Original Paper

Excluding Those Who Report Having "Syndomitis" or "Chekalism" Improves Health Survey Data Quality

Ron D. Hays, PhD,[1] Nabeel Qureshi, MPH, MPhil,[2] Patricia M. Herman, ND, PhD, [2]
Anthony Rodriguez, PhD,[3] Arie Kapteyn,PhD[4] Maria Orlando Edelen, PhD[5]

[1]Division of General Internal Medicine and Health Services Research, UCLA Department of Medicine, Los Angeles, CA
[2]RAND Corporation, Behavioral and Policy Sciences, 1776 Main Street, Santa Monica, CA
[3]RAND Corporation, Behavioral and Policy Sciences, 20 Park Plaza #920, Boston, MA
[4]Center for Economic and Social Research, University of Southern California, Los Angeles, CA, USA
[5]Patient Reported Outcomes, Value and Experience (PROVE) Center, Department of Surgery, Brigham and Women's Hospital, Boston, MA

Corresponding author

Ron D. Hays, PhD
Division of General Internal Medicine and Health Services Research
UCLA Department of Medicine
1100 Glendon Avenue Suite 850
Los Angeles, CA, USA
310-794-2294
drhays@ucla.edu
https://orcid.org/0000-0001-6697-907X

Abstract

Background: Researchers have implemented a variety of approaches to increase data quality from existing online panels such as Amazon's Mechanical Turk (MTurk).

Objective: This study extends prior work by examining improvements in data quality and effects on mean estimates of health status by excluding respondents who endorse either or both of two fake health conditions ("Syndomitis" and "Chekalism").

Methods: Data were collected in 2021 from MTurk study participants, 18 years or older, with an internet protocol address in the United States who had completed a minimum of 500 previous MTurk "human intelligence tasks." The survey included questions about demographic characteristics, health conditions (including two fake conditions), and the Patient Reported Outcomes Measurement Information System (PROMIS®)-29+2 v2.1.

Results: Fifteen percent (n = 996 out of 6832) of the sample endorsed at least one of the two fake conditions at baseline. Those who endorsed a fake condition at baseline were more likely to be male, non-White, younger, report more health conditions, and take longer to complete the survey than those who did not endorse a fake condition. They also had lower score reliability and reported significantly worse self-reported health scores than those who did not endorse a fake condition. Excluding those who endorsed a fake condition reduced the overall mean PROMIS-29+2 v2.1 T-scores by 1-2 points and the PROPr preference-based score by 0.04.

Conclusions: This study provides evidence that asking about fake health conditions can help to screen out respondents who may be dishonest or careless respondents.

Key words

Misrepresentation; survey; data quality

Introduction

Amazon's Mechanical Turk (MTurk; www.MTurk.com) is a crowdsourcing platform including a pool of "workers" willing to complete tasks for low levels of compensation [1]. The extent to which MTurk and other convenience-based samples are representative of the general population is a concern in many studies [2]. Most MTurk participants are young, White, male, and highly educated, but report relatively poor mental health [3-4]. In addition to questions about representativeness, problems with data integrity among MTurk respondents have been identified [5]. Chandler et al. [6] found relatively low reliability of data provided by MTurk respondents who scored poorly on a test of comprehension and ability to respond to questions. Ophir et al. [7] reported that the estimated prevalence of depression was about 50% higher when inattentive responders were included.

Researchers have implemented a variety of approaches to increase data quality from existing online panels such as removing those who have an average item response of one second or less, adding screener questions before the main survey, doing internet protocol (IP) address verification, and conducting test-retest comparisons on demographic variables [8-9]. Excluding people who endorse bogus (fake) health conditions has also been employed [4]. This study extends the work of Qureshi et al, [4] by examining improvements in data quality and effects on mean estimates of health status by excluding respondents who endorse either or both of two fake health conditions ("Syndomitis" and "Chekalism").

