Crowdsourcing Qualitative Thematic Analysis

Lara G. Hilton¹,² and Tarek Azzam³

Abstract
Evaluations that include stakeholders aim to understand their perspectives and to ensure that their views are represented. This article offers a new approach to gaining stakeholder perspectives through crowdsourcing. We recruited a sample of individuals with chronic low back pain through a crowdsourcing site. This sample coded textual data describing pain, provided themes, and provided feedback on constructs and procedures. The results generated by the crowdsourced participants were compared to results generated by experts. We found that crowdsourcing to develop code and textual responses was feasible, rapid, and inexpensive, offering the potential to enhance patient stakeholder engagement in evaluation. Crowdsourcing has broad implications for evaluation science beyond the health sector.

Keywords
crowdsourcing, qualitative methods, stakeholder engagement, health, evaluation practice

Most evaluations seek to include stakeholders in the evaluative process. Inclusion can take multiple forms, such as meeting with stakeholders to better understand their needs and interests, engaging them in the logic modeling process to attain a shared vision of program activities and goals, and conducting member checks using informant feedback and respondent validation to help ensure that stakeholder perspectives are represented authentically. This article offers a new approach to gaining stakeholder perspectives. We accessed hundreds of potential stakeholders through the web and engaged them in a qualitative analysis task that directly reflected their experiences and perspectives. This process is a novel adaptation of a concept called crowdsourcing—the process of recruiting hundreds, even thousands, of individuals to work on a specific task or set of tasks.

Contemporary crowdsourcing first emerged in 2006¹ as an online labor market where services, ideas, or content were obtained for a fee from a large group of people (Howe, 2006). Crowdsourcing can take many forms including crowdfunding (e.g., donate money to fund a project), crowd labor

¹ Deloitte Consulting LLP, Los Angeles, CA, USA
² RAND Corporation, Santa Monica, CA, USA
³ Division of Behavioral and Organizational Sciences, Claremont Evaluation Center, Claremont Graduate University, Claremont, CA, USA

Corresponding Author:
Lara G. Hilton, 555 West 5th Street, Suite 2700, Los Angeles, CA 90013, USA.
Email: lhilton@deloitte.com
(e.g., many people are asked to transcribe audio files), and crowd research (e.g., respond to surveys; Parvanta, Roth, & Keller, 2013).

The integration of crowdsourcing into research as a scientific methodology has been slowly developing over the past decade. There is literature demonstrating the utility of crowdsourcing for conducting social, psychological, and clinical research. For example, mental health populations have been successfully accessed to study conditions such as attention deficit hyperactivity disorder and psychopathology (Shapiro, Chandler, & Mueller, 2013; Wymbs & Dawson, 2015). Studies on populations with chronic disease such as diabetes and low back pain have been conducted using crowdsourcing (Bartek et al., 2017; Harris, Mart, Moreland-Russell, & Caburnay, 2015). Political science experiments have been replicated using crowdsourcing (Berinsky, Huber, & Lenz, 2012). In the Berinsky, Huber, and Lenz’s (2012) study, the researchers were able to replicate the effects of word choice (welfare vs. assistance to the poor) on participants’ support for a political issue (government funding of poverty reduction programs) in a crowdsourcing environment. This experiment illustrated that the crowd can produce similar experimental findings to well-established experimental studies conducted with a noncrowdsourced sample.

The crowd can now be accessed through multiple websites such as clickworker.com or crowdflower.com; however, the most commonly used crowdsourcing platform in research contexts is Amazon Mechanical Turk (MTurk; Amazon Mechanical Turk, 2018). There are many reasons why MTurk has dominated the crowd research arena, including (a) a large subject pool with access to over 500,000 participants from across the world, (b) participants who are slightly more demographically diverse than standard Internet samples or in-person convenience samples and are significantly more diverse than the typical American college samples often used in social science research,2 (c) participants who can be recruited rapidly and inexpensively, and (d) others have also observed and claimed that data obtained from this source are as reliable as what is obtained via other data collection methods (Berinsky et al., 2012; Buhrmester, Kwang, & Gosling, 2011; Mason & Suri, 2011). For example, multiple studies have used samples drawn from MTurk to code psychological constructs in a reliable, accurate, and efficient way, as well as to successfully code qualitative texts of Twitter comments related to diabetes (Harris et al., 2015; Tosti-Kharas & Conley, 2016).

The utility of crowdsourcing for evaluation science has also been gaining attention. For example, Azzam and Jacobson (2013) explored the potential for creating a matched comparison group from a sample recruited from MTurk for situations where a control/comparison is not possible due to logistical or ethical reasons. Crowdsourcing has also been used to help validate program theories, by having the crowd identify program theory components (either activities or outcomes) within interviews conducted with program participants (Azzam & Harman, 2016).

