Title
Comparing Data Collected on Amazon's Mechanical Turk to National Surveys

Permalink
https://escholarship.org/uc/item/8p43z11f

Journal
American Journal of Health Behavior, 46(5)

ISSN
1087-3244

Authors
Qureshi, Nabeel
Edelen, Maria
Hilton, Lara
et al.

Publication Date
2022-10-17

DOI
10.5993/ajhb.46.5.1

Peer reviewed

# Comparing Data Collected on Amazon's Mechanical Turk to National Surveys

**Nabeel Qureshi, MPH, MPhil**
**Maria Edelen, PhD**
**Lara Hilton, PhD, MPH**
**Anthony Rodriguez, PhD**
**Ron D. Hays, PhD**
**Patricia M. Herman, ND, PhD**

**Objective:** In this study, we examined the impact of a range of methods to improve data quality on the demographic and health status representativeness of Amazon Mechanical Turk (MTurk) samples. **Methods:** We developed and field-tested a general survey of health on MTurk in 2017 among 5755 participants and 2021 among 6752 participants. We collected information on participant demographic characteristics and health status and implemented different quality checks in 2017 and 2021. **Results:** Adding data quality checks generally improves the representativeness of the final MTurk sample, but there are persistent differences in mental health and pain conditions, age, education, and income between the MTurk population and the broader US population. **Conclusion:** We conclude that data quality checks improve the data quality and representativeness.

**Keywords:** Mechanical Turk; representativeness; data quality; online data collection

***Am J Health Behav.*™ 2022;46(5):497-502**
**DOI: https://doi.org/10.5993/AJHB.46.5.1**

Amazon's Mechanical Turk (MTurk; www.MTurk.com) is a crowdsourcing platform hosted by Amazon that connects individuals creating tasks (requesters) to workers willing to complete those tasks for compensation.[1] MTurk offers a low-cost, fast-turnaround option for completing tasks that can be done online. MTurk has been used increasingly in research, particularly in studies related to health.[2,3] Previous shows that MTurk workers are younger, more likely to be white, male, and have a college degree, and less likely to be in good health than the general population. Workers on Mechanical Turk (herein called "Turkers") are less likely to be vaccinated for influenza, exercise, and smoke than the general population.[4] Current recommendations to increase the representativeness of an MTurk sample include post hoc recruitment for specific under-represented demographic groups,[5] or using new tools such as curated panels of participants that allow access to harder-to-reach groups on MTurk.[6,7]

In addition to questions about representativeness, there is a concern about the quality of data obtained from MTurk samples.[8] Previous work shows that there are issues with worker inattentiveness to tasks. For example, Chandler et al.[6] studied the impact on MTurk participation when workers were subjected to a pre-screening survey that tested their ability to comprehend and correctly respond to questions. Those who failed the screener were allowed to complete the assessment and were found to score worse on attention checks and have lower reliability of responses than those who passed.

Researchers have implemented a variety of approaches to increase the response quality of Turkers, including removing those who have an average item response of one second or less, adding screener questions before the main survey, adding IP address verification, and conducting test-retest comparisons on key demographic variables.[9-11] Previous studies show that surveys that do not implement data quality checks can yield spurious results. For example, Ophir et al.[12] estimated the prevalence of depression in 2 surveys of Turkers,

*Nabeel Qureshi, Assistant Policy Researcher, RAND Corporation, Santa Monica, CA, United States. Maria Edelen, Associate Director, PROVE Center, Boston, MA, United States. Lara Hilton, Director, USC WorkWell Center, Los Angeles, CA, United States. Anthony Rodriguez, Behavioral and Social Scientist, RAND Corporation, Boston, MA, United States. Ron D. Hays, Professor, UCLA Department of Medicine, Los Angeles, CA, United States. Patricia M. Herman, Senior Behavioral and Social Scientist, RAND Corporation, Santa Monica, CA, United States. Correspondence Mr Qureshi; nqureshi@rand.org*

one without attention checks and one with attention checks. The authors created a scale to identify likely inattentive workers by combining a reading speed task, an internal consistency check, and 3 checks of response strings to identify long strings of the same response. Workers were grouped into categories based on how inattentive they were. In both samples, MTurk workers were found to have a higher prevalence of depression than the general population, a finding consistent with other studies,[4] but they also found that prevalence was about 50% higher when inattentive responders were included. It is recommended that a combination of the above approaches be used to produce the highest quality results.[10]

The goal of this paper is to evaluate the national representativeness of 2 samples of MTurk workers and evaluate how different recruitment strategies impact the representativeness of the samples. We also describe how a screening approach using fake conditions impacts the quality and representativeness of the MTurk sample.