Methods

Data were collected in 2021. Eligible MTurk study participants were 18 years or older with an IP address in the United States who had completed a minimum of 500 previous MTurk "human intelligence tasks" (surveys, writing product descriptions, coding, or identifying content in images or videos) with a completion rate of at least 95%. All participants provided electronic consent at the start of the survey. Those who completed a general health survey and reported currently having back pain were asked to complete a back pain survey. Those who completed the general health and back pain survey were paid $4 for participation. All procedures were reviewed and approved by the research team's institutional review

board (XXXX Human Subjects Research Committee FWA00003425; IRB00000051).

Survey

The survey included questions about demographic characteristics and health conditions. Thirteen bona fide health conditions were assessed: Have you EVER been told by a doctor or other health professional that you had 1) hypertension; 2) high cholesterol; 3) heart disease; 4) angina; 5) heart attack; 6) stroke; 7) asthma; 8) cancer; 9) diabetes; 10) chronic obstructive pulmonary disease; 11) arthritis; 12) anxiety disorder; and 13) depression. In addition, the survey asked respondents if they were ever told they had "Syndomitis" (a fake condition). Further, participants were asked if they currently have 9 other bona fide conditions: 1) allergies or sinus trouble; 2) back pain; 3) sciatica; 4) neck pain; 5) trouble seeing; 6) dermatitis; 7) stomach trouble; 8) trouble hearing; and 9) trouble sleeping. They were also asked if they have "Chekalism" (a fake condition).

The Patient Reported Outcomes Measurement Information System (PROMIS®)-29+2 v2.1 (PROPr) was also administered [10]. The PROMIS-29+2 v2.1 includes 7 multi-item scales with 4 items each (physical function, pain interference, fatigue, sleep disturbance, depression, anxiety, ability to participate in social roles and activities), a 2-item cognitive function scale, and a single 0-10 pain intensity item. All items within 7 of the multi-item scales are worded in the same direction (e.g., represent better health) but 2 of the items in the sleep disturbance scale were worded in the direction of less disturbance and the other 2 items were worded to indicate more disturbance. In addition to scores for the 8 scales and the single pain intensity item, the PROMIS-29+2 v.2.1 yields a physical health and mental health summary scores and a preference-based score (PROPr) [11-12].

Analysis Plan

We compute at baseline and at the 3-month follow-up estimates of internal consistency reliability [13] product-moment correlations among scales, and mean scores for the PROMIS-29+2 v2.1 separately for those who did not versus did endorse a fake health condition. We hypothesized that those who endorse a fake condition provide less reliability information, smaller correlations among scales, and worse mean

scores than those who do not endorse a fake condition.

Results

Fifteen percent (n = 996 out of 6832) of the sample endorsed at least one of the two fake conditions at baseline. Characteristics of those who did versus did not endorse a fake condition are shown in Table 1.

Table 1. Characteristics of Those Endorsing and Not Endorsing a Fake Health Condition

Variable	Did not Endorse Fake Health Condition	Endorsed Fake Health Condition
Gender		
Female	46%	32%
Male	53%	67%
Transgender or do not identify as female, male or transgender	1%	0%
Non-White	18%	28%
Age	40 years old	38 years old
Number of conditions	4	15
Minutes to complete	19	27

Those who endorsed a fake condition at baseline were more likely to be male, non-White, younger, report more health conditions, and take longer to complete the survey than those who did not endorse a fake condition. Those who endorsed a fake condition at baseline were not asked to complete a 3-month survey. Even though they did not endorse a fake condition at baseline, 6% (n = 59) endorsed at least one of the fake conditions on the 3-month survey (n = 972, 94%, did not endorse a fake condition). There, the estimated proportion of fakers in the sample is 25% (see Appendix), within the range of 20-30% of detected fraud reported in other web-based studies [14].

Baseline Survey PROMIS-29+2 v2.1 Data Quality and Mean Scores

Internal consistency reliabilities for the PROMIS-29+2 v2.1 scales were uniformly larger at baseline for those who did not endorse a fake condition than for those who did (Table 2).

