The Turker Blues: Hidden Factors behind Increased Depression Rates in Amazon's Mechanical Turk

Yaakov Ophir[1,2], Itay Sisso[1], Christa Asterhan[1], Refael Tikochinski[1] & Roi Reichart[2]

[1]The Hebrew University of Jerusalem, [2]Technion – Israel Institute of Technology

In press for *Clinical Psychological Science*

June 2019

Author Notes

Yaakov Ophir and Itay Sisso contributed equally to this research. Correspondence should be addressed to Yaakov Ophir, School of Education, Hebrew University of Jerusalem, Mount Scopus, Jerusalem, Israel, 91905. E-mail: yaakov.ophir@mail.huji.ac.il

Abstract

Data collection from online platforms, such as Mechanical Turk (MTurk), has become popular in clinical research. However, there are also concerns about the representativeness and the quality of this data for clinical studies. The present work explores these issues in the specific case of major depression. Analyses of two large data sets gathered from MTurk (N_1 = 2,692 and N_2 = 2,354) revealed two major findings: First, failing to screen for inattentive and fake respondents inflates the rates of major depression artificially and significantly (to 18.5% to 27.5%). Second, after cleaning the data sets, depression in MTurk is still 1.6 to 3.6 times higher than general population estimates. Approximately half of this difference can be attributed to differences in the composition of MTurk samples and the general population (i.e., socio-demographics, health and physical activity lifestyle). Several explanations for the other half are proposed and practical data-quality tools are provided.

Keywords: Depression, crowdsourcing, Mechanical Turk, prevalence of depression, data quality measures

Research practices in clinical psychology are changing. An increasing number of researchers have discovered the advantages of online data collection platforms, which provide researchers with unprecedented accessibility to thousands of registered research participants from diverse demographical backgrounds (Buhrmester, Kwang, & Gosling, 2011), including 'hard-to-reach' clinical populations, such as individuals who suffer from major depression. To date, the most commonly used online data collection platform is Amazon's Mechanical Turk (MTurk). A Google Scholar search (January, 2019) for publications using the phrases "Mechanical Turk / MTurk" and "depression" resulted in over 7500 hits. Moreover, recent reports suggest that depression rates are substantively higher in MTurk, compared to the general population (Arditte et al., 2016; Bunge et al., 2018; McCredie & Morey, 2018; Walters, Crhistakis & Wright, 2018). To the extent that this reflects a genuine difference, increased prevalence makes MTurk an even more convenient and attractive recruitment platform for clinical researchers.

However, concerns have been raised regarding the representativeness and the quality of data collected from MTurk and alike platforms for clinical studies (Chandler & Shapiro, 2016; McCredie & Morey, 2018). If higher rates of depression in MTurk prove to be false or inflated, than this would raise serious questions about the reliability of clinical research outcomes that are based on such data sets. If, on the other hand, it proves to be genuine, then knowledge about the reasons behind increased prevalence rates is not only of interest in and by itself, but also necessary for accurate interpretations of research results, in particular concerning the external validity of findings that are derived from MTurk samples.

In the present work, we examine whether depression is indeed more common in MTurk and explore several reasons that could account for such differences. In particular, we argue that failure to control for random, poor quality and fake responses from MTurk respondents artificially inflates depression estimates and aim to quantify this confounding effect. Finally, we

also explore the extent to which differences between MTurk and the general population on variables that are known to be associated with depression can account for remaining differences in the prevalence rates of depression.

Existing research on depression rates in MTurk

According to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), the twelve-month prevalence of major depressive disorder in the US is approximately 7% (American Psychiatric Association, 2013; Kessler, Petukhova, Sampson, Zaslavsky, & Wittchen, 2012). In the first study to focus on MTurk depression rates, Shapiro and colleagues (2013) used on the Beck Depression Inventory (Beck, Steer, & Brown, 1996) to assess clinical levels of depression. Results showed that prevalence of clinical depression in MTurk was equivalent to the epidemiology of major depressive disorder in the general population. More recent studies, however, consistently report that depression in MTurk is more common than in the general population: McCredie and Morey (2018) used the 344-item Personality Assessment Inventory (PAI; Morey, 1991, 2007) and found that MTurk scores on the PAI sub-scale of depression were moderately higher than in a representative sample of 1,000 US citizens (d = .57). Walters et al. (2018) administered a two-item screening tool for depression (PHQ-2; Arroll et al., 2010) to a sample of 591 young MTurk workers (age < 50). They found that they were two to three times more likely to screen for depression than participants from a National Health and Nutrition Examination Survey (NHANES). Finally, Arditte et al. (2016) found that MTurk workers scored significantly higher on the depression subscale of the Depression, Anxiety, and Stress Scales (DASS-21) than subjects from a nonclinical sample (Osman et al., 2012) (d = .94). Thus, even though most of the abovementioned studies did not include screening tools with formal cutoffs for major depression (and, therefore, do not offer a definite estimate for the prevalence of the disorder in MTurk), taken together they reveal a range of outcomes, from lack of differences, to moderate and even large differences between depression in MTurk and the general population.

A possible explanation for this pattern of mixed findings may be the different types of data quality assurance methods that were applied in each study. Poor quality responses can be the result of workers' inattentiveness, boredom, fatigue, or carelessness. Moreover, MTurk researchers have recently evidenced illicit automated activity that is operated by fake MTurk workers (also known as "bots/cyborgs") (APS, 2018; Bai, 2018; Dennis et al. 2018; Kennedy et al., working paper). Fake or inattentive workers pose a serious threat to the quality of the collected data, especially in highly skewed clinical distributions such as depression: Whereas, most of the responses in such distributions should be concentrated around the zero point (i.e., no depression) (Tomitaka et al., 2018), random false responses bias the clinical distribution towards the center, as fake or inattentive workers assume uniform distributions or distributions that are centered around the middle of the scale. These false responses may even improve artificially the overall scale reliability (Fong, Ho & Lam, 2010). Essentially, this bias would create an inflated misrepresentation of the data according to which artificially increased and allegedly reliable levels of depression would be observed.

Although MTurk operates an internal rating system of their workers, researchers are strongly advised to embed various attention and validity checks in their data collection tools to verify and assure the quality of the collected data (Dunn et al., 2018; Chandler, Shapiro & Sisso, working paper). Our own review of the methodological studies on MTurk has revealed various types of data quality measures, including *infrequency items* (which have only one correct/highly probable answer), *time measurements* (with a minimum reading speed threshold), *person-total correlation*s (capturing the subject's internal consistency relative to the expected patterns based on all other participants), *long string analyses* (which flags participants with long string of identical answers), and *exclusion of non-US IP addresses* and ones that are suspected to be indicative of automated activity.

A close inspection of the data quality assurance methods used in the studies addressing

depression rates in MTurk reveals considerable differences: Arditte et al. (2016) used one criterion of time measurement and excluded 17.3% of their sample. Shapiro et al., (2013) used two criterions of infrequency items and filtered non-US IP addresses, resulting in an exclusion of 15.6% of their sample. Walters et al. (2018) applied one criterion of infrequency items and removed duplicate IP addresses, excluding 7% of their sample. McCredie & Morey (2018) did not exclude any participant from their sample. It is, therefore, reasonable to assume that the different data quality assurance methods and the different exclusion rates would produce different estimates of depression. Moreover, it is well established that different methods spot different types of inattentive participants (Jones et al. 2015; Dunn et al. 2018), which implies that any use of one type of method might fail to catch all or most of the inattentive or fake participants.

The present work

The specific goals of the current research are twofold, namely (1) to estimate the prevalence of depression among MTurk workers in comparison to the general population; and (2) to explore possible reasons behind increased depression estimates among MTurk workers such as poor-quality responses and socio-demographic differences between the MTurk and the general population. Two consecutive studies are presented, each conducted on a large sample of MTurk workers (N_1 = 2,692 and N_2 = 2,354) with a time interval of four months between studies. Together these two samples cover a significant portion of the available MTurk participant pool[1]. In both studies, we used the Patient Health Questionnaire (PHQ-9; Kroenke, Spitzer, & Williams, 2001), the most common and most often validated screening tool for depressive disorders (El-Den, Chen, Gan, Wong, & O'Reilly, 2018). The PHQ-9 consists of nine items,

[1] Previous estimates by Stewart et al. (2015) suggest that an average lab could reach about 7,300 participants within a month time. More recent efforts to estimate the size of the entire MTurk population (using updated methods) range between 100,000-180,000 workers (Difallah et al. 2018).

each targeting one of the DSM-defined symptoms (American Psychiatric Association, 2013). The sensitivity and the specificity of the PHQ-9 cut-off point for major depression have been documented using the "gold standard" criterion structured clinical interview and the tool is recommended over other self-report screening tools (Löwe et al.,2004). Both studies also included a multilayered data quality assurance methodology. De facto, we have applied a procedure that monitored fake workers ("bots/cyborgs") and created a solid *inattentiveness index* that ensures the validity of unsupervised data collection in MTurk. Using four complementary data quality assurance methods specifically calibrated to MTurk populations and clinical assessment tools, the inattentiveness index produced different hierarchical levels of workers' attentiveness estimation (see Method, Study 1).

