Using Mechanical Turk to Study Clinical Populations

Danielle N. Shapiro, Jesse Chandler and Pam A. Mueller

Clinical Psychological Science published online 31 January 2013
DOI: 10.1177/2167702612469015

Review Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Clinical Psychological Science
XX(X) 1–8

DOI: 10.1177/2167702612469015
http://cpx.sagepub.com
SAGE

Using Mechanical Turk to Study Clinical Populations

Danielle N. Shapiro[1], Jesse Chandler[2], and Pam A. Mueller[2]
[1]Department of Psychology, University of Michigan, Ann Arbor and [2]Department of Psychology, Princeton University

Abstract

Although participants with psychiatric symptoms, specific risk factors, or rare demographic characteristics can be difficult to identify and recruit for participation in research, participants with these characteristics are crucial for research in the social, behavioral, and clinical sciences. Online research in general and crowdsourcing software in particular may offer a solution. However, no research to date has examined the utility of crowdsourcing software for conducting research on psychopathology. In the current study, we examined the prevalence of several psychiatric disorders and related problems, as well as the reliability and validity of participant reports on these domains, among users of Amazon's Mechanical Turk. Findings suggest that crowdsourcing software offers several advantages for clinical research while providing insight into potential problems, such as misrepresentation, that researchers should address when collecting data online.

Keywords

crowdsourcing, Mechanical Turk, clinical methods, Internet research

Received 8/21/12; Revision accepted 10/28/12

Psychologists, public health researchers, psychiatrists, and other social and behavioral scientists rely on access to diverse samples of clinical and subclinical populations to conduct their research. However, identifying and recruiting people with psychiatric symptoms can be challenging and complex (e.g., Patel, Doku, & Tennakoon, 2003; Rinck & Becker, 2005). Symptoms of interest are often rare, and structural barriers make access to some demographic groups, such as low socioeconomic or racial minority populations, especially difficult (Snowdon & Cheung, 1990; Yancey, Ortega, & Kumanyika, 2006). Further complicating matters, people with some disorders may be less likely to volunteer for clinical research or participate in the mental health care system more generally, either as a direct consequence of their symptoms (e.g., social anxiety) or as a result of the stigma associated with psychiatric illness (for discussions, see Corrigan, 2004; Sartorius, 1998).

One solution to the difficulty of identifying participants with psychiatric symptoms or unique demographic characteristics is to conduct research online. Some clinical researchers have taken advantage of the tendency for people with certain clinical disorders to congregate online (e.g., Lambert, Senior, Phillips, & David, 2000) and they have targeted specific difficult-to-reach populations through online communities (e.g., Shapiro & Stewart, 2011). Recently, Amazon's Mechanical Turk (MTurk; http://www.mturk.com) has been adopted by clinical researchers for recruiting healthy control participants (Aharoni, Sinnott-Armstrong, & Kiehl, 2012), as a means of investigating clinically relevant variables (Jones & Paulhus, 2011) and as a potential method of clinical intervention (Morris & Picard, 2012). However, on the whole, adoption of online data collection methods has been slower in the clinical sciences than in other quantitative social and behavioral sciences, perhaps because of concerns about privacy, data quality, and the administrative difficulties of managing online recruitment and payment (Binik, Mah, & Kiesler, 1999).

MTurk is becoming a popular method for recruiting large samples at relatively low cost online. MTurk was created as an online labor market that allows "requesters" to flexibly recruit large numbers of "workers" to complete tasks that are difficult to automate. On MTurk, workers browse Human Intelligence Tasks (HITs) by title, keyword, reward, availability, and so on, and complete HITs of interest. They are paid by requesters upon successful completion of tasks (for an introduction to using MTurk, see Mason & Suri, 2012). This format lends itself well to the collection of survey, experimental, and online intervention data.

There are a number of compelling technical advantages that make MTurk a potentially useful clinical research tool. First,

Corresponding Author:
Danielle N. Shapiro, University of Michigan, 325 Eisenhower Parkway, Suite 300, Ann Arbor, MI 48108
E-mail: dnshap@umich.edu

data can be collected quickly for a minimal cost (Horton & Chilton, 2010). Second, a diverse range of participants can be recruited from across the United States and around the world (e.g., Paolacci, Chandler, & Ipeirotis, 2010). Third, requesters can discretionally reject work, and worker reputation has a direct impact on the future HITs that workers can complete. As a result, the MTurk community is governed by strong norms of honesty and accuracy (Rand, 2012; Suri, Goldstein, & Mason, 2011).