Table 2. Internal Consistency Reliability of PROMIS Scales at Baseline

Scale	Did not Endorse Fake Health Condition (n = 5836)	Endorsed Fake Health Condition (n = 996)
Physical function	0.89	0.69
Pain interference	0.94	0.80
Fatigue	0.92	0.80
Depression	0.92	0.78
Anxiety	0.90	0.78
Sleep disturbance	0.84	-.27
Ability to participate in social roles/ activities	0.92	0.77
Cognitive function	0.77	0.65

Coefficient alpha for sleep disturbance (the scale with 2 items worded in the direction of less sleep disturbance and the other 2 worded in the opposite direction) was negative. Consistent with the difference in reliability estimates, most of the product-moment correlations among the PROMIS-29 v2.1 scales were lower for those who endorsed a fake condition than for those who did not (Table 3).

Table 3. Correlations Among PROMIS Scales at Baseline

	PF	**PIter**	**PIten**	**FAT**	**DEP**	**ANX**	**SLPD**	**SOC**	**CF**
PF		-.12	-.14	-.20	-.21	-.26	-.11	0.15	-.01
PIter	-.72		0.26	0.73	0.60	0.56	0.01	-.78	0.64*
PIten	-.59	0.72		0.32	0.29	0.29	0.01	-.27	0.22*
FAT	-.47	0.54	0.48		0.72	0.66	0.06	-.70	0.50*
DEP	-.43	0.50	0.45	0.71		0.77	0.09	-.68	0.40*
ANX	-.43	0.51	0.46	0.70	0.82		0.05	-.64	0.39*
SLPD	-.30	0.37	0.37	0.61	0.53	0.52		-.03	-.17
SOC	0.64	-.72	-.56	-.68	-.66	-.66	-.49		-.54*
CF	0.33	-.31	-.29	-.30	-.37	-.37	-.31	0.39	

Note: Did not endorse fake health condition <u>below</u> diagonal; endorsed fake condition <u>above</u> diagonal.
* Correlation is in the "wrong" direction.
PF = physical function; PIter = Pain interference; PIten = Pain intensity; FAT = Fatigue; DEP = Depression; ANX = Anxiety; SLPD = Sleep disturbance; SOC = Ability to participate in social roles and activities; CF = Cognitive function.

Those who endorsed a fake condition had significantly worse self-reported health scores for all scales except for the sleep disturbance scale. Except for sleep disturbance, excluding those who endorsed a fake condition changed the mean PROMIS-29+2 v2.1 T-scores by 1-2 points and the PROPr preference-based score by 0.04 towards better self-reported health. The sleep disturbance scale means did not differ between those who endorsed versus did not endorse a fake condition because the former provided inconsistent answers to the positively and negatively worded items (Table 4).

Table 4. PROMIS Scale Means at Baseline

Scale	Did not Endorse Fake Health Condition (n = 5836)	Endorsed Fake Health Condition (n = 996)	Overall Sample (n = 6832)
Physical function	49	41	48
Pain interference	51	63	53
Pain intensity	52	64	54
Fatigue	50	58	51
Depression	53	63	54
Anxiety	54	63	56
Sleep disturbance	50	51	50
Ability social roles/ activities	53	43	52
Cognitive function	50	47	49
P-29 Physical Health Summary	49	40	48
P-29 Mental Health Summary	50	39	48
PROPr	0.45	0.20	0.41

Note: P-29 = PROMIS®-29; PROPr = PROMIS preference-based score.

3-Month Survey PROMIS-29+2 v2.1 Data Quality and Mean Scores

Differences between those who reported on the 3-month survey having versus not having a fake condition were like what was observed on the baseline survey. Internal consistency reliabilities for the PROMIS-29+2 v2.1 scales were uniformly larger at for those who did not endorse a fake condition than for those who did (Table 5).

Table 5. Internal Consistency Reliability of PROMIS Scales at 3 Months

Scale	Did not Endorse Fake Health Condition (n = 972)	Endorsed Fake Health Condition (n = 59)
Physical function	0.92	0.53
Pain interference	0.95	0.76
Fatigue	0.94	0.77
Depression	0.93	0.81
Anxiety	0.92	0.80
Sleep disturbance	0.88	-.21
Ability to participate in social roles and activities	0.94	0.78
Cognitive function	0.70	0.44

Coefficient alpha for sleep disturbance was negative among those who endorsed a fake condition because this subgroup answered all 4 questions similarly despite the wording of 2 items indicating less sleep disturbance and 2 items indicating more sleep disturbance. Most of the product-moment correlations among the PROMIS-29+2 v2.1 scales were smaller for those who endorsed a fake condition than for those who did not (Table 6).

