Evaluation and Program Planning 54 (2016) 63–73

Contents lists available at ScienceDirect

Evaluation and Program Planning

journal homepage: www.elsevier.com/locate/evalprogplan

Crowdsourcing for quantifying transcripts: An exploratory study

Tarek Azzam *, Elena Harman

Claremont Graduate University, United States

ARTICLE INFO

Article history:
Received 12 December 2014
Received in revised form 31 August 2015
Accepted 16 September 2015
Available online 9 October 2015

Keywords:
Crowdsourcing
Qualitative analysis
Stability
Transcript coding
Transcript rating
Mechanical Turk
MTurk

ABSTRACT

This exploratory study attempts to demonstrate the potential utility of crowdsourcing as a supplemental technique for quantifying transcribed interviews. Crowdsourcing is the harnessing of the abilities of many people to complete a specific task or a set of tasks. In this study multiple samples of crowdsourced individuals were asked to rate and select supporting quotes from two different transcripts. The findings indicate that the different crowdsourced samples produced nearly identical ratings of the transcripts, and were able to consistently select the same supporting text from the transcripts. These findings suggest that crowdsourcing, with further development, can potentially be used as a mixed method tool to offer a supplemental perspective on transcribed interviews.

1. Introduction

The advent of recent technological developments has brought with it the potential for new qualitative coding methods that can, in certain instance, be cost effective and require less time to conduct. One important technological development is crowdsourcing, or the "paid recruitment of an independent global workforce for the objective of working on a specifically defined task or set of tasks" (Behrend, Sharek, Meade, & Wiebe, 2011, p. 800). In the evaluation context, crowdsourcing has emerged in recent years as a viable research subject pool (Berinsky, Huber, & Lenz, 2012; Paolacci, Chandler, & Ipeirotis, 2010), and as a potential source for comparison groups in quasi-experimental evaluations (Harman & Azzam, Revise & Resubmit). However, relatively little is known about the viability of this tool as a mixed method approach for quantifying qualitative data. This paper expands the research on crowdsourcing in evaluation to include the rating and coding of transcripts by members of the "crowd." The research attempts to test the stability of quantitative ratings and the selection of supporting quotes produced by different samples of crowdsourced raters across multiple transcribed interviews. Stability is the main focus of this paper because it is the foundation on which validity can be built. If the crowdsourced ratings changed from sample to sample then it would be difficult to get a sense of how the transcript is being interpreted and this would make it difficult to reach an understanding of transcript.

The concept of stability of findings is closely related to reliability, which has traditionally been defined as the degree of consistency between multiple raters (e.g., Hayes & Krippendorff, 2007). However, the commonly used measures of reliability are not applicable when examining the stability of crowdsourced findings in this type rating and text selection. This is due to differences in units of analysis. When assessing reliability between raters the unit of analysis is each individual rater (Saal, Downey, & Lahey, 1980), however in crowdsourcing the unit of analysis is the average of the sample, because that is the best measure of the crowd's consensus (Surowiecki, 2004). Conceptually this would mean that each sample's average would be analogous to a single rater.[1]

Crowdsourcing occurs when many people work on a common task, and thanks to recent technological advances, it is now possible to call upon hundreds—if not thousands—of people simultaneously. As technology contributes to improvements in both quantitative and qualitative inquiry, it is worthwhile to continue to develop and refine new methods of inquiry and analysis. The potential of crowdsourcing in social science research is not limited to data collection. Mason and Suri (2012), for

[1] It should be noted that there could be exceptions to this general guidance if the distribution is bimodal or highly skewed, in such cases we recommend using a more appropriate measure of centrality such as the median or mode. If the results demonstrate stability, we hope to utilize such findings as bases for future studies that explore the potential strengths and limitations of this approach.

* Corresponding author.
E-mail address: tarek.azzam@cgu.edu (T. Azzam).

http://dx.doi.org/10.1016/j.evalprogplan.2015.09.002
0149-7189/

example, found that respondents from a popular crowdsourcing website[2] (called: Amazon's Mechanical Turk, MTurk) can substitute for expert judgments in tasks such as language processing, audio transcription, and document comparison. More specific to evaluation, (Harman & Azzam, Revise & Resubmit) demonstrated the viability of MTurk as a virtual comparison group in the evaluation of a college retention program. The present study seeks to explore and assess the viability of crowdsourcing for quantifying transcripts through ratings and the selection of supporting quotes.

