTurk has over other convenience samples in this regard is that it is a shared pool and therefore knowledge about its characteristics has accumulated over time. This shared body of knowledge makes it easier to develop informed hypotheses about the potential problems with using MTurk samples to study specific research questions.

Mechanical Turk Studies Are Nonprobability Samples of the Mechanical Turk Worker Population

Interest in the representativeness of MTurk workers has focused on the ways in which the pool of available workers differs from the general population. It is often overlooked that participants within individual studies are also nonprobability samples of the MTurk population as a whole, leading to large differences in sample characteristics across studies. For example, Greenblatt (2014) recruited 3,010 workers in a study of the prevalence of "miscellaneous refrigeration products" in US households, of which 55% (95% CI: 53%, 57%) of respondents were female. In contrast, Cabrera and colleagues (2014) recruited 2,776 participants in a study on perceived ethicality of drugs that enhance cognitive performance, of which only 43% (95% CI: 41%, 45%) were female. Despite large sample sizes, these samples differ substantially in terms of their gender composition, $z = 9.4$, $p < 0.001$.

Little is known about the specific causes of selection bias. Some bias may result from incidental design decisions that are uncorrelated with study content. For example, one set of studies found suggestive evidence that workers recruited during the daytime were older, more likely to be female, and less likely to use a computer mouse to complete the survey (suggesting that they were using mobile devices) relative to participants recruited in the evening (Komarov et al. 2013).

Idiosyncratic events can also influence selection. For example, in one study that was composed of an unusually high number of men, Chandler and colleagues (2014) observed that many of the respondents reported that the study information had been posted on Reddit, a site frequented more often by men. In another study, Higgins and colleagues (2010) noted that posting a well-paying HIT boosted the completion rate of other lower-paying HITs from the same account, suggesting that workers who found one lucrative HIT searched for other HITs posted by specific requesters.

Evidence for these kinds of incidental variation in sample composition is scattered and largely anecdotal. Although this remains a potentially fruitful area of future investigation, it is likely that

differences of the magnitude of that described above between the observations of Cabrera and colleagues (2014) and of Greenblatt (2014) reflect preferences for studies on different topics. This is particularly concerning for the generalizability of study findings because factors directly related to willingness to complete studies on a specific topic (e.g., engagement with or expertise in that topic) are also likely to moderate how people think about and answer questions related to it.

Recruiting Specific Populations

Many clinical scientists are interested in using MTurk to recruit specific, hard-to-find populations. Although the large size of the worker pool can make MTurk an effective recruitment tool, prescreening based on participant self-report is not without its challenges. Some proportion of individuals will be misidentified as belonging to a desired group either because they misunderstand or incorrectly respond to prescreening questions. Even for simple demographic questions like sex, up to 2% of survey respondents provide incorrect responses (Voracek et al. 2001), and the same is true of MTurk (Rand 2012, Shapiro et al. 2013). This is a particularly challenging problem for researchers recruiting rare groups, where the number of false-positive group members (i.e., people who make mistakes when completing the survey) will approach or exceed the number of true group members.

Complicating matters further, although MTurk workers are no more dishonest than other people (Beramendi et al. 2014, Cavanagh 2014, Farrell et al. 2014), they will act deceptively if incentivized to do so, which includes lying to gain access to paid research studies. For example, in one large market research panel, approximately 17% of panelists claimed to own a Segway, many times higher than the actual rate of private ownership (Downes-Le Guin et al. 2006). Similar problems can arise within MTurk samples. To illustrate the impact of careless recruiting on MTurk, at the end of a recent large study, half of participants (who were mostly parents) were asked whether they were the parent or guardian of a child with autism, and 4.3% indicated that they were—a proportion that is already suspiciously high. Crucially, the other half of parents were first told that that we were trying to determine their eligibility for another study; in this condition, the proportion of respondents who reported being the parent or guardian of a child with autism increased to 7.6% (J. Chandler, unpublished data). This suggests that if a researcher were to recruit parents of children with autism through an explicit request for this sample, approximately half of the sample would consist of people who have misrepresented themselves.

Malingering

Several researchers have noted that workers score unusually high on measures of malingering relative to established norms (Arch et al. 2015, Carr 2014, Shapiro et al. 2013). Measures of malingering consist of items that are rarely endorsed among the general population or among people with genuine psychopathological symptoms but are frequently endorsed by people attempting to fake psychopathology. Although these measures are used to identify individuals who are faking symptoms, there are several other possible contributors to elevated malingering scores. Careless responding is one (unlikely) cause of inflated malingering scores. As a more probable cause, norms for existing malingering scales are old and likely outdated, failing to account for cohort differences in willingness to endorse specific items (as has occurred in the past; Arbisi & Ben-Porath 1995). For example, on the malingering scale used by Shapiro and colleagues, disagreement with the statement "I believe in law enforcement" is keyed as malingering, yet it may also tap into increasing skepticism toward police among younger Americans (Pew Res. Cent. 2010). Likewise, the statement, "I have often wished I were a member of the opposite sex" is keyed as malingering, but it may also be sensitive to more fluid attitudes toward gender among younger generations (Wong

2015). The presumed rarity of responses keyed as diagnostic of malingering makes them especially sensitive to errors or cultural change: Endorsing just one item incorrectly is sufficient to increase a respondent's score by more than two-thirds of a standard deviation (Arbisi & Ben-Porath 1995).

Although benign explanations of elevated malingering scores are plausible, the presence of malingerers in a sample is a potentially serious issue: They will display elevated scores across most other clinical measures, thereby inflating observed relationships between these measures. A reanalysis of data reported in Shapiro and colleagues (2013) provides inconclusive but suggestive evidence about the impact of potential malingerers on data quality: Excluding respondents defined as scoring high on malingering (as originally reported) decreased the correlation between anxiety and depression from 0.64 to 0.57, a significant difference, $z = 1.67$, $p < 0.05$, one-tailed. Similarly, the correlations between anxiety and depression and social anxiety dropped from around 0.53 to around 0.47 (although these differences were not significant) when malingerers were excluded. These findings suggest that it will be important for future research to investigate the causes of, better identification of, and remedies for elevated malingering scores among online respondents.

Nonnaïveté of Participants

MTurk workers can complete as many studies as they want, and investigators have discovered that a small number of workers produce a large proportion of survey responses. Chandler and colleagues (2013) found that the most productive 10% of individuals on MTurk produced 41% of all completed responses to experiments. Other researchers examining the distribution of HIT completion by workers have found similar results (Berinsky et al. 2012, Fort et al. 2011, Stewart et al. 2015). This issue is typically not encountered within college student samples, where participation is frequently capped, but it is a problem common within online panels more generally, where it has been estimated that as many as one-third of all responses are provided by the most active 1% of panel members (Miller 2006).

Workers who participate in many studies become familiar with study materials, including scales and measures. Practice effects are a problem of repeated participation that is well known to clinical psychologists: Performance on measures of ability, such as IQ tests, tends to improve over successive attempts. Practice effects have been documented several times on MTurk. Over recent years, the rate at which workers correctly answer "trick" questions that check for attentiveness has increased from a level similar to that observed among undergraduate samples (Goodman et al. 2013, Paolacci et al. 2010) to far above that of undergraduate samples (Hauser & Schwarz 2015a). Likewise, worker performance on the standard version of the cognitive reflection task (a problem-solving task with factually correct answers; Frederick 2005) is correlated with the number of HITs a worker has completed, but performance on novel but logically identical problems is not, suggesting that workers have learned the correct answers to the familiar items over time (Chandler et al. 2014). Repeated exposure to research studies can also influence measurements not obviously related to ability, as suggested by research on panel conditioning effects that finds that exposure to attitudinal measures provides participants with an opportunity to elaborate on and clarify their beliefs and changes subsequent responses (Sturgis et al. 2009).