Table 6. Correlations Among PROMIS Scales at 3-Months

	PF	**PIter**	**PIten**	**FAT**	**DEP**	**ANX**	**SLPD**	**SOC**	**CF**
PF		-.31	-.38	-.33	-.51	-.55	-.25	0.32	0.12
PIter	-.73		0.40	0.71	0.62	0.65	0.22	-.69	0.43*
PIten	-.58	0.73		0.34	0.33	0.47	0.20	-.23	0.16*
FAT	-.46	0.52	0.39		0.65	0.75	0.16	-.71	0.33*
DEP	-.33	0.40	0.33	0.61		0.82	0.26	-.68	0.27*
ANX	-.34	0.43	0.35	0.62	0.81		0.26	-.71	0.15*
SLPD	-.31	0.37	0.32	0.58	0.50	0.49		-.26	-.18
SOC	0.64	-.66	-.51	-.69	-.61	-.64	-.53		-.17*
CF	0.29	-.33	-.30	-.42	-.46	-.49	-.38	0.49	

Note: Did not endorse fake health condition below diagonal; endorsed fake condition above diagonal.
* Correlation is in the "wrong" direction.
PF = physical function; PIter = Pain interference; PIten = Pain intensity; FAT = Fatigue; DEP =

Depression; ANX = Anxiety; SLPD = Sleep disturbance; SOC = Ability to participate in social roles and activities; CF = Cognitive function.

As was the case at baseline, those who endorsed a fake condition at 3 months had significantly worse health scores for all scales except for the sleep disturbance scale where they provided inconsistent answers to the positively and negatively worded items (Table 7).

Table 7. PROMIS Scale Means at 3-Months

Scale	Did not Endorse Fake Health Condition (n = 972)	Endorsed Fake Health Condition (n = 59)	Overall Sample (n = 1031)
Physical function	46	41	46
Pain interference	54	62	55
Pain intensity	56	62	56
Fatigue	54	57	54
Depression	55	62	55
Anxiety	56	63	56
Sleep disturbance	53	51	54
Ability social roles/ activities	51	44	51
Cognitive function	50	46	50
P-29 Physical Health Summary	47	40	46
P-29 Mental Health Summary	46	41	46
PROPr	0.37	0.22	0.37

Note: P-29 = PROMIS®-29; PROPr = PROMIS preference-based score.

Because of the small sample size, the means including and excluding those who endorsed a fake condition were similar. Note that our estimates are conservative because we estimate there are about 4% "fake" respondents still undetected in the second data wave (25%-15%-6%) = 4%.

Discussion

This study provides evidence that asking about fake health conditions can help to screen out respondents who may be either dishonest or careless respondents. The 15% rate of endorsing a fake health condition is consistent with prior estimates of about a 12% rate of careless responding in crowdsourced samples [9]. The association of reporting a fake condition with a greater number of self-reported health conditions

parallels research documenting that those who report using a non-existent recreational drug ("bindro") tend to self-report more use of actual drugs [15].

Those who endorsed a fake health condition provided less reliable information and tended to have more negative reports about their health. Moreover, the internal consistency reliability estimates in this study were likely overly optimistic because the wording of most of the items was in the same direction so that consistently answering in one direction of the response scale may bias reliability estimates upward [16].

The one scale (sleep disturbance) where changing the direction of responding was needed to be consistent in self-reports had zero reliability (negative alpha) among those who endorsed a fake condition. Note that we found a similar pattern in this dataset for correlations of a PROMIS cognitive function item ("I have had trouble shifting back and forth between different activities that require thinking") not included in PROMIS-29+2 V2.1 with the 2 items that are included (results not presented).