Using these data-quality measures and other convergent validity tests, Study 1 aims to offer a first, yet reliable estimate of the prevalence of depression in MTurk relative to its acknowledged prevalence in the general population. Study 1 also presents a comparison between the different groups of the inattentiveness index, so as to estimate the effects of poor responses on depression rate estimates in MTurk. In Study 2, we aim to replicate the Study 1 findings in a new sample, while controlling for several additional explanations for any differences in depression rates between MTurk and the general population. In this study, we compare the depression rates in MTurk with an existing database from a recent national representative survey by the Center for Disease Control and Prevention (CDC, 2018) that used the *same* assessment tool for depression (the PHQ-9), but in a face-to-face format. We test also whether the differences in the prevalence of depression between the two samples could be explained by differences in socio-demographic variables (e.g., gender, age, and occupational status), general health factors and physical activity (e.g., sleep and amount of time sitting down). Finally, we also explored whether similar patterns would be attained when using a different indication of depression, namely the use of depression-related psychiatric medications.

Study 1

A large sample of MTurk workers completed a survey battery, including the Patient Health Questionnaire (PHQ-9) for depressive disorders. Convergent validity of the depression scale was established with three additional questionnaires that each target a well-established, close predictors of depression: generalized anxiety disorder (Sartorius, Üstün, Lecrubier, & Wittchen, 1996), depressive rumination (Nolen-Hoeksema, Wisco, & Lyubomirsky, 2008), and loneliness (Cacioppo, Hughes, Waite, Hawkley, & Thisted, 2006). Data quality was assured by screening for suspicious IP addresses that are suspected to be operated by illicit semi-automated MTurk workers ("cyborgs") and a comprehensive inattentiveness index that included multiple attention checks integrated into the survey battery.

Method

Participants and procedure. The procedure of the study has been approved by the Ethics for Research on Human Subjects Committees at the Hebrew University and the Technion – Israel Institute of Technology. Subjects were eligible to participate were US-based MTurk workers who had completed at least 100 HITs with a minimum of 95% success rate, and who owned a Facebook account at the time of recruitment (a requirement resulting from a different study). A total of 2,719 MTurk adult workers (36% female) participated in the study, which ran in several batches during May-July 2018. Twenty-two workers (0.8%) dropped out before completing the full survey, thus reducing the final sample to 2692 MTurk workers. The median completion time was 8.3 minutes (mean time – 11.8 min.), and participants were payed $2 for completion of the survey. The average age of the participants was 34.80 yrs (*SD* = 11.05). The average income per household was $58,400 a year (*SD* = $38,900, *MD* = $55,000). Only 10.8% of the sample did not receive higher education. The majority of MTurk workers (74.3%) had studied or completed a BA college degree, 12.4% completed a master's degree, 1% a doctoral degree, and 1.5% had a

professional degree (JD, MD). Altogether, the online survey included six self-report research questionnaires and eight attention checks/indicators. These measures are described below.

Depression. Depression was measured using the Patient Health Questionnaire-9 (PHQ-9; Kroenke et al., 2001), which consists of 9 items that assess the presence and severity of the nine DSM-based symptoms of depression in the previous two weeks, scored from 0 to 3 (“not at all”, ”several days”, “more than half the days” and “nearly every day”). In the current work we use the term "*major depression*" to describe participants who met the threshold criteria for major depression according to the scoring system described by the developers of the tool (Spitzer, Kroenke & Williams, 1999). According to the developers, the cut-off point for major depression follows the DSM diagnostic criteria. Major depression can therefore be calculated when five or more symptoms receive a score of 2 (“more than half the days”) or 3 (“nearly every day”). Other depressive disorders (e.g., dysthymia) can be diagnosed if 2-4 symptoms are reported with at least a 2 score (“more than half the days”). Corresponding with the DSM criteria, both diagnoses are valid only when one of the key symptoms of depression, low interest and depressed mood, are reported for at least more than half the days (Spitzer et al., 1999). In this scale, the term 'any depression' is used to describe individuals who have either one of the depressive disorders.

Total PHQ-9 scores (range = 0–27) can also be analyzed as a continuous variable that measures the severity of the depressive disorder (Kroenke et al., 2001). The cut-off point of 10 and 15 indicate moderate and severe depression, respectively (Kroenke et al., 2001). A further psychometric analysis of the PHQ-9 has revealed that the this cut-off point for moderate depression (i.e., a total score of 10 points or higher) may actually detect more cases of major depressive disorder than the original method of scoring (Arroll et al., 2010). The reliability and validity of the PHQ-9 have been extensively documented in the literature (El-Den et al., 2018). In the current sample, the internal consistency of the scale was high ($\alpha = .88$).

Three self-report measures were administered to check the convergent validity of the

depression scale, namely the Generalized Anxiety Disorder scale (GAD-7;),α = .91 the Brooding sub-scale from the Ruminative Responses Scale (RRS; Nolen-Hoeksema & Morrow, 1991), and a short version of the UCLA-Loneliness Scale (Russell, Peplau, & Ferguson, 1978; Russell, 1996; α = .92). Detailed descriptions of these measures are provided in the supplementary materials of the study.

Suspicious IP addresses. As described in the introduction, the past months were characterized with growing suspicious automated/semi-automated activities on MTurk originated from bogus workers ("Bots/cyborgs", respectively) (APS, 2018; Bai, 2018; Kennedy et al., working paper). Dennis et al. (2018) found that this new surge of low quality workers is characterized by worker IP addresses that can be traced to a certain type of Internet Service Providers (ISP) known as Virtual Private Servers. Furthermore, there are serious suspicions that these accounts are operated by people from outside the United States, mainly from India (Moss & Litman, 2018) and Venezuela (Kennedy et al., working paper). Therefore, participation in this study was limited to US workers only and applied a newly designed tool to flag MTurk workers whose IP address is suspicious of malicious activity and/or traced to a non-US location (Prims et al., 2018).

Inattentiveness index. To ensure the quality of the unsupervised self-reported data, a designated scale was created. This scale consisted of eight checks based on four different methods. Two checks were based on *infrequency items* (Huang et al., 2015) which have only one correct/highly probable answer. Two checks were based on *time measurements* with reading speed threshold of 10 words per second (Sisso, working paper). One attention check was based on the *person-total correlation* that captures the person internal consistency relative to the expected patterns generated by all other participants (Curran, 2016). Finally, three attention checks relied on *long string analysis*, which flags participants with long string of identical answers. The number of failed attention checks for each participant was counted and a unified 3-

point scale of attentiveness was created: "Attentive workers" who did not fail any attention check received an attentiveness score of zero. "Questionable workers" with one mistake only received one point and "Inattentive workers" who had two or more errors received two points.

Results

The final sample included 2,692 MTurk workers. To ensure the quality of the unsupervised self-reported data the data set was screened for suspicious/non-US IP addresses and the attentiveness scores were calculated for each participants. A total of 236 workers (8.8% of the sample) were identified as suspicious/non-US workers. Not surprisingly, 35.6% of these suspicious IP's also failed our attentiveness test (scored 2 on the inattentiveness scale), compared to 7.4% of non-suspicious IP's, $\chi^2(1) = 193.3$, $p < .001$. From the remaining 2,456, non-suspicious MTurk workers, 181 "Inattentive workers" (7.4%) had failed two or more attention checks, 427 "Questionable workers" (17.4%) had failed one attention check only, and a total of 1848 "Attentive workers" (75.2%) did not fail any of the eight attention checks.

The full, filtered dataset, which included attentive workers only and excluded suspicious workers, demonstrated good psychometric qualities. All measures used in the current study achieved good internal consistency. Descriptive statistics and zero-order correlations of the variables are presented in Table A in the supplementary materials. Consistent with the literature, psychopathology (depression and anxiety) and distress (depressive rumination and loneliness) measures demonstrated high convergent validity (Table A, supplementary materials).

To establish the prevalence of depression among MTurk workers and to examine the effects of poor responses, the depression rates for each one of the sub-group of the study ("Inattentive", "Questionable", and "Attentive" MTurk workers) was calculated separately. Table 1 presents the means scores of the PHQ-9 along with the prevalence of PHQ major depression and of any depression (that includes both major depression and other depressive disorders), according to the validated cut-off points of the PHQ-9.

The initial prevalence of PHQ major depression in the entire sample, including inattentive MTurk workers was 18.5%, that is: almost one in five MTurk workers. Less than a third of the sample (27.2%) scored 0-2 total points (skewness = .77) and only 13.8% were free of depressive symptoms (PHQ-9 = 0). Significant differences were documented between the various groups of the inattentiveness index, $F(3,2688) = 40.80$, $p < .001$. A clear pattern was observed: The more attentive the user, the less frequent the depression estimates in the sample.