Fourth, although identifiable via worker IDs, workers are anonymous to requesters, which protects respondent anonymity and therefore increases response rates (O'Neil & Penrod, 2001). Identifying information is securely stored in a high-security server, entirely separate from collected responses. Finally, and most important, each worker ID is unique, making it possible to seek out or avoid workers that meet specific criteria (e.g., nationality or worker reputation) and to prevent any individual worker from participating in a HIT more than once, while maintaining participant anonymity (for a discussion of how to prescreen workers, see Chandler, Mueller, & Paolacci, 2013).

Cross-sample investigations comparing MTurk to other methods of data collection have demonstrated that data obtained from its workers are similar to data collected from more traditional subject pools (e.g., college undergraduates or community samples derived from college towns) in a variety of research domains, including political orientation (Berinsky, Huber, & Lenz, 2012), personality characteristics (Buhrmester, Kwang, & Gosling, 2011), and basic biases in decision making (Paolacci et al., 2010). Demographic surveys reveal that MTurk workers are very similar to the national populations from which they are drawn (mostly India and the United States), though they are typically younger and have lower income than the national average (Paolacci et al., 2010).

Despite the potential utility of crowdsourcing sites[1] as a clinical research tool, no work to date has assessed the prevalence of clinical symptoms among users of these sites or the feasibility of attracting participants to complete clinical measures. To fill this gap, we investigated the viability of using MTurk as a research tool for studying mental health. To this end, we measured the descriptive characteristics of several measures of potential clinical relevance (e.g., measures of depression and anxiety) and the frequency of clinically relevant life events (e.g., trauma and drug and alcohol consumption) among MTurk workers. Additionally, we extended research on the viability of MTurk as a research tool by assessing the honesty of workers on subjective measures and their comfort in disclosing responses to sensitive measures online. Although workers tell the truth about seemingly innocuous details, such as age or location, they may be inclined to either underreport or inflate estimates of psychological distress, the latter because they infer that requesters are interested in clinical populations. Therefore, we assessed misrepresentation and inconsistencies in basic demographic information and clinical symptom reporting.

Method

Participants and procedures

Participants were recruited from Amazon's MTurk under the restriction that they were U.S. residents (based on ownership of a U.S. bank account) and had at least a 90% task approval rate for their previous HITs. Non-U.S. residents were excluded because the measures that we used in this study may not be valid for non-English-speaking, non-American samples. Participants provided their informed consent to participate in a study on well-being. The first wave of participants included 530 participants, who were paid $0.75 for approximately 20 minutes. This rate of pay ($2.25 per hour) is above average for MTurk HITs; the median hourly wage for tasks performed on MTurk is $1.38 (Horton & Chilton, 2010). Of the original 530 participants, 33 responded from non-American Internet protocol addresses[2] or provided inconsistent demographic information; these participants were excluded from analyses and were not recruited for participation in Wave 2. Of the remaining 497 participants, 50 scored highly on a measure of malingering (discussed later) and 4 did not complete the measure of malingering; these participants were also excluded from analyses, resulting in a final Wave 1 sample of 443 participants.

In Wave 1, participants completed a mental health survey, which included the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), the Beck Anxiety Index (Beck, Epstein, Brown, & Steer, 1988), the Satisfaction With Life Scale (Diener, Emmons, Larsen, & Griffin, 1985), and the Liebowitz Social Anxiety Scale (Liebowitz, 1987). Participants also completed the Infrequency-Psychopathology Scale (Arbisi & Ben-Porath, 1995) of the Minnesota Multiphasic Personality Inventory–2 (Butcher, Dahlstron, Graham, Tellegen, & Kaemmer, 1989) to assess the extent to which they report implausible psychological symptoms (discussed in the reliability section in Results) and indicated their level of comfort disclosing mental health information to researchers online or in a hypothetical in-person interview (1 = *strongly disagree*, 5 = *strongly agree*). Participants also provided detailed demographic information.

One week after completing the first wave, participants were recontacted to complete a second survey in exchange for $0.80 (Wave 2; for details on how to recontact workers, see Mueller & Chandler, 2012); 397 (80%) participants responded to this request. Wave 1 participants who participated in Wave 2 did not differ from those who did not on demographic or mental health variables, with the exception that they were slightly older (M = 32.80, SD = 10.82, and M = 29.53, SD = 11.60, respectively), $t(489) = 2.58$, $p = .01$. In Wave 2, participants completed the BDI again and the Altman Self-Rating Scale for Mania (Altman, Hedeker, Peterson, & Davis, 1997). Participants also reported whether they had experienced 17 potentially traumatic events (e.g., fire or explosion), using the Life Events Checklist (Blake et al., 1995). Participants were also asked how many times in the average week they consume alcohol, recreational drugs, and prescription drugs (e.g., Vicodin,

Ritalin) for recreational purposes, and they completed a screener for potential substance abuse, the CAGE-AID (Brown & Rounds, 1995). An endorsement of one of the four items on the CAGE-AID results in a positive screen. Finally, participants were asked to provide demographic information again as a mechanism to identify potential data validity issues (discussed in the next section).

