Title

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.

Journal

Journal of manipulative and physiological therapeutics, 44(8)

ISSN

0161-4754

Authors

Hilton, Lara G
Coulter, Ian D
Ryan, Gery W
et al.

Publication Date

2021-10-01

DOI

10.1016/j.jmpt.2022.02.004

Peer reviewed

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

Lara G. Hilton, PhD, MPH, [a] Ian D. Coulter, PhD, [b] Gery W. Ryan, PhD, [c] and Ron D. Hays, PhD [d]

Abstract

Objective: The purpose of this study was to compare the crowdsourcing platform Amazon Mechanical Turk (MTurk) with in-person recruitment and web-based surveys as a method to (1) recruit study participants and (2) obtain low-cost data quickly from chiropractic patients with chronic low back pain in the United States.

Methods: In this 2-arm quasi-experimental study, we used in-person clinical sampling and web-based surveys from a separate study (RAND sample, n = 1677, data collected October 2016 to January 2017) compared with MTurk (n = 310, data collected November 2016) as a sampling and data collection tool. We gathered patient-reported health outcomes and other characteristics of adults with chronic low back pain receiving chiropractic care. Parametric and nonparametric tests were run. We assessed statistical and practical differences based on *P* values and effect sizes, respectively.

Results: Compared with the RAND sample, the MTurk sample was statistically significantly younger (mean age 35.4 years, SD 9.7 vs 48.9, SD 14.8), made less money (24% vs 17% reported less than $30,000 annual income), and reported worst mental health than the RAND sample. Other differences were that the MTurk sample had more men (37% vs 29%), fewer White patients (87% vs 92%), more Hispanic patients (9% vs 5%), fewer people with a college degree (59% vs 68%), and patients were more likely to be working full time (62% vs 58%). The MTurk sample was more likely to have chronic low back pain (78% vs 66%) that differed in pain frequency and duration. The MTurk sample had less disability and better global health scores. In terms of efficiency, the surveys cost $2.50 per participant in incentives for the MTurk sample. Survey development took 2 weeks and data collection took 1 month.

Conclusion: Our results suggest that there may be differences between crowdsourcing and a clinic-based sample. These differences range from small to medium on demographics and self-reported health. The low incentive costs and rapid data collection of MTurk makes it an economically viable method of collecting data from chiropractic patients with low back pain. Further research is needed to explore the utility of MTurk for recruiting clinical samples, such as comparisons to nationally representative samples. (J Manipulative Physiol Ther 2021;44;601-611)

Key Indexing Terms: *Crowdsourcing; Back Pain; Chiropractic; Spine*

Introduction

Traditional methods of participant recruitment and data collection in health services research involve in-person enrollment, face-to-face interviews, telephone interviews, mail, and web-based surveys. Each of these methods can be expensive and time consuming. Innovations in online access to the American public hold promise for overcoming the hurdles of logistics, cost, and time associated with including patients and other stakeholders in research.

One alternative to costly and time intensive traditional methods of recruiting patients from clinical settings is crowdsourcing.[1,2] Crowdsourcing is the process of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, and especially from

[a] University of Southern California, Los Angeles, California.
[b] RAND Corporation, Health Division, Los Angeles, California.
[c] Kaiser Permanente Tyson Medical School, Los Angeles, California.
[d] Department of Medicine, University of California Los Angeles, Los Angeles, California.
Corresponding author: Ian Coulter, PhD. 6 Leslie Frost Lane, Lindsay, Ontario, Canada, K9V 4R6.
(e-mail: *coulter@rand.org*).

Paper submitted September 17, 2021; in revised form February 21, 2022; accepted February 22, 2022.
0161-4754

https://doi.org/10.1016/j.jmpt.2022.02.004

an online community. Crowdsourcing provides access to a large pool of participants and collects data faster and more cheaply than typical data collection methods.[1,2] The crowdsourcing tool has been used in behavioral and psychological research,[3,4] but it has not yet been tested in those with chronic pain. Assessing the comparability of crowdsourcing to traditional survey methods has the potential to advance pain research and has implications for accessing other clinical populations.

Crowdsourcing as a Potential Solution to Address Gaps in Literature

The notion of the *wisdom of crowds* was established in the late 1800s when John Galton found that layperson estimates of livestock weights were "more creditable to the trustworthiness of a democratic judgment than might have been expected."[5] Contemporary web-based crowdsourcing first emerged in 2006 as an online labor market where services, ideas, or content were obtained for a fee from a large group of people, and especially from an online community.[6] Crowdsourcing is "the paid recruitment of an online, independent global workforce for the objective of working on a specifically defined task or set of tasks."[7] There are various kinds of crowdsourcing: *crowdfunding* (eg, donate money to fund a project), *crowd labor* (eg, transcribe audio files), and *crowd research* (eg, respond to surveys).[8]

One of the most used crowdsourcing platform in research contexts is Amazon Mechanical Turk (MTurk).[3] MTurk operates a marketplace for work that requires human intelligence. The MTurk web service enables organizations to access this marketplace and an on-demand workforce of over 500,000 participants by posting Human Intelligence Tasks (HITs) that workers can browse and choose.[9]

An emerging literature demonstrates the utility of crowdsourcing for conducting social, behavioral, and clinical research.[4,10] Participants tend to be slightly more demographically diverse than internet-based or in-person convenience samples. Furthermore, participants can be recruited rapidly and inexpensively, and the quality of the data is similar to that of traditional methods.[3,11,12]

MTurk has been used to code psychological constructs in a reliable, accurate, and efficient way, as well as to code qualitative texts of Twitter comments related to diabetes.[13,14] Recent studies have examined the utility of crowdsourcing to advance patient engagement in health and medicine. Weiner [15] argued for crowdsourcing to support patient-centered care, based on empirical literature showing that patient participation has the potential to shape health policy and clinical decision-making. Hogg[16] tested crowdsourcing to engage patients in priority setting and strategic planning within a primary care organization. He recommended crowdsourcing as a novel opportunity to capture many voices throughout the research process. Other biomedical studies used crowdsourcing for mining big data, annotating biomedical language in PubMed, and testing the ability of the crowd to detect errors in biomedical ontologies. These studies suggest that crowdsourcing may augment existing medical research methods.[17-19] In addition, Shapiro et al.[4] established the reliability and validity of MTurk data for studying clinical populations and gave guidance for maximizing these attributes when using crowdsourcing software.

Need for Research on Crowdsourcing for Patients With Low Back Pain

The Institute of Medicine reports that over 116 million US adults experience chronic pain at a cost to society of at least $560 to 635 billion annually.[20] Collecting information about chronic pain from a representative sample of individuals who experience it provides insights that can inform policy and clinical pain management. Research on people with chronic pain is especially needed considering the vast need to find nonpharmacologic approaches to treat pain and mitigate the current opioid epidemic. However, the costs in time and money of recruitment and data collection through surveys and interviews present barriers that may be reduced using new technologies.

At present, there are no studies that compare crowdsourcing with in-person surveys as a method to obtain data from individuals with chronic low back pain in the United States. Therefore, the purpose of this study was to compare the crowdsourcing platform MTurk to access, recruit, screen, and survey a sample of people with chronic pain with data collected from a clinical sample of patients with chronic low back pain who had been recruited for studies by the RAND Center of Excellence in Research on Complementary and Alternative Medicine (RAND Study).[21] The research questions for this study were as follows: (1) Can we access and recruit people who are like the clinic-based sample? (2) How do the 2 methods compare in terms of efficiencies in time and cost?

Method

We conducted a cross-sectional comparison of 2 independent samples, 1 sample from a RAND clinic-based study and the other an MTurk sample of patients receiving care from chiropractic clinics (Fig 1).

Procedure

Using microbatching (ie, a process that allows the researcher to automatically release a pre-set number of tasks per hour), we released 9 human intelligence tasks (HITs) every hour, 24 hours per day, for one month for a total target of 6,048 participants. To accomplish microbatching, we used an application called TurkPrime. TurkPrime connects with the MTurk programming interface and

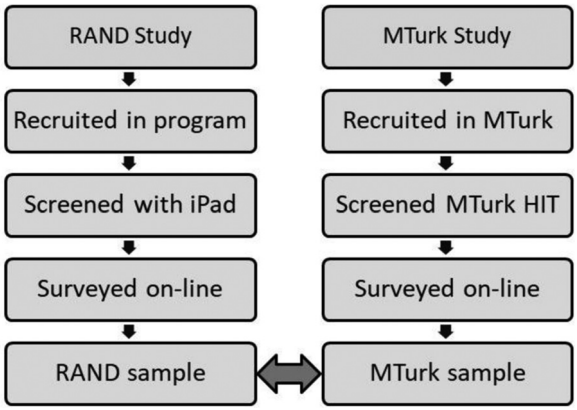

Fig. 1. *Overview of procedures - RAND clinic-based study and the MTurk sample.*

enables greater control over the survey process.[22] TurkPrime has a tool that breaks a larger survey into microbatches of fewer than 10 participants each and excludes people who have already completed the study.

Microbatching provided 2 benefits. First, the microbatches were launched at different times of the day, which increased the potential representativeness of the sample. Because many workers enter and leave MTurk during a day, collecting data at different times of the day reduced the bias that could result from collecting data from workers who happened to be online on a specific day or at a specific time (eg, Monday at 9 AM). Second, microbatching reduced Amazon's fees by 50%. When fewer than 10 HITs were requested, the fees were limited to 20% of the total cost of the survey incentives, instead of the standard rate of 40%.

It has been previously reported that few people on MTurk appear to be untrustworthy and deceitful.[23] However, even a few people being dishonest about eligibility to participate in high-paying HITs may threaten a study's validity. To minimize issues of untrustworthy or deceitful MTurk participants signing up, we did not overtly state eligibility requirements, but instead selected workers using a screener.[24,25] We recruited MTurk participants using screening questions built into the survey we fielded in MTurk and included only participants with chronic low back pain and chiropractic use. After screening, participants were asked if they would continue to answer a series of questions for a bonus payment.