## METHODS

### Data Collection

We developed and programmed 2 web-based surveys to collect data from MTurk participants in 2017[13] (funded by a grant from the National Institutes of Health/ National Center for Complementary and Integrative Health, Grant Number 1R21AT009124-01) and 2021 (funded by a grant from the National Institutes of Health/National Center for Complementary and Integrative Health, Grant Number 1R01AT010402-01A1). In both instances, we used the online platform CloudResearch (formerly TurkPrime) to field the survey. CloudResearch allows users to host tasks on MTurk and supports advanced features to customize survey implementation.[14]

Our general health survey was designed first to collect participants' demographic characteristics and health conditions and then use this information to target a follow-up survey to those who endorsed back pain from the list of conditions presented. Use of this conditions list made it possible to identify those with back pain without revealing that a particular answer would lead to more work (and compensation). Conditions included 14 that required a doctor's diagnosis (ie, "Have you EVER been told by a doctor or other health professional that you had…") and 10 self-attested issues (ie, "Do you currently have..."). The question about back pain was in this second group. The surveys were limited to respondents aged 18 years or older with an IP address in the United States (US).

Surveys were fielded continuously in batches of 9 surveys every hour for 2 months. MTurk charges a fee for tasks that require more than 9 individuals to complete. Using features native to CloudResearch, we were able to create batches of tasks for 9 individuals at a time while ensuring that no individual could take the same task twice. This approach also allowed individuals who complete MTurk tasks at different times of the day and days of the week to have an equal chance to participate.

All participants provided electronic consent starting the survey. Those who completed the general health survey were offered $1.50 for participation, and those who qualified and went on to complete the back pain survey were offered an additional $2.50. Payments were determined by approximating the amount of time needed to complete the survey and offering the equivalent of US federal minimum wage for completion of the general health survey and a slight bonus for completing the subsequent back pain survey. The average time to complete was 22 minutes.

### Data Sources

The data collected in 2017 ("MTurk 2017") were developed and programmed on Qualtrics using several native MTurk features and CloudResearch advanced features to increase the quality of the MTurk sample. First, each participant had to complete a minimum of 100 previous human intelligence tasks (HITs) on MTurk with a successful completion rate of at least 90%. Second, the data collection process used a native US IP address check on CloudResearch to verify whether individuals were accessing the survey with an IP address from a US location.

The second dataset was collected in 2021 ("MTurk 2021") and programmed on SelectSurvey. Like the previous survey, we used the same CloudResearch advanced features to increase the quality of our responses. This time participants had to complete a minimum of 500 previous HITs on MTurk with a successful completion rate of at least 95%. The 95% threshold was selected as it is shown to improve response quality[15] and based on a time to complete analysis on a pilot study to test the survey.

In addition, we included 2 conditions that were fake in the condition checklist (ie, Syndomitis, Checkalism) to identify and screen out individuals who may be gaming the survey to qualify for a potential follow-up survey. Any individual who identified as having back pain and did not endorse a fake condition was then offered the opportunity to take a targeted survey on back pain. We also identified the subset of the sample who did not endorse a fake condition ("MTurk 2021A").

## Data Analysis

We conducted univariate analyses to describe the composition of our general health survey samples by age, gender, race/ethnicity, income, education, and self-reported health conditions. We compared the 3 MTurk samples to national estimates of demographic characteristics and health conditions based on the National Health and Nutrition Examination Survey and the National Health Interview Survey. We also compared MTurk estimates of demographics to estimates from the US population using the US Census Current Population Survey (CPS) and American Community Survey (ACS). Relative comparisons between MTurk 2017 and MTurk 2021 reflect the differences in representativeness based on time (2017 vs 2021 during a pandemic) as well as an increased threshold of worker experience (100 vs 500 previous tasks) and quality (90% vs 95% completion rate) of work. Comparisons to national rates for MTurk 2021 versus MTurk 2021A reflect the improvements in representativeness due to exclusion of those who endorsed fake conditions.