Careless responding and acceptance acquiescent response patterns are problematic because they introduce error in the measurement of the concept of interest [17-18]. In the current study, consistently selecting extreme responses that represent worse health for most items may have been a strategy adopted by some members of MTurk. Like gaming demographic questions to be study eligible, the consistent reporting of negative health may maximize the likelihood of qualifying for study participation [19]. Use of balanced scales has been advocated to address responding the same way to items regardless of content, but this means that those with the problematic response patterns receive "middling scores on the scale regardless of their true attitudes" [20].

One caveat about the value of including bogus health conditions to screen out respondents is that its usefulness will fade over time if information about it spreads among potential survey respondents. For example, the urban dictionary warns readers not to select "Bindro" on surveys of drug use because selecting it "voids the whole test" (https://www.urbandictionary.com/define.php?term=Bindro). If

potential survey respondents become aware of the fake health conditions, it may be necessary to rely on consistency checks using person fit indices for items within scales that are worded in opposite directions to identify careless respondents [21-22].

Conflicts of Interests

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgements

The author(s) disclosed the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Center for Complementary and Integrative Health (NCCIH) [Grant Number 1R01AT010402-01A1]. NCCIH had no role in the design; data collection, analysis, or interpretation; or writing of this manuscript.

Abbreviations

IP: internet protocol

MTurk: Amazon's Mechanical Turk

PROPr: Patient Reported Outcomes Measurement Information System preference-based score

PROMIS®: Patient Reported Outcomes Measurement Information System

References

1. Mason W, Suri S. Conducting behavioral research on Amazon's Mechanical Turk. Behav Res Methods. 2012 Mar;44(1):1-23. doi: 10.3758/s13428-011-0124-6. PMID: 21717266.

2. Hays RD, Liu H, Kapteyn A. Use of Internet panels to conduct surveys. Behav Res Methods. 2015 Sep;47(3):685-90. doi: 10.3758/s13428-015-0617-9. PMID: 26170052; PMCID: PMC4546874.

3. Hilton LG, Coulter ID, Ryan GW, Hays RD. Comparing the Recruitment of Research Participants With Chronic Low Back Pain Using Amazon Mechanical Turk With the Recruitment of Patients From Chiropractic Clinics: A Quasi-Experimental Study. J Manipulative Physiol Ther. 2021 Oct;44(8):601-611. doi: 10.1016/j.jmpt.2022.02.004. Epub 2022 Jun 18. PMID: 35728997.

4. Qureshi N, Edelen M, Hilton L, Rodriguez A, Hays RD, Herman PM. Comparing Data Collected on Amazon's Mechanical Turk to National Surveys. Am J Health Behav. 2022 Oct 17;46(5):497-502. doi: 10.5993/AJHB.46.5.1. PMID: 36333833.

5. Siegel, JT, Navarro, M. A conceptual replication examining the risk of overtly listing eligibility criteria on Amazon's Mechanical Turk. *J Appl Soc Psychol*. 2019; 49: 239– 248.

6. Chandler J, Rosenzweig C, Moss AJ, Robinson J, Litman L. Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. Behav Res Methods. 2019 Oct;51(5):2022-2038. doi: 10.3758/s13428-019-01273-7. PMID: 31512174; PMCID: PMC6797699.

7. Ophir, Y., Sisso, I., Asterhan, C. S., Tikochinski, R., Reichart, R. The Turker blues: Hidden factors behind increased depression rates among Amazon's Mechanical Turkers. Clinical Psychological Science 2020; 8(1): 65-83. doi.org/10.1177/2167702619865973.

8. Göritz, AS, Borchert K, Hirth M. Using attention testing to select crowdsourced workers and research participants. Social Science Computer Review 2021;39(1): 84–104. doi.org/10.1177/0894439319848726

9. Jones A, Earnest J, Adam M, Clarke R, Yates J, Pennington CR. Careless responding in crowdsourced alcohol research: A systematic review and meta-analysis of practices and prevalence.

Exp Clin Psychopharmacol. 2022 Aug;30(4):381-399. doi: 10.1037/pha0000546. Epub 2022 Feb 7. PMID: 35130007.

10. Cella D, Choi SW, Condon DM, Schalet B, Hays RD, Rothrock NE, et al. PROMIS® Adult Health Profiles: Efficient Short-Form Measures of Seven Health Domains. Value Health 2019 May;22(5):537-544. doi: 10.1016/j.jval.2019.02.004. PMID: 31104731; PMCID: PMC7201383.

11. Dewitt B, Feeny D, Fischhoff B, Cella D, Hays RD, Hess R, et al. Estimation of a Preference-Based Summary Score for the Patient-Reported Outcomes Measurement Information System: The PROMIS®-Preference (PROPr) Scoring System. Med Decis Making. 2018 Aug;38(6):683-698. doi: 10.1177/0272989X18776637. Epub 2018 Jun 26. PMID: 29944456; PMCID: PMC6502464.

12. Hays RD, Spritzer KL, Schalet BD, Cella D. PROMIS®-29 v2.0 profile physical and mental health summary scores. Qual Life Res. 2018 Jul;27(7):1885-1891. doi: 10.1007/s11136-018-1842-3. Epub 2018 Mar 22. PMID: 29569016; PMCID: PMC5999556.

13. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16(3), 297–334.

14. Glazer JV, MacDonnell K, Frederick C, Ingersoll K, Ritterband LM. Liar! Liar! Identifying eligibility fraud by applicants in digital health research. Internet Interv. 2021 May 9;25:100401. doi: 10.1016/j.invent.2021.100401. PMID: 34094883; PMCID: PMC8164029.

15. Johnson TP. Sources of Error in Substance Use Prevalence Surveys. Int Sch Res Notices. 2014 Nov 5;2014:923290. doi: 10.1155/2014/923290. PMID: 27437511; PMCID: PMC4897110.

16. Zimmerman DW, Zumbo BD, Lalonde C. Coefficient alpha as an estimate of test reliability under violation of two assumptions. Educational and Psychological Measurement 1993;53(1), 33–49. doi.org/10.1177/0013164493053001.

17. Bentler PM, Jackson DN, Messick S. Identification of content and style: a two-dimensional interpretation of acquiescence. Psychol Bull. 1971 Sep;76(3):186-204. doi: 10.1037/h0031474. PMID:

4399323.

18. Meade AW, Craig SB. Identifying careless responses in survey data. Psychol Methods. 2012 Sep;17(3):437-55. doi: 10.1037/a0028085. Epub 2012 Apr 16. PMID: 22506584.

19. Chandler JJ, Paolacci G. Lie for a dime: When most prescreening responses are honest, but most study participants are impostors. Social Psychological and Personality Science 2017;8(5): 500-508. doi.org/10.1177/1948550617698203.

20. Kuru O, Pasek J. Improving social media measurement in surveys: Avoiding acquiescence bias in Facebook research. Computers in Human Behavior 2016;57: 82-92.

21. Hays RD. Response 1 to Reeve's chapter: Applying Item response theory for questionnaire evaluation. In J. Madans, K. Miller, A. Maitland, & G. Willis (Eds.). Question Evaluation Methods: Contributing to the Science of Data Quality. New York, NY: Wiley & Sons, Inc; 2011; p. 125-135.

22. Reise SP. Using Multilevel Logistic Regression to Evaluate Person-Fit in IRT Models. Multivariate Behav Res. 2000 Oct 1;35(4):543-68. doi: 10.1207/S15327906MBR3504_06. PMID: 26811204.

Appendix: Estimated Proportion of Respondents Who Report Fake Conditions

Assume there is proportion φ of fake respondents in the sample. Assume that the probability to get caught (i.e., endorsing one or both fake conditions) is equal to η. Then in the first wave a proportion $\varphi\eta$ of respondents will get caught. These will get removed from the sample, so in the second wave the true proportion of fake respondents will be equal to $\varphi(1-\eta)$ and the proportion that will get caught in the second wave is then $\varphi(1-\eta)\eta$. From the numbers on slides 10 and 14 we then get

$$\eta\varphi = \frac{996}{6832} = .146$$
$$\varphi(1-\eta)\eta = \frac{59}{972} = .061$$

Solving these equations, one gets $\eta = .58,\ \varphi = .25$

Thus, we find that one quarter of the respondents are fake respondents. We have assumed that the three months follow-up sample is a random draw from the non-fake respondents in the first wave.