Results

Reliability of participant reporting

We assessed participants' honesty in reporting innocuous information by comparing their responses to the demographic items across the two surveys. Survey responses were separated by a week; thus, participants' ability to maintain consistent deception was unlikely. In general, demographics remained consistent across time periods. Only 10 MTurk workers provided demographic information at Wave 2 that differed from that they provided at Wave 1; these participants were excluded from further analyses.

We also examined whether workers tended to fabricate psychiatric symptoms. The Infrequency-Psychopathology Scale consists of Minnesota Multiphasic Personality Inventory–2 items that are rarely endorsed by healthy or clinical populations but are endorsed by those attempting to fake a psychiatric disorder. It has been established as a valid and reliable measure of symptom malingering, even among sophisticated test takers. Arbisi and Ben-Porath (1995) recommended that a *T* score corresponding to five standard deviations above the normed mean designates a malingered response. Three percent (n = 10) of our sample scored above this cutoff. However, we adopted a more conservative cutoff of three standard deviations above the mean (n = 50, 10.1%); participants whose responses fell above this cutoff were excluded from analyses to provide a more conservative measure of symptom prevalence.

Results do not differ substantively when participants who reported inconsistent demographic data or high Infrequency-Psychopathology Scale scores are included in analyses. These participants did, however, report higher symptom levels than other participants on all symptom domains (all $t > 3.55$, all $p > .001$).[3] Participants excluded from analysis did not differ in their demographic characteristics except that they were more likely to be Asian (n = 6 of 50) than other participants (n = 26 of 443), $\chi^2(1, n = 497) = 26.99$, $p < .001$, a finding that should be interpreted with caution given the low number of Asian participants.

Table 1. Sample Demographics and Psychiatric/Health History (in Percentages)

Demographic	Percentage
Age[a]	32.64 (11.63)
Race/ethnicity	
Caucasian or White	83.5
African American or Black	5.2
Hispanic or Latino	4.1
Asian or Asian American	5.9
Native American or other	1.3
Gender (female)	54.0
Sexual orientation	
Heterosexual	91.0
Lesbian or gay	3.6
Bisexual or other	5.4
Currently in a romantic relationship	64.8
Currently married	48.1
Have biological children	37.2
Have stepchildren	5.6
Education	
High school degree or less	13.1
Some college	38.8
College degree	31.8
Some graduate school	5.9
Postgraduate degree	10.4
Employment status	
Employed full-time	38.0
Employed part-time by choice	14.3
Underemployed	8.6
Unemployed by choice	14.7
Unemployed but would prefer not to be	24.4
Household income	
Less than $20,000	20.1
$20,000–$40,000	26.9
$41,000–$60,000	20.5
$61,000–$80,000	15.3
More than $80,000	17.2
Psychiatric and health history	
Ever diagnosed with a psychiatric or psychological condition	21.0
Currently in talk therapy	5.6
Currently taking medication for a psychiatric or psychological condition	12.2
Ever sought treatment for a substance abuse problem	4.3
Ever diagnosed with a chronic illness or physical disability (nonpsychiatric)	16.3

[a] *M* (*SD*).

Demographic characteristics of MTurk

The demographic features of the MTurk sample are summarized in Table 1. As other studies have suggested (Paolacci et al., 2010), MTurk workers are younger and more educated than the general U.S. population and are predominantly Caucasian and middle class. We also measured relationship and employment histories. Approximately 24% of MTurkers reported that they are unemployed but would prefer to be employed (compared to 8% of Americans overall; U.S. Bureau of Labor Statistics, 2012), and an additional 8.6% reported that they were employed part-time but would prefer full-time

employment (i.e., are underemployed). With regard to family and relationships, about 91% of respondents identified themselves as heterosexual, in line with national estimates (Gates, 2011), and many were married and had children (Table 1). Parents had mean of 2.14 (*SD* = 1.17) children.

Psychiatric and health history

Exposure to traumatic events. Sixty-six percent of MTurk participants reported experiencing at least one traumatic event in their lifetimes, and 23% had experienced four or more such events (*M* = 2.07, *SD* = 2.22). The frequency of trauma exposure in this sample is comparable to estimates of trauma exposure in college populations (e.g., Bernat, Ronfeldt, Calhoun, & Arias, 1998; Vrana & Lauterbach, 1994).