Participants

After internal review board review and approval from the RAND Human Subjects Research Committee (HSPC Project ID: 2015-1061; FWA00003425), crowd-sourced participants were recruited through MTurk. Participants had to live in the United States and have at least 90% approval ratings on past completed tasks. MTurk participants were compensated $1.00 for the general health screener survey and an additional $1.50 if they experienced recent or current low back pain and answered additional survey questions related to their condition. The clinical sample of patients used for comparison in this study was recruited through a practice-based network of providers within a separate RAND study. A total of 125 clinics were recruited into the study across 6 states: California, Florida, Minnesota, New York, Oregon, and Texas. Data were collected between October 2016 and January 2017 through an iPad-based prescreening questionnaire in the clinic, and links to full screening and baseline online questionnaires were emailed. Participants in the RAND sample received up to $200 in incentives for completing all questionnaires in the study.

Inclusion criteria for both samples were age (21 and older), condition (low back pain), provider utilization (chiropractic care), chronicity (at least 3 months of pain or self-reported chronic pain), and no workers' compensation or personal injury litigation associated with the condition. Participant characteristics and differences between groups are shown in Tables 1-4.

This study has been registered, reviewed, and made public on ClinicalTrials.gov as RAND Protocol Record 1R21AT009124-01, Patient Engagement Via Crowdsourcing.

Measures

To test comparability of the samples, the MTurk survey included demographic items; pain intensity; functional disability; general, physical, and mental health; overall quality of life; and measures of chronicity which include pain duration, frequency, and self-identified chronicity. To gauge efficiency of time and cost, we compared time for data collection and total cost of participant incentives from both samples. The survey measures are described briefly herein.

Demographic variables included age, sex, race and ethnicity, education, employment status, income, and household number. Other characteristics captured for comparison were insurance status, coverage for chiropractic care, cost burden of chiropractic care, last time the participant received chiropractic care, and the number of chiropractic visits in the last 3 months.

Pain intensity was measured using Numeric Rating Scale (NRS) items that ask, "In the last 7 days, how would you rate your pain on average?" and "In the last 7 days, how would you rate your pain at its worst?" The response scale is 0 (no pain) to 10 (worst imaginable pain).

The Oswestry Disability Index (ODI) is the most commonly used tool to measure a functional disability.[26] It consists of 10 items assessing pain intensity, personal care, lifting, walking, sitting, standing, sleeping, sex life, social life, and traveling. The scale is scored across all items to determine a disability rating: <20% minimal disability, 21% to 40% moderate, 41% to 60% severe, 61% to 80% crippled, and 81% to 100% bedbound.

Table 1. *Participant Characteristics: Sex, Race, Ethnicity, and Age*

	RAND (n = 1677) Frequency (%)	MTurk (n = 310) Frequency (%)	Difference
Sex			
Female	1184 (70.6)	196 (63.2)	7.4
Male	477 (28.4)	114 (36.8)	-8.3
Not reported	16 (1.0)	0 (0.0)	1.0
Race			
White	1460 (87.1)	267 (86.1)	0.9
Asian	33 (2.0)	15 (4.8)	-2.9
African American	32 (1.9)	12 (3.9)	-2.0
American Indian/Alaskan Native	6 (0.4)	2 (0.6)	-0.3
Native Hawaiian/Pacific Islander	5 (0.3)	0 (0.0)	-0.3
Multiple race	52 (3.1)	11 (3.5)	-0.4
Other	5 (0.3)	0 (0.0)	0.3
Not reported	84 (5.0)	1 (0.3)	4.7
Ethnicity			
No, not Hispanic nor Latino	1544 (92.1)	281 (90.6)	1.4
Yes, Hispanic or Latino	74 (4.4)	29 (9.4)	-4.9
Not reported	59 (3.5)	0 (0.0)	3.5
Age			
Mean (standard deviation)	48.9 (14.8)	35.4 (9.7)	13.5 y
Ranges	21-95	21-77	

The Patient-Reported Outcomes Measurement Information System measures included general health, physical health, mental health, and overall quality of life items administered using an excellent to poor response scale.[27-30]

Chronicity of pain was captured with survey items on pain duration, frequency, and self-identified chronicity—that is, whether the participants considered themselves to have a chronic pain condition. Duration was measured by asking participants how long they have had pain before seeing a chiropractor. Frequency was measured by asking how often in the last 6 months had pain been a problem before they saw a chiropractor. Duration and frequency were recommended by the NIH Task Force as a way to define chronicity.[31] We extended this definition by creating items of self-reported chronicity by asking patients if they considered their pain to be chronic (Yes, No, and Don't know).

Analysis

A power analysis was conducted using G-Power software to provide guidance on the sample size needed to detect a medium effect size difference.[32] For the comparison of ODI scores between the 2 samples at 95% power with a significance level of .05 and in order to detect a medium effect, a sample size of 210 total participants was required. To account for potential nonresponse or drop out, we increased the target sample to 300.