The prevalence of major depression among valid, attentive MTurk workers only, was 12.9%. This prevalence increases dramatically among questionable workers (26.0%), inattentive workers (31.5%) and workers with a suspicious IP address (38.6%), $\chi^2(3) = 138.1$, $p<.001$. The prevalence of major depression in the current sample of attentive, valid MTurk workers (12.9%) was significantly higher than the prevalence of major depressive disorder in the general population (7%) according to the DSM (American Psychiatric Association, 2013), $\chi^2 = 99.01$, $p < .001$.

Discussion

The findings from Study 1 demonstrate the importance of including stringent data quality assurance methods in MTurk-based clinical research on psychopathologies. Based on their response patterns to the screening tool, between 26% to 39% of inattentive and suspicious Mturk workers would be screened for major depression in this sample. These exceptionally inflated scores corroborate with our concern that random-false responses might bias clinical skewed distribution and create a misrepresentation of extremely high prevalence of major depression. Second, even after the exclusion of inattentive and suspicious MTurk workers, the prevalence of major depression in this large MTurk sample seems substantively higher (13%) compared to general population estimates (7%). Although this comparison should be carefully interpreted, it corroborates with previous research reporting higher depression scores among samples of MTurk workers (Arditte, Çek, Shaw, & Timpano, 2016; McCredie & Morey, 2018; Walters et al., 2018),

as well as higher rates of other psychopathologies (Goodman, Cryder, & Cheema, 2013; Kosara & Ziemkiewicz, 2010).

Study 1 has several limitations. The recruitment procedure in Study 1 excluded MTurk workers without active Facebook accounts and the depression estimates in MTurk were based on a self-report screening tool (PHQ-9) whereas the DSM estimates are based on face-to-face, structured clinical interviews (Kessler et al., 2012). Difference in prevalence rates could thus be the result of using different screening tools. Indeed, previous large studies that evidenced significantly lower rates of PHQ major depression (5% ~) in the general population (Eisenberg, Gollust, Golberstein & Hefner, 2007; Tomitaka et al., 2018) support the current findings in which MTurk is characterized with unusual high rates of depression. Yet, a direct comparison based on the same screening tool is preferable. Study 2 then seeks to replicate the findings from Study 1 in a new MTurk sample, while addressing these two alternative explanations. In addition, Study 2 was designed to explore the extent to which observed differences in prevalence rates could be attributed to differences in the composition of the MTurk population on a number of individual variables that are known to be associated with depression, such as such as age, income, and education level (American Psychiatric Association, 2013).

Study 2

To address these issues, a new data set was collected from MTurk. We compared its characteristics to a recently collected data set from the National Health and Nutrition Examination Survey (NHANES, hereafter) (CDC, 2018). The NHANES survey is administered biennially to a large, representative sample of the US population, and includes (among other measures) the PHQ-9, as well as many additional socio-demographic, health-related and physical activity characteristics.

In addition, the following changes were made to the MTurk data set collection procedure: First, the inclusion criterion of owning a Facebook account was removed in Study 2. Second,

measures of socio-demographic, health-related and physical activity lifestyle characteristics that were found to be associated with depression in the NHANES study were added to the MTurk survey battery. Third, we added a second indicator of depression to cross-validate the outcomes of the self-reported behavioral measures of depression, namely self-reported use of depression-related psychiatric medications.

In order to test whether measured differences between the prevalence of depression in MTurk and the general population could be attributed to differences in sociodemographic and/or health/physical activity–related variables between the two, we conducted a statistical comparison of the depression rates in each, after controlling for the former.

Method

Participants and procedure. The data collection in Study 2 was conducted 4 months after the completion of Study 1 (October, 2018). The procedures were similar to the procedures of Study 1, with some minor modifications that increased the generalizability of results. These modifications included opening the data collection 24 *hrs* a day for a wide-range of workers' classifications (reputation and experience). A total of 2,444 US-based MTurk workers participated in Study 2 (46.5% female, M_{age} = 35.4, *SD* = 11.3). The median completion time was 7.9 *min* (*M* = 9.5 *min*). Participants were payed $1 upon completion. Ninety workers (3.7%) dropped out before completing the full survey. The final sample of Study 2 comprised 2,354 MTurk workers, including 187 workers (7.9%) who also participated in Study 1. The average income per household was $58,300 a year (*SD* = $42,500, *MD* = $50,000). The majority of the workers (74.3%) had studied or completed a BA college degree, 12.4% completed a master's degree, 1% a doctoral degree, and 1.5% had a professional degree (JD, MD).

In addition to measuring self-reported depression, similar to the NHANES questionnaire, participants were asked to indicate if they had used any prescribed psychiatric medications (i.e., medications that are used to treat mental conditions such as: depression or anxiety) in the past 30

days. Participants who responded positively then indicated the type of the medication, while choosing one or more options from the following: Antidepressants (e.g., Prozac, Zoloft, Effexor, Cymbalta), mood stabilizers (e.g., Lithium carbonate, Tegretol, Lamictal), Benzodiazepine/Anxiolytics (e.g., Xanax, Valium, Ambien, Stilnox), stimulants (e.g., Ritalin, Concerta, Focalin, Adderall), and antipsychotics (e.g., Risperdal, Zyprexa, Seroquel, Thorazine). A positive response to at least one of the first two categories was classified as an indication of the participant using depression-related medications (a dichotomous variable).

CDC - National Health Survey. To conduct a comparison with representative data from a national survey, we analyzed published data from 5,134 individuals who participated in a recent NHANES survey, which was collected in 2015-2016 (CDC, 2018). It was selected as an anchored reference to the current study, because it concerns a representative sample from the US population and is based on the exact same screening tool for depression (i.e., the PHQ-9). The NHANES survey is conducted every two years. Data is collected by a face-to-face interview in the survey participants' homes. The data collected targets information about the physical and mental health of adults and children in the US, as well as their nutritional habits and physical activity.

We extracted significant predictors of depression from the NHANES survey report (CDC, 2018) and included these variables in our own MTurk-based data collection procedure, using identical prompts: (1) socio-demographic variables (Gender, Age, Income, Education, and Work status); and (2) health/lifestyle-related variables (Poor health, Weight status, Physical activity, Time sitting down during the day, and Hours of sleep during the night).

Results

The final sample included 2,354 MTurk workers. As described in Study 1, to ensure the quality of the unsupervised, self-reported data we monitored the data for suspicious IP addresses

and calculated the remaining participants' attentiveness scores using six[2] attention checks. Using the same detection method from Study 1, a total of 614 (26.1%) IP's were flagged as suspicious/non-US addresses. Compared to the first sample, was collected between May-July 2018, this increase in the percentage of suspicious/non-US addresses in October 2018 is dramatic. This increase is not particular to the current study and has been documented in the literature (Kennedy et al., 2018). From the remaining 1,731 MTurk workers, 118 "Inattentive workers" (6.8%) had failed two or more attention checks, 161 "Questionable workers" (9.3%) had failed one attention check only, and a total of 1,461 "Attentive workers" (84.4%) did not fail any of the six attention checks (only 18.7% of the suspicious participants did not fail any of the attention checks). All further analyses and comparisons are conducted on the valid "Attentive workers" only, unless specified otherwise.

Replication of Study 1. First, the test-retest reliability of the PHQ-9 scale was tested by calculating the correlation between PHQ-9 scores of the participants that particpated in both samples. The observed reliability of the PHQ-9 over a period of 4 months among 108 attentive returning workers (who were also attentive in Study 1) was good ($r = 0.79$, $p < .001$). In this sample, the proportion of attentive MTurk workers that met the PHQ-9 criterion of major depression was 11.0%. Despite the modifications in the procedures of the two studies, the observed difference between the prevalence of major depression in Study 1 (12.9%) and Study 2 (11%) was not significant, $\chi^2(1) = 2.78$, $p = .095$. Thus, the findings of Study 2 replicate the Study 1 finding showing increased rates of major depression among MTurk workers, when using the PHQ-9 criterion for major depression. Moreover, similar patterns were found about the effects of including inattentive and invalid MTurk workers. The prevalence of PHQ major

[2] The attention checks applied in Study 2 were similar to the ones used in Study 1. The reason why we used only six in Study 2 vs. eight in Study 1 is that we included only one measure of straight-lining in Study 2 (on the PHQ-9) vs. three in Study 1, as these measures require the use of reasonably long and/or reverse coded scales, which were not included in Study 2 beyond the PHQ-9.

depression in this sample was considerably higher among questionable (26.1%) and inattentive workers (55.1%), $\chi^2(2) = 179.5$, $p < .001$. The prevalence of major depression among suspicious IP participants was even slightly (though not significantly) higher than that of non-suspicious, inattentive participants (61.9%, $\chi^2(1) = 1.92$, $p = .166$).