Such studies demonstrate promising applications of crowdsourcing as a valued tool in the evaluation tool kit. This study contributes to this growing body of knowledge by offering another practical application of how crowdsourcing can be used in social sciences research.

The study stemmed from an opportunity to compare results produced from a crowdsourced task to results produced from (1) a panel of experts and (2) a nationally representative sample of participants. The study focused on the criteria used to classify patients as chronic low back pain sufferers. In previous years, the criteria were determined by the National Institutes of Health (NIH) Task Force for Research Standards in Chronic Low Back Pain (Deyo et al., 2014), which is a panel of experts on the topic. The criteria and standards developed by this panel are used to develop clinical practice guidelines that help providers, funders, and payers decide on effective programs and reimbursements and so may result in determination of benefits for potentially millions of people. However, patient perspectives and experiences were not directly included this process, as there were no patients or patient advocates engaged in the research guideline development. We saw crowdsourcing as a way to incorporate the patient perspective.
Several recent studies have focused on patient access and participation based on crowdsourcing. Shapiro, Chandler, and Mueller (2013) established the reliability and validity of MTurk data for studying clinical populations and gave guidance on maximizing these dimensions of data when using crowdsourcing software. Weiner (2014) argues for the use of crowdsourcing in support of patient-centered care, based on empirical literature that shows patient participation helps shape health policy and clinical decision-making and that patient perspectives are the best predictors of health outcomes. He further argues that crowdsourcing has the potential to prevent policy makers from generating views that are too narrow or exclusive. Crowdsourcing would allow decision makers to gain broader input rapidly and at low cost (Ranard et al., 2013; Weiner, 2014).

In the current study, we investigated the utility of the crowdsourcing platform, MTurk, for stakeholder engagement in health evaluations. We addressed four research questions in two study phases:

1. Phase 1
   - Can MTurk be used to engage participants in revising program inclusion criteria for treatment of chronic low back pain?

2. Phase 2
   - How stable are the results of crowdsourced coding?
   - To what extent do crowdsourced data align with expert-derived data?
   - Can evaluators use crowdsourcing for thematic coding?

**Method**

We conducted a mixed-methods study using a within- and between-group design as displayed in Figure 1 (Creswell & Clark, 2017). The study uses crowdsourcing as a method to access, measure, and engage patient stakeholders in evaluation. We sampled two groups of MTurk participants and used two different methods supported by MTurk for data collection: web-based surveys and qualitative coding tasks.

**Procedures**

To assess whether the crowd would engage in thematic code development, we used MTurk to identify people with chronic low back pain and tested their ability to revise domains of chronic
pain that were developed by an NIH Task Force on Low Back Pain (Deyo et al., 2014). The task force recommended four dimensions of chronic pain: duration, frequency, intensity, and function. We added an “other” category to provide an opportunity for MTurk participants to add themes.

We screened MTurkers for chronic low back pain and consented them. Before undertaking the coding task, participants completed a code training and quiz to get a sense of the code descriptions. MTurk participants who had chronic low back pain were asked to apply these expert-derived codes (duration, frequency, intensity, and function) to textual descriptions of chronic pain. Then, participants were randomized into two conditions and instructed to “pick one code” or “check all that apply.” This procedure helped to refine the coding process by gathering feedback on whether the crowd preferred to singly code quotes or the option of tagging quotes with multiple codes. Past literature successfully tested the “pick one” condition (Jacobson, Whyte, & Azzam, 2017). In the context of this study, some quotes seemed to warrant an option of multiple codings.

Participants were also randomized across four blocks of 30 quotations each to minimize burden. Figure 2 displays the survey flow. The quotes participants coded were generated from pilot MTurk surveys in response to the question, “What does chronic pain mean to you?” The sample of 90 quotes used throughout this study is shown in Supplemental Appendix.

Participants ended the survey by answering a set of feedback questions about the codes, coding process, and a general question about their overall experience with the human intelligence task (HIT). We paid participants US$2.50 for an initial screener to identify those with chronic low back
pain and another US$5 for the coding task. The revised set of codes and procedures that were refined by the crowd were used in the study’s subsequent phase.

In the second phase, a new sample of participants were screened, 316 qualified and 250 (80.0% response rate) consented, participated, and were compensated US$2 for the screener and US$4 for the coding task. An e-mail was generated through MTurk/TurkPrime that invited the qualified individual MTurk workers to participate in a “back pain categorizing task.” TurkPrime is an intermediary platform that helps scientists to use MTurk for research purposes. It provides options such as limiting survey responses to single respondents and allowing for anonymous e-mails to go out to workers (Litman, Robinson, & Abberbock, 2016). An HIT, visible only to qualified workers, was posted to MTurk describing the task.