Exogenous and unmeasured influences (like familiarity with experimental materials) can influence the probability of observing true relationships between variables of theoretical interest. Rand and colleagues (2014) used MTurk to demonstrate that when people are under time pressure they are more likely to cooperate at the cost of maximizing personal financial gain. Importantly, they further observed that the effect of time pressure declined in later experiments, presumably because the proportion of experienced participants grew and experienced participants became better (faster) at maximizing individual returns (see also Mason et al. 2014).

In a more direct demonstration of the consequences of repeated participation, Chandler and colleagues (2015) asked MTurk workers to complete nine short psychology experiments at two different time points. Reductions in effect sizes were observed across most dependent measures, particularly among participants assigned to different experimental conditions at each time point, suggesting that information learned in the first exposure to the experiments contaminated subsequent responses (Chandler et al. 2015). Sometimes workers can become aware of the contents or hypothesis of a study through discussions posted on worker forums, which can produce effects similar to those caused by prior participation in studies. In practice this is rare: The primary purpose of worker forums is to share information about lucrative HITs and the people who post them (e.g., problems securing payment). Workers rarely directly discuss the hypotheses of research studies, with less than 15% of workers reporting having ever seen a forum post about the contents of a HIT (Chandler et al. 2014). However, important information is sometimes inadvertently revealed. For example, a worker may encounter a technical glitch and inadvertently reveal information about the contents of an experimental condition while trying to troubleshoot the problem with other workers. In these cases, it is possible that exposure to forum posts may have a similar attenuating effect as prior direct exposure to research materials.

BEST RESEARCH PRACTICES

The flexibility of the MTurk platform is one of its appealing features, but it can also cause problems: The platform will not prevent researchers from making mistakes, nor will it suggest better ways to structure research studies. Below are concrete suggestions that address commonly observed problems with HITs and studies that report MTurk data.

Pay a Fair Wage

Larger financial incentives generally increase the speed of data collection (Berinsky et al. 2012, Heer & Bostock 2010, Mason & Watts 2010) and increase workers' willingness to engage in and persevere through difficult tasks (Crump et al. 2013). Data quality is generally unaffected by pay when US workers are sampled and asked only to provide opinions or other self-report answers (Buhrmester et al. 2011, Mason & Watts 2010), probably because it is not much harder to report a true opinion than it is to make one up. In general, we consider payment to be more of an ethical issue than a data-quality issue and suggest that researchers should pay participants at a rate they consider to be fair and in line with the ethical standards of the field (for guidance, see section below titled Ethics of Using Crowdsourced Research Participants).

Disguise the Purpose of the Study Until the Task Is Accepted

The payment amount, an accurate estimate of the time needed to complete the task (Brawley et al. 2016), and a general sense of the task contents (e.g., that it is a survey) are essential information for workers deciding whether to accept a task and should be included in the description of the HIT. Topical details of the study are not necessary and can lead to selection bias. These details should be explained in a consent form within the survey platform that is used to collect data (for an example, see Rose & Segrist 2014). By placing information about study content in the study materials themselves, any potential unmeasurable selection bias caused by the topic of the study will be converted into measurable study attrition.

Reduce and Measure Attrition

Researchers should always measure and report study attrition (for a discussion, see Crump et al. 2013). If necessary, robustness to attrition can be estimated by imputing the highest and lowest possible values for those who attrite and using the result of these analyses to place upper and lower bounds around a potential effect (Gerber & Green 2012). The potential for selective attrition within experimental studies can be addressed by demonstrating equivalence on relevant demographic or psychological characteristics among those who remain across conditions (Jurs & Glass 1971, Schleider & Weisz 2015; for an example, see Kazai et al. 2012).

Selective attrition in experiments can be minimized by ensuring that demands placed on participants are highest before the critical independent and dependent variables (e.g., by placing a difficult task first; Horton et al. 2011) or that participants in the less burdensome condition complete the burdensome task following the dependent measures so that those who drop out following the burdensome task can be excluded regardless of its position in the research study (Rand 2012).

Prescreen Unobtrusively

In many cases prescreening is unnecessary. If a desired population is relatively common, it may be easier to allow all respondents to complete the survey and then remove undesired participants after the fact, or to treat the variable of interest as a moderator in the analysis rather than as an inclusion criterion. In many cases, the latter option is particularly generative because it allows researchers to ask the more nuanced question of whether relationships between variables are larger or smaller in a particular subpopulation than in the population as a whole.

In cases where prescreening is topically or theoretically necessary, or where the population is rare enough that allowing the entire population to participate in the research study is not cost effective, steps can be taken to minimize the likelihood that people will misrepresent themselves in order to be eligible to participate. We suggest unobtrusively prescreening for relevant characteristics within an initial questionnaire and restricting access to the longer survey to workers who meet the desired criteria. In a particularly clever example of unobtrusive prescreening, Wiens & Walker (2015) had participants complete an initial survey on beverage preferences, which was actually a screening survey for a research study on alcoholism (see also Reese & Veilleux 2016).

When prescreening, relevant screening characteristics should be measured again when workers are recontacted so that inconsistent responders can be excluded from analysis. For example, Carr (2014) asked cancer survivors to specify the type of cancer they were diagnosed with at two different time points and excluded inconsistent responders from analysis. Alternatively, factual knowledge that correlates highly with the desired prescreening characteristics can be used as an additional safeguard against workers who have likely misrepresented themselves. For example, Lynn (2014) asked participants who claimed to be veterans to order military insignia by rank.

Prevent Duplicate Workers

We recommend that researchers take steps to minimize repeated participation by workers across related studies. Specific workers can be prevented from participating in a study by assigning them a Qualification value and specifying that workers who posess this Qualification value are denied access to the HIT. Qualifications can only be assigned to workers who have previously completed work for a requester, but if researchers are aware of labs conducting similar research, worker lists can be shared and duplicate workers blocked from both labs using external software such as

TurkGate (Goldin & Darlow 2013) or Qualtrics (Peer et al. 2012). Alternatively, if data can be linked to worker identification numbers, duplicate responses can be deleted after the fact (e.g., Tylka & Wood-Barcalow 2015).

HIT acceptance ratio (HAR): the proportion of a worker's completed human intelligence tasks that has been approved by their requester

Instructional manipulation check: a task that implies one response but instructs a different response. Used to measure whether respondents read and follow directions

Avoid Obtrusive Attention Checks

Many researchers measure worker attentiveness by using trick questions that have an obvious correct choice (e.g., "While watching the television, have you ever had a fatal heart attack?"; Paolacci et al. 2010) or by presenting a superficially easy question along with detailed instructions that ask the participant to do something not implied by the question structure (e.g., a multiple choice question with instructions to click the question title; Oppenheimer et al. 2009).

These measures are probably less effective at improving true data quality obtained from MTurk samples than researchers may assume. Workers expect attention checks and are especially likely to recognize those that are copied verbatim from earlier research studies or even those that share a structurally similar format (Hauser & Schwarz 2015a). Further, performance on different attention checks in the same survey shows only a modest correlation, and associations between attentiveness items are equally strong regardless of their proximity, suggesting that attention to one question cannot be assumed to guarantee attention to measures administered in close proximity and that they probably measure an individual difference in attentiveness rather than current attentiveness (Berinsky et al. 2014). Thus, attention checks exclude some unknown combination of novice MTurk workers who are distracted or who have characteristics related to their ability to pass these items (such as reading ability). A final argument against attention checks is that when participants notice them, they will adopt a more careful and deliberative processing style, which can change responses to subsequent questions (Hauser & Schwarz 2015b).

As an alternative to attention checks, MTurk workers can be selected on the basis of how many HITs they have successfully completed in the past [their HIT acceptance ratio (HAR)]. HAR alone does an excellent job discriminating attentive from inattentive workers (Peer et al. 2014). Alternatively, if data quality at the time of data collection is crucial, researchers would be better served by more relevant and less obtrusive approaches to quality control, such as instructional manipulation checks (for experiments) and measures of split-half reliability on synonymous or antonymous scale items (e.g., Behrend et al. 2011; for a detailed discussion, see Huang et al. 2012).

Use Novel Research Materials When Appropriate

Good reasons sometimes exist for using standardized methods. For example, it may be important to compare questionnaire responses to normed data, ensure that the measurement of a construct is reliable and valid, or directly replicate an earlier research finding. At other times, standardization is less important and materials (particularly experimental paradigms) are reused because doing so is easy. Because workers may be familiar with common measures, which can affect data quality, researchers should ensure that their decision to reuse materials is deliberate and theoretically justified rather than a matter of convenience.

Monitor Cross Talk

As previously noted, cross talk is relatively rare, but it can occur. We recommend that requesters ask participants whether they discovered the HIT somewhere other than MTurk and to paste a link to the site that referred them. This way researchers can learn what was discussed about their HIT during data collection.

Pilot Test Studies and Provide an Outlet for Worker Comments

MTurk HITs can be accepted and completed by many workers very quickly, making any mistakes costly to researchers, both in terms of participant payments and in terms of exposing the limited population of naïve participants to study materials. Studies should always be piloted on a small number of workers before they are posted to the entire sample to ensure that all random assignment, branching, and piping work properly. If the study will focus on a relatively rare subpopulation, researchers should consider pilot-testing materials on a more common population first. Including an open-ended question at the end of the survey, in which workers can volunteer information, makes it easier to identify troublesome aspects of the survey that were overlooked in, or developed after, the pilot study.

Reporting Methods and Results

Transparent methods and results are essential to generating reproducible research and interpreting research findings and have recently become a topic of renewed interest among research methodologists (e.g., Nosek et al. 2015). Sample demographic characteristics should be collected and reported: Because MTurk workers self-select into studies, the demographic characteristics of one MTurk study cannot be assumed to generalize to another. Information about recruitment procedures should be provided to enable the replication of results by other researchers. Essential details include the minimum HAR, minimum number of completed HITs, worker nationality, and any other qualifications used to restrict worker eligibility. Researchers should also report measurements of attrition. If there are different experimental or quasi-experimental conditions, attrition should be reported for each. Finally, if workers are excluded (e.g., for providing poor-quality data or failing to fulfill screening criteria), then researchers should disclose the number of workers excluded, the criteria used to exclude these workers, and how the exclusion criteria were determined, including whether they were determined before or after beginning data collection.

ETHICS OF USING CROWDSOURCED RESEARCH PARTICIPANTS

Research conducted on MTurk can raise questions not addressed by ethical standards tailored to more traditional forms of recruitment and data collection. For example, the current version of the American Psychological Association (APA) Code of Ethics, which was drafted in 2002 and most recently amended in 2010 (Am. Psychol. Assoc. 2010), does not reflect recent developments in Internet technology. In a recent review of the state of Internet research in psychology, Gosling & Mason (2015) highlight that researchers have less control over the conditions under which their studies are completed when data are collected online. As with data collected online from other sources, researchers cannot directly verify that MTurk participants have thoroughly reviewed and understand consenting materials. Nor can researchers intervene and provide debriefing materials when participants encounter technical problems and cannot complete a study. In online research studies on clinical populations, it is also difficult to provide adequate access to local mental health resources in the event that participants are distressed by research materials, experience a psychiatric emergency during the course of the study, or endorse suicidal ideation.

Compensation and anonymity have been identified by social scientists and workers themselves as two particularly concerning ethical issues in MTurk research. In practice, incentives provided to workers are smaller than those offered to traditional paid subject pools but are higher than those offered to participant pools that rely on volunteer participants (e.g., undergraduate subject pools and research funded by federal agencies). Horton & Chilton (2010) identified an average wage of $1.38/hour on MTurk, although this study included international workers. Wages among US workers are likely somewhat higher. Recently, workers have collectively written a document

recommending a pay rate of 10 cents per minute (**http://wiki.wearedynamo.org/index.php?title=Fair_payment**). It is also possible to pay too much. When one is considering appropriate payment, fairness relative to wage rates external to MTurk must be counterbalanced by APA ethical guidelines that prevent researchers from offering coercively high incentives. Internal review boards also often specify that compensation should align with community norms.

Beyond the actual amount of payment, researchers must make efforts to ensure that the estimated time to complete a study is correct and that payment procedures (e.g., time to payment and reasons for denying payment) are clearly articulated before workers accept the HIT (Martin et al. 2014). Researchers should also be mindful of the power imbalance between requesters and workers on MTurk. Requesters have the final say about who gets paid, and there is no formal mechanism to moderate disputes between requesters and workers (Silberman et al. 2010). Requesters' payment decisions also directly influence worker reputation and access to work, whereas workers are forced to informally track and report on requester reputation in forums outside of MTurk (e.g., Turkopticon; Irani & Silberman 2013).

Although MTurk can be confidential, it is not technically anonymous; worker IDs also serve as Amazon IDs and may be linked to personally identifying information disclosed in user profiles or product reviews on Amazon.com (Lease et al. 2013). Further, unlike in a controlled lab setting, some workers may be observed while they are working or may complete studies on unsecured Internet connections. MTurk workers are concerned about their privacy and anonymity and are perhaps more sensitive to issues of privacy online than is the general population (Kang et al. 2014, Schnorf et al. 2014).

Researchers cannot protect participants completely from a breach of confidentiality, but they should take care to explain the relevant risks to participants, use accurate language when describing the protections available (e.g., avoid the term "anonymous" in consent materials), and, of course, never seek out or share potentially identifying information (e.g., by making a data set that includes worker IDs publicly available or by listing the IDs of workers who participated in a study with specific inclusion criteria).

While MTurk introduces unique ethical problems, it also provides some advantages for conducting ethical research. Although identities are difficult to verify online (Buchanan & Williams 2010), the independent registration process required by MTurk compels participants to certify basic information of high ethical relevance—for example, that they are adults—both directly (e.g., providing their birthdate or checking a box indicating that they are at least 18 years of age) and indirectly (e.g., providing bank account information, which is less likely to be available to minors). Also, as is true of Internet research more broadly (Buchanan & Williams 2010, Gosling & Mason 2015), MTurk affords relatively more anonymity than do traditional face-to-face methods of data collection because studies can be completed at home without any direct interaction between the research team and participant. Perhaps as a result, MTurk users report greater comfort reporting on psychological variables online than in person (Shapiro et al. 2013). Online research may also be perceived as less coercive because it reduces barriers to exiting a study, which requires only that participants close a browser window.

CONCLUSION

Numerous techniques and platforms have been developed that use the Internet to connect researchers to participants, including forums, online communities, and social media channels where researchers can post links to studies; commercial online panels; and researcher-administered websites, such as yourmorals.org or Project Implicit. Amazon's Mechanical Turk is a specific example of one kind of resource—an online microtask labor market—that offers readily available

samples and a simple and secure payment infrastructure. As a result, MTurk and similar platforms are as convenient as online panels but cost less and offer more control (for better or worse) in sample selection.

The widespread popularity of MTurk has led to equally widespread questions about its soundness as a subject pool. Numerous investigations of the characteristics of MTurk workers have made it one of the most well understood convenience samples available to researchers, while raising awareness among psychologists of the kinds of issues faced by online convenience samples in general. One of the major, if unappreciated, benefits of MTurk is that its flaws are made more readily detectible by the particular way that MTurk tracks participants and the degree of researcher interest in this platform.

It is possible and—based on the historical trajectory of technology—even likely that new and unimagined methods of recruiting participants will replace MTurk. Any replacement platform is likely to possess many of the attributes that have made MTurk so popular. On the balance, a replacement is also likely to retain at least some of MTurk's flaws, as many of these are inherent to online convenience samples in general.

In sum, methodological evaluations of MTurk suggest that it is a potentially valuable tool for clinical researchers—or at least as valuable as the other nonprobability samples that are used routinely and often without scrutiny. Moreover, researchers have demonstrated that MTurk can be used in innovative ways to conduct a wide range of projects of clinical relevance, including longitudinal and intervention research. For many situations, the advantages of using MTurk seem to outweigh the disadvantages. However, and representing a significant caveat, like any other tool the quality of MTurk data is determined by how the data are used. MTurk is not appropriate for all research questions, populations, or circumstances and researchers should fully understand its drawbacks and limitations before they elect to use MTurk. There is ultimately no substitute for careful study design and documentation.

SUMMARY POINTS

1. Mechanical Turk (MTurk) is a fast and cost-effective way to collect nonprobability samples that are more diverse than those typically used by psychologists.
2. MTurk samples differ from the population as a whole in several important ways. They are younger, more liberal, and more educated, and they include more Caucasians and Asians. Of particular relevance to clinical psychologists, they also have a unique profile of clinically relevant symptoms, particularly elevated levels of social anxiety, difficulty with emotion regulation, and characteristics associated with ASDs.
3. Individual samples recruited from MTurk are nonprobability samples of the available pool of workers. Studies may differ in the demographic characteristics of those they recruit, and selection bias is a potential threat to experimental validity.
4. MTurk workers are about as honest as participants drawn from other convenience samples and are usually consistent in their self-reported demographic and psychological trait information. However, MTurk workers will use deception when financially incentivized to do so.
5. Like those recruited from other web panels, participants recruited on MTurk are likely to have completed numerous other experiments. All available evidence suggests that this attenuates the relationship between measured variables.

6. Unlike other online samples, MTurk respondents can be tracked through their unique identification numbers, which allows workers to be recontacted (permitting unobtrusive prescreening and other longitudinal methods) and makes it possible to allow or prevent specific workers from participating in research studies.

FUTURE ISSUES

1. All available evidence suggests that known threats to the internal validity of research studies on MTurk attenuate true relationships. It is a pressing concern to understand whether there are threats to internal validity that can lead to spurious results. Selection effects are one particularly pressing area of future research.
2. Malingering scores on MTurk are high, but these measures have not been normed on the dominant cohort on MTurk. Unobtrusive, open access measures of malingering that are suitable for web-based research are needed as an alternative to existing measures.
3. Evidence of intertemporal differences in the population available on MTurk is limited, but intuitively, respondents recruited at different times of the day or on different days of the week should differ in terms of place of residence, type of occupation, and perhaps psychological profile. Additional research is needed to clarify whether such differences exist.
4. Some workers are much more productive than others. Little is known about whether workers with different levels of productivity differ in terms of their psychological profiles and whether such differences are of interest to clinical psychologists.
5. Repeated participation in studies is known to change self-reported attitudes, presumably because asking people questions gives them an opportunity to carefully consider topics that might not otherwise occur to them. It is not known whether clinical self-report items are subject to similar effects, and if so, whether this represents a threat to internal validity or simply the reduction of measurement error.
6. English-language studies consistently demonstrate that the quality of Indian data is quite poor. It is not known whether deficiencies in data quality can be overcome by using culturally appropriate or native-language tasks and instruments.
7. It is not currently known to what extent research findings—particularly on data quality—generalize from MTurk to other similar online labor markets. Benchmarking studies are needed to provide insight into the viability of alternative platforms.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Allan NP, Norr AM, Macatee RJ, Gajewska A, Schmidt NB. 2015. Interactive effects of anxiety sensitivity and emotion regulation on anxiety symptoms. *J. Psychopathol. Behav.* 37:663–72

Am. Psychol. Assoc. 2010. 2010 Amendments to the 2002 "Ethical principles of psychologists and code of conduct." *Am. Psychol.* 65:493

Andersson G. 2016. Internet-delivered psychological treatments. *Annu. Rev. Clin. Psychol.* 12:157–79

Andover MS. 2014. Non-suicidal self-injury disorder in a community sample of adults. *Psychiatry Res.* 219:305–10

Arbisi PA, Ben-Porath YS. 1995. An MMPI-2 infrequent response scale for use with psychopathological populations: the Infrequency-Psychopathology Scale, F(p). *Psychol. Assess.* 7:424–31

Arch JJ. 2014. Cognitive behavioral therapy and pharmacotherapy for anxiety: treatment preferences and credibility among pregnant and non-pregnant women. *Behav. Res. Ther.* 52:53–60

Arch JJ, Twohig MP, Deacon BJ, Landy LN, Bluett EJ. 2015. The credibility of exposure therapy: Does the theoretical rationale matter? *Behav. Res. Ther.* 72:81–92

Arditte KA, Çek D, Shaw AM, Timpano KR. 2015. The importance of assessing clinical phenomena in Mechanical Turk research. *Psychol. Assess.* doi:10.137/pas0000217

Austin EJ. 2005. Personality correlates of the broader autism phenotype as assessed by the Autism Spectrum Quotient (AQ). *Personal. Individ. Differ.* 38:451–60

Azzam T, Jacobson MR. 2013. Finding a comparison group: Is online crowdsourcing a viable option? *Am. J. Eval.* 34:372–84

Barnhoorn JS, Haasnoot E, Bocanegra BR, van Steenbergen H. 2014. QRTEngine: an easy solution for running online reaction time experiments using Qualtrics. *Behav. Res. Methods* 530:1–12

Behrend TS, Sharek DJ, Meade AW, Wiebe EN. 2011. The viability of crowdsourcing for survey research. *Behav. Res. Methods* 43:800–13

Benoit K, Conway D, Lauderdale BE, Laver M, Mikhaylov S. 2015. Crowd-sourced text analysis: reproducible and agile reproduction of political data. *Am. Polit. Sci. Rev.* In press

Beramendi P, Duch RM, Matsuo A. 2014. When lab subjects meet real people: comparing different modes of experiments. Presented at *Asian Polit. Methodol. Conf., Taipai*

Berinsky AJ, Huber GA, Lenz GS. 2012. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Polit. Anal.* 20:351–68

Berinsky AJ, Margolis MF, Sances MW. 2014. Separating the shirkers from the workers? Making sure respondents pay attention on self administered surveys. *Am. J. Polit. Sci.* 58:739–53

Bernal DR. 2014. *Acculturation, acculturative stress, social status, and well-being among English language proficient immigrants.* PhD Thesis, Univ. Miami, Coral Gables

Black D, Gates G, Sanders S, Taylor L. 2000. Demographics of the gay and lesbian population in the United States: evidence from available systematic data sources. *Demography* 37:139–54

Blumberg SJ, Luke JV. 2007. Coverage bias in traditional telephone surveys of low-income and young adults. *Public Opin. Q.* 71:734–49

Boynton MH, Richman LS. 2014. An online daily diary study of alcohol use using Amazon's Mechanical Turk. *Drug Alcohol. Rev.* 33:456–61

Brawley AM, Pury CL. 2016. Work experiences on MTurk: job satisfaction, turnover, and information sharing. *Comput. Hum. Behav.* 54:531–46

Buchanan T, Williams JE. 2010. Ethical issues in psychological research on the Internet. In *Advanced Methods for Conducting Behavioral Research*, ed. SD Gosling, JA Johnson, pp. 255–71. Washington, DC: Am. Psychol. Assoc.