Comparison with the NHANES survey. To compare the obtained depression scores with the representative national survey that used the same screening tool, an independent sample *t*-test (equality of variances not assumed[3]) was conducted comparing mean PHQ-9 scores in the MTurk (N = 1,461) and the NHANES (N = 5,134) databases. On average, attentive MTurk workers scored higher on depression (M = 6.05, SD = 5.60) than participants in the NHANES survey (M = 3.24, SD = 4.22), $t(1954.2) = 17.76$, $p < .001$, with a medium-large effect size, Cohen's d = 0.615 (95% CI: 0.556 - 0.647). The prevalence of PHQ major depression among attentive/non-suspicious MTurk workers was 11%, which is three times higher than the 3.6% prevalence in the 2015-2016 NHANES sample, $\chi^2(1) = 125.9$, $p < .001$.

Controlling for socio-demographic and health/lifestyle-related differences. A comparison between the two data sets on the depression-related socio-demographic and health/lifestyle-related characteristics revealed significant differences between them on all variables, except for gender (see Table 2). These variables (including gender) were then controlled for in a hierarchical regression analysis. As the distribution of the continuous PHQ-9 scores was heavily (positively) skewed, a logistic regression model was chosen to predict major depression as a dichotomous dependent variable. Two additional regression models with a continuous depression variable (an OLS model, and a left censored Tobit model[4]) are provided in Table B in the supplementary materials, as well as an additional logistic regression on a lower

[3] White's test for heteroscedasticity revealed a significant difference in variance between the groups ($\chi^2(1) = 86.12$, $p < .001$)

[4] The left censored Tobit model addresses the unique shape of the PHQ-9 score distribution (typical in a non-clinical population), in which 26.6% of participants got the lowest score (0 out of 27).

cutoff of the PHQ9, which represents 'any depression'. The first model included the sample group (MTurk or NHANES) only. In the second model, the socio-demographic variables associated with depression (i.e., Gender, Age, Income, Education, and Work) were added. Finally, in the third model, the health/lifestyle-related variables (Poor health, Weight, Physical activity, Sleep, and Time sitting down during the day).

Table 2 presents the frequencies of the independent variables in each sample along with their associations with major depression. The first model, including the sample group only, explained approximately 4% of the variance in PHQ major depression (R^2_p = .0396)[5]. Not surprisingly, the stepwise addition of the socio-demographic variables ($\chi^2(36) = 320.9$, $p < .001$) and the health/lifestyle-related variables ($\chi^2(49) = 581.4$, $p < .001$) improved the prediction of major depression, beyond the group effect of the first model. Even so, the group effect (MTurk/NHANES) remained significant in each model. Similar results were obtained using the continuous analyses, as well as the different cutoff analysis ('any depression'), in all of which the group effect remained highly significant.

To compare the level of variance between the samples that is left unexplained after controlling for the socio-demographic and health/lifestyle-related variables, we calculated the R^2_p and $\Delta R^2 p$ on nested models ($R^2_p{}^{(nes)}$ and $\Delta R^2_p{}^{(nes)}$, respectively), which only includes the 5,624 participants that are in all three models (i.e., those who answered all the questions). The remaining unexplained variance in the last model ($\Delta R^2_p{}^{(nes)}$ = .026) suggests that socio-demographic and health/lifestyle-related differences cannot fully explain the increased prevalence of major depression in MTurk. In Figure 1, we present the extent to which each of the variables explained depression prevalence differences between the two samples as well as the

[5] R^2_p (Pseudo R^2) = A proxy for the explained variation of the dependent variable in logistic regressions. ΔR^2_p = A proxy for the percentage of variance between the samples that remains unexplained after the inclusion of the new model.

combined contribution per cluster (socio-demographic, physical health and physical activity). For each variable and cluster, we calculated this contribution by dividing the $\Delta {R^2}_p^{(nes)}$ of each by the ${R^2}_p^{(nes)}$ of the first model (including only the sample variable). For example, since age is negatively correlated with depression (see Table 2) and MTurk participants are on average much younger than those in the NHANES data set, this age difference explains to some extent (16.6%) why MTurk participants score higher on depression. Level of education is also negatively correlated with depression. However, MTurk workers are on average more educated than the NHANES participants, which works against the increased rates of depression on MTurk, hence the negative contribution. Figure 1 shows that the socio-demographic variables combined explain only about 4% of the difference between samples, and the two physical health indicators combined explain an additional 4%. Indicators of physical activity (Physical activity, Sleep, and Time sitting down during the day), on the other hand, explain more than an additional third of the difference in depression prevalence between samples (that is: after controlling for the demographic and health variables). Even though the composition of the two samples differed on a range of individual variables that are known to be associated with depression, these variables together explained only less than half (42.7%) of the difference in depression prevalence between the MTurk and the NHANES sample. Similar patterns of results were obtained in statistical analyses that aimed to predict alternative different dependent variables instead of major depression: a less severe dichotomous variable ('any depression') and a continuous total PHQ-9 scores variable (using OLS and a Tobit model) that measures the severity of the depression (Table B, supplementary materials).

Psychiatric medications. Altogether, 277 (19%) attentive workers reported that they have used prescribed psychiatric medications in the past 30 days. The prevalence of medication use among MTurk workers was as follows: Antidepressants (15.4%), Benzodiazepine/Anxiolytics (5.4%), Mood stabilizers (2.3%), Stimulants (3.0%), and

Antipsychotics (1.3%), with some workers reporting more than one type of medication. Not surprisingly, the prevalence of psychiatric medications among MTurk workers was associated with the PHQ-9 depression categories: Use of depression-related medications was 34.4%, 23.9%, and 15.8% for MTurk workers classified as major depression, other depressive disorders, and non-depressed MTurk workers, respectively.

These results were then compared to the NHANES dataset which included data on prescribed medication. Whereas only 7.1% of the NHANES participants reported using depression-related medications (i.e., Antidepressants and/or Mood stabilizers), a total of 15.5% of attentive workers in our sample reported using such medications, χ^2 (1) = 96.2, $p < .001$. Furthermore, a logistic regression model showed that the differences in depression-related medication between the two samples remained significant after controlling for socio-demographic and health/lifestyle-related variables (Table B, supplementary materials).

Finally, whereas the PHQ-9 scores were affected by the inattentiveness level classification of MTurk workers, reports of depression-related psychiatric medications remained relatively stable throughout the different inattentiveness groups: The prevalence of medication use among attentive, questionable and inattentive workers was 18.5%, 14.9%, and 24.6%, respectively. A Chi square test revealed that these differences are not significant, $\chi^2(2) = 4.25$, $p = .119$. Thus, self-reported use of depression-related medications seem to be much less vulnerable to bad data threats.

Discussion

Similar to Study 1, extremely high (bogus) rates of depression were observed among inattentive and suspicious MTurk workers in Study 2, further emphasizing the importance of implementing strict and multiple data quality assurance tests. The obtained estimate of major depression among attentive and valid MTurk respondents in this second sample (11%) was statistically identical to that in Study 1, which is still substantially higher than its reported

prevalence in the general population (American Psychiatric Association, 2013).

More importantly, a comparison with data from a comprehensive national survey (NHANES; CDC, 2018) that employed the same screening tool for depression (PHQ-9) showed that, compared to the general population (3.6%), major depression is estimated to be three times higher in MTurk. Thus, the difference in depression rates between MTurk and the general population becomes apparent also when the identical screening tool is used. This difference was further validated by a second, alternative indicator of depressive mood disorders that was measured in Study 2, namely the self-reported use of depression-related medications (7% in the national survey versus 16% in the MTurk sample). Although medication use is not comparable to measures of depression, it provides another perspective on the workers' mental state, which may be less susceptible to self-report biases.

Even though the composition of the two samples (MTurk and NHANES) differed on a number of socio-demographic and health/lifestyle-related characteristics that are known to predict depression (e.g., age, education, and poor health, physical activity, hours sitting, and hours sleeping), the analyses reported here showed that these differences explain less than half of the increased prevalence of depression in MTurk. In the next section, we suggest several explanations that could account for the remaining difference in (higher) depression rates in MTurk samples.

General discussion

There is a growing trend in clinical research to collect data from online crowdsourcing platforms, such as Amazon’s Mechanical Turk (MTurk). The use of these platforms offers several advantages for the study of clinical populations. However, this trend also raises questions about the reliability of the data collected in these unsupervised conditions and, consequently, about the conclusions that may be drawn from research based on such data sets. In the present work, we addressed these challenges by examining depression rates in MTurk, comparing them

to conventional depression rates in the general population, and exploring possible reasons for increased levels of depression in MTurk.

The main contributions of this research are twofold: First, the present work contributes to the literature on clinical research using unsupervised, crowdsourcing data collection platforms, such as MTurk. We developed a procedure to detect suspicious (bots/cyborgs) and inattentive respondents. The multilayered inattentiveness index comprised four different data quality assurance methods that are specifically calibrated to internet-based surveys and skewed clinical assessments. The results of both studies reported here showed a consistent pattern of reverse associations between data quality and depression rates: Increases in data quality (i.e., more attentive and less suspicious users) yielded lower depression rates. The inclusion of inattentive and suspicious MTurk workers artificially increased the prevalence estimates of major depression in both studies to extremely high rates, as 26% (Study 1) and 55% (Study 2) of them were found to be meet the PHQ criterion of major depression.