Before proceeding to the coding task, participants were asked to review the codes and code descriptions, then take a set of quizzes for training purposes. They received feedback on the quizzes and the correct answers. They were then randomly assigned to one of three blocks of 30 quotes of chronic pain definitions, the same quotes used in the study’s initial phase. The codes they applied to these texts were the newly revised pain domains generated by the first set of MTurkers.

After the coding task, participants answered a set of closed-ended and open-ended survey questions to gather feedback. These items replicated questions from the initial phase including most important domain to them personally, easiest to use, hardest to use, new codes, feedback on coding procedures, and general comments about the HIT. The flow of this phase of the study is displayed in Figure 3.

---

Figure 3. Coding implementation for chronic low back pain participants.
Analysis

We conducted a mixed-methods analysis on crowdsourced quantitative survey data and qualitative text responses (Creswell & Clark, 2017). Numeric responses to survey items “What is the most important domain of chronic pain for you personally?” and “Which is the least important domain of chronic pain for you?” were examined using frequencies to show the level of importance. Additional themes that participants endorsed were analyzed through an inductive pile sorting procedure (Ryan & Bernard, 2003), where comments were compiled in an excel database, sorted into similar groups, and given a label. Next, the comments and labels were loaded into a qualitative text analysis software (Dedoose), and interrater reliability was conducted to measure agreement between coders on those labels (Dedoose, 2017). Disagreements were discussed and rectified by consensus, the codes and code descriptions were adjusted, and the coders proceeded to tag all comments. The same inductive process was conducted with text responses from the question, “Was there any additional feedback you would like to give us on the coding procedure?” We then revised the original domains based on input from crowd comments.

Stability of crowdsourced coding. Assessing the extent to which the coding selections of MTurk participants would replicate across subsamples of crowdsourced data is best thought of as establishing the “stability” of coding (Azzam & Harman, 2016). Coding stability across samples can be measured using statistical significance, and the analytic approach in this current study is informed by previous studies of crowdsourcing qualitative coding (Harman, 2016; Jacobson et al., 2017).

The stability of crowdsourced coding results was measured with the question: To what extent did the subsamples within MTurk produce the same distribution of codes for each quote? As shown in Figure 3, the sample of 250 coders was randomly assigned into three blocks of approximately 80 MTurk participants per block, who received 30 quotes to code for a total of 90 quotes. For the stability analysis across subsamples of MTurk coders, the three blocks of coders were then randomized, so each block had half its coders (~40) in each subsample. For each quote, Fisher’s exact test (two-tailed) was conducted to assess whether there was a statistically significant difference in the distribution of codes between the two groups. As some cells had expected value less than 5, $\chi^2$ tests were not appropriate for this analysis. Figure 4 shows our analytic framework for assessing stability and reliability.
Comparing crowdsourced and expert coding. Once stability is analyzed between multiple coding passes by different crowd samples, the consistency of the findings can be compared with expert coding through interrater reliability. In this case, the expert opinion is not the predominant view, but rather these procedures are in support of engaging patients in health policy. Therefore, the aim is to highlight the differences or similarities in coding, with an emphasis on valuing the patient perspective as one that has historically been underrepresented.

We assessed the extent to which crowdsourced coding aligns with expert-derived coding. First, the crowd codes needed to be established for each quote. There is typically a distribution across all the code options in crowdsourced coding because there are many people coding a single quote with differing perspectives and expertise. Past crowdsourcing work has identified a method for picking a modal response by analyzing the difference in the proportion that selected the top two codes using a $\chi^2$ goodness-of-fit test between the two most frequently selected codes for each quote (Jacobson et al., 2017). If there was a statistically significant difference between the top two selected codes, then the assumption was that the code most common was the mode. If there was not a statistically significant difference, then there were two modes. In this current study, the crowd could pick more than one code and there appeared to be quotes with three modes. To systematically and statistically select codes that might be multimodal, we conducted $\chi^2$ goodness-of-fit tests for the top three most frequently selected quotes based on distributions.

Second, an acceptable threshold of agreement was established across expert coders using interrater reliability exercises in Dedoose software (Dedoose, 2017). Two experts in chronic pain research were tasked with using the same code list and descriptions that MTurk participants used for their coding task. The experts coded the same set of 90 quotes, and then, interrater reliability was conducted using Cohen’s $\kappa$ at the code level.