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

Bui DC, Myerson J, Hale S. 2015. Age-related slowing in online samples. *Psychol. Record.* doi: 10.1007/s40732-015-0135-2

Burke SE, Wang K, Dovidio JF. 2014. Witnessing disclosure of depression: Gender and attachment avoidance moderate interpersonal evaluations. *J. Soc. Clin. Psychol.* 33:536–59

Cabrera LY, Fitz NS, Reiner PB. 2014. Empirical support for the moral salience of the therapy-enhancement distinction in the debate over cognitive, affective and social enhancement. *Neuroethics* 8:243–56

Carr A. 2014. An exploration of Mechanical Turk as a feasible recruitment platform for cancer survivors. Undergrad. honors thesis: Univ. Colo., Boulder

Casey L, Chandler J, Levine AS, Proctor A, Strolovitch D. 2015. Demographic characteristics of a large sample of US workers. *Mathematica Policy Res. Rep.* In press

Casler K, Bickel L, Hackett E. 2013. Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Comput. Hum. Behav.* 29:2156–60

Cavanagh TM. 2014. Cheating on online assessment tests: prevalence and impact on validity. PhD Thesis, Colo. State Univ., Ft. Collins

Chandler J, Mueller P, Paolacci G. 2014. Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. *Behav. Res. Methods* 46:112–30

Chandler J, Paolacci G, Mueller P. 2013. Risks and rewards of crowdsourcing marketplaces. In *Handbook of Human Computation*, ed. P Michelucci, pp. 377–92. New York: Springer

Chandler J, Paolacci G, Peer E, Mueller P, Ratliff KA. 2015. Using nonnaive participants can reduce effect sizes. *Psychol. Sci.* 26:1131–39

Clifford S, Jerit J. 2014. Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies. *J. Exp. Polit. Sci.* 1:120–31

Cooper EA, Farid H. 2014. Does the Sun revolve around the Earth? A comparison between the general public and online survey respondents in basic scientific knowledge. *Public Underst. Sci.* doi: 10.1177/0963662514554354

Corrigan PW, Bink AB, Fokuo JK, Schmidt A. 2015. The public stigma of mental illness means a difference between you and me. *Psychiatry Res.* 226:186–91

Cougle JR, Hawkins KA, Macatee RJ, Zvolensky MJ, Sarawgi S. 2014. Multiple facets of problematic anger in regular smokers: exploring associations with smoking motives and cessation difficulties. *Nicotine Tob. Res.* 16:881–85

Crump MJ, McDonnell JV, Gureckis TM. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE* 8:e57410

Daly TM, Nataraajan R. 2015. Swapping bricks for clicks: crowdsourcing longitudinal data on Amazon Turk. *J. Bus. Res.* 68:2603–9

Davenport SW, Bergman SM, Bergman JZ, Fearrington ME. 2014. Twitter versus Facebook: exploring the role of narcissism in the motives and usage of different social media platforms. *Comput. Hum. Behav.* 32:212–20

de Leeuw JR. 2014. jsPsych: a JavaScript library for creating behavioral experiments in a Web browser. *Behav. Res. Methods* 47:1–12

Devlin HC, Johnson SL, Gruber J. 2015. Feeling good and taking a chance? Associations of hypomania risk with cognitive and behavioral risk taking. *Cogn. Ther. Res.* 39:473–79

DeWall CN, Pond RS Jr, Carter EC, McCullough ME, Lambert NM, et al. 2014. Explaining the relationship between religiousness and substance use: Self-control matters. *J. Personal. Soc. Psychol.* 107:339–51

Downes-Le Guin T, Mechling J, Baker R. 2006. Great results from ambiguous sources: cleaning Internet panel data. *ESOMAR World Res. Conf. Panel Res.* Amsterdam: ESOMAR

Eriksson K. 2013. Autism-spectrum traits predict humor styles in the general population. *Humor* 26:461–75

Farrell AM, Grenier JH, Leiby J. 2014. Scoundrels or stars? Theory and evidence on the quality of workers in online labor markets. SSRN. doi: 10.2139/ssrn.2424718

Feitosa J, Joseph DL, Newman DA. 2015. Crowdsourcing and personality measurement equivalence: a warning about countries whose primary language is not English. *Personal. Individ. Differ.* 75:47–52

Fergus TA. 2014a. The Cyberchondria Severity Scale (CSS): an examination of structure and relations with health anxiety in a community sample. *J. Anxiety Disord.* 28:504–10

Fergus TA. 2014b. Are "not just right experiences" (NJREs) specific to obsessive-compulsive symptoms? Evidence that NJREs span across symptoms of emotional disorders. *J. Clin. Psychol.* 70:353–63

Fergus TA, Bardeen JR. 2014. Emotion regulation and obsessive-compulsive symptoms: a further examination of associations. *J. Obsessive-Compuls. Relat. Disord.* 3:243–48

Fergus TA, Rowatt WC. 2015. Uncertainty, God, and scrupulosity: Uncertainty salience and priming God concepts interact to cause greater fears of sin. *J. Behav. Ther. Exp. Psychiatry* 46:93–98

Fergus TA, Valentiner DP, McGrath PB, Gier-Lonsway SL, Kim HS. 2012. Short forms of the Social Interaction Anxiety Scale and the Social Phobia Scale. *J. Personal. Assess.* 94:310–20

Fort K, Adda G, Cohen KB. 2011. Amazon Mechanical Turk: gold mine or coal mine? *Comput. Linguist* 37:413–20

Frederick S. 2005. Cognitive reflection and decision making. *J. Econ. Perspect.* 19:25–42

Gates GJ. 2014. *LGBT demographics: comparisons among population-based surveys.* Williams Inst., UCLA Sch. Law

Gerber AS, Green DP. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: Norton

Goldin G, Darlow A. 2013. TurkGate, Version 0.4.0. *Software.* **http://www.gideongoldin.github.com/TurkGate/**

Goodman JK, Cryder CE, Cheema A. 2013. Data collection in a flat world: the strengths and weaknesses of Mechanical Turk samples. *J. Behav. Decis. Mak.* 26:213–24

Gootzeit JH. 2014. *ACT process measures: specificity and incremental value*. PhD Thesis, Univ. Iowa, Iowa City

Gosling SD, Mason W. 2015. Internet research in psychology. *Annu. Rev. Psychol.* 66:877–902

Green AE, Kenworthy L, Mosner MG, Gallagher NM, Fearon EW, et al. 2014. Abstract analogical reasoning in high-functioning children with autism spectrum disorders. *Autism Res.* 7:677–86