These results confirmed our concern that highly skewed clinical distributions, such as in the case of major depression, are particularly vulnerable to random-false responses. According to the findings presented here, fake and inattentive respondents bias the clinical distribution toward the center. Moreover, these workers are not necessarily detected in traditional reliability and convergent validity checks, as these also tend to inflate under such conditions (Fong, Ho & Lam, 2010). Even though in this study we specifically focused on depression, our findings on the effects of failing to screen for inattentive and suspicious respondents are relevant for any crowdsourcing-based clinical research on psychopathologies whose distributions are highly skewed by definition. We recommend a critical reappraisal of research reporting on extremely high rates of other mental health conditions in MTurk (or other crowdsourcing platforms) to examine whether reported estimates on pathologies such as social anxiety (Arditte et al., 2016) may prove to have been artificially inflated as well.

Second, the findings presented here contribute to the current debate concerning differences between depression rates in MTurk and the general population. As described in the introduction, the findings from previous studies have been equivocal (Arditte et al., 2016; McCredie & Morey, 2018; Shapiro et al., 2013; Walters et al., 2018). Based on the findings presented here, it is likely that a portion of the higher depression rates reported in previous MTurk studies should be attributed to insufficient data quality assurance procedures. In the current work, we implemented a number of methodological improvements to arrive at a more reliable estimate of depression in MTurk. Two large samples were included (reaching about 5,000 MTurk workers altogether) and we used a common and validated screening tool for depression (El-Den et al., 2018) to compare findings with that of a national survey (NHANES) that used the exact same screening tool (PHQ-9). Finally, a rigorous, multi-layered data quality assurance procedure was implemented. The results of this research, including the replication of the findings, the comparison with the national representative database, and the comparison of depression-related, psychiatric medications, suggest that actual depression rates in MTurk are still higher compared to estimates based on representative, comprehensive surveys on the general population.

In addition to these methodological improvements, we empirically explored several plausible reasons behind the actual, increased MTurk depression rates. A comparison with data from the NHANES showed that the composition of the MTurk-based sample differed on a number of individual variables that are known to be associated with depression, such as education, age, income, occupation, physical health and physical activity. Together, these differences in individual background variables explained almost half of the difference in depression rates. In particular, the physical activity lifestyle variables (hours of sleep, amount of physical activity and hours sitting in a day) accounted for a substantive amount of this difference (37%). Yet, more than half of the difference in depression rates between the MTurk and the

representative NHANES survey could not be attributed to such individual differences in group composition. In addition, the comparison with the NHANES data set also shows that higher depression rates cannot (only) be attributed to the particular screening tool that was used in the current work (the PHQ-9). Importantly, the significance of the difference between the samples (after controlling for all other variables) is robust to: the two cutoff points for depression (major depression and any depression), the continuous variable (using OLS, and Tobit models to predict total PHQ-9 scores), and the less explicit proxy for depression (depression-related medications). Taken together, these findings then suggest that further research is needed to explore additional reasons that could explain the remaining difference in higher depression rates in MTurk.

We speculate here on three possible explanations for the increased prevalence of depression in MTurk: The first explanation continues the abovementioned rationale and suggests that working in MTurk attracts particular groups of individuals that are already more vulnerable and more prone to suffer from depression. It is possible that paid participation in online research panels from one's own home draws a particular subgroup of individuals that already suffer from depression. They may also differ on depression-relevant characteristics, in addition to the abovementioned demographic, health and physical activity related variables. One such subgroup may be individuals who suffer from social anxiety, which is a dominant risk factor for major depression (odds ratio 2.9) (Kessler, Stang, Wittchen, Stein, & Walters, 1999) and often precedes the onset of the depressive episode (Beesdo et al., 2007). Interestingly, previous studies have reported that approximately 50% of MTurk users suffer from clinical levels of social anxiety (Arditte et al., 2016; Shapiro et al., 2013). In contrast, the 12-month prevalence estimates in the general population are around 7% - 8% (American Psychiatric Association, 2013; Connor, Kobak, Churchill, Katzelnick, & Davidson, 2001). Social anxiety has been found to be associated with preferences of computer-mediated over face-to-face interactions (Lee & Stapinski, 2012; Prizant-Passal, Shechner, and Aderka, 2016). MTurk allows workers from all

socio-demographic strata to work from their homes without having to confront and navigate the external social world and this type of work may, therefore, attract individuals who suffer from social anxiety.

A second, more provocative possibility is that excessive use of MTurk *triggers* depressive feelings. This claim corresponds with warnings against overuse of screens and Internet. Based on findings from a recent study which included over half a million participants, Twenge and colleagues (Twenge, Joiner, Rogers, & Martin, 2018) concluded that increases in new media screen time contributed to more depressive symptoms. Similarly, in a recent systematic review, Elhai, Dvorak, Levine and Hall (2017) reported that problematic smartphone use is consistently associated with depressive symptoms (Elhai, Dvorak, Levine, & Hall, 2017). In contrast to other screen activities that are more social in nature (e.g., Facebook), the work at MTurk typically does not include social interactions. It is possible that individuals who work for online data collection forums for extended periods during a day are at increased risk of developing feelings of loneliness, which is a well-documented risk factor for developing depressive symptoms (Cacioppo et al., 2006). Moreover, MTurk workers typically do not receive feedback regarding the impact or the consequences of their work, which can lead to a sense of purposelessness. Worryingly, this type of work might impair subjective experiences of leading a meaningful life and cause a reduction in a person's overall well-being (Zika & Chamberlain, 1992).

Finally, an even more provocative explanation for the observed high depression rates in MTurk is that the existing DSM-based estimates of depression are in fact underestimates and that the actual depression rate in the general population is higher and closer to the MTurk-based figures. Anonymous, computer-mediated administration of depression screening tools may enable more honest response patterns. Despite growing awareness about mental illnesses such as depression, they are still strongly associated with a negative stigma (Menke & Flynn, 2009). The stigma could inhibit depressed individuals from sharing personal information in face-to-face

interactions (de Leeuw, 1992; Tourangeau & Yan, 2007), which is the standard method of data collection in the NHANES surveys. Sharing personal information about mental conditions or about psychiatric medications may prove to be easier in anonymous, online surveys (the so-called disinhibition effect of computer-mediated communication), especially when there is no expectation of follow-up interactions. In addition, in a personal interview set-up, the physical presence of the interviewer may (temporarily) alleviate negative feelings, which could temper responses on the depression assessment tools administered.

Even though the current work cannot provide any decisive answers on these issues, these potential explanations should nevertheless be considered. Future studies are recommended, preferably with similar procedures to the current research to determine which of these hypotheses explains better the obtained increased prevalence of depression.

Limitations of the current research

We highlight several limitations of the present research. First, due to the large scale of the present work and due to the strict data collection anonymity policy enforced by MTurk, we relied on a self-report scale to assess depression. Even though the use of self-report assessment tools is commonplace in large scale mental health surveys (e.g., CDC, 2018), we acknowledge that they cannot replace formal diagnoses of major depressive disorder that are determined by trained mental health professionals in face-to-face, clinical interviews. This is partially because different self-report questionnaires may result in different clusters of symptoms (Fried, 2017) and fail to encompass the heterogenic psychosocial nature of depression (Fried & Nesse, 2015).

To address these concerns to the best of our ability given the aforementioned constraints of large scale, MTurk based studies, we chose the PHQ-9 as the preferred self-report survey : In contrast to other self-report depression scales, the PHQ-9 items correspond directly with the DSM criteria for major depressive disorder. Major depression was determined using a cut-off point that has been validated psychometrically and deemed a good proxy of major depression. A

large study that tested several self-report depression scales, found that, compared to the "gold standard" criterion of the Structured Clinical Interview for DSM (SCID), the sensitivity and the specificity of the PHQ-9 cut-off point for major depressive disorder were 98% and 78%, respectively (Löwe et al.,2004). Lowe et al. (2004) demonstrated the superiority of the PHQ-9 over two other self-report scales and therefore recommended its use in clinical research. Another, more recent, systematic review of the available screening tools for depression showed that between 1995 and 2015, the PHQ-9 was the most common and most often validated screening tool for depressive disorders (El-Den, Chen, Gan, Wong, & O'Reilly, 2018). In sum, the PHQ-9 was the choice of self-reported screening tool for depression because of its well-established validity, sensitivity and specificity, compared to other screening tools for which less evaluative information is available.

Furthermore, in Study 2, we conducted a direct comparison between PHQ-9 scores in MTurk and PHQ-9 scores in the general population (CDC national survey). This comparison, which included multiple statistical controls of confounding variables, supports the claim that the higher rates of depression in MTurk are unlikely to have resulted from differences in the screening tools that were used.

A second limitation of the present work is the lack of information on the ethical or racial background of the MTurk workers. Previous surveys on depression suggest that rates of depression differ by race/ethnicity (Riolo, Nguyen, Greden, & King, 2005).Although this study controlled for multiple variables that relate to depression, future studies should examine and control for race/ethnicity as well.