Drug abuse and addiction. A large proportion of MTurkers screened positive for possible substance abuse problems; 37.1% answered at least one screening item positively on the CAGE-AID, although relatively few (4.3%) reported ever seeking treatment for substance abuse. Despite the high frequency of positive screens, MTurkers reported only light to moderate average weekly consumption of alcohol and recreational drugs. Just over half (51%) reported having at least one alcoholic beverage per week (*M* = 3.41, *SD* = 5.95; range = 0–40). Heavy drinking can be defined as an average of more than 14 drinks per week for men and more than 7 drinks per week for women (Tsai, Ford, Li, Pearson, & Zhao, 2010); 8.5% of male respondents and 8.7% of female respondents reported consumption above this level. Only 10.6% of MTurkers reported using marijuana during the average week (range, 0–25 times; *M* = 0.62, *SD* = 2.58), and fewer still (3.1% and 1.1%, respectively) reported using prescription drugs for recreational purposes or using other recreational drugs (e.g., cocaine, ecstasy).

Prior diagnoses. Participants were asked to provide information about their psychiatric and health histories (Table 1). Workers reported being diagnosed with a variety of conditions, including relatively common disorders (e.g., major depressive disorder, bipolar mood disorder, posttraumatic stress disorder, attention-deficit/hyperactivity disorder, and generalized anxiety disorder) and a few relatively rarer disorders (e.g., personality disorders and gender dysphoria). Some respondents are currently in talk therapy (5.6%) or taking medication (12.2%) for a psychiatric or psychological condition. When asked if they have a chronic health condition or disability, 16.3% reported that they did; most of these reported chronic pain disorders, such as fibromyalgia or back pain.

Prevalence of clinically relevant symptoms

The prevalence of clinical levels of depression among participants on MTurk was consistent with prevalence in the general population. Approximately 5% of respondents reported clinical levels of depressive symptoms at both Wave 1 (*M* = 11.51, *SD* = 9.15) and Wave 2 (*M* = 10.31, *SD* = 9.02); the 12-month prevalence rate of major depressive disorder in the general population is approximately 7% (Kessler, Chiu, Demler, & Walters, 2005). The levels of general anxiety reported in this sample also paralleled rates of general anxiety identified in the population; 2.9% of MTurkers reported clinical levels of anxiety (*M* = 9.66, *SD* = 10.21) similar to the 3.1% 12-month prevalence rate identified in representative samples (Kessler et al., 2005).

A remarkably high proportion of MTurk users reported clinically significant social anxiety symptoms. Just over half (50.5%) of respondents met clinical criteria for social anxiety (*M* = 37.40, *SD* = 27.37), 7 times the 6.8% 12-month prevalence rate in the general population (Kessler et al., 2005). Though surprising, this finding mirrors other research on Internet usage patterns suggesting that people with social anxiety disorder have a greater online presence than people without social anxiety (Bargh & McKenna, 2004; Shepherd & Edelmann, 2005; Stevens & Morris, 2007; Weidman et al., 2012). With respect to satisfaction with life, MTurkers reported a mean of 19.89 (*SD* = 7.78), somewhat lower than the reported means of 23.5 and 25.8 in Diener and colleagues' (1985) two validation studies for the Satisfaction With Life Scale. Intercorrelations among mental health variables can be found in Table 2.

Data quality

Psychometrics. Consistent with earlier research, the quality of data produced by MTurk workers is high. Responses on the BDI were found to be reliable at both Waves 1 and 2 (α values > .9), and BDI scores across the waves were highly correlated

Table 2. Intercorrelations Among Mental Health Variables

	Depressive symptoms	Generalized anxiety symptoms	Satisfaction with life	Social anxiety
Generalized anxiety symptoms	.57**			
Satisfaction with life	–.59**	–.27**		
Social anxiety	.47**	.48**	–.30**	
Trauma exposure	.12*	.14*	–.16*	.00

*$p < .01$. **$p < .001$.

$r(350) = .87$, suggesting adequate test-retest reliability (Nunnally, 1978). The Beck Anxiety Inventory ($\alpha = .93$) and Liebowitz Social Anxiety Scale ($\alpha = .97$) were also reliable. However, the Altman Self-Rating Scale for Mania was only moderately reliable ($\alpha = .68$), and an improbably high proportion of participants met the clinical cutoff (24%). It may be that manic symptoms cannot easily be measured using self-report, particularly online, or that high ratings on this particular scale (e.g., "I feel happier or more cheerful than usual all of the time") are too easily misinterpreted as nonpathological without an interviewer present to explain them. In any case, given this evidence of potential reliability and validity problems, we excluded manic symptoms from this report and will not discuss them further.

Criterion-related validity of self-reported clinical symptoms. Previously established associations between demographic characteristics and clinical symptoms were also observed in this sample. Mirroring research on the psychiatric correlates of unemployment (e.g., Dooley, Catalano, & Wilson, 1994; Rodriguez, Allen, Frongillo, & Chandra, 1999), unemployed participants reported higher levels of depressive symptoms, $t(440) = -3.65$, $p < .001$, general anxiety, $t(439) = -3.13$, $p = .001$, and social anxiety, $t(437) = -3.27$, $p = .001$, as well as lower levels of life satisfaction, $t(440) = 5.52$, $p < .001$.