Greenblatt JB. 2013. *US residential consumer product information: validation of methods for post-stratification weighting of Amazon Mechanical Turk surveys.* Lawrence Berkeley Natl. Lab., LBNL Pap. LBNL-6163E, Berkeley, CA

Greenblatt JB. 2014. *US residential miscellaneous refrigeration products: results from Amazon Mechanical Turk surveys.* Lawrence Berkeley Natl. Lab., LBNL Pap. LBNL-6537E, Berkeley, CA

Groves RM. 2004. *Survey Errors and Survey Costs*, Vol. 536. New York: Wiley

Gupta N, Martin D, Hanrahan BV, O'Neill J. 2014. Turk-life in India. In *Proc. Int. Conf. Support. Group Work, 18th, Sanibel Island*, pp. 1–11. New York: ACM

Hauser DJ, Schwarz N. 2015a. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behav. Res. Methods.* In press

Hauser DJ, Schwarz N. 2015b. It's a trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *SAGE Open* 5:1–5

Hawkins RXD. 2014. Conducting real-time multiplayer experiments on the web. *Behav. Res. Methods.* doi: 10.3758/s13428-014-0515-6

Heer J, Bostock M. 2010. Crowdsourcing graphical perception: using Mechanical Turk to assess visualization design. In *Proc. SIGCHI Conf. Hum. Factors Comput. Syst., 30th, Atlanta*, pp. 203–12. New York: ACM

Higgins C, McGrath E, Moretto L. 2010. MTurk crowdsourcing: a viable method for rapid discovery of Arabic nicknames. In *Proc. NAACL HLT 2010 Workshop on Creat. Speech and Language Data with Amazon's Mechanical Turk, Los Angeles*, pp. 89–92. Stroudburg, PA: Assoc. Comput. Linguist.

Hillygus DS, Jackson N, Young M. 2014. Professional respondents in non-probability online panels. In *Online Panel Research: A Data Quality Perspective*, ed. M Callegaro, R Baker, J Bethlehem, AS Göritz, JA Krosnick, PJ Lavrakas, pp. 219–37. Chichester, UK: Wiley

Horton JJ, Chilton LB. 2010. The labor economics of paid crowdsourcing. In *Proc. ACM Conf. Electron. Commer., 11th, Cambridge, MA*, pp. 209–18. New York: ACM

Horton JJ, Rand DG, Zeckhauser RJ. 2011. The online laboratory: conducting experiments in a real labor market. *Exp. Econ.* 14:399–425

Howard A, MacCalla J. 2014. Pilot study to evaluate the effectiveness of a mobile-based therapy and educational app for children. In *Proc. Workshop on Mob. Med. Appl., 1st, Memphis*, pp. 12–15. New York: ACM

Huang JL, Curran PG, Keeney J, Poposki EM, DeShon RP. 2012. Detecting and deterring insufficient effort responding to surveys. *J. Bus. Psychol.* 27:99–114

Huff C, Tingley D. 2014. "*Who are these people?": evaluating the demographic characteristics and political preferences of MTurk survey respondents*. Work. Pap., Dep. Gov., Harvard Univ.

Ipeirotis P. 2010. *Demographics of Mechanical Turk*. Work. Pap., Stern Sch. Bus., New York Univ.

Irani LC, Silberman M. 2013, April. Turkopticon: interrupting worker invisibility in Amazon Mechanical Turk. In *Proc. SIGCHI Conf. Hum. Factors Comput. Syst., 31st, Paris*, pp. 611–20. New York: ACM

Jahnke S, Imhoff R, Hoyer J. 2015. Stigmatization of people with pedophilia: two comparative surveys. *Arch. Sex. Behav.* 44:21–34

Jiang L, Wagner C. 2014. *Participation in micro-task crowdsourcing markets as work and leisure: the impact of motivation and micro-time structuring.* Presented at Collect. Intell. Conf., June 10–12, MIT, Cambridge, MA

Johnson DR, Borden LA. 2012. Participants at your fingertips using Amazon's Mechanical Turk to increase student-faculty collaborative research. *Teach. Psychol.* 39:245–51

Johnson PS, Herrmann ES, Johnson MW. 2015. Opportunity costs of reward delays and the discounting of hypothetical money and cigarettes. *J. Exp. Anal. Behav.* 103:87–107

Jones DN, Olderbak SG. 2014. The associations among dark personalities and sexual tactics across different scenarios. *J. Interpers. Violence* 29:1050–107

Jones DN, Paulhus DL. 2011. The role of impulsivity in the dark triad of personality. *Personal. Individ. Differ.* 51:679–82

Jurs SG, Glass GV. 1971. The effect of experimental mortality on the internal and external validity of the randomized comparative experiment. *J. Exp. Educ.* 40:62–66

Kang R, Brown S, Dabbish L, Kiesler S. 2014. *Privacy attitudes of Mechanical Turk workers and the US public.* Presented at SOUPS Workshop on Privacy Personas and Segmentation, 10th, Menlo Park, CA, July 9–11, pp. 37–47. Berkeley, CA: USENIX Assoc.

Kazai G, Kamps J, Milic-Frayling N. 2012. The face of quality in crowdsourcing relevance labels: demographics, personality and labeling accuracy. In *Proc. Int. Conf. Inf. Knowl., 21st, Maui*, pp. 2583–86. New York: ACM

Khanna S, Ratan A, Davis J, Thies W. 2010. Evaluating and improving the usability of Mechanical Turk for low-income workers in India. In *Proc. Symp. Comput. Dev., 1st, London*, pp. 453–56. New York: ACM

Kim J. 2014. *Does stigma against smokers really motivate cessation? A moderated mediation model on the effect of anti-smoking campaigns promoting smoker-related stigma on cessation intentions.* Master's thesis, Univ. Wis., Milwaukee

Kittur A, Chi EH, Suh B. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proc. SIGCHI Conf. Hum. Factors Comput. Syst., 26th, Florence*, pp. 453–56. New York: ACM

Klein RA, Ratliff KA, Vianello M, Adams RB Jr, Bahník Š, et al. 2014. Investigating variation in replicability: a "many labs" replication project. *Soc. Psychol.* 45:142–52

Komarov S, Reinecke K, Gajos KZ. 2013. Crowdsourcing performance evaluations of user interfaces. In *Proc. SIGCHI Conf. Hum. Factors Comput. Syst., 31st, Paris*, pp. 207–16. New York: ACM

Konstam V, Tomek S, Celen-Demirtas S, Sweeney K. 2015. Volunteering and reemployment status in unemployed emerging adults a time-worthy investment? *J. Career Assess.* 23:152–65

Kosara R, Ziemkiewicz C. 2010. Do Mechanical Turks dream of square pie charts? In *Proc. BELIV'10 Workshop. Beyond Time and Errors: Novel Eval. Methods Inf. Vis., 3rd, Atlanta*, pp. 63–70. New York: ACM

Krische SD. 2014. Who is the average individual investor? Numerical skills and implications for accounting research. SSRN. **http://www.dx.doi.org/10.2139/ssrn.2426570**

Lanaj K, Johnson RE, Barnes CM. 2014. Beginning the workday yet already depleted? Consequences of late-night smartphone use and sleep. *Organ. Behav. Hum. Dec.* 124:11–23

Lang ASID, Rio-Ross J. 2011. Using Amazon Mechanical Turk to transcribe historical handwritten documents. *Code4Lib J.* 15:1

Lease M, Hullman J, Bigham JP, Bernstein M, Kim J, et al. 2013. Mechanical Turk is not anonymous. SSRN. doi: 10.2139/ssrn.2228728