However, this approach, as applied to qualitative data, should not be considered qualitative analyses and may not replace a deep analysis of the meanings and perspectives that are expressed in qualitative data. An evaluator may conduct the interviews, review them for content, themes, and attempt to make sense of them and how they relate to the program, and generally follow the practices of quality qualitative inquiry. As part of this process, the evaluator can potentially utilize crowdsourcing to provide an overall quantitative rating of the persons' experience and identify supporting quotes to help with the interpretation of those ratings. Our proposed approach is closer to a mixed methods approach that attempts to quantify qualitative to provide a broader sense of the experiences discussed in the transcripts. The process of quantification has been discussed multiple times within the mixed method literature (Greene, 2007; Teddlie & Tashakkori, 2006), and is described as the "...the process of assigning numerical (nominal or ordinal) values to data conceived as not numerical" (Sandelowski, Voils, & Knafl, 2009, p. 209). This process can take on many forms that include counting the frequencies of word or phrase appearances, recording the presence or absence of specific content, or (in the case of this study) attempting to place a numerical value to represent the general reactions or experiences of an interviewee.

This quantification process has limitations connected to the interpretation and potential oversimplification of the experiences present in the qualitative data (Sandelowski et al., 2009). This is why we do not believe that this approach can replace genuine qualitative analysis, but we think that it can supplement it by offering another interesting and broader perspective on the experiences described or observed in qualitative data. Although a number may not be sufficient to capture these experiences, it may, however, help in describing an overall pattern of experiences across multiple qualitative sources and provide additional information to aid in the interpretation of qualitative themes. The goal in this article is to test the stability of findings from this crowdsourcing process.

1.1. Accessing the "Crowd"

Amazon's Mechanical Turk (MTurk) represents one of the largest and most accessible crowdsourcing websites (Berinsky et al., 2012). MTurk was designed to facilitate the completion of human computation tasks that are difficult for a computer to do accurately. The platform allows "requesters" to crowdsource the completion of "human intelligence tasks" (HITs) using "workers"—participants who are paid a small sum of money for each completed task. The tasks range in size, topic, and complexity, and MTurk is increasingly being used as a subject pool for social science research. Recent studies conducted using MTurk workers as participants have successfully replicated the results of classic laboratory experiments (Berinsky et al., 2012; Paolacci et al., 2010).

In terms of demographics, many studies have shown that MTurk samples tended to be more closely representative of the US population when compared to subject pools at universities and colleges and other internet samples (Berinsky et al., 2012; Buhrmester, Kwang, & Gosling, 2011; Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010). However, it is important to note that samples collected from MTurk do tend to have a larger proportion of respondents with higher levels of education, and they also tend to be younger than the general US population (Ross et al., 2010). A common concern with MTurk is the quality of the responses given the low pay rates. Studies were conducted to examine this concern, the first set of research were a series of replication studies that compared the findings from MTurk studies to commonly known patterns or findings from studies in the social sciences. These studies found that the data gathered from MTurk participants generally produced the same results as was found with other methods and across different disciplines (Berinsky et al., 2012; Paolacci et al., 2010). Other studies examined the motivation of the MTurk populations for completing the tasks. Overall the findings from these studies suggest that money, and the perception of MTurk tasks as a form of mental engagement were major driving motivation for participating in MTurk and that money paid did not affect the quality of responses (Chandler & Kapelner, 2010; Horton, Rand, & Zeckhauser, 2010; Paolacci et al., 2010).

The present paper examines the viability of crowdsourcing to quantify transcripts. It includes two phases that address the research question: Can multiple samples of crowdsourced raters/coders produce the same average ratings and consistently identify supporting phrases after reviewing interview transcripts? It is important to note that our task in this paper is to begin the process of establishing the stability of the findings from this approach, while briefly touching on the issue of validity and the implications of the findings.

2. Methods

The study was divided into two phases. The procedures for each of the phases were similar, with some variations in the tasks requested from participants. In each phase, participants were recruited through Amazon's Mechanical Turk (MTurk) website and asked to read a transcribed interview and then respond to questions about the interviews. The MTurk participants were given the role of raters of the transcripts. As part of the study design multiple samples of MTurk participants (i.e. raters) were recruited and the results from each sample were compared to determine the stability of findings from sample to sample. Phase I was used to test the mechanics of the survey (e.g. highlighting text), clarity of the instructions, and as an initial proof of concept. At the end of the survey we asked participants to give us feedback on the survey and task and to let us know if any improvements were needed. Through the feedback and examination of the data we found that the mechanics of the survey and instructions worked very well and we ended up with no revisions to the task and instructions for Phase II. The second phase added a bit more complexity and direction to the coding task to test the stability of the coders' analysis across varying samples and times.

MTurk participants were accessed via the site by following instructions for how to post a Human Intelligence Task (HIT) (www.mturk.com). The HIT required a brief description of the task, the length of time needed to complete the task (in our case approx. 15 min), the payment rate ($0.50 per HIT), and a link to the survey containing the transcript and rating questions. Once the description was finalized, the HIT was included with other HITs from different people on the MTurk site. After participants selected and completed the survey they were given a random number at the end of the survey. This random number was inputted into the MTurk site to verify completion of the HIT, and once verified the participants were paid.

MTurk also allows requesters to select individuals with specific qualifications by creating a qualification test through the system. If

[2] Amazon's Mechanical Turk (MTurk) is a website that allows users to post paid tasks that can be performed by a large number of people (i.e. the crowd).