Turning to gender, although we did not find gender differences on the continuous measure of depressive symptoms, $t(441) < 1$, women were more likely than men to report symptoms exceeding the clinical cutoff, $\chi^2(1, n = 443) = 4.53$, $p < .05$, mirroring findings in representative samples comparing American men and women (Kessler, Berglund, Demler, Jin, & Walters, 2005). Women also reported more symptoms of general anxiety, $t(440) = 2.00$, $p < .05$, and social anxiety, $t(438) = 3.03$, $p < .01$, and were more likely to meet clinical criteria for social anxiety disorder, $\chi^2(1, n = 440) = 12.42$, $p < .001$, again reflecting findings in the literature (e.g., Weinstock, 1999).

Similar to other research on gender and substance use and abuse (Byrnes, Miller, & Schafer, 1999; Lex, 1991), men reported consuming more alcoholic beverages, $t(347) = 4.05$, $p < .001$, and marijuana, $t(347) = 2.39$, $p < .001$, than women and were more likely to screen positive for potential substance abuse problems, $\chi^2(1, n = 348) = 10.88$, $p = .001$.

Comfort disclosing symptoms online

Workers reported greater comfort disclosing clinical information in an online format ($M = 4.02$, $SD = 0.98$) than they predicted for an in-person interview ($M = 3.57$, $SD = 1.23$), $t(442) = 8.49$, $p < .001$. This was particularly true of people with clinical levels of social anxiety, who were as comfortable disclosing information online as were those without clinical levels of social anxiety, $t(438) = 1.85$, $p < .1$, but who reported less comfort disclosing information in an in-person interview than those without clinical levels of social anxiety, $t(438) = 3.42$, $p = .001$. This suggests that participants may more readily disclose information to mental health researchers when approached online.

Discussion

This study suggests that MTurk (and other crowdsourcing tools, should they emerge) might be a useful resource for accessing and studying clinical and subclinical populations. Not only is data collection fast (it took 2 and 5 days to collect the first and second wave of participants, respectively), but the quality of the data is relatively high. Furthermore, MTurk can be used to complete sophisticated research designs, including longitudinal or survey research, as demonstrated here, as well as experimental or intervention research.

Consistent with earlier research on the psychometric properties of personality scales on MTurk (Buhrmester et al., 2011), mental health measures were found, overall, to demonstrate satisfactory internal reliability and test-retest reliability. Extending this work, we demonstrate the criterion validity of these measures on the MTurk population by replicating associations between psychopathology and established demographic predictors (e.g., unemployment).

Of particular importance to researchers interested in clinical and subclinical populations, the prevalence of depression, general anxiety, and trauma exposure among MTurk workers matches or exceeds the prevalence of these issues in the general population. Thus, researchers can access participants expressing the full range of symptoms of these disorders, as they would in the general population. Moreover, substantial numbers of workers report unemployment, social anxiety, or a positive screen for potential substance abuse problems, suggesting that MTurk might be particularly useful for recruiting participants with these characteristics. These prevalence rates must also be considered in conjunction with the size of the sample available (somewhere in the tens to hundreds of thousands; Tamir, 2011) and the possibility to prescreen and target workers with specific characteristics of interest (for instructions, see Chandler et al., 2013).

Despite not being fully anonymous (in that responses are linked to an ID number), participants reported feeling more comfortable disclosing mental health information online, consistent with earlier research that suggests that visual anonymity is important for comfort in self-disclosure when reporting on sensitive topics (Joinson, 2001). In particular, as a function of their symptoms, people with social anxiety may be especially reluctant to participate in in-person research interviews, making online recruitment an especially useful strategy with this group. Recent intervention research, for example, suggests that online treatment modalities may be useful for people with social anxiety as a result of their reluctance to attend face-to-face therapy sessions or groups (e.g., Andersson et al., 2006; Carlbring et al., 2007), a pattern that our findings suggest might extend to research participation.

Despite the overall honesty and consistency in participant reports and the potential utility of collecting data through crowdsourcing sites, some concerns about data quality did

arise in this investigation. A surprisingly large proportion of workers endorsed items consistent with malingering (i.e., they reported a high frequency of symptoms that should be exceedingly rare), suggesting that a segment of this population may be motivated to fake distress. One possibility is that these participants perceived distress to be of interest to the researcher and thus reported high levels of distress for a variety of reasons that range from selfish (e.g., gaining access to future surveys) to altruistic (e.g., being a cooperative research participant; for a discussion of these issues, see Rosenthal & Rosnow, 2009). Alternatively, the anonymity afforded by MTurk, though in many ways an advantage, may also facilitate malingering or exaggerating symptoms among some participants. Although providing poor data quality affects worker ratings and is therefore disincentivized, data are not linked to participants' names or other identifiers, which may remove a layer of accountability for providing falsified information.