Finally, the crowd and expert coding were compared using Cohen’s $\kappa$ test statistic and percent agreement. Cohen’s $\kappa$—a coefficient of agreement for nominal scales—is a widely used measure to evaluate interrater agreement as compared to the rate of agreement expected by chance, based on the coding behavior of each rater (Cohen, 1960). Thresholds for assessing the “significance” of the Cohen’s $\kappa$ values were set at standard cut points of $<.20 = \text{poor agreement}$, .21 to .4 = fair agreement, .41 to .6 = moderate agreement, .61 to .8 = good agreement, and .81 to 1.0 = very good agreement (Landis & Koch, 1977).

Participants

In Phase I, $N = 226$ MTurkers met inclusion criteria and consented to do the coding task. In Phase II, $N = 250$ MTurkers met criteria and consented to do the coding task. Basic demographic characteristics of MTurk participants are presented in Table 1. The majority of the participants were employed, young, non-Hispanic, White females. Only 40.0% had at least a bachelor’s degree and just under half made less than US$50,000 per year.

Results

Thematic Code Development

The participants were first presented with a code training exercise in which they were asked to read the code descriptions and examples before completing four coding quizzes. The code training revealed that approximately 60.0% missed one of the four quiz questions or got all four correct. However, 35.4% missed two quiz questions. Only 6.2% missed more than half of the answers. While this performance is adequate for code training, it is not excellent. It means that the MTurk participants did not read the code descriptions carefully, the codes are ambiguous, or their descriptions
were not adequate. These issues with code applications in training foreshadow the feedback that participants provided about the codes later in this phase.

**Most important theme.** In response to the question, “What was the most important pain domain to you personally?” the crowd overwhelmingly endorsed “severity” (50.0%) over the other codes. Only about a quarter thought “frequency” (25.2%) was most important, followed by “function” (16.4%). “Duration” was endorsed by 8.8% of the crowd as the most important domain of chronic pain to them personally. In comparison to expert definitions of chronic pain, duration of pain greater than 3 months constitutes chronicity as defined by experts in health evaluation. This is an important finding in that the domain of least importance to the crowd is the one experts value and use as the gold standard definition of chronic pain.

**Easiest theme to apply.** The distribution of responses to the question, “Which code was the easiest for you to apply?” has less range. The crowd found “severity” to be the easiest code to apply (26.6%), followed closely by “function” (27.4%), “frequency” (22.6%), and finally duration (20.4%).

**New themes.** To collaborate with the crowd on refining these domains of chronic pain, we asked, “Is there a new category that you would like to add?” The question yielded 216 responses from the 226 participants surveyed; responses were sorted and analyzed using inductive pile sorting (Ryan & Bernard, 2003). The bulk of the responses offered no additional themes; however, participants noted a few new themes that warranted inclusion in the next round of qualitative coding. “Treatment” by itself and in conjunction with “pain management” was mentioned by more than 10.0% (29 of the 216) of the respondents as a theme not captured by expert codes. There were comments about treating and managing pain, the kinds of providers patients used, medication, therapy, and self-care coping strategies such as stretching. A second theme mentioned in 10.0% (N = 21 of the 216 responses) of comments was “diagnosis” or “cause” of pain. Respondents mentioned issues of root causes and conditions, such as arthritis or injury, that were underlying the pain. A few respondents mentioned issues like location of pain (N = 6), getting relief (N = 5), psychological components of pain (N = 3), and assorted other themes.

**Revised themes.** Other coding comments included the uncertainty participants experienced when applying several of the codes, which they viewed as overlapping constructs. They observed that “duration” and “frequency” were too closely related to differentiate between and either applied both or made a best guess at the most appropriate. They noted that the definition for duration of “how long

| Demographics       | Phase I MTurk Frequency (%; N = 226) | Phase II MTurk Frequency (%; N = 250) |
|--------------------|--------------------------------------|---------------------------------------|
| Age mean (SD)      | 40.1 (11.4)                          | 35.8 (9.5)                            |
| Female             | 129 (56.9)                            | 160 (64.0)                            |
| White              | 194 (85.5)                            | 211 (84.4)                            |
| Hispanic or Latino | 15 (6.6)                              | 18 (7.2)                              |
| ≥Bachelor’s degree | 90 (39.6)                             | 90 (39.6)                             |
| ≥US$50,000         | 100 (44.1)                            | 100 (44.1)                            |
| Working full time  | 126 (55.5)                            | 151 (60.4)                            |
| Retired            | 7 (3.1)                               | 7 (2.8)                               |
| Working part time  | 46 (20.3)                             | 43 (17.2)                             |

Note. MTurk = Mechanical Turk.
has it lasted” and definition for frequency of “how much does it happen” are not clearly mutually exclusive. Based on this feedback, we combined these two domains into a single code called “duration” and expanded the code definition to include “frequency.”