## RESULTS

MTurk 2017 included 5755 participants, MTurk 2021 6752 participants, and MTurk 2021A 5760 individuals. That is, 992 individuals (15%) endorsed at least one fake condition. On average, those who endorsed one fake condition endorsed 12.2 other conditions, those who endorsed both fake conditions endorsed 15.0 other conditions, and those who did not endorse a fake condition endorsed 3.7 other conditions.

**Table 1**
**Demographic Characteristics of MTurk Compared to US Population**

| **Gender** | **MTurk 2017 (%)<br>N=5755** | **MTurk 2021 (%)<br>N=6752** | **MTurk 2021A (%)<br>N=5760** | **US (%)** | **Source for US comparison** |
|---|---|---|---|---|---|
| Male | 47.31 | 43.35 | 45.18 | 49.2 | US Census Bureau, Current Population Survey, Annual Social and Economic Supplement, 2019 |
| Female | 52.14 | 55.87 | 53.9 | 50.8 | |
| Other | 0.54 | 0.78 | 0.92 | N/A | |
| **Age** | **MTurk 2017 (%)<br>N=5755** | **MTurk 2021 (%)<br>N=6752** | **MTurk 2021A (%)<br>N=5760** | **US (%)** | |
| Average age | 35.9 | 39.4 | 39.6 | 47.1 | US Census Bureau, Current Population Survey, Annual Social and Economic Supplement, 2019 |
| **Ethnicity** | **MTurk 2017 (%)<br>N=5755** | **MTurk 2021 (%)<br>N=6752** | **MTurk 2021A (%)<br>N=5760** | **US (%)** | |
| Hispanic | 8.19 | 20.25 | 14.55 | 18.5 | US Census Bureau, Current Population Survey, Annual Social and Economic Supplement, 2019 |
| Not Hispanic | 91.81 | 79.75 | 85.45 | 81.5 | |
| **Race** | **MTurk 2017 (%)<br>N=5755** | **MTurk 2021 (%)<br>N=6752** | **MTurk 2021A (%)<br>N=5760** | **US (%)** | |
| White | 80.22 | 78.33 | 79.72 | 72.49 | 2019 American Community Survey |
| Black or African American | 8.20 | 12.67 | 10.52 | 12.7 | |
| Asian | 6.16 | 5.49 | 5.97 | 5.52 | |
| Native Hawaiian or Pacific Islander | 0.22 | 0.16 | 0.15 | 0.18 | |
| American Indian or Alaska Native | 0.79 | 3.34 | 3.65 | 0.85 | |
| Other | 1.08 | 0 | 0 | 4.94 | |
| **Income** | **MTurk 2017 (%)<br>N=5755** | **MTurk 2021 (%)<br>N=6752** | **MTurk 2021A (%)<br>N=5760** | **US (%)** | |
| Less than $10,000 | 5.18 | 3.98 | 4.16 | 5.05 | US Census Bureau, Current Population Survey, 2019 Annual Social and Economic Supplement (CPS ASEC) |
| $10,000 - $19,999 | 8.81 | 6.16 | 6.72 | 8.03 | |
| $20,000 - $29,999 | 12.68 | 10.77 | 10.96 | 8.03 | |
| $30,000 - $39,999 | 13.67 | 11.13 | 11.78 | 7.91 | |
| $40,000 - $49,999 | 12.23 | 14.18 | 13.33 | 8.06 | |
| $50,000 - $59,999 | 10.90 | 16.46 | 15.12 | 7.2 | |
| $60,000 - $79,999 | 15.71 | 14.9 | 14.5 | 12.1 | |
| $80,000 - $99,999 | 8.86 | 11.07 | 10.73 | 9.54 | |
| $100,000 - $199,999 | 10.72 | 10.19 | 11.42 | 23.84 | |
| $200,000 or more | 1.24 | 1.17 | 1.29 | 10.25 | |
| **Education** | **MTurk 2017 (%)<br>N=5755** | **MTurk 2021 (%)<br>N=6752** | **MTurk 2021A (%)<br>N=5760** | **US (%)** | |
| No high school diploma | 0.46 | 0.28 | 0.33 | 10.6 | US Census Bureau, Current Population Survey, 2019 Annual Social and Economic Supplement (CPS ASEC) |
| High school graduate or GED | 11.05 | 6.93 | 8.08 | 28.32 | |
| Some college, no degree | 24.52 | 12.19 | 14.14 | 17.97 | |
| Occupational/technical/vocational program | 2.74 | 1.83 | 2.12 | 4.14 | |
| Associate degree: academic program | 11.83 | 6.62 | 7.68 | 5.65 | |
| Bachelor's degree | 36.84 | 51.16 | 49.78 | 21.28 | |
| Master's degree | 9.79 | 18.69 | 15.39 | 8.96 | |
| Professional school degree | 1.31 | 1.29 | 1.37 | 1.26 | |
| Doctoral degree | 1.44 | 0.93 | 1.02 | 1.82 | |
| Other | 0.03 | 0.08 | 0.1 | 0 | |

