The State of Pilot Study Reporting in Crowdsourcing: A Reflection on Best Practices and Guidelines

Jonas Oppenlaender[1*], Tahir Abbas[2] and Ujwal Gadiraju[2]

[1*]University of Jyväskylä, Seminaarinkatu 15, Jyväskylä, 40014, Finland.

[2]Delft University of Technology, Mekelweg 5, Delft, 2628 CD, Netherlands.

*Corresponding author(s). E-mail(s): joppenlu@jyu.fi;
Contributing authors: t.abbas-1@tudelft.nl;
u.k.gadiraju@tudelft.nl;

Abstract

Pilot studies are an essential cornerstone of the design of crowdsourcing campaigns, yet they are often only mentioned in passing in the scholarly literature. A lack of details surrounding pilot studies in crowdsourcing research hinders the replication of studies and the reproduction of findings, stalling potential scientific advances. We conducted a systematic literature review on the current state of pilot study reporting at the intersection of crowdsourcing and HCI research. The literature review included 171 articles published between 2012 and 2021 in the proceedings of the Conference on Human Computation and Crowdsourcing (AAAI HCOMP) and the ACM Digital Library. We found that pilot studies in crowdsourcing research (i.e., crowd pilot studies) are often under-reported in the literature. On the basis of our findings, we formulate a set of guidelines for reporting pilot studies in crowdsourcing research. We also reflect on the current state of practice and provide design implications for crowdsourcing platforms.

1 Introduction

Crowdsourcing is an empirical research area that involves human subjects. The very ingredients that make crowdsourcing a powerful paradigm – diversity in the background of participating individuals and independence in their opinion (Surowiecki, 2005) – also lead to a wide range of behavior and a high variance in performance. It is therefore no surprise that a majority of work in the realms of crowdsourcing research over the last two decades has focused on addressing challenges related to quality (Kittur et al, 2008, 2013; Gadiraju et al, 2015a; Chang et al, 2017). This well-documented variability in human behavior and performance while carrying out crowdsourcing tasks interacts with other task parameters to shape outcomes, such as the task reward (Yin and Chen, 2015), task complexity (Yang et al, 2016), task clarity (Gadiraju et al, 2017c), batch size (Difallah et al, 2015), and reward schemes (Fan et al, 2020). Many of such influential configuration parameters of a crowdsourcing campaign are not known before the campaign is launched. Due to this, researchers and practitioners turn to pilot studies to inform their design choices and fine-tune such parameters. Pilot studies are a vital part of crowdsourcing research. For instance, researchers often launch one or several small-scale studies to estimate the average completion time of crowdsourced tasks with the aim of appropriately setting the monetary rewards for a larger main study. In fact, to help researchers accurately structure and price their work, objective measures like ETA (error-time area) have been proposed. ETA empirically models the relationship between time and error rate by manipulating the time that workers have to complete a task (Cheng et al, 2015). The measure proposes that requesters rapidly iterate on task designs and measure whether the changes improve the performance of workers and task outcomes. In this work, we refer to these small preliminary studies which are often used to calibrate crowdsourcing task design parameters or inform main studies in one or more ways as **crowd pilot studies**.

Despite the important role that crowd pilot studies play in configuring and thereby shaping crowdsourcing studies, from a preliminary review of the literature, we found that it is common for authors to mention crowd pilot studies only in passing. Crowd pilot studies are often conducted in an ad-hoc manner and details about the crowd pilot studies are seldom reported. This is undesirable, since a lack of detail hinders future reproduction and replication (Echtler and Häußler, 2018). This also lies in stark contrast to one of the fundamental tenets of *open science* and recent frameworks such as the Open Science Framework (Foster and Deardorff, 2017) – to make knowledge transparent and accessible (Vicente-Saez and Martinez-Fuentes, 2018). For instance, readers can glean little from reading that authors '*iterated extensively in pilot studies with crowd workers to strike a balance between simplicity (avoid complex or numerous instructions) and effectiveness (make the layout better)*' — a quote from literature reviewed in this work. Researchers or practitioners who may want to replicate such a process to identify the right balance between simplicity and effectiveness for their own crowdsourcing study based on such a crowd

pilot study description, will arguably be left guessing. As a research community that is still evolving (Kittur et al, 2013), it is important to set the right precedents and establish good practices. It is worth noting that pilot studies are just as likely to be flawed as any other (main) studies which are expected to withstand the scrutiny of peer-review as a means to ensure quality, reliability, good practice, and a sound scientific method. Such flaws in pilot studies can go unnoticed if they are not reported in sufficient detail. While there are guidelines and checklists for running and reporting crowdsourcing studies (Ramírez et al, 2021b, 2020; Redish and Laskowsk, 2009; Simperl, 2021; Draws et al, 2021a), there is a gap in the scholarly literature on pilot studies in crowdsourcing research.

In this paper, we aim to address this gap and synthesize the best practices in reporting crowd pilot studies. To this end, we first conducted a systematic literature review. Our screening of 513 articles downloaded from the ACM Digital Library (ACM-DL) and the proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP) – a premier venue for crowdsourcing related research – resulted in a corpus of 171 articles. We systematically analyze this corpus to capture the current state of crowd pilot study reporting in the scholarly literature. Our aim is to report and reflect on the current state of pilot study reporting at the intersection of the HCI and crowdsourcing literature. To this end, we identified whether and to what extent the following information is being reported in articles:

RQ1: Why are crowd pilot studies typically conducted?
RQ2: How are crowd pilot studies typically reported?
RQ3: What do crowd pilot studies report?

While pilot studies are very common in crowdsourcing research, little is known and reported about them in the scholarly literature (Ramírez et al, 2020, 2021b). Therefore, much of the knowledge from running pilot studies is bound in researchers with experience in crowdsourcing. An experienced researcher may, for instance, decide to not even conduct a crowd pilot study because the researcher's experience will tell what parameters of a crowdsourcing campaign will work best. We therefore complemented our literature review with a survey study with experienced crowdsourcing researchers to fill the aforementioned gap. The survey study investigated broader topics not explicitly reported in the scholarly literature:

RQ4: What makes a "good" crowd pilot study?
RQ5: What are the factors that promote or obstruct reporting crowd pilot studies?
RQ6: How can crowd pilot studies be facilitated with platform-specific features?

To the best of our knowledge, our work is the first to provide a detailed investigation on the current state of practice of crowd pilot study reporting in the crowdsourcing and HCI literature. Based on the findings of our literature

review and survey study, we provide a set of guidelines for reporting crowd pilot studies. We reflect on the trade-offs around running pilot studies and discuss implications for the design of crowdsourcing platforms. All data pertaining to our work in this paper are made publicly available for the benefit of the research community and in the spirit of open science.[1]

Our work is structured as follows. We first provide a brief review of related literature in Section 2. We then describe our methodological approach for conducting the systematic literature review and the complementary survey study in Section 3. In Section 4, we present the results of our analysis, followed by a reflection and discussion of our findings in Section 5. We discuss caveats and limitations of our work in Section 6 and conclude in Section 7.

2 Related literature

2.1 Pilot studies in crowdsourcing-based research

The crowdsourcing paradigm has seen vast adoption in academia and industry. Crowdsourcing is a cost-effective method for conducting online experiments (Paolacci et al, 2010) and user studies (Kittur et al, 2008). However, designing an effective crowdsourcing campaign is not an easy task and there are many pitfalls for requesters when designing crowdsourcing campaigns. For instance, task clarity is one important determinant of work quality (Gadiraju et al, 2017c). Many other factors can potentially affect the work quality, such as a task's complexity (Borromeo et al, 2016), usability, and accessibility (Uzor et al, 2021).

Crowd pilot studies are typically conducted to address these challenges. Crowd pilot studies aim to iteratively design a task and empirically determine design parameters of a crowdsourcing campaign, such as an estimate of the average completion time per task. This estimate can then be used to calculate the price per task for the main study. Before running a pilot study, the average completion time is unknown. Therefore, trial and error is needed to determine an accurate task pricing for microtasks (Whiting et al, 2019a). Determining the amount of pay is part of the design of every crowdsourcing campaign involving monetary incentives.

Tools and methods have been developed to support requesters in determining the above parameters. As mentioned in the introduction, error-time-area (ETA) is an approach to empirically model the relationship between time and error rate (Cheng et al, 2015). The ETA measure is derived by manipulating the time that workers have to complete a task. Requesters may use the ETA measure to rapidly and iteratively test different task designs and measure whether the changes improve the performance of workers and task outcomes. Besides ETA, other tools supporting requesters in designing tasks and crowdsourcing campaigns have been developed. Manam et al (2022) developed a linting tool that automatically uncovers ambiguities in task instructions and

[1] https://osf.io/46fxj/?view_only=0eac3aaf2c734a6096e33f9734f62902

supports requesters in writing task instructions with greater clarity. Nouri et al (2021) proposed methods to computationally assess the clarity of tasks and designed a tool to help requesters improve tasks iteratively. Nobre et al (2021) presented a system for running and monitoring pilot studies.

However, in practice, the most typical remedy to the above challenges is running informal, small-scale studies with prototypical tasks. These small-scale studies are often conducted iteratively to rapidly uncover issues in the design of the task or to empirically derive estimates of important determinants of the crowdsourcing campaign (such as the task pricing).

2.2 Guidelines for conducting and reporting crowdsourcing studies

Several best practices and guidelines have been developed for requesters to design crowdsourcing campaigns. These guidelines are motivated with two primary concerns. Some authors take the workers' perspective and aim to provide guidelines for requesters to conduct fair and responsible crowd work. Other authors provide guidelines from the requester's point of view, aiming to optimize the efficiency, cost, quality, and accuracy of crowdsourced work.

From the requester's perspective, Cobanoglu et al (2021) presented a guide and best practices for using crowdsourcing platforms. These guidelines are primarily meant as a beginner's guide to crowdsourcing. Simperl (2021) also provided guidelines and examples on using crowdsourcing effectively. The guidelines take a system development perspective aiming to provide "design and participation best practices" guiding the development of crowdsourcing systems. Alonso (2009) provided a short list of guidelines for designing crowdsourcing studies. The article is scoped to practical aspects when conducting relevance evaluations. Redish and Laskowsk (2009) presented guidelines for writing clear instructions for voters and poll workers. While this report is not written for the crowdsourcing domain, the report provides takeaways for writing clear instructions to crowd workers. Gadiraju et al (2015b) explore the different ways in which tasks can be exploited by unreliable workers in surveys and propose task design guidelines to thwart such behavior and ensure quality control. Whiting et al (2019b) introduced a means to help requesters in automatically paying workers a minimum wage by adding a one-line script tag to their task HTML on Amazon Mechanical Turk (MTurk). Draws et al (2021b) proposed a checklist as a practical tool that requesters can use to improve their task designs by mitigating cognitive biases of workers and appropriately describe potential limitations of collected data.

Guidelines written with the workers' perspective in mind are fewer in number. For instance, Dynamo by Salehi et al (2015) provided worker-generated "Guidelines for Academic Requesters" for ethical research on Amazon Mechanical Turk (Dynamo Contributors, 2014). The guidelines aim to provide guidance for requesters on "how to be a good requester," fair payment, and other aspects of fair crowd work. Schäfer et al (2017) formulated key principles for effective communication with workers in crowdsourcing contests. In a