Another change to the code list was to revise the code descriptions. MTurk participants suggested revising the descriptions of the codes “function” and “severity.” For “function,” several participants suggested adding “impact,” while several others responded that they would like to see words like “limiting” added to the code description for “function.” For “severity,” they mentioned including “intensity” right at the beginning of the code description. The comparison of initial expert domains, the crowd suggestions for revisions, and the final code set are shown in Table 2.

### Table 2. Revised Thematic Codes and Descriptions.

| Revision | Category | Description | Example |
|----------|----------|-------------|---------|
| Added    | Diagnosis| What caused pain. medical diagnoses, injury, accident, birth defects | “Due to arthritis in my spine,” “disk disease,” “car accident,” “I was born with it” |
| Merged   | Duration | Pain frequency and duration described as ongoing, constant, long-lasting, reoccurring | “Constant pain reoccurring daily,” “always there,” “comes and goes” |
| No change| Function | Impact and limitation on daily functioning, quality of life, social or family life, limits activities | “It hurts when I rake the leaves,” “I can’t sleep or play with my kids,” “affecting life in all areas” |
| No change| Severity | Intensity of pain level, descriptions of the pain symptoms | “Debilitating pain,” “low dull ache,” “stabbing pain,” “severe pain affects me in my legs” |
| Added    | Treatment| Comments about treating or managing pain, either by self-care, providers, medication, or anything else | “Pain you have to manage with drugs and therapy,” “only relief is when on medication or from stretching” |
| No change| Other    | Check this option if you want to add a new category and enter text | No examples |

Procedural revisions to coding. Another key component to coding procedures is the issue of whether applying a single theme to a textual response or more than one theme is more appropriate. This work was replicating a previous study on coding open-ended responses using crowdsourcing (Jacobson et al., 2017). The authors asked the crowd was asked to select one code, as they believed that, “... it would offer a clear indication of the presence or absence of different themes and which code was more readily or easily identifiable...” (Jacobson et al., 2017). In piloting this study, replication of this approach garnered feedback that there was more than one key idea in some quotes and the crowd wanted the option of picking more than one code. Therefore, in this study, we systematically tested whether single or double coding would be preferred by randomizing the quote blocks, asking half the crowd to “pick one” and the other half to “check all that apply.”

In the “pick one” group, the participants overwhelming felt that “It was difficult to choose one when more than one category was mentioned.” The feedback about the difficulty of choosing only one theme was not surprising. The descriptions of chronic pain that participants were coding covered complex pain experiences that often encompassed more than one of the thematic codes. Based on this test and feedback, the code implementation task in the next phase allowed the MTurk coders to check all codes that apply. This refinement is another example of the way in which the crowd can collaborate to improve procedures.
**Code Implementation**

*Code training.* Results from the coding training were excellent, with between 75.0% and 92.0% of the participants getting correct answers on each of the five quizzes. This improvement in code training results for Phase II as compared to Phase I is likely a consequence of the revisions to the codes, code definitions, and training procedures recommended by MTurk participants. The quotes were held constant, so improvement is not attributable to different quotes being coded.

*Feedback on coding and procedures.* To confirm that this coding structure worked well for this sample of crowdsourced participants and that the changes made to the code list in Phase I were not idiosyncratic of that sample, we asked the Phase II participants for feedback on the codes with the question, “Is there a new category that you would like to add?” Of the 250 participants surveyed, almost 88.8% (N = 222) had no new categories to add. The remaining participants mentioned psychological and emotional issues that are a part of living life with chronic pain (N = 8, 3.2%), the location of the pain (N = 8, 3.2%), the results and outcomes of certain treatments (N = 3, 1.2%), and a handful of other miscellaneous comments (N = 9, 3.6%).

As in the previous phase, we gathered feedback on the coding process with an open-ended question, “What did you think about the process of categorizing quotes?” More than 90.0% of the MTurk participants offered comments about the coding process. Overall, most comments were positive and no major revisions to procedures emerged during the qualitative analysis of these comments.

One of the study’s unexpected findings was the quality of the feedback on procedures. There were comments about how the participants experienced the task, including, “The training made it really easy. Figuring out the proper categories was basically common sense. The more I did it, the easier and more natural it became.” Other comments were about how participants decided on single or multiple codes. The study had an unintended positive consequence: Almost 9.0% of the participants wrote that this task was a good experience and helpful because they identified and could relate with the descriptions of chronic low back pain. They felt they were “no longer alone” and could use these domains to discuss care with their provider.