Lebowitz MS, Ahn WK, Oltman K. 2015. Sometimes more competent, but always less warm: perceptions of biologically oriented mental-health clinicians. *Int. J. Soc. Psychiatry* 61:668–76

Lebowitz MS, Pyun JJ, Ahn WK. 2014. Biological explanations of generalized anxiety disorder: effects on beliefs about prognosis and responsibility. *Psychiatr. Serv.* 65:498–503

Lebreton P, Mäki T, Skodras E, Hupont I, Hirth M. 2015. Bridging the gap between eye tracking and crowdsourcing. In *Proc. SPIE Conf. Hum. Vis. Electron. Imaging, 20th, San Francisco*. Bellingham, WA: Int. Soc. Opt. Photonics. doi: 10.1117/12.2076745

Lintott C, Reed J. 2013. Human computation in citizen science. In *Handbook of Human Computation*, ed. P Michelucci, pp. 153–62. New York: Springer

Litman L, Robinson J, Rosenzweig C. 2015. The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behav. Res. Methods* 47:519–28

Lynn BMD. 2014. *Shared sense of purpose and well-being among veterans and non-veterans*. PhD Thesis, North Carolina State Univ., Raleigh

Marge M, Banerjee S, Rudnicky A. 2010. Using the Amazon Mechanical Turk for transcription of spoken language. In *Int. Conf. Acoust. Speech Signal Process.* (*ICASSP*), *2010 IEEE, Dallas*, pp. 5270–73. New York: IEEE

Marietta G, Viola J, Ibekwe N, Claremon J, Gehlbach H. 2014. *Improving relationships through virtual environments: how seeing the world through victims' eyes may prevent bullying*. Work. Pap., Grad. Sch. Educ., Harvard Univ.

Martin D, Hanrahan BV, O'Neill J, Gupta N. 2014. Being a Turker. In *Proc. ACM Conf. Comput. Support. Coop. Work Social Comput., 17th, San Francisco*, pp. 224–35. New York: ACM

Mason W, Suri S. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behav. Res. Methods* 44:1–23

Mason W, Suri S, Watts DJ. 2014. Long-run learning in games of cooperation. In *Proc. ACM Conf. Econ. Comput., 15th, Palo Alto*, pp. 821–38. New York: ACM

Mason W, Watts DJ. 2010. Financial incentives and the performance of crowds. *ACM SigKDD Explor. Newsl.* 11:100–8

Mason W, Watts DJ. 2012. Collaborative learning in networks. *PNAS* 109:764–69

McAllister Byun TM, Halpin PF, Szeredi D. 2014. Online crowdsourcing for efficient rating of speech: a validation study. *J. Commun. Disord.* 53:70–83

Miller J. 2006. Online marketing research. In *The Handbook of Marketing Research*, ed. R Grover, M Vriens, pp. 110–32. Thousand Oaks, CA: SAGE

Miller JD, Few LR, Wilson L, Gentile B, Widiger TA, et al. 2013. The Five-Factor Narcissism Inventory (FFNI): a test of the convergent, discriminant, and incremental validity of FFNI scores in clinical and community samples. *Psychol. Assess.* 25:748–58

Mitchell GE, Locke KD. 2015. Lay beliefs about autism spectrum disorder among the general public and childcare providers. *Autism* 19:553–61

Morris RR, Picard R. 2014. Crowd-powered positive psychological interventions. *J. Posit. Psychol.* 9:509–16

Mullinix K, Drukman J, Freese J. 2014. *The generalizability of survey experiments.* Inst. Policy Res., Work. Pap Ser., WP-14–19, Northwest. Univ., Evanston, IL

Nichols AL, Webster GD. 2015. Designing a brief measure of social anxiety: psychometric support for a three-item version of the Interaction Anxiousness Scale (IAS-3). *Personal. Individ. Differ.* 79:110–15

Norr AM, Allan NP, Boffa JW, Raines AM, Schmidt NB. 2015. Validation of the Cyberchondria Severity Scale (CSS): replication and extension with bifactor modeling. *J. Anxiety Disord.* 31:58–64

Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, et al. 2015. Promoting an open research culture. *Science* 348(6242):1422–25

Nouri E, Traaum D. 2014. Cultural differences in playing repeated ultimatum game online with virtual humans. In *Hawaii Int. Conf. System Sci.* (*HICSS*), 47th, Waikoloa, Hawaii, pp. 1213–20. New York: IEEE

O'Connell BH, O'Shea D, Gallagher S. 2016. Enhancing social relationships through positive psychology activities: a randomised controlled trial. *J. Posit. Psychol.* 11:149–62

Oppenheimer DM, Meyvis T, Davidenko N. 2009. Instructional manipulation checks: detecting satisficing to increase statistical power. *J. Exp. Soc. Psychol.* 45:867–72

Palmer CJ, Paton B, Enticott PG, Hohwy J. 2015. "Subtypes" in the presentation of autistic traits in the general adult population. *J. Autism Dev. Disord.* 45:1291–301

Paolacci G, Chandler J. 2014. Inside the Turk: understanding Mechanical Turk as a participant pool. *Curr. Dir. Psychol. Sci.* 23:184–88

Paolacci G, Chandler J, Ipeirotis PG. 2010. Running experiments on Amazon Mechanical Turk. *Judgm. Decis. Mak.* 5:411–19

Papa A, Lancaster NG, Kahler J. 2014. Commonalities in grief responding across bereavement and non-bereavement losses. *J. Affect. Disord.* 161:136–43

Parent J, McKee LG, Rough JN, Forehand R. 2015. The association of parent mindfulness with parenting and youth psychopathology across three developmental stages. *J. Abnorm. Child Psychol.* doi: 10.1007/s10802-015-9978-x