We used *t* tests for comparisons of pain and function measures (ODI, last 7 days average pain, last 7 days worst pain, and pain if no chiropractic care) and demographics (age, number in household). We estimated effect sizes for each using Cohen's *d*. Cohen suggested that $d = 0.2$ be considered a "small" effect size, 0.5 represents a "medium" effect size and 0.8 a "large" effect size. If groups' means

Table 2. *Participant Characteristics: Education and Income*

	RAND (n = 1677) Frequency (%)	MTurk (n = 310) Frequency (%)	Difference
Education			
No high school diploma	5 (0.3)	1 (0.3)	0.0
High school graduate or GED	124 (7.4)	43 (13.9)	-6.5
Some college, no degree	297 (17.7)	68 (21.9)	-4.2
Occupational/technical/vocational	115 (6.9)	16 (5.2)	1.7
Associate degree	216 (12.9)	41 (13.2)	-0.3
Bachelor's degree	571 (34.0)	97 (31.3)	2.8
Master's degree	288 (17.2)	39 (12.6)	4.6
Professional school degree	34 (2.0)	4 (1.3)	0.7
Doctoral degree	24 (1.4)	1 (0.3)	1.1
Not reported	3 (0.2)	0 (0.0)	0.2
Income			
Less than $10 000	25 (1.5)	16 (5.2)	-3.7
$10 000-$19 999	50 (3.0)	20 (6.5)	-3.5
$20 000-$29 999	105 (6.3)	39 (12.6)	-6.3
$30 000-$39 999	105 (6.3)	42 (13.5)	-7.3
$40 000-$49 999	116 (6.9)	29 (9.4)	-2.4
$50 000-$59 999	160 (9.5)	36 (11.6)	-2.1
$60 000-$79 999	232 (13.8)	78 (25.2)	-11.3
$80 000-$99 999	195 (11.6)	22 (7.1)	4.5
$100 000-$199 999	393 (23.4)	24 (7.7)	15.7
$200 000 or more	76 (4.5)	4 (1.3)	3.2
Not reported	220 (13.1)	0 (0.0)	13.1

don't differ by 0.2 standard deviations or more, the difference may be considered trivial, even if it is statistically significant.[33]

We used χ^2 tests of independence with nominal independent variables, dependent variables of categorical pain and function outcomes (eg, chronicity duration and frequency, global health, physical health, mental health, and overall quality of life), and demographics variables (eg, sex, race/ethnicity, education, income, and employment status). Effect sizes were calculated with Cramer's V to indicate the strength of association. Thresholds of practical significance were assessed using cut points of Cramer's V: 0.1 small, 0.3 medium, and 0.5 large.

Numeric Rating Scales. We utilized the Numeric Rating Scale (NRS) to measure differences between groups related to pain intensity. The NRS response range is 0 to 10, and improvements in low back pain should be seen as irrelevant if a difference pre/post is ≤1.5 NRS points.[34] A difference of >1.5 points was considered a clinically meaningful difference between the groups.

ODI. A threshold of >50% improvement on the ODI was used as a clinically significant difference between groups for patients with low back pain.[35,36]

Table 3. *Participant Characteristics: Employment Status*

	RAND (n = 1677) Frequency (%)	MTurk (n = 310) Frequency (%)	Difference
Employment Status			
Working full time	969 (57.8)	193 (62.3)	-4.5
Retired	289 (17.2)	8 (2.6)	14.7
Working part time	190 (11.3)	40 (12.9)	-1.6
Keeping house/caring for dependent	97 (5.8)	24 (7.7)	-2.0
Not working owing to health problems	56 (3.3)	20 (6.5)	-3.1
Unemployed	29 (1.7)	15 (4.8)	-3.1
Student	28 (1.7)	9 (2.9)	-1.2
Maternity/paternity leave	4 (0.2)	0 (0.0)	0.2
Not reported	15 (0.9)	1 (0.3)	0.6

Table 4. *Participant Characteristics: Pain Duration, Pain Frequency, and Self-defined Chronic*

	RAND (n = 1677) Frequency (%)	MTurk (n = 310) Frequency (%)	Difference
Duration of pain			
Less than 1 month	295 (17.6)	31 (10.0)	7.6
At least 1 month and less than 3 months	234 (14.0)	42 (13.5)	0.4
At least 3 months and less than 6 months	198 (11.8)	42 (13.5)	-1.7
At least 6 months and less than a year	194 (11.6)	46 (14.8)	-3.3
1 year to 5 years	492 (29.3)	96 (31.0)	-1.6
More than 5 years	262 (15.6)	52 (16.8)	-1.2
Not reported	2 (0.1)	1 (0.3)	-0.2
Frequency of pain in last 6 months			
Every day/nearly every day in 6 months	600 (35.8)	126 (40.6)	-4.9
At least half of the days in 6 months	551 (32.9)	135 (43.5)	-10.7
Less than half the days in 6 months	520 (31.0)	49 (15.8)	15.2
Not reported	6 (0.4)	0 (0.0)	0.4
Self-defined chronic low back pain			
Yes	1098 (65.5)	242 (78.1)	-12.6
No	99 (5.9)	25 (8.1)	-2.2
Don't know	476 (28.4)	43 (13.9)	14.5
Not reported	4 (0.2)	0 (0.0)	0.2

Results

Recruiting Participants

Our first goal was to investigate whether MTurk could be used as a recruitment pool for persons receiving chiropractic care for chronic low back pain. The flow of participants in the MTurk sample is displayed in Figure 2.

Out of the 6,048 people targeted using a HIT entitled "Brief Health Survey," 5,930 respondents accepted (Fig 2) and 5,755 MTurk participants (99% response rate) advanced to the web-based survey administration via Qualtrics. From the total survey sample of 5,755, 26% (n = 1,500) reported low back pain, which is comparable to national estimates of the condition in the US population, where the estimated prevalence is 23%.[37] A small percentage (2%, n = 25) declined to continue with a longer survey on their pain condition; another small percentage (2%, n = 31) were excluded because they did not have at least 3 months of pain or identify themselves as having chronic pain.

Participants advanced in the survey only if they used chiropractic care for their chronic low back pain. Of the initial sample of 5755 (11%, n = 629) said they had chronic low back pain and had used a chiropractor to treat their pain. This finding falls within the range of national rates of chiropractic use by region, estimated between 6% (West South Central) and 16% (West North Central) and a national average of 8%.[38]

One of the key inclusion criteria to match between the samples was that the participants were current chiropractic patients. The MTurk sample was further culled to 5% (n = 310) of the initial sample of 5755 because over half of the participants had not seen a chiropractor for at least a year. With the prevalence of chronic, impairing, low back pain in the United States estimated at 23% and with an estimated 8% of individuals treated with chiropractic care, we expected approximately 3% of the MTurk sample to meet these criteria and 5% did.

Demographics

The samples had trivial to small differences, based on thresholds noted earlier for effect size, in sex, race/ethnicity, education, and employment status as shown in Table 5. There were, however, significant differences and a medium effect size between the samples with respect to income (χ^2 (10) = 165.35, $P < .001$, Cramer's $V = 0.29$). More than three-quarters of the RAND sample made an average of at least $50,000 (76%) compared with slightly more than half of the MTurk sample (53%). The largest statistically significant difference in the demographic characteristics assessed in this study was in age. The MTurk sample was younger

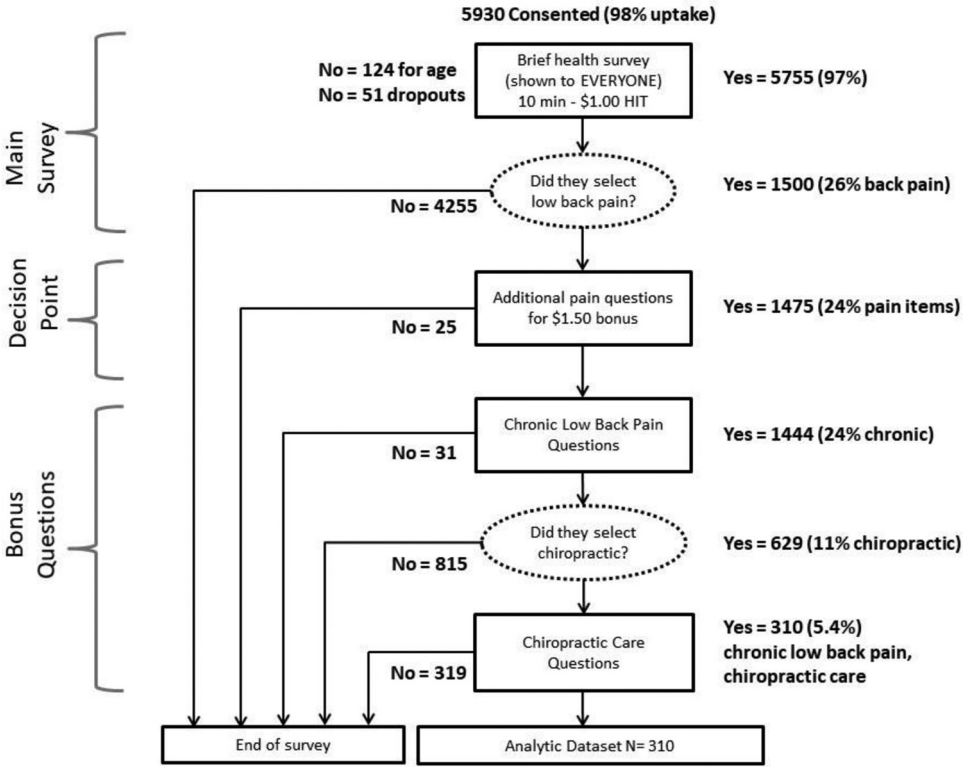

Fig. 2. *Brief health survey procedures and participant flow.*

Table 5. *Demographic Differences in Samples*

Demographic	X^2	df	Cramer's V	Practical Significance
Sex	10.54[a]	2	0.07	*Trivial*
Race	29.46[b]	7	0.12	*Small*
Ethnicity	23.29[b]	2	0.11	*Small*
Education	23.99[a]	8	0.11	*Small*
Employment status	61.36[b]	7	0.18	*Small*
Income	165.35[b]	10	0.29	*Medium*

[a] $P < .01$.
[b] $P < .001$. Cramer's V threshold cut points: 0.1 small, 0.3 medium, and 0.5 large.

(M = 35, SD = 9.7) than the RAND sample (M = 49, SD = 14.8, t [613] = 20.61, $P < .001$, $d = 0.96$). The effect size was large, reflecting the important practical difference in the mean age between groups.

Self-Rated Health

Table 6 shows statistical findings for continuous primary outcomes. Because of the large sample size, there were statistically significant differences on most pain and function outcomes. "Average pain in last 7 days" (NRS) was the only primary outcome that was not significant at $P < .05$ level; however, effect sizes were trivial to small. For ODI (scale range is 0-100), the mean difference between the RAND and MTurk samples was 4.41, and the effect size of the observed difference was small (Cohen's $d = 0.34$) and did not meet the threshold of a clinically minimal important difference of 6.53 points.[35,36]

The 3 NRS items had trivial mean differences (MD) at a threshold of 1.5 NRS points between the RAND and MTurk samples: "Average pain in last 7 days" (MD = 0.21); "Worst pain in last 7 days" (MD = 0.30); "Pain if no chiropractic care" (MD = 0.83).

Other Characteristics

We performed χ^2 tests of independence to compare the 2 samples on the global general health, overall quality of life, physical and mental health, pain duration, pain frequency, and self-identified chronicity. The results are displayed in Table 7.

The relation between these variables was significant ($P < .05$); however, effect sizes were small and ranged from Cramer's V = 0.15 for global health to Cramer's V = 0.18 for both general physical health and overall quality of life. General mental health was the only health outcome that approached a medium effect in the difference between samples (Cramer's V = 0.23). In terms of length of pain and frequency with which the participants experienced their pain, the distribution of responses across the categories for duration and frequency was similar and effect sizes were small. The higher morbidity aligns with the higher functional disability scores on the ODI found in the MTurk participants as compared with the RAND clinic-based sample.

Efficiency

Results of efficiency in terms of time and cost for each method were the second research question. In the MTurk

Table 6. *Comparison of Samples on Continuous Primary Outcomes*

	MTURK	RAND					
Outcome	M (SD)	M (SD)	t test	df	M Diff(95% CI)	d	Practical Significance
Oswestry Disability Index	20.45 (12.75)	24.85 (14.77)	−4.92[b]	399	−4.41 (−6.16 to −2.65)	0.34	Small
Avg pain, 7-day NRS	3.87 (2.05)	4.08 (2.10)	−1.64 *ns*	1982	−0.21 (−0.46 to 0.04)	0.10	Trivial
Worst pain, 7-day NRS	5.64 (2.44)	5.94 (2.47)	−2.00[a]	1964	−0.30 (−0.60 to −0.01)	0.12	Trivial
Pain if no chiro care, NRS	6.77 (2.34)	5.94 (2.23)	5.58[b]	1430	0.83 (0.54-1.12)	0.36	Small

Oswestry Disability Index (0-100). NRS Numeric Rating Scale (0-10). Cohen's d effect sizes are in absolute values. Practical significance assessed using 50% change for ODI, >1.5 points on NRS measures.
ns, nonsignificant; *95% CI*, 95% confidence interval of the difference.
[a] $P < .05$.
[b] $P < .0001$

Table 7. *Comparison of Samples on Categorical Primary Outcomes*

Outcome	χ^2	df	Cramer's V	Practical Significance
Global health	43.93[b]	4	0.15	*Small*
Global quality of life	64.14[b]	4	0.18	*Small*
Global physical health	66.12[b]	4	0.18	*Small*
Global mental health	106.39[b]	4	0.23	*Small*
Pain duration	12.67[a]	5	0.08	*Small*
Pain frequency	31.39[b]	2	0.13	*Small*

Cramer's V threshold cut points: 0.1 small, 0.3 medium, and 05 large.
[a] $P < .05$.
[b] $P < .001$.

sample the incentives cost $2.50 per participant and data collection took 1 month. In the RAND sample, data were collected in 4 months and participants received up to $200 in incentives for completing all questionnaires in the study. Therefore, the RAND study took 4 times as long to collect data and cost 80 times more in incentives than the MTurk study.

Discussion

Differences between the MTurk and RAND samples on characteristics of sex, race/ethnicity, education, and employment were minimal. However, MTurkers were younger and made less money than participants in the RAND sample. The age difference is consistent with past studies comparing MTurk to the general US population.[39] Previous studies have reported an average age of 42 for chiropractic patients.[40,41] Differences between the MTurk and RAND samples were trivial to small on self-reported health measures of chronic low back pain, general health, physical health, and overall quality of life. There were differences that approached a medium effect in general mental health, echoing findings of recent studies of MTurk population health status.[42]

The crowdsourced participants we sampled match previous descriptions of MTurk participants—they were typically younger and had less income than the US national average.[43] Furthermore, this study replicated an earlier study that established the reliability and validity of MTurk data for studying clinical populations.[4]

MTurk offered a recruitment pool that gave access to the population of interest and provided credible data compared with primary data collected within the RAND Center of Excellence in Research on CAM. It appears MTurk can be used to overcome logistical challenges of accessing and recruiting a special clinical population such as this chronic pain patient sample.

But ultimately MTurk's utility depends on the research question. If the question asked is about how a population with chronic back pain manages care, in which age is not a primary variable of interest, then MTurk may be able to provide an easily recruited, inexpensive sample. Data collection can be quick and also can be repeated. Access to younger individuals might even be an advantage for some research questions, particularly if the research focuses on back pain history and the movement from acute to chronic pain. But if the research interest requires access to patient files or measures that are not self-reported, then a practice-based network of patients may be required. Further, if the research question requires a nationally representative sample for generalizability, a US probability sample would be a better-suited data source. Both the MTurk and RAND samples assessed self-reported health, but the RAND clinic-based sample provided access to patient files. These data were merged with identifiable survey data, making it possible to investigate questions such as the impact of appropriate/inappropriate care on patient outcomes not dependent on self-reports.