*Stability.* The stability analysis involved randomly assigning the MTurk coders into one of two groups for each block and then comparing their code applications by conducting a total of 90 individual statistical tests using Fisher’s exact test, one for each quote (Supplemental Appendix). Of the 90 quotes, only 3 (3.3%) had a statistically significant difference between the two samples at \( p < .05 \) \( (p = .029, p = .033, p = .049) \). Given the number of tests conducted, these may be due to chance. It should also be noted that a statistical significance test is not a fully accurate way to test for the extent to which subsamples within MTurk produce the same distribution of codes for each quote. A statistical significance test tells, if there were no differences between the subsamples, that we would find the difference that we did, or a greater one, by chance sampling error alone. We have attempted to mitigate this flaw by providing the actual differences we found and the \( p \) values for all 90 quotes (Supplemental Appendix). We hope that this allows the readers to judge for themselves if these are adequate for them to use.

*Reliability between expert coders and crowd coders.* In terms of comparison of crowd versus expert coding on the number of codes applied to each quote, there was over 90.0% agreement across all three potential levels of coding, as displayed in Table 3.

Results of the alignment of crowd and expert coders on domains revealed that the percent agreement ranged from 93.3% for “diagnosis” to 96.7% for “duration.” The frequencies of agreement, disagreement, and percentages for each domain are shown in Table 4.
In terms of reliability measured by Cohen’s \( \kappa \), the overall \( \kappa \) for the expert coders was high at .84. The overall \( \kappa \) comparing expert coders versus crowd coders was similarly high at .82, as shown in Table 5. Reliability scores over .80 are considered excellent (Landis & Koch, 1977).

These results show that the crowd can provide high-quality qualitative coding at nearly the same level as experts. The code with the lowest individual \( \kappa \) (“diagnosis”) was the same in both analyses, which may be a signal that the code description is ambiguous and needs further refinement. However, the code with the highest reliability was different for each group: “duration” for expert coders and “function” for the comparison of expert and crowd coders. This could be a signal of the bias that experts have toward the NIH Task Force definition of chronic pain as duration of pain lasting longer than 3 months, while the crowd did not use the code “duration” as frequently as the experts.

**Discussion**

This study illustrated the potential utility of crowdsourcing as a way to gain insights from stakeholders through the process of collaborative thematic code development. New themes arose that reflected the patient perspectives on the domains for inclusion into national chronic low back pain treatment programs. This coding experiment yielded a new code set of chronic pain domains:
diagnosis, duration, function, severity, and treatment. It also revised coding procedures to allow for multiple code applications on a single textual quote.

The findings revealed that the crowdsourced patient stakeholders overwhelmingly prioritized the themes of severity of pain, pain frequency, and functional status over duration of pain. This is an important finding: The domain of least importance to the crowd is the one experts value and use as the gold standard definition of chronic pain. This finding supports the notion that there is a gap between what expert researchers studying chronic pain think and what the actual people who suffer from the condition think. This method has the potential to bridge that gap and maybe even out the power structure between researchers and people.

The study offers an innovative methodological contribution to the evaluation process, which could reduce costs and expedite data collection and analysis. It may also be useful to the evaluation field in myriad other ways. With a large, diverse participant pool, MTurk provided a fast and inexpensive way to get the special population of interest to participate in a textual coding task. Coding stability between MTurk samples and reliability between MTurk and expert coders were very high, providing strong empirical evidence for the utility of MTurk as a thematic coding tool. During the process of the current study, a special population was accessed, participants gave feedback on survey items and research procedures, they were sampled into a panel, they were used to pilot coding tasks, and they acted as collaborators to develop and refine code lists.

The results of this study are consistent with previous research on evaluation, which found that qualitative coding tasks can be conducted successfully on MTurk (Harman & Azzam, 2018; Jacobson et al., 2017). Also in the evaluation literature, MTurk was effectively used to create matched samples for control groups (Azzam & Jacobson, 2013), used for validating program theory (Harman & Azzam, 2018), and tested for prioritizing research questions for evaluation of low back pain programs (Bartek et al., 2017). These studies are a fraction of the research procedures and experiments conducted on MTurk that are accelerating and democratizing science (Paolacci & Chandler, 2014).

The current study contributes to the field of evaluation by developing and testing a new tool for including and engaging participants. The findings show that evaluators can use crowdsourcing to collaborate with program participants for thematic code development and application. This innovation provides a valid, replicable, and resource-efficient tool. Although this study was conducted in the health sector, the study findings have broad practical and theoretical implications for participant engagement in evaluation more generally.