Table 1 details demographic characteristics. The MTurk samples tend to be more female, less Hispanic, and more likely to identify as white, have more than a high school degree, and more likely report annual household incomes in the $20,000 to $80,000 range than the general US population. Increasing the number of previous tasks completed (from MTurk 2017 to MTurk 2021) and conducting the survey 4 years later during the COVID-19 pandemic increased the proportion of individuals who endorsed being female, Hispanic, black/African-American, American Indian or Alaskan Native, and having a 4-year or higher degree. Exclusion of those who endorsed fake conditions (MTurk 2021A) improved the representativeness of gender and education but reduced the representativeness of ethnicity and race compared to those in MTurk 2021.

MTurk samples were younger than the general population, over-representing those from ages 25-44, and underrepresenting those age 50+ (Figure 1). Increasing the number of previous tasks completed from 100 to 500 in conjunction with the difference in when data were collected and impact of the COVID-19 pandemic increased the age of participants from the MTurk 2017 average of 35.9 to 39.6 years old for participants in MTurk 2021A, slightly higher than previous estimates of the age of MTurk workers.[16]

**Figure 1**
**Age Distribution in MTurk Compared to US Population**

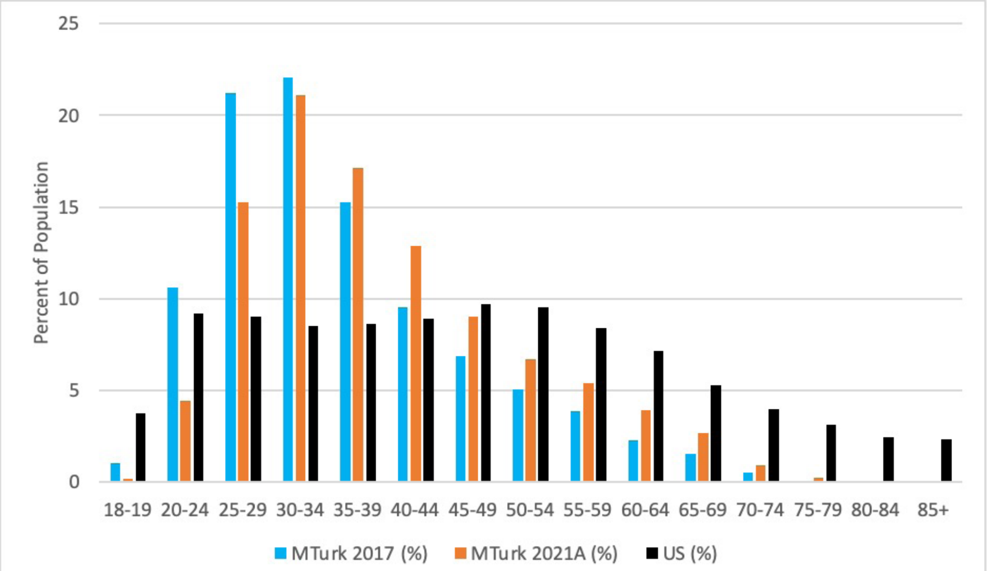

**Table 2**
**Estimates of Disease Prevalence in MTurk Compared to US Population**

| Condition | MTurk 2017 (%)<br>N=5755 | MTurk 2021 (%)<br>N=6752 | MTurk 2021A (%)<br>N=5760 | US (%) | Source |
|---|---|---|---|---|---|
| Anxiety | 42 | 34 | 28 | 19 | National Comorbidity Study Replication (NCS-R) |
| Depression | 30 | 40 | 34 | 7 | 2017 National Survey on Drug Use and Health (NSDUH) |
| Back pain | 26 | 45 | 40 | 29 | 2013-2015 National Health Interview Survey (NHIS) Data |
| Hypertension | 15 | 36 | 27 | 50 | National Health and Nutrition Examination Survey (NHANES) 2017-2018 |
| Neck Pain | 13 | 31 | 24 | 15 | 2013-2015 National Health Interview Survey (NHIS) Data |
| Asthma | 11 | 23 | 15 | 8 | 2018 National Health Interview Survey (NHIS) Data |
| Diabetes | 4 | 21 | 12 | 13 | National Health and Nutrition Examination Survey (NHANES) 2017-2018 |
| Heart Disease | 1 | 15 | 5 | 12 | 2018 National Health Interview Survey (NHIS) Data |
| COPD | 1 | 14 | 5 | 7 | 2017 Behavioral Risk Factor Surveillance System (BRFSS) |
| Cancer | 1 | 14 | 5 | 9 | 2018 National Health Interview Survey (NHIS) Data |
| Stroke | 1 | 14 | 4 | 3 | 2018 National Health Interview Survey (NHIS) Data |

Table 2 details the rates of endorsement for the healthcare conditions. We found a higher proportion of individuals with mental health conditions such as anxiety and depression in the MTurk samples than the national averages. Compared to the national averages, both MTurk 2017 and MTurk 2021 samples were less likely to report hypertension, but MTurk 2021 had higher rates of other conditions related to age (cancer, COPD, stroke, heart disease, and diabetes), and asthma. MTurk 2021 respondents reported more health conditions than MTurk 2017 except for anxiety. When we removed the fake conditions (MTurk 2021A), we found that that prevalence of health conditions decreased and was more similar to the national averages. Between 65% and 89% of individuals who endorsed a fake condition also endorsed every other condition and diagnosis included in the survey (not displayed).