A substantial proportion of respondents also reported that they were currently in the United States but completed the survey from a foreign Internet protocol address. As a result, we recommend screening workers' addresses to ensure that they are from the nation of interest, especially when using measures that have been developed and normed on U.S. populations.

Regarding the suspiciously high malingering scores, two considerations must be counterbalanced if researchers wish to avoid this population. On one hand, efforts to ensure data quality through a priori restrictions are ineffective if workers become aware of a direct link between a specific response and a reward. On the other hand, efforts to ensure data quality by excluding workers post hoc are sometimes abused, either intentionally or unintentionally, by researchers as a way to remove participants whose responses do not align with research hypotheses (for a discussion, see Simmons, Nelson, & Simonsohn, 2011) and in many cases are used without consideration for whether they actually improve data quality (Downs, Holbrook, & Peel, 2012).

Current best practices for ensuring data quality on MTurk are discussed elsewhere (Chandler et al., 2013). However, in sum, we believe that the best approach to gathering high-quality data is to passively prescreen workers for relevant variables (including variables related to data reliability) and to contact only those who meet predefined criteria to complete the measures of interest because this keeps workers blind to inclusion criteria and forces researchers to commit to a data quality strategy. Additional participants, such as those who misrepresent themselves, could be excluded post hoc, but all decisions involving sample selection should be reported transparently. Moreover, careful survey design can go a long way toward improving data quality (for a review, see Couper, 2008). In particular, on MTurk, including questions with factually verifiable answers improves the reliability of subjective judgments (Kittur, Chi, & Suh, 2008). This strategy may have a similar effect on clinically relevant self-report measures.

In addition to these concerns about data quality and as with all online research methods, there are research areas and questions for which data collected on MTurk may be ill equipped to answer. For example, data from populations that may not have insight into their symptoms (such as those experiencing mania or psychotic symptoms) or who do not have reliable access to the Internet (such as those in poverty) may not be feasible to collect online. We encourage researchers to critically evaluate the appropriateness of MTurk for their population or question of interest.

Taken together, the findings reported here suggest that crowdsourcing software in general and Amazon's MTurk in particular might be a useful resource for conducting research on clinical populations, providing that the proper precautionary measures are taken. Researchers who have limited budgets or are in areas of the country that preclude access to diverse populations might benefit from using online modalities such as crowdsourcing as a way to conduct fast, efficient, and high-quality clinical research.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Notes

1. For a discussion on the definition of crowdsourcing, see Estellés-Arolas & Ladrún-de-Guevara, 2012.
2. Although it is possible that some of these workers were American, U.S. workers' earnings can be directly deposited into their bank account by Amazon, while non-U.S. workers are paid in Amazon credit, creating a strong incentive for non-U.S. workers to claim U.S. residency. Furthermore, the distribution of non-U.S. Internet protocol addresses does not correspond with the distribution of U.S. expatriates (most responses were from Eastern Europe, the former USSR, and India yet none from Canada, Mexico, or the United Kingdom).
3. This pattern of responding is atypical because of its indiscriminate nature.

References

Aharoni, E., Sinnott-Armstrong, W., & Kiehl, K. A. (2012). Can psychopathic offenders discern moral wrongs? A new look at the moral/conventional distinction. *Journal of Abnormal Psychology*, *121*, 484–497. doi:10.1037/a0024796.

Altman, E. G., Hedeker, D., Peterson, J. L., & Davis, J. M. (1997). The Altman Self-Rating Mania Scale. *Biological Psychiatry*, *42*, 948–955. doi:10.1016/S0006-3223(96)00548-3

Andersson, G., Carlbring, P., Holmström, A., Sparthan, E., Furmark, T., Nilsson-Ihrfelt, E., . . . Ekselius, L. (2006). Internet-based self-help with therapist feedback and in vivo group exposure for social phobia: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *74*, 677–686. doi:10.1037/0022-006X.74.4.677

Arbisi, P. A., & Ben-Porath, Y. S. (1995). An MMPI-2 infrequent response scale for use with psychopathological populations: The

Infrequency-Psychopathology Scale, F(p). *Psychological Assessment*, *7*, 424–431. doi:10.1037/1040-3590.7.4.424

Bargh, J. A., & McKenna, K. Y. A. (2004). The Internet and social life. *Annual Review of Psychology*, *55*, 573–590. doi:10.1146/annurev.psych.55.090902.141922