Limitations

While study findings were promising, MTurk is not the panacea for all evaluation contexts. There are practical issues that are part of the system that cannot be overlooked or altered. For instance, the participants are 18 and over in age and typically more than a quarter of workers at any given time live outside of the United States (Ipeirotis, 2010, 2018). There is an option to limit samples in MTurk to U.S.-based participants. As MTurk data are anonymous, credibility is of utmost concern. The most efficient and rigorous way to sample participants for scientific studies is to screen them for the condition of interest as was done in the current study (Siegel, Navarro, & Thomson, 2015). Lastly, there is a slight learning curve to using the MTurk platform as a scientific tool. The evaluator needs to have some technical acumen for programming or use an intermediate platform such as TurkPrime that provides some of the functionality needed for research and evaluation (e.g., limiting a person from taking a survey more than once; Litman et al., 2016). Start-up information is available through the MTurk website, journal articles, and blogs. Articles are published weekly on the use of MTurk for research and an emerging literature base is growing in evaluation science.
Another limitation to this method is possibility of reducing the participants’ experiences into anonyms and brief statements that does not truly reflect the context of their experiences, particularly since they are virtual. More attention should be paid to acknowledge the humanity of these participants.

**Beyond Program Evaluation: Method for Researchers to Enhance Patient Engagement**

While this study was focused on the application of evaluation theory to address issues in program evaluation, it has implications for the research endeavor more generally. Currently, we have limited tools for recruiting, soliciting input, and further engaging large numbers of patients in research. The gold standard data collection approaches are costly and time-consuming and, therefore, cannot be repeated very often. Crowdsourcing promises to be a more efficient and less resource-intensive method for researchers as well as evaluators to add to their tool kits.

Gaining access to patients and including them in program evaluation does not necessarily create meaningful engagement opportunities. Inclusion is a necessary but insufficient element of engagement, and many evaluators aspire to include participants beyond the role of survey subjects. This study explored how to increase stakeholder inclusion and participation in evaluation through crowdsourced qualitative analysis. Designed to involve participants in a meaningful way, this research identified a valid, reliable, and efficient approach for incorporating patient perspectives into healthcare evaluation. The study is significant in that it explored and developed new strategies for patient participation in developing inclusion criteria for national pain treatment programs.

Researchers in a variety of disciplines use crowdsourcing as a way to collect data and engage people in science; over half of the top 30 U.S. universities are conducting research on the MTurk platform (Goodman, Dryder, & Cheema, 2013). Researchers are beginning to test MTurk in evaluation science. This study adds to the empirical evidence base by demonstrating the validity, reliability, and cost efficiency of crowdsourcing as a resource for evaluators to broaden stakeholder engagement.

**Authors’ Note**

This work represents a subset of findings from a dissertation defended at Claremont Graduate University as part of a doctoral program in the School of Social Science, Policy, and Evaluation.

**Acknowledgment**

We would like to thank Drs. Ian Coulter and Gery Ryan of RAND Corporation for their subject matter expertise and acquisition of funding.

**Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by a grant from the National Institutes of Health/National Center for Complementary and Integrative Health (Grant Number 1R21AT009124-01).

**ORCID iD**

Lara G. Hilton https://orcid.org/0000-0003-2375-8810
Supplemental Material

The Supplemental Appendix is available with the online article at http://journals.sagepub.com/home/aje.

Notes

1. Crowdsourcing has been in existence since the 1800s (e.g., the Oxford English Dictionary was crowd-sourced); however, the invention of the web has made access to the “crowd” easier.

2. Please note that the samples were not representative of the U.S. population, but when compared to other common participant sources such as colleges and internet samples, it was better in terms of diversity. The Mechanical Turk samples tend to be more educated and younger when compared to the U.S. population.

3. The amount of compensation was based on federal minimum wage standards. We divided the amount of time it takes to complete the survey by the hourly rate. On average, the screening surveys took approx. 10 min to complete, this is approx. 16% of an hour, and with the federal minimum wage of US$7.25, this amounted to US$1.2 for the task, and the participants were paid US$2 for this task. Those that passed the screening (chronic pain participants) were invited to participate in a 30-min coding refinement survey and were paid US$5 (the federal minimum wage for that time would have been US$3.63). Following that those who had chronic pain and were chiropractic users were invited to participate in another coding task which lasted 30 min and were paid US$4 for completing it (the federal minimum wage was US$3.63). In all these cases, the participants received compensation that was higher than the federal minimum wage.

References

Amazon Mechanical Turk. (2018). Amazon Mechanical Turk, Inc. An Amazon Company. Retrieved from https://www.mturk.com

Azzam, T., & Harman, E. (2016). Crowdsourcing for quantifying transcripts: An exploratory study. Evaluation and Program Planning, 54, 63–73.

Azzam, T., & Jacobson, M. (2013). Finding a comparison group: Is online crowdsourcing a viable option? American Journal of Evaluation, 34, 372–384.