Another use of MTurk might be in developing instruments to be used in surveys. In the RAND study, we made use of a pilot phase, not only to devise a pragmatic approach for going into chiropractic clinics and collecting data but also as a way to test our research instruments. MTurk would provide an efficient way of testing procedures, as long as the age differences are not thought to be a significant factor in how participants answer the instruments.

Limitations

Although innovative and timely, this study has some potential limitations that may threaten the validity of its findings. The most obvious methodological issue that is problematic but not fatal to the use of MTurk for sampling participants was that MTurkers were not sampled from within particular treatment programs, as they were in the comparison group. This logistic issue impacted a few factors such as differences in provider utilization and satisfaction ratings, which were higher in the RAND sample. Although past studies found MTurk is appropriate for conducting research on a specific clinical condition or special population,[4] previous studies have also noted that MTurkers are likely to be much younger than the traditional pain population of interest in this study.[1,43] Indeed, this study replicated previous results and found the largest difference between the MTurk and RAND samples was in the age category; therefore, the age distribution of the MTurk population was different from those recruited from clinical treatment programs of chronic low back pain. In

this study, we did not attempt to match samples based on demographic characteristics, opting instead to make comparisons based on the raw datasets. As such, the MTurk sample contained younger and lower income participants than did the RAND sample. It is important to recognize this lack of similarity between the samples, but the fact that similar experiences of pain and function were demonstrated despite these demographic differences provides stronger evidence for the utility of this unconventional recruitment technique. This study has demonstrated MTurk may provide a reasonable route through which to obtain targeted populations from the large sampling pool.

Practical Applications

- The results of this study suggest that there may be differences between crowdsourcing and a clinic-based sample.
- Differences range from small to medium on demographics and self-reported health.
- Low incentive costs and rapid data collection of MTurk makes it an economically viable method of collecting data from chiropractic patients with low back pain

Conclusion

This study adds to the empirical evidence base by providing information on the comparability of data collected from MTurk with a large national sample of patients recruited through clinical settings. This study has implications for the practical application of MTurk as a recruitment and data collection tool. Researchers in a variety of disciplines use crowdsourcing as a way to collect data; over half of the top 30 US universities are conducting research on the MTurk platform.

Funding Sources and Conflicts of Interest

This dissertation research was supported by a National Institutes of Health grant, funded as an R21 project entitled, "Feasibility of Crowdsourcing for Eliciting Patient Experiences of Chronic Pain" through grant number 1R21AT009124-01. No conflicts of interest were reported for this study.

Contributorship Information

Concept development (provided idea for the research): L.H., I.C., G.R., R.H.
Design (planned the methods to generate the results): L.H., I.C., G.R., R.H.
Supervision (oversight, organization and implementation, writing of the manuscript): L.H., I.C.
Data collection/processing (experiments, organization, or reporting data): L.H.
Analysis/interpretation (statistical analysis, evaluation, and presentation of the results): L.H., R.H.
Literature search (performed the literature search): L.H.
Writing (responsible for writing a substantive part of the manuscript): L.H.
Critical review (revised manuscript for intellectual content, not spelling, grammar): R.H., I.C., G.R.

References

1. Sheehan KB. Crowdsourcing research: data collection with Amazon's Mechanical Turk. *Comm Monogr*. 2018;85(1):140-156.
2. Hays RD, Liu H, Kapteyn A. Use of Internet panels to conduct surveys. *Behav Res Methods*. 2015;47(3):685-690.
3. Mason W, Suri S. Conducting behavioral research on Amazon's Mechanical Turk. *Behav Res Methods*. 2011;30:1-23.
4. Shapiro DN, Chandler J, Mueller PA. Using Mechanical Turk to study clinical populations. *Clin Psychol Sci*. 2013;1(2):213-220.
5. Galton F. Vox populi. *Nature*. 1907;75(7):450-451.
6. Howe J. The rise of crowdsourcing. Wired. Available at: https://www.wired.com/2006/06/crowds/. Accessed 10 February 2018.
7. Behrend TS, Sharek DJ, Meade AW, Wiebe EN. The viability of crowdsourcing for survey research. *Behav Res Methods*. 2011;43(3):800-813.
8. Parvanta C, Roth Y, Keller H. Crowdsourcing 101: A few basics to make you leader of the pack. *Health Promotion Practice*. 2013;14(2):163-167.
9. Amazon Mechanical Turk. Amazon Mechanical Turk, Inc. An Amazon Company. Available at: https://www.mturk.com. Accessed 6 February 2018.
10. Goodman JK, Dryder CE, Cheema A. Data collection in a flat world: the strengths and weaknesses of Mechanical Turk. *J Behav Decis Mak*. 2013;26:213-224.
11. Berinsky AJ, Huber GA, Lenz GS. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Polit Anal*. 2012;20:351-368.
12. 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.
13. Harris JK, Mart A, Moreland-Russell S, Caburnay CA. Diabetes topics associated with engagement on Twitter. *Prev Chronic Dis*. 2015;12(E62):1-9.
14. Tosti-Kharas J, Conley C. Coding psychological constructs in text using Mechanical Turk: a reliable, accurate, and efficient alternative. *Front Psychol*. 2016;7(741):1-9.
15. Weiner M. The potential of crowdsourcing to improve patient-centered care. *Patient*. 2014;7:123-127.
16. Hogg WE. Crowdsourcing and patient engagement in research. *Can Fam Physician*. 2015;61(3):283-284.