A third limitation derives from our focus on one particular crowdsourcing platform. It is uncertain whether the obtained results are confined to MTurk specifically or whether they are characteristic of crowdsourcing platforms in general. Yet, the findings from the present research highlight the vulnerability of highly skewed clinical distributions (such as depression) to false or

bad data. We therefore expect that similar trends will be found in research that collects from other crowdsourcing platforms and / or focuses on other psychopathologies.

Practical implications for research

In addition to the aforementioned, we highlight three practically oriented implications that arise from the present work. Knowledge about increased depression prevalence in MTurk is not only relevant for clinical researchers, but may also have implications for MTurk-based research in general. Depression, affects many aspects of human experiences, including emotions, cognitions and daily choices (Greenberg, Vazquez, & Alloy, 1988). It is characterized by cognitive distortions and irrational-negative beliefs about the self, as well as about life in general (Beck, 1967). Depressed individuals tend to perceive the outside world in depressive colors and to engage in excessive rumination about negative experiences and past events (Nolen-Hoeksema et al., 2008). These depression-specific characteristics may affect how individuals act and respond to particular cues, surveys and activities presented to them. In such cases, unusual high rates of depression could, therefore, pose a threat to the generalizability of empirical findings from MTurk-based data.

Second, the current research has produced a concrete contribution for researchers who collect data using MTurk and similar crowdsourcing platforms. The methodological procedures that were described and applied in the two studies reported here provide specific and detailed data quality measures that can help researchers overcome the obstacles inherent to these platforms and ensure the quality of the unsupervised self-report data collection.

Finally, from a research participant recruitment viewpoint, increased depression rates among MTurk workers (major depression: 11% - 13%, any depression: 21%) may also be considered as an advantage for clinical researchers who are specifically interested in psychopathologies, such as depression and anxiety. Recruitment of clinical subjects is a major challenge in psychopathology research. Even depression, which is considered a rather common

mental condition, is still relatively rare in random samples recruited from the general population (Katon & Ciechanowski, 2002), and using MTurk can facilitate recruitment efforts and provide researchers with easy access to peripheral populations. However, and as aforementioned in detail, it is imperative that strict data quality assurance tests are implemented to ensure the reliability of research results and avoid artificially inflated depression rates.

Acknowledgement

The research reported here received financial support from the Israeli Innovation Authority (grants # 60561 and 60560).

Bibliography

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

APS. (2018). Researchers Investigate Problems with MTurk Data. Retrieved from https://www.psychologicalscience.org/publications/observer/obsonline/researchers-investigate-problems-with-mturk-data.html.

Arditte, K. A., Çek, D., Shaw, A. M., & Timpano, K. R. (2016). The importance of assessing clinical phenomena in Mechanical Turk research. *Psychological Assessment, 28*(6), 684.

Arroll, B., Goodyear-Smith, F., Crengle, S., Gunn, J., Kerse, N., Fishman, T., . . . Hatcher, S. (2010). Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. *The Annals of Family Medicine, 8*(4), 348-353 .

Bai, H. (2018). Evidence that A Large Amount of Low Quality Responses on MTurk Can Be Detected with Repeated GPS Coordinates. Retrieved from https://www.maxhuibai.com/blog/evidence-that-responses-from-repeating-gps-are-random

Beck, A. T. (1967). *Depression: Clinical, experimental, and theoretical aspects* (Vol. 32): University of Pennsylvania Press.

Beck, A. T., Steer, R .A., & Brown, G. K. (1996). *Beck depression inventory-II*. San Antonio, 78(2), 490-498.

Beesdo, K., Bittner, A., Pine, D. S., Stein, M. B., Höfler, M., Lieb, R., & Wittchen, H.-U. (2007). Incidence of social anxiety disorder and the consistent risk for secondary depression in the first three decades of life. *Archives of General Psychiatry*, 64(8), 903-912.

Bunge, E., Cook, H. M., Bond, M., Williamson, R. E., Cano, M., Barrera, A. Z., Leykin, Y. & Muñoz, R. F. (2018). Comparing Amazon Mechanical Turk with unpaid internet resources in online clinical trials. *Internet interventions*, 12, 68-73

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

Burns, D. D. (1999). *The feeling good handbook (rev. ed.)*. New York, NY: Plume/Penguin.

Cacioppo, J. T., Hughes, M. E., Waite, L. J., Hawkley, L. C., & Thisted, R. A. (2006). Loneliness as a specific risk factor for depressive symptoms: cross-sectional and longitudinal analyses. *Psychology and Aging, 21*(1), 140.

Center for Disease Control and Prevention (CDC) & National Center for Health Statistics (2018). *National Health and Nutrition Examination Survey data, 2015–2016.* Hyattsville, MD: US Department of Health and Human Services.

Chandler, J., Shapiro, D., & Sisso, I. (working paper). Best Practices and Pitfalls when Recruiting Rare Groups Online.

Connor, K. M., Kobak, K. A., Churchill, L. E., Katzelnick, D., & Davidson, J. R. T. (2001). Mini-SPIN: A brief screening assessment for generalized social anxiety disorder . *Depression and anxiety, 14*(2), 137-140.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4-19.

De Leeuw, E. D. (1992). *Data quality in mail, telephone and face to face surveys*. Amsterdam, Netherlands: TT Publikaties.

Dennis, S. A., Goodson, B. M., & Pearson, C. (2018). Mturk Workers' Use of Low-Cost "Virtual Private Servers" to Circumvent Screening Methods: A Research Note. http://dx.doi.org/10.2139/ssrn.3233954

Difallah, D., Filatova, E., & Ipeirotis, P. (2018, February). Demographics and dynamics of mechanical turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 135-143). ACM. https://doi.org/10.1145/3159652.3159661

Dunn, A. M., Heggestad, E. D., Shanock, L. R. and Theilgard, N. (2018). Intra-individual Response Variability as an Indicator of Insufficient Effort Responding: Comparison to Other Indicators and Relationships with Individual Differences. *Journal of Business and Psychology*, 33(1), 105–121.

Eisenberg, D., Gollust, S. E., Golberstein, E., & Hefner, J. L. (2007). Prevalence and correlates of depression, anxiety, and suicidality among university students. *American Journal of Orthopsychiatry, 77*(4), 534 .

El-Den, S., Chen, T. F., Gan, Y.-L., Wong, E., & O'Reilly, C. L. (2018). The psychometric properties of depression screening tools in primary healthcare settings: A systematic review. *Journal of Affective Disorders, 225*, 503-522. doi:https://doi.org/10/1016.j.jad.2017.08.060

Elhai, J. D., Dvorak, R. D., Levine, J. C., & Hall, B. J. (2017). Problematic smartphone use: A conceptual overview and systematic review of relations with anxiety and depression

psychopathology. *Journal of Affective Disorders*, 207, 251-259. doi:10.1016/j.jad.2016.08.030

Elphinstone, B. (2018). Identification of a Suitable Short-form of the UCLA-Loneliness Scale. *Australian Psychologist, 53*(2), 107-115.

Fong, D. Y., Ho, S. Y., & Lam, T. H. (2010). Evaluation of internal reliability in the presence of inconsistent responses. *Health and Quality of Life Outcomes*, 8(1), 27.

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, 208, 191-197. doi:https://doi.org/10.1016/j.jad.2016.10.019

Fried, E. I., & Nesse, R. M. (2015). Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR* D study. *Journal of Affective disorders*, 172, 96-102.

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

Greenberg, M. S., Vazquez, C. V., & Alloy, L. B. (1988). Depression versus anxiety: Differences in self- and other-schemata. In L. B. Alloy (Ed.), *Cognitive processes in depression* (pp. 109-142). New York, NY, US: Guilford Press.

Huang, J. L., Bowling, N. A., Liu, M. and Li, Y. (2015). Detecting Insufficient Effort Responding with an Infrequency Scale: Evaluating Validity and Participant Reactions. *Journal of Business and Psychology*, 30(2), 299–311.

Ibrahim, A. K., Kelly, S. J ,.Adams, C. E., & Glazebrook, C. (2013). A systematic review of studies of depression prevalence in university students. *Journal of Psychiatric Research, 47*(3), 391-400.

Katon, W., & Ciechanowski, P. (2002). Impact of major depression on chronic medical illness. *Journal of Psychosomatic Research*, 53(4), 859-863.

Kees, J., Berry, C., Burton, S. and Sheehan, K. (2017). An Analysis of Data Quality: Professional Panels, Student Subject Pools, and Amazon's Mechanical Turk. *Journal of Advertising*, 46(1), 141–155.

Kessler, R. C., Petukhova, M., Sampson, N. A., Zaslavsky, A. M., & Wittchen, H. U. (2012). Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. *International Journal of Methods in Psychiatric Research, 21*(3), 169-184.

Kessler, R. C., Stang, P., Wittchen, H. U., Stein, M., & Walters, E. E. (1999). Lifetime co-morbidities between social phobia and mood disorders in the US National Comorbidity Survey. *Psychological Medicine*, 29(3), 555-567.