Bartek, M., Truitt, A. R., Widmer-Rodriguez, S., Baur, Z. A., Comstock, B. A., Edwards, T. C., ... Lavallee, D. C. (2017). The promise and pitfalls of using crowdsourcing in research prioritization for back pain: Cross-sectional surveys. Journal of Medical Internet Research, 19(10), e341. doi:10.2196/jmir.8821

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon’s Mechanical Turk: A new source of inexpensive, yet high-quality data? Perspectives on Psychological Science, 6(1), 3–5. doi:10.1177/1745691610393980

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

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37–46.

Creswell, J. W., & Clark, V. L. P. (2017). Designing and conducting mixed methods research. Newberry Park, CA: Sage.

Dedoose. (2017). Dedoose Version 7.0.23, web application for managing, analyzing, and presenting qualitative and mixed method research data. Los Angeles, CA: SocioCultural Research Consultants, LLC. Retrieved from www.dedoose.com

Deyo, R. A., Dworkin, S. F., Amtmann, D., Andersson, G., Borenstein, D., Carragee, E., ... Weiner, D. K. (2014). Report of the NIH Task Force on research standards for chronic low back pain. Journal of Pain, 15, 569–585. doi:10.1016/j.jpain.2014.03.005

Goodman, J. K., Dryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk. Journal of Behavioral Decision Making, 26, 213–224. doi:10.1002/bdm.1753
Harman, E. (2016). Incorporating public values into evaluative criteria: Using crowdsourcing to identify criteria and standards. Claremont, CA: Claremont Graduate University.
Harman, E., & Azzam, T. (2018). Validating program theories: A new role for qualitative data using crowdsourcing. *Evaluation and Program Planning, 66*, 183–194. doi:10.1016/j.evalprogplan.2017.08.008
Harris, J. K., Mart, A., Moreland-Russell, S., & Caburnay, C. A. (2015). Diabetes topics associated with engagement on Twitter. *Preventing Chronic Disease, 12*, 1–9. doi:10.5888/pcd12.140402
Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine, 14*, 1–4.
Ipeirotis, P. G. (2010). Analyzing the Amazon Mechanical Turk Marketplace. *ACM XRDS, 17*, 16–21. doi:10.1145/1869086.1869094
Ipeirotis, P. G. (2018). MTurk Tracker. Retrieved from http://www.mturk-tracker.com/
Jacobson, M. R., Whyte, C. E., & Azzam, T. (2017). Using crowdsourcing to code open-ended responses: A mixed methods approach. *American Journal of Evaluation, 1–17*. doi:10.1177/1098214017717014
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 378–382.
Litman, L., Robinson, J., & Abberbock, T. (2016). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 1–10*. doi:10.3758/s13428-016-0727-z
Mason, W., & Suri, S. (2011). Conducting behavioral research on Amazon’s Mechanical Turk. *Behavior Research, 30 June 2011*. doi:10.3758/s13428-011-0124-6
Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23*, 183–188. doi:10.1177/0963721414531598
Parvanta, C., Roth, Y., & Keller, H. (2013). Crowdsourcing 101: A few basics to make you leader of the pack. *Health Promotion Practice, 14*(2), 163–167.
Ranard, B. L., Ha, Y. P., Meisel, Z. F., Asch, D. A., Hill, S. S., Becker, L. B., … Merchant, R. M. (2013). Crowdsourcing—Harnessing the masses to advance health and medicine, a systematic review. *Journal General Internal Medicine, 29*, 187–203. doi:10.1007/s11606-013-2536-8
Ryan, G. W., & Bernard, H. R. (2003). Techniques to identify themes in qualitative data. *Field Methods, 15*, 85–109.
Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychology Science, 1*, 213–220.
Siegel, J. T., Navarro, M. A., & Thomson, A. L. (2015). The impact of overtly listing eligibility requirements on MTurk: An investigation involving organ donation, recruitment scripts, and feelings of elevation. *Social Science & Medicine, 142*, 256–260. doi:10.1016/j.soscimed.2015.08.020
Tosti-Kharas, J., & Conley, C. (2016). Coding psychological constructs in text using Mechanical Turk: A reliable, accurate, and efficient alternative. *Frontiers in Psychology, 7*(741), 1–9. doi:10.3389/fpsyg.2016.00741
Weiner, M. (2014). The potential of crowdsourcing to improve patient-centered care. *Patient, 7*, 123–127. doi:10.1007/s40271-014-0051-6
Wyms, B. T., & Dawson, A. E. (2015). Screening Amazon’s Mechanical Turk for adults with ADHD. *Journal of Attention Disorders, 1–10*. doi:10.1177/1087054715597471