Pearl RL, Puhl RM, Dovidio JF. 2014. Differential effects of weight bias experiences and internalization on exercise among women with overweight and obesity. *J. Health Psychol.* doi: 10.1177/1359105313520338
Peer E, Paolacci G, Chandler J, Mueller P. 2012. Selectively recruiting participants from Amazon Mechanical Turk using Qualtrics. SSRN. **http://www.ssrn.com/abstract=092100631**
Peer E, Vosgerau J, Acquisti A. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behav. Res. Methods* 46:1023–31
Pew Res. Cent. 2010. *Blacks upbeat about black progress, prospects: a year after Obama's election.* **http://www.pewsocialtrends.org/2010/01/12/blacks-upbeat-about-black-progress-prospects/**
Raihani NJ, Bshary R. 2012. A positive effect of flowers rather than eye images in a large-scale, cross-cultural dictator game. *Proc. R. Soc. B* 279:3556–64
Raines AM, Boffa JW, Allan NP, Short NA, Schmidt NB. 2015. Hoarding and eating pathology: the mediating role of emotion regulation. *Compr. Psychiatry* 57:29–35
Ranard BL, Ha YP, Meisel ZF, Asch DA, Hill SS, et al. 2014. Crowdsourcing—harnessing the masses to advance health and medicine, a systematic review. *J. Gen. Intern. Med.* 29:187–203
Rand DG. 2012. The promise of Mechanical Turk: how online labor markets can help theorists run behavioral experiments. *J. Theor. Biol.* 299:172–79
Rand DG, Peysakhovich A, Kraft-Todd GT, Newman GE, Wurzbacher O, et al. 2014. Social heuristics shape intuitive cooperation. *Nat. Commun.* 5:3677
Reese ED, Veilleux JC. 2016. Relationships between craving beliefs and abstinence self-efficacy are mediated by smoking motives and moderated by nicotine dependence. *Nicotine Tob. Res.* 18:48–55
Reidy DE, Berke DS, Gentile B, Zeichner A. 2014. Man enough? Masculine discrepancy stress and intimate partner violence. *Personal. Individ. Differ.* 68:160–64
Reips UD. 2002. Standards for Internet-based experimenting. *Exp. Psychol.* 49:243–56
Rojas SL, Widiger TA. 2014. Convergent and discriminant validity of the Five Factor Form. *Assessment* 21:143–57
Rose P, Segrist DJ. 2012. Difficulty identifying feelings, distress tolerance and compulsive buying: analyzing the associations to inform therapeutic strategies. *J. Ment. Health Addict.* 10:927–35
Rose P, Segrist DJ. 2014. Negative and positive urgency may both be risk factors for compulsive buying. *J. Behav. Addict.* 3:128–32
Ruzich E, Allison C, Smith P, Watson P, Auyeung B, et al. 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. *Mol. Autism* 6:2
Schleider JL, Weisz JR. 2015. Using Mechanical Turk to study family processes and youth mental health: a test of feasibility. *J. Child Fam. Stud.* 24:3235–46
Schnorf S, Sedley A, Ortlieb M, Woodruff A. 2014. *A comparison of six sample providers regarding online privacy benchmarks.* Presented at SOUPS Workshop on Privacy Personas and Segmentation, 10th, Menlo Park, CA, July 9–11. Berkeley, CA: USENIX Assoc.
Shao W, Guan W, Liu T, Clark MA, Merchant RC, et al. 2015. Variations in recruitment yield, costs, speed and participant diversity across Internet platforms in a global study examining the efficacy of an HIV/AIDs and HIV testing animated and live-action video among English- or Spanish-speaking Internet or social media users. *Digit. Cult. Educ.* 7(2)
Shapiro DN, Chandler J, Mueller PA. 2013. Using Mechanical Turk to study clinical populations. *Clin. Psychol. Sci.* 1:213–20
Shaw AD, Horton JJ, Chen DL. 2011. Designing incentives for inexpert human raters. In *Proc. ACM 2011 Conf., Comput. Support. Coop. Work, Hangzhou*, pp. 275–84. New York: ACM
Silberman M, Irani L, Ross J. 2010. Ethics and tactics of professional crowdwork. *XRDS Crossroads ACM Mag. Stud.* 17:39–43
Simcox T, Fiez JA. 2014. Collecting response times using Amazon Mechanical Turk and Adobe Flash. *Behav. Res. Methods* 46:95–111
Simons DJ, Chabris CF. 2012. Common (mis) beliefs about memory: a replication and comparison of telephone and Mechanical Turk survey methods. *PLOS ONE* 7:e51876
Sina S, Kraus S, Rosenfeld A. 2014. Using the crowd to generate content for scenario-based serious-games. arXiv:1402.5034 [cs.AI]

Standish AJ, Benfield JA, Bernstein MJ, Tragesser S. 2014. Characteristics of borderline personality disorder and disgust sensitivity. *Psychol. Rec.* 64:869–77
Stewart N, Ungemach C, Harris AJL, Bartels DM, Newell BR, et al. 2015. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgm. Decis. Mak.* 10:479–91
Sturgis P, Allum N, Brunton-Smith I. 2009. Attitudes over time: the psychology of panel conditioning. In *Methodology of Longitudinal Surveys*, ed. P Lynn, pp. 113–26. Hoboken, NJ: Wiley
Suri S, Goldstein DG, Mason WA. 2011. Honesty in an online labor market. Presented at *AAAI Workshops, 25th, San Francisco*, pp. 61–66. Menlo Park, CA: AAAI Press
Suri S, Watts DJ. 2011. Cooperation and contagion in web-based, networked public goods experiments. *PLOS ONE* 6:e16836
ten Brinke L, Black PJ, Porter S, Carney DR. 2015. Psychopathic personality traits predict competitive wins and cooperative losses in negotiation. *Personal. Individ. Differ.* 79:116–22
Tobin SJ, Loxton NJ, Neighbors C. 2014. Coping with causal uncertainty through alcohol use. *Addict. Behav.* 39:580–85
Tylka TL, Wood-Barcalow NL. 2015. The Body Appreciation Scale-2: item refinement and psychometric evaluation. *Body Image* 12:53–67
Usinger T. 2014. *Effect of internet administered mindfulness training on anxiety and sleep quality.* Undergraduate thesis, Univ. Colo., Boulder
Veilleux JC, Salomaa AC, Shaver JA, Zielinski MJ, Pollert GA. 2015. Multidimensional assessment of beliefs about emotion: development and validation of the Emotion and Regulation Beliefs Scale. *Assessment* 22:86–100
Veilleux JC, Skinner KD, Reese ED, Shaver JA. 2014. Negative affect intensity influences drinking to cope through facets of emotion dysregulation. *Personal. Individ. Differ.* 59:96–101
Victor SE, Klonsky ED. 2014. Daily emotion in non-suicidal self-injury. *J. Clin. Psychol.* 70:364–75
Vlahovic TA, Wang YC, Kraut RE, Levine JM. 2014. Support matching and satisfaction in an online breast cancer support community. In *Proc. ACM Conf. Hum. Factors Comput. Syst., 32nd, Toronto*, pp. 1625–34. New York: ACM
Voracek M, Stieger S, Gindl A. 2001. Online replication of evolutionary psychology evidence: sex differences in sexual jealousy in imagined scenarios of mate's sexual versus emotional infidelity. In *Dimensions of Internet Science*, ed. U Reips, M Bosnjak, pp. 91–112. Lengerich: Pabst Sci.
Weatherly JN, Cookman ML. 2014. Investigating several factors potentially related to endorsing gambling as an escape. *Curr. Psychol.* 33:422–33
Weinberg JD, Freese J, McElhattan D. 2014. Comparing data characteristics and results of an online factorial survey between a population-based and a crowdsource-recruited sample. *Sociol. Sci.* 1:292–310
Wickham RE, Reed DE, Williamson RE. 2015. Establishing the psychometric properties of the Self and Perceived-Partner Authenticity in Relationships Scale-Short Form (AIRS-SF): measurement invariance, reliability, and incremental validity. *Personal. Individ. Differ.* 77:62–67
Wiens TK, Walker LJ. 2015. The chronic disease concept of addiction: helpful or harmful? *Addict. Res. Theory* 23:309–21
Winer ES, Veilleux JC, Ginger EJ. 2014. Development and validation of the Specific Loss of Interest and Pleasure Scale (SLIPS). *J. Affect. Disord.* 152:193–201
Wong CM. 2015. 50 Percent of millennials believe gender is a spectrum, Fusion's Massive Millennial Poll finds. *The Huffington Post.* **http://www.huffingtonpost.com/2015/02/05/fusion-millennial-poll-gender_n_6624200.html**
Wurtele SK, Simons D, Moreno T. 2014. Sexual interest in children among an online sample of men and women: prevalence and correlates. *Sex. Abuse* 26:546–68
Wymbs BT, Dawson AE. 2015. Screening Amazon's Mechanical Turk for adults with ADHD. *J. Attention Disord.* doi: 10.1177/1087054715597471
Yang K, Friedman-Wheeler DG, Pronin E. 2014. Thought acceleration boosts positive mood among individuals with minimal to moderate depressive symptoms. *Cogn. Ther. Res.* 38:261–69