Kosara, R., & Ziemkiewicz, C. (2010, April). Do Mechanical Turks dream of square pie charts?. In *Proceedings of the 3rd BELIV'10 Workshop: Beyond time and errors: Novel evaluation methods for information visualization* (pp. 63-70). ACM.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a Brief Depression Severity Measure. *Journal of General Internal Medicine, 16*(9), 606-613. doi:10.1046/j.1525-1497.2001.016009606.x

Löwe, B., Spitzer, R. L., Gräfe, K., Kroenke, K., Quenter, A., Zipfel, S., ... & Herzog, W. (2004). Comparative validity of three screening questionnaires for DSM-IV depressive disorders and physicians' diagnoses. *Journal of affective disorders*, 78(2), 131-140.

McCredie, M. N., & Morey, L. C. (2018). Who Are the Turkers? A Characterization of MTurk Workers Using the Personality Assessment Inventory .*Assessment, Online First*, 1-8. 1073191118760709.

Menke, R., & Flynn, H. (2009). Relationships between stigma, depression, and treatment in white and African American primary care patients. *The Journal of nervous and mental disease, 197*(6), 407-411.

Mor, N., Hertel, P., Ngo, T. A., Shachar, T., & Redak, S. (2014). Interpretation bias characterizes trait rumination. *Journal of Behavior Therapy and Experimental Psychiatry*, *45*(1), 67-73. doi:http://dx.doi.org/10.1016/j.jbtep.2013.08.002

Moss, A. J., & Litman, L. (2018). After the bot scare: Understanding what's been happening with data collection on MTurk and how to stop it [blog post]. Retrieved from https://blog.turkprime.com/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it.

Nolen-Hoeksema, S., & Morrow, J. (1991 .(A prospective study of depression and posttraumatic stress symptoms after a natural disaster: The 1989 Loma Prieta earthquake. *Journal of Personality and Social Psychology, 61*(1), 115-121. doi:10.1037/0022-3514.61.1.115

Nolen-Hoeksema, S., Parker, L. E & ,.Larson, J. (1994). Ruminative coping with depressed mood following loss. *Journal of Personality and Social Psychology, 67*(1), 92-104. doi:10.1037/0022-3514.67.1.92

Nolen-Hoeksema, S., Wisco, B. E., & Lyubomirsky, S. (2008). Rethinking rumination. *Perspectives on Psychological Science, 3*(5), 400-424. doi:10.1111/j.1745-6924.2008.00088.x

Riolo, S. A., Nguyen, T. A., Greden, J. F., & King, C. A. (2005). Prevalence of depression by race/ethnicity: findings from the National Health and Nutrition Examination Survey III. *American Journal of Public Health*, 95(6), 998.

Russell, D., Peplau, L. A., & Ferguson, M. L. (1978). Developing a measure of loneliness. *Journal of Personality Assessment, 42*(3), 290-294 .

Russell, D. W. (1996). UCLA Loneliness Scale (Version 3): Reliability, validity, and factor structure. *Journal of Personality Assessment, 66*(1), 20-40 .

Sartorius, N., Üstün, T. B., Lecrubier, Y., & Wittchen, H.-u. (1996). Depression comorbid with anxiety: results from the WHO study on psychological disorders in primary health care. *The British journal of psychiatry, 168*(S30), 38-43 .

Schoofs, H., Hermans, D., & Raes, F. (2010). Brooding and reflection as subtypes of rumination: Evidence from confirmatory factor analysis in nonclinical samples using the Dutch Ruminative Response Scale. *Journal of Psychopathology and Behavioral Assessment, 32*(4), 609-617 .

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science, 1*(2), 213-220.

Sisso, I. (working paper). Best Practices in Online Experiments - 42 Ways to Measure and Increase Data Quality.

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine, 166*(10), 1092-1097. doi:10.1001/archinte.166.10.1092

Spitzer, R. L., Kroenke, K., & Williams, J. B. W. (1999). Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *Jama, 282*(18), 1737-1744.

Stewart, N., Ungemach, C., Harris, A. J., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision making, 10*(5), 479-491.

Tomitaka, S., Kawasaki, Y., Ide, K., Akutagawa, M., Yamada, H., Ono, Y., & Furukawa, T. A. (2018). Distributional patterns of item responses and total scores on the PHQ-9 in the general population: data from the National Health and Nutrition Examination Survey. *BMC psychiatry, 18(1),* 108.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological bulletin*, *133*(5), 859.

Twenge, J. M., Joiner, T. E., Rogers, M. L., & Martin, G. N. (2018). Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates Among U.S. Adolescents After 2010 and Links to Increased New Media Screen Time. *Clinical Psychological Science*, *6*(1), 3-17.

Vésteinsdóttir, V., Reips, U. D., Joinson, A., & Thorsdottir, F. (2017). An item level evaluation of the Marlowe-Crowne Social Desirability Scale using item response theory on Icelandic Internet panel data and cognitive interviews. *Personality and Individual Differences*, *107*, 164-173.

Walters, K., Christakis, D. A., & Wright, D. R. (2018). Are Mechanical Turk worker samples representative of health status and health behaviors in the US? *PloS one*, *13*(6), e0198835.

Watkins, E. R. (2008). Constructive and unconstructive repetitive thought. *Psychological Bulletin*, *134*(2), 163.

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*(3), 191-195.

Zika, S., & Chamberlain, K. (1992). On the relation between meaning in life and psychological well-being. *British Journal of Psychology*, *83*(1), 133-145.

Table 1.

Prevalence of major depression in MTurk, Study 1

	Number of workers (*N*)	PHQ-9 *Means* (*SD*)	Moderate-Severe depression	Any depression Prevalence	Major depression Prevalence
Whole sample	2692	7.48 (6.30)	890 (33.1%)	712 (26.4%)	497 (18.5%)
Attentive workers (IS=0)	1848	6.60 (5.48)	493 (26.7%)	381 (20.6%)	238 (12.9%)
Questionable workers (IS=1)	427	8.76 (7.29)	179 (41.9%)	140 (32.8%)	111 (26.0%)
Inattentive workers (IS≥2)	181	9.51 (7.75)	87 (48.1%)	72 (39.8%)	57 (31.5%)
Suspicious IP	236	10.54 (7.35)	131 (55.5%)	119 (50.4%)	91 (38.6%)

Note: Moderate-Severe Depression = PHQ-9 > 10. Major depression = When five or more PHQ-9 items receive a score of 2 or 3; Any depression = Either major depression or Other depressive disorders that are diagnosed if 2-4 PHQ-9 items receive a score of 2 or 3 (both diagnoses are valid only when one of two first key symptoms are reported). IS = Inattentiveness Score (i.e., the number of failed attention checks); Attentive workers = MTurk workers that did not fail any of the attention checks and do not have a suspicious IP; Questionable workers = MTurk workers that failed only one attention check and do not have a suspicious IP; Inattentive workers = MTurk workers that failed two or more attention checks and do not have a suspicious IP; Suspicious IP = IP address recognized as possibly malicious or from outside the US.

Table 2.

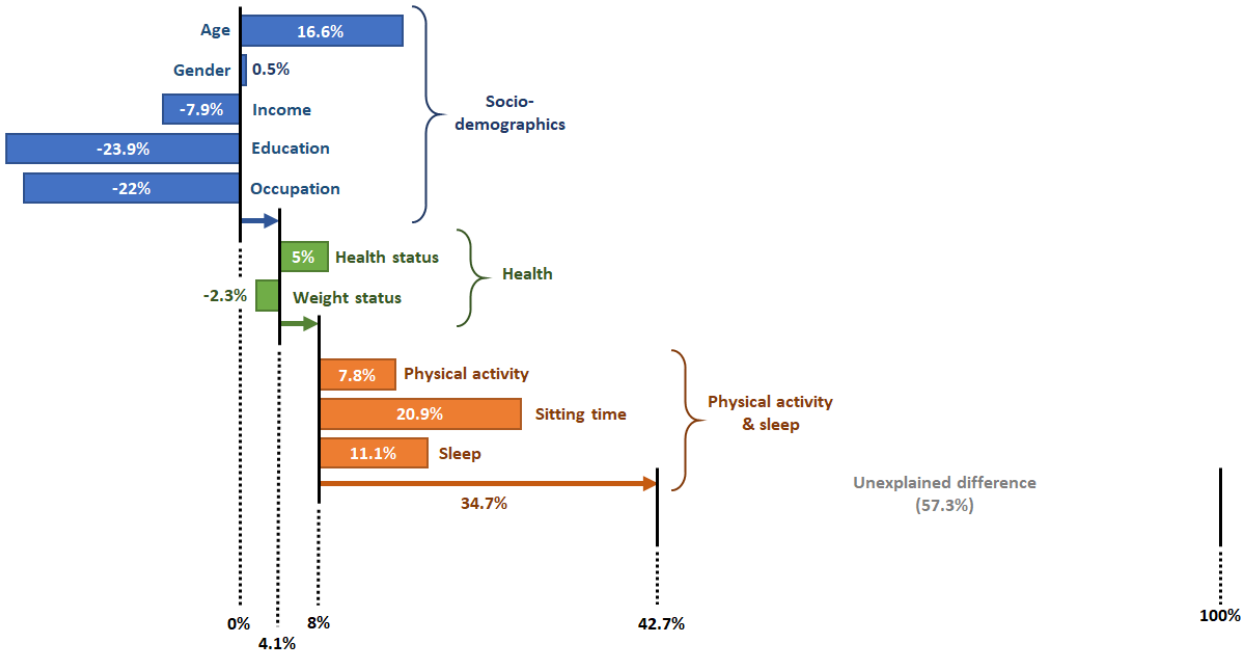
Frequencies of all independent variables and their associations with major depression