## DISCUSSION

The representativeness of MTurk samples and the quality of responses is a concern for some researchers. Whereas various methods to improve quality exist, how those approaches impact the sample and its representativeness is not well understood. This study compares MTurk workers at 2 time-points and evaluates the impact of various methods for improving data quality on sample representativeness. It seems that requiring more previous MTurk tasks helped with the representativeness of the MTurk sample. Weeding out those who endorsed fake conditions also improved representativeness, but there are still several issues with creating representative samples in MTurk. Like previous studies, MTurk samples tend to be younger, "whiter," and more educated than national samples. Increasing the number of previous HITs completed (and the passage of time) did increase the average age of the MTurk sample and generally improved the representativeness by race and ethnicity. However, these approaches had little impact on national representativeness for income. Additionally, increasing the number of previous HITs completed improved the representativeness of the sample in terms of self-reported conditions and diagnoses, particularly among low prevalence conditions. When we removed those who endorsed fake conditions, the estimated prevalence of conditions and diagnoses moved closer to national estimates. However, we still see higher prevalence for anxiety, depression, asthma, back pain, and neck pain after removing those who endorsed a fake condition. These results may indicate that MTurk may be a good source for samples with high rates of anxiety, depression, asthma back and neck pain. We also saw that those who endorsed fake conditions endorsed a large percentage of self-reported conditions and diagnoses, potentially indicating that Turkers might attempt to game the survey for additional work.

Although our data quality approaches generally improve representativeness, the MTurk population may not be nationally representative. However, there are other approaches that could be used to improve representativeness in a targeted manner. One approach would be to oversample individuals with certain conditions or demographic groups and apply weights to create more representative samples. Researchers could consider targeting individuals with specific conditions or from particular demographic groups using methods like ours that reduce survey gaming. Whereas these approaches would incur additional costs, they could improve sample representativeness if the appropriate data quality approaches are also used.

Our study has several limitations. First, there was a 4-year gap between our first sample (MTurk 2017) and our more recent samples (MTurk 2021 and MTurk 2021A). In addition to the time difference, our later data were collected during the COVID-19 pandemic, which has had impacts on time availability and income for many workers. The number and types of individuals choosing to work on MTurk may have changed due to both conditions. Second, we are only using self-reported measured for conditions without a way to validate those responses. However, this same approach is used for national estimates and reasonably good correspondence between self-reports of conditions and medical records has been reported.[17]

Approaches to ensuring high quality MTurk data can support creating more representative samples of workers from MTurk. Our study shows the impact of various approaches to improving data quality and the subsequent effects on representativeness by comparing estimates to US population estimates. Future work is needed to continue to develop research use cases in which MTurk is the appropriate venue.