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology*, *56*, 893–897. doi:10.1037/0022-006X.56.6.893

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561–571. doi:10.1001/archpsyc.1961.01710120031004

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

Bernat, J. A., Ronfeldt, H. M., Calhoun, K. S., & Arias, I. (1998). Prevalence of traumatic events and peritraumatic predictors of posttraumatic stress symptoms in a nonclinical sample of college students. *Journal of Traumatic Stress*, *11*, 645–664. doi:10.1023/A:1024485130934

Binik, Y. M., Mah, K., & Kiesler, S. (1999). Ethical issues in conducting sex research on the Internet. *Journal of Sex Research*, *36*, 82–90. doi:10.1080/00224499909551971

Blake, D. D., Weathers, F. W., Nagy, L. M., Kaloupek, D. G., Gusman, F. D., Charney, D. S., & Keane, T. M. (1995). The development of a Clinician-Administered PTSD Scale. *Journal of Traumatic Stress*, *8*, 75–90. doi:10.1007/BF02105408

Brown, R. L., & Rounds, L. A. (1995). Conjoint screening questionnaires for alcohol and other drug abuse: Criterion validity in a primary care practice. *Wisconsin Medical Journal*, *94*, 135–140.

Buhrmester, M., Kwang, T., & Gosling, S. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5. doi:10.1177/1745691610393980

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI–2: Manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.

Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin*, *125*, 367–383. doi:10.1037/0033-2909.125.3.367

Carlbring, P., Gunnarsdóttir, M., Hedensjö, L., Andersson, G., Ekselius, L., & Furmark, T. (2007). Treatment of social phobia: Randomised trial of Internet-delivered cognitive-behavioural therapy with telephone support. *British Journal of Psychiatry*, *190*, 123–128. doi:10.1192/bjp.bp.105.020107

Chandler, J., Mueller, P., & Paolacci, G. (2013). *Methodological concerns and advanced uses of Amazon Mechanical Turk in psychological research*. Manuscript submitted for publication.

Corrigan, P. (2004). How stigma interferes with mental health care. *American Psychologist*, *59*, 614–625. doi:10.1037/0003-066X.59.7.614

Couper, M. (2008). *Designing effective Web surveys*. New York, NY: Cambridge University Press.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, *49*, 71–75. doi:10.1207/s15327752jpa4901_13

Dooley, D., Catalano, R., & Wilson, G. (1994). Depression and unemployment: Panel findings from the Epidemiologic Catchment Area Study. *American Journal of Community Psychology*, *22*, 745–765. doi:10.1007/BF02521557

Downs, J. S., Holbrook, M., & Peel, E. (2012, October). *Screening participants on Mechanical Turk: Techniques and justifications.* Paper presented at the annual conference of the Association for Consumer Research, Vancouver, BC, Canada.

Estellés-Arolas, E., & Ladrün-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science, 38*, 189–200.

Gates, G. (2011). *How many people are lesbian, gay, bisexual, and transgender?* Retrieved from http://williamsinstitute.law.ucla.edu/wp-content/uploads/Gates-How-Many-People-LGBT-Apr-2011.pdf

Horton, J. J., & Chilton, L. B. (2010). The labor economics of paid crowdsourcing. In *Proceedings from EC '10: The 11th ACM Conference on Electronic Commerce* (pp. 209–218). New York, NY: ACM. doi:10.1145/1807342.1807376

Joinson, A. N. (2001). Self-disclosure in computer mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology*, *31*, 177–192. doi:10.1002/ejsp.36

Jones, D. N., & Paulhus, D. L. (2011). The role of impulsivity in the dark triad of personality. *Personality and Individual Differences*, *51*, 679–682. doi:10.1016/j.paid.2011.04.011

Kessler, R. C., Berglund, P. A., Demler, O., Jin, R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of *DSM-IV* disorders in the National Comorbidity Survey Replication (NCS-R). *Archives of General Psychiatry*, *62*, 593–602. doi:10.1001/archpsyc.62.6.593

Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of twelve-month *DSM-IV* disorders in the National Comorbidity Survey Replication (NCS-R). *Archives of General Psychiatry*, *62*, 617–627. doi:10.1001/archpsyc.62.6.617

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the ACM Conference on Human Factors in Computing Systems* (pp. 453–456). New York, NY: ACM.