Predictors	Descriptives of sample MTurk	Descriptives of sample NHANES	Sample only	+ Demographics	+ Health factors
Sample = MTurk	N=1461 (22.2%)	N=5134 (77.8%)	3.33*** (2.67 - 4.15) $R^2_p{}^{(nes.)}$ = 0.0453	4.38*** (3.26 - 5.90) $\Delta R^2_p{}^{(nes.)}$ =.0434 $\chi^2(1)$ = 100.8***	3.91*** (2.75 - 5.55) $\Delta R^2_p{}^{(nes.)}$ =.0260 $\chi^2(1)$ = 60.3***
Gender = Female	51.2%	53.4%	$\chi^2(1)$ = 2.28	1.35* (1.05 - 1.73)	1.25 (.95 - 1.65)
Age					
(ref: 18-29)	30.2%	20.7%		*(ref)*	*(ref)*
30-39	37.2%	15.7%		1.01 (0.71 - 1.43)	0.88 (0.60 - 1.28)
40-49	17.1%	15.8%		0.94 (0.63 - 1.40)	0.66 (0.43 - 1.03)
50-59	10.1%	15.4%	$\chi^2(6)$ = 648.5***	0.76 (0.50 - 1.17)	0.58* (0.36 - 0.93)
60-69	4.5%	16.7%		0.43** (0.25 - 0.73)	0.35*** (0.19 - 0.62)
70-79	0.9%	9.9%		0.34** (0.16 - 0.72)	0.27** (0.12 - 0.60)
80+	0%	5.8%		0.43 (0.17 - 1.07)	0.26* (0.09 - 0.76)
Annual household income					
(ref:0-$5K)	2.5%	2.7%		*(ref)*	*(ref)*
$5K-$10K	1.8%	4.3%		1.13 (0.53 - 2.40)	0.91 (0.40 - 2.08)
$10K-$15K	3.6%	6.6%		1.02 (0.50 - 2.08)	0.84 (0.39 - 1.83)
$15K-$20K	4.9%	7.4%		1.14 (0.57 - 2.30)	0.81 (0.37 - 1.76)
$20K-$25K	8.0%	6.7%		1.04 (0.52 - 2.08)	0.98 (0.46 - 2.07)
$25K-$35K	13.0%	12.4%	$\chi^2(11)$ = 102.0***	0.76 (0.39 - 1.49)	0.65 (0.31 - 1.36)
$35K-$45K	10.5%	10.6%		0.87 (0.43 - 1.74)	0.83 (0.39 - 1.76)
$45K-$55K	11.3%	8.9%		0.95 (0.47 - 1.89)	0.85 (0.40 - 1.81)
$55K-$65K	9.9%	6.9%		0.72 (0.34 - 1.55)	0.82 (0.37 - 1.85)
$65K-$75K	8.8%	5.5%		0.72 (0.33 - 1.56)	0.75 (0.33 - 1.73)
$75K-$100K	12.5%	10.3%		0.41* (0.19 - 0.91)	0.43 (0.18 - 1.02)
>$100K	13.2%	17.8%		0.60 (0.29 - 1.25)	0.72 (0.33 - 1.58)
Education					
(ref: <9th grade)	0.2%	11.2%		*(ref)*	*(ref)*
9-11th grade	0.7%	11.9%		0.94 (0.54 - 1.65)	0.84 (0.46 - 1.55)
High school	14.4%	22.1%	$\chi^2(4)$ = 564.1***	0.72 (0.43 - 1.19)	0.85 (0.50 - 1.47)
Some college or AA deg.	35.9%	30.0%		0.76 (0.46 - 1.23)	0.86 (0.50 - 1.46)
College grad. +	48.8%	24.8%		0.51* (0.30 - 0.88)	0.79 (0.44 - 1.42)
Work status last week					
Not at work	2.9%	2.1%		1.61 (0.77 - 3.38)	1.69 (0.75 - 3.79)
Looking for work	8.0%	4.2%		2.99*** (1.92 - 4.67)	3.00*** (1.84 - 4.91)
Work 1-19 hours	3.1%	5.2%		0.81 (0.34 - 1.91)	0.85 (0.34 - 2.13)
Work 20-39 hours	13.7%	13.8%		0.86 (0.54 - 1.38)	0.95 (0.58 - 1.55)
(ref: Work 40-59 hours)	47.9%	31.3%		*(ref)*	*(ref)*
Work 60-79 hours	8.4%	4.3%		1.12 (0.62 - 2.02)	1.01 (0.53 - 1.92)
Work 80+ hours	1.1%	0.9%	$\chi^2(13)$ = 427.4***	2.52 (0.94 - 6.81)	2.48 (0.82 - 7.52)
Taking care of house/family	6.2%	6.2%		1.20 (0.69 - 2.08)	1.38 (0.77 - 2.48)
School	1.7%	2.1%		0.97 (0.29 - 3.26)	1.23 (0.35 - 4.32)
Retired	2.3%	17.2%		2.54** (1.34 - 4.82)	2.51** (1.26 - 5.00)
Sick leave	1.6%	3.3%		8.46*** (5.03 - 14.2)	3.68*** (2.04 - 6.65)
Layoff	0.3%	0.3%		1.13 (0.14 - 8.95)	1.14 (0.13 - 9.93)
Disabled	1.7%	5.8%		7.66*** (4.80 - 12.22)	4.21*** (2.48 - 7.14)
No work - other	1.1%	3.3%		2.34* (1.12 - 4.90)	1.95 (2.48 - 7.15)
Poor health status[1]	2.67(0.95)	2.86(0.96)	t(6593) = 6.64*** Cohen's d = 0.20 (0.14 - 0.26)		2.76*** (2.33 - 3.27)
Weight status					
(ref: About the right weight)	46.6%	43.2%			*(ref)*
Overweight	48.9%	50.8%	$\chi^2(2)$ = 8.3*		1.13 (0.84 - 1.53)
Underweight	4.5%	6.0%			2.36*** (1.44 - 3.84)
Physical activity					
(ref: none)	24.6%	50.8%			*(ref)*
0-1h a week	54.3%	16.6%			0.87 (0.62 - 1.20)
1-2h a week	16.1%	16.9%	$\chi^2(4)$ = 915.2***		0.90 (0.57 - 1.43)
2-3h a week	2.7%	8.4%			0.83 (0.40 - 1.74)
>3h a week	2.3%	7.3%			0.66 (0.25 - 1.74)
Hours sitting down a day[1]	8.12(3.50)	6.19(3.32)	t(6505) = 19.17*** Cohen's d = 0.57 (0.51-0.63)		1.08*** (1.04 - 1.12)
Average sleep time per night					
<5h	3.1%	3.1%			2.32** (1.26 - 4.27)
5-6.4h	28.9%	12.4%			1.73** (1.19 - 2.52)
6.5-7.9h	33.5%	32.6%	$\chi^2(5)$ = 311.6***		1.41 (0.99 - 2.02)
(ref: 8-9.4h)	31.4%	39.1%			*(ref)*
9.5-10.9h	2.2%	9.0%			1.18 (0.69 - 2.02)
11h+	0.9%	3.8%			2.80*** (1.57 - 5.03)
N			6,595	5,743	5,624
R^2_p			0.0396	0.133	0. 251
χ^2			106.6	320.9	581.4
df			1	36	49

Note: values in the model columns represent odds ratios (CI in parentheses). *(ref)* – reference value. R^2_p – McFadden pseudo R^2. $R^2_p{}^{(nes.)} - R^2_p$ for the nested model (which includes participants who are in all the models). $\Delta R^2_p{}^{(nes.)}$ – the increase in $R^2_p{}^{(nes.)}$ attributed to the sample variable. [1] Continuous variable - mean and standard deviation (in parenthesis) are reported in the descriptive statistics, as well as a *t*-test and effect size for the difference between samples. *** p <.001, ** p< .01, * p < .05.

Figure 1.

Contribution of each variable to the explained difference in the prevalence of PHQ major depression between the NHANES and MTurk samples



Note: Each bar represents a specific variable's ability to explain the observed difference in the prevalence of major depression between the MTurk and NHANES sample. The values are calculated by the change in $R^2_p{}^{(nes.)}$ resulted by the inclusion of each variable. A non-zero starting point for the bar implies that all the previous (above) variables are controlled for.