### Human Subjects Statement

The study was funded by the National Center for Complementary and Integrative Health (NCCIH), grant numbers 1R01AT010402-01A1 and

1R21AT009124-01. All procedures were reviewed and approved by the research team's Institutional Review Board (RAND Human Subjects Research Committee FWA00003425; IRB00000051). All individuals surveyed were given information about the study, including risk and benefits, and required to voluntarily consent before participation.

## Conflict of Interest Disclosure Statement

The authors have no relevant financial or non-financial interests to disclose.

## Acknowledgement

We acknowledge Rebecca Weir for providing support for data cleaning and processing.

## References

1. Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk a New Source of Inexpensive, Yet High-Quality, Data? Perspect Psychol Sci. 2011;6(1):3-5. https://doi.org/10.1177/1745691610393980
2. Mason W, Suri S. Conducting behavioral research on Amazon's Mechanical Turk. Behav Res Methods. 2012;44(1):1-23. https://doi.org/10.3758/s13428-011-0124-6
3. Mortensen K, Hughes TL. Comparing Amazon's Mechanical Turk platform to conventional data collection methods in the health and medical research literature. J Gen Intern Med. 2018;33(4):533-38. https://doi.org/10.1007/s11606-017-4246-0
4. Walters K, Christakis DA, Wright DR. Are Mechanical Turk worker samples representative of health status and health behaviors in the US? PLoS One. 2018;13(6):e0198835. https://doi.org/10.1371/journal.pone.0198835
5. Huff C, Tingley D. "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. Research & Politics. 2015;2(3). https://doi.org/10.1177/2053168018822174
6. Chandler J, Rosenzweig C, Moss AJ, et al. Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. Behav Res Methods. 2019;51(5):2022-38. https://doi.org/10.3758/s13428-019-01273-7
7. Hay JW, Gong CL, Jiao X, et al. A US population health survey on the impact of COVID-19 using the EQ-5D-5L. J Gen Intern Med. 2021;36(5):1292-301. https://doi.org/10.1007/s11606-021-06674-z
8. Chmielewski M, Kucker SC. An MTurk crisis? Shifts in data quality and the impact on study results. Soc Psychol Personal Sci. 2020;11(4):464-73. https://doi.org/10.1177/1948550619875149
9. Paolacci G, Chandler J, Ipeirotis PG. Running experiments on amazon mechanical turk. Judgment and Decision making. 2010;5(5):411-19. https://doi.org/10.1037/t69659-000
10. Agley J, Xiao Y, Nolan R, et al. Quality control questions on Amazon's Mechanical Turk (MTurk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7. Behav Res Methods. 2022;54(2):885-97. https://doi.org/10.3758/s13428-021-01665-8
11. Kennedy R, Clifford S, Burleigh T, et al. The shape of and solutions to the MTurk quality crisis. Political Sci Res Methods. 2020;8(4):614-29. https://doi.org/10.1017/psrm.2020.6
12. Ophir Y, Sisso I, Asterhan CS, et al. The Turker blues: Hidden factors behind increased depression rates among Amazon's Mechanical Turkers. Clin Psychol Sci. 2020;8(1):65-83. https://doi.org/10.1177/2167702619865973
13. Hilton LG. Advancing democratic evaluation: using crowdsourcing to include and engage program participants. Doctoral dissertation. The Claremont Graduate University; 2018. Available from: https://www.proquest.com/openview/09c46de10dff88ca46111fcbd4325adc
14. Litman L, Robinson J, Abberbock T. TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. Behav Res Methods. 2017;49(2):433-42. https://doi.org/10.3758/s13428-016-0727-z
15. Peer E, Vosgerau J, Acquisti A. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. Behav Res Methods. 2014;46(4):1023-31. https://doi.org/10.3758/s13428-013-0434-y
16. Arechar AA, Kraft-Todd GT, Rand DG. Turking overtime: How participant characteristics and behavior vary over time and day on Amazon Mechanical Turk. J Econ Sci Assoc. 2017;3(1):1-11. https://doi.org/10.1007/s40881-017-0035-0
17. Miller DR, Rogers WH, Kazis LE, et al. Patients' self-report of diseases in the Medicare Health Outcomes Survey based on comparisons with linked survey and medical data from the Veterans Health Administration. J Ambul Care Manage. 2008;31(2):161-77. https://doi.org/10.1097/01.JAC.0000314707.88160.9c