Lambert, M. V., Senior, C., Phillips, M. L., & David, A. S. (2000). Depersonalization in cyberspace. *Journal of Nervous and Mental Disease*, *188*, 764–771. doi:10.1097/00005053-200011000-00007

Lex, B. W. (1991). Some gender differences in alcohol and polysubstance users. *Health Psychology*, *10*, 121–132. doi:10.1037/0278-6133.10.2.121

Liebowitz, M. R. (1987). Social phobia. *Modern Problems of Pharmacopsychiatry*, *22*, 141–173.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*, 1–23. doi:10.3758/s13428-011-0124-6

Morris, R., & Picard, R. (2012). Crowdsourcing collective emotional intelligence. In T. W. Malone & L. von Ahn (Eds.), *Proceedings of Collective Intelligence 2012*. arXiv:1204.2991

Mueller, P., & Chandler, J. (2012). *Emailing workers using Python*. Available at SSRN: http://ssrn.com/abstract=2100601

Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

O'Neil, K. M. & Penrods, D. (2001). Methodological variables in web-based research that may affect results: Sample type, monetary incentives, and personal information. *Behavioral Research Methods, Instruments, & Computers, 33*, 226–233. doi:10.3758/BF03195369

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

Patel, M. X., Doku, V., & Tennakoon, L. (2003). Challenges in recruitment of research participants. *Advances in Psychiatric Treatment*, *9*, 229–238. doi:10.1192/apt.9.3.229

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology, 299*, 172–179. doi:10.1016/j.jtbi.2011.03.004

Rinck, M., & Becker, E. S. (2005). A comparison of attentional biases and memory biases in women with social phobia and major depression. *Journal of Abnormal Psychology*, *114*, 62–74. doi:10.1037/0021-843X.114.1.62

Rodriguez, E., Allen, J. A., Frongillo, E. A., & Chandra, P. (1999). Unemployment, depression, and health: A look at the African-American community. *Journal of Epidemiology and Community Health*, *53*, 335–342. doi:10.1136/jech.53.6.335

Rosenthal, R., & Rosnow, R. L. (2009). *Artifacts in behavioral research*. New York, NY: Oxford University Press.

Sartorius, N. (1998). Stigma: What can psychiatrists do about it? *Lancet*, *352*, 1058–1059. doi:10.1016/S0140-6736(98)08008-8

Shapiro, D. N., & Stewart, A. J. (2011). Parenting stress, perceived child regard, and depression in stepmothers and biological mothers. *Family Relations*, *60*, 533–544. doi:10.1111/j.1741-3729.2011.00665.x

Shepherd, R. M., & Edelmann, R. J. (2005). Reasons for Internet use and social anxiety. *Personality and Individual Differences*, 39, 949–958. doi:10.1016/j.paid.2005.04.001

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi:10.1177/0956797611417632

Snowden, L. R., & Cheung, F. K. (1990). Use of inpatient mental health services by members of ethnic minority groups. *American Psychologist*, *45*, 347–355. doi:10.1037/0003-066X.45.3.347

Stevens, S. B., & Morris, T. L. (2007). College dating and social anxiety: Using the Internet as a means of connecting to others. *CyberPsychology & Behavior*, *10*, 680–688. doi:10.1089/cpb.2007.9970

Suri, S., Goldstein, D. G., & Mason, W. A. (2011, August). Honesty in an online labor market. In L. von Ahn, & P. G. Ipeirotis (Eds.), *Papers from the 2011 AAAI Workshop.* Menlo Park, CA: AAAI Press.

Tamir, D. (2011). 50,000 worldwide mechanical Turk workers. *Techlist*. Retrieved from http://techlist.com/mturk/global-mturk-worker-map.php

Tsai, J., Ford, E. S., Li, C., Pearson, W. S., & Zhao, G. (2010). Binge drinking and suboptimal self-rated health among adult drinkers. *Alcoholism: Clinical and Experimental Research*, *34*, 1465–1471. doi:10.1111/j.1530-0277.2010.01231.x

U.S. Bureau of Labor Statistics (2012). *Labor force statistics from the Current Population Survey*. Retrieved from http://data.bls.gov/timeseries/LNS14000000

Vrana, S., & Lauterbach, D. (1994). Prevalence of traumatic events and post-traumatic psychological symptoms in a nonclinical sample of college students. *Journal of Traumatic Stress*, *7*, 289–302. doi:10.1007/BF02102949

Weidman, A. C., Fernandez, K. C., Levinson, C. A., Augustine, A. A., Larsen, R. J., & Rodebaugh, T. L. (2012). Compensatory Internet use among individuals higher in social anxiety and its implications for well-being. *Personality and Individual Differences*, *53*, 191–195. doi:10.1016/j.paid.2012.03.003

Weinstock, L. S. (1999). Gender differences in the presentation and management of social anxiety disorder. *Journal of Clinical Psychiatry*, *60*, 9–13. doi:10.1007/BF02102949

Yancey, A. K., Ortega, A. N., & Kumanyika, S. K. (2006). Effective recruitment and retention of minority research participants. *Annual Review of Public Health*, *27*, 1–28. doi:10.1146/annurev.publhealth.27.021405.102113