17. Good BM, Nanis M, Wu C, Su AI. Microtask crowdsourcing for disease mention annotation in PubMed abstracts. *Pac Symp Biocomput*. 2015:282-293.
18. Khare R, Good BM, Leaman R, Su AI, Lu Z. Crowdsourcing in biomedicine: challenges and opportunities. *Brief Bioinformatics*. 2016;17(1):23-32.
19. Mortensen JM, Minty EP, Januszyk M, et al. Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of SNOMED CT. *J Am Med Inform Assoc*. 2014;0:1-7.
20. Institutes of Medicine. *Relieving Pain in America: A Blueprint for Transforming Prevention, Care, Education, and Research*. Washington, DC: The National Academies Press; 2011.
21. Herman PM, Kommareddi M, Sorbero ME, et al. Characteristics of chiropractic patients being treated for chronic low back and chronic neck pain. *J of Manipulative Physiol Ther*. 2018;41(6):445-455.
22. Litman L, Robinson J, Abberbock T. TurkPrime.com: a versatile crowdsourcing data acquistion platform for the behavioral sciences. *Behav Res Methods*. 2016:1-10.
23. Chandler J, Mueller P, Paolacci G. Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. *Behav Res Methods*. 2014;46 (1):112-130.
24. Hydock C. Assessing and overcoming participant dishonesty in online data collection. *Behav Res Methods*. 2018;50 (4):1563-1567.
25. Siegel JT, Navarro MA, Thomson AL. The impact of overtly listing eligibility requirements on MTurk: an investigation involving organ donation, recruitment scripts, and feelings of elevation. *Soc Sci Med*. 2015;142:256-260.
26. Fairbank JCT, Pynsent PB. The Oswestry Disability Index. *Spine*. 2000;25(22):2940-2953.
27. Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. *Qual Life Rese*. 2009;18 (7):873-880.
28. Hays RD, Schalet BD, Spritzer KL, Cella D. Two-item PROMIS® global physical and mental health scales. *J Patient Rep Outcomes*. 2017;1(1):2.
29. Hays RD, Spritzer KL, Thompson WW, Cella DUS. General Population Estimate for "Excellent" to "Poor" Self-Rated Health Item. *J General Internal Medicine*. 2015;30(10):1511-1516.
30. Schalet BD, Rothrock NE, Hays RD, et al. Linking Physical and mental health summary scores from the Veterans RAND 12-Item Health Survey (VR-12) to the PROMIS(R) Global Health Scale. *J General Internal Medicine*. 2015;30 (10):1524-1530.
31. Deyo RA, Dworkin SF, Amtmann D, et al. Report of the NIH Task Force on research standards for chronic low back pain. *J Pain*. 2014;15(6):569-585.
32. Faul F, Erdfelder E, Lang A-G, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods*. 2007;39:175-191.
33. Cohen J. *Statistical Power Analysis for the Behavioral Sciences (Revised Edition)*. New York: Routledge Academic Press; 1988.
34. Kovacs FM, Abraira V, Royuela A, et al. Minimal clinically important change for pain intensity and disability in patients with nonspecific low back pain. *Spine*. 2007;32(25):2915-2920.
35. Fritz JM, Hebert J, Koppenhaver S, Parent E. Beyond minimally important change: Defining a successful outcome of physical therapy for patients with low back pain. *Spine*. 2009;34(25):2803-2809.
36. Lauridsen HH, Hartvigsen J, Manniche C, Korsholm L, Grunnet-Nilsson N. Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients. *BMC Musculoskelet Disord*. 2006;7. 82-82.
37. Balagué F, Mannion AF, Pellisé F, Cedraschi C. Non-specific low back pain. *Lancet*. 2011;379(9814):482-491.
38. Peregoy JA, Clarke TC, Jones LI, Stussman BJ, Nahin RL. Regional variation in use of complementary health approaches by U.S. adults. *NCHS Data Brief*. 2014:(146):1-8.
39. Paolacci G, Chandler J. Inside the Turk: understanding Mechanical Turk as a participant pool. *Curr Dir Psychol Sci*. 2014;23(3):183-188.
40. Coulter ID, Shekelle PG. Chiropractic in North America: a descriptive analysis. *J Manipulative Physiol Ther*. 2005;28 (2):83-89.
41. Hurwitz EL, Coulter ID, Adams AH, Genovese BJ, Shekelle PG. Use of chiropractic services from 1985 through 1991 in the United States and Canada. *Am J Public Health*. 1998;88 (5):771-776.
42. Walters K, Christakis DA, Wright DR. Are Mechanical Turk worker samples representative of health status and health behaviors in the U.S.? *PLoS One*. 2018;13(6):e0198835.
43. Paolacci G, Chandler J, Stern LN. Running experiments on Amazon Mechanical Turk. *Judgm Decis Mak*. 2010;5 (5):411-419.