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

Oppenlaender, Jonas; Abbas, Tahir; Gadiraju, Ujwal

DOI
10.1145/3641023

Publication date
2024

Document Version
Final published version

Published in
Proceedings of the ACM on Human-Computer Interaction

Citation (APA)
Oppenlaender, J., Abbas, T., & Gadiraju, U. (2024). The State of Pilot Study Reporting in Crowdsourcing: A Reflection on Best Practices and Guidelines. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), Article 184. https://doi.org/10.1145/3641023

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

JONAS OPPENLAENDER, Elisa Corporation, Finland
TAHIR ABBAS, Delft University of Technology, Netherlands
UJWAL GADIRAJU, Delft University of Technology, Netherlands

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. Our review of ten years of literature included 171 articles published 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. Important details, such as the number of workers and rewards to workers, are often not reported. Based on our findings, we reflect on the current state of practice and formulate a set of best practice guidelines for reporting crowd pilot studies in crowdsourcing research. We also provide implications for the design of crowdsourcing platforms and make practical suggestions for supporting crowd pilot study reporting.

CCS Concepts: • **Information systems** → **Crowdsourcing**.

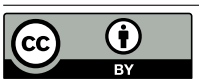
Additional Key Words and Phrases: pilot studies, crowdsourcing, systematic literature review

ACM Reference Format:
Jonas Oppenlaender, Tahir Abbas, and Ujwal Gadiraju. 2024. The State of Pilot Study Reporting in Crowdsourcing: A Reflection on Best Practices and Guidelines . *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 184 (April 2024), 45 pages. https://doi.org/10.1145/3641023

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 [203] – 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 [29, 65, 112, 113]. 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 [240], task complexity [238], task clarity [68], batch size [47], and reward schemes [59]. 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 and researchers often launch one or several small-scale

Authors' addresses: Jonas Oppenlaender, jonas.oppenlander@elisa.fi, Elisa Corporation, Ratavartijankatu 5, Helsinki, Finland, 00520; Tahir Abbas, t.abbas-1@tudelft.nl, Delft University of Technology, Mekelweg 5, 2628 CD, Delft, Netherlands; Ujwal Gadiraju, u.k.gadiraju@tudelft.nl, Delft University of Technology, Mekelweg 5, 2628 CD, Delft, Netherlands.



ACM 2573-0142/2024/4-ART184
https://doi.org/10.1145/3641023

studies before launching the main study. One typical reason, among others, is to estimate the average completion time of crowdsourced tasks to appropriately set the monetary rewards for the larger main study. In this work, we refer to these small preliminary studies which are often used to calibrate crowdsourcing task design parameters or inform main studies 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 find 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.

As a research community that is still evolving [113], it is important to set the right precedents and establish good practices. One obstacle to establishing good practices is that crowdsourcing research is conducted in many research communities from different disciplines. The prevailing practices in these different research communities may be highly diverse. A simple psychological model of why pilot studies are not more often reported in the literature is the availability heuristic [210]. The more researchers report crowd pilot studies in an opaque way, the more frequently other researchers from their community (or other communities) will observe such reporting and read about it in the literature. Practitioners and researchers in the communities, therefore, may receive signals from the community that the details of the pilot study are not important. A further signal is sent by reviewers and editors who accept papers with sparse details on pilot studies. Therefore, researchers and practitioners may not consider details about pilot studies as important or useful for their own work, and may not place importance on reporting their own pilot studies.

This is undesirable for several reasons. Opaque reporting of pilot studies lies in stark contrast to one of the fundamental tenets of *open science* and recent frameworks such as the Open Science Framework [62] – to make knowledge transparent and accessible [217]. A lack of transparency on key design parameters of a crowdsourcing campaign hinders future reproduction and replication [54]. 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 the literature reviewed in this work. Based on such a description of a crowd pilot study, researchers or practitioners, who may want to learn more about how to achieve a balance between simplicity and effectiveness for their own crowdsourcing study, will arguably be left guessing. 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.

Another reason why more transparency around pilot studies is warranted is motivated by the crowd workers' perspective. It is not uncommon for researchers to underestimate the price of a task (we speak from our own experience in this regard). Yet in our literature review, we found it is not very common to make up for these estimation errors (e.g., with bonus payments to workers). This may contribute to the systematic underpayment of workers [45, 80, 133]. Further, crowd pilot studies are very common, yet they are relatively unattractive to many workers as they are small-scale – that is, done in only small batches of tasks –, and underpaid. While some workers may be motivated to participate in pilot studies [151], other workers may want to avoid them.

Given the prevailing practices and highly interdisciplinary nature of crowdsourcing research, we believe it is not likely for common reporting standards to emerge on their own from within the different research communities unless the community is alerted about the state of pilot study reporting and incentivized to change their prevailing practices. Our literature review serves this end, by providing guidelines and practical suggestions on how the current state of crowd pilot study reporting could be improved. We aim toward the development of common reporting standards for crowd pilot studies. Pilot studies are highly diverse and this, of course, is one reason contributing

to the fact that there is no consensus on reporting them in the academic community. Researchers use different "terms" to denote pilot studies (i.e., pilot studies or pilot tests), and also report them in different sections or different levels of detail. Yet in our literature review, we show that there are many commonalities between crowd pilot studies across a multitude of different fields that would allow the development of common reporting standards. But while there are guidelines and checklists for running and reporting crowdsourcing studies [50, 168, 170, 174, 196], 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 [168, 170]. 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 the caveats and limitations of our work in Section 5.5 and conclude in Section 6.

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

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 [156] and user studies [112]. 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 [68]. Many other factors can potentially affect the work quality, such as a task's complexity [21], usability, and accessibility [211].

Crowd pilot studies are typically conducted to address these challenges. Crowd pilot studies aim to iteratively design a task and empirically determine the 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 [229]. 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. Objective measures like ETA (error-time area) task [32] have been proposed to help researchers accurately structure and price their work. ETA empirically models the relationship between time and error rate by manipulating the time that workers have to complete a task [32]. The measure proposes that requesters rapidly iterate on task designs and measure whether the changes improve the performance of workers and task outcomes. 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. [131] developed a linting tool that automatically uncovers ambiguities in task instructions and supports requesters in writing task instructions with greater clarity. Nouri et al. [148] proposed methods to computationally assess the clarity of tasks and designed a tool to help requesters improve tasks iteratively. Nobre et al. [147] 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 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. [35] presented a guide and best practices for using crowdsourcing platforms. These guidelines are primarily meant as a beginner's guide to crowdsourcing. Simperl [196] 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 [6] provided a short list of guidelines for designing crowdsourcing studies. The article is scoped to practical aspects when conducting relevance evaluations. Redish and Laskowsk [174] 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. [66] 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. [229] 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. [50] 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. [186] provided worker-generated "Guidelines for Academic Requesters" for ethical research on Amazon Mechanical Turk [53]. 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. [191] formulated key principles for effective communication with workers in crowdsourcing contests. In a similar vein, the "ground rules" hosted at Crowdsourcing-code.com aim to provide guidance "for a prosperous and fair cooperation between crowdsourcing companies and crowdworkers" [39]. Besides the above documents, guidance and feedback for requesters can also be found on worker-focused websites and in online forums, such as Turkopticon [103] and Turker Nation [133].

However, when it comes to reporting crowdsourcing studies, little guidance is available in the scholarly literature [168, 170]. To the best of our knowledge, there are only two papers providing guidance on how to report crowdsourcing studies and experiments. Ramírez et al. [168] proposed DREC, a datasheet for reporting experiments in crowdsourcing. Based on an analysis of a sample of 15 scientific articles, the authors provided a glimpse into the state of reporting on crowdsourcing studies. The authors found that details of crowdsourcing studies are often not being reported in scientific articles. The authors created a taxonomy of attributes relevant to crowdsourcing studies aiming to support requesters in reporting crowdsourcing experiments. Ramírez et al. later extended the DREC taxonomy in scope and provided a checklist for reporting crowdsourcing experiments, based on an analysis of the literature [170]. The article examines the state of reporting on crowdsourcing experiments and offers guidance for requesters.

In relation to these two studies, there is an overlap with our work in that the authors make recommendations for the reporting of key statistics of a crowdsourcing campaign, such as the number of participants. However, the checklist is clearly scoped to report the results of the main crowdsourcing experiment. The section on pilot studies (Item No. 6) in Ramírez et al. [170] is very short and only recommends to "[d]escribe if pilot studies were performed before the main experiment" (p. 30).

Our work connects to the above two guidelines by providing an extensive and in-depth investigation of the state and practice of pilot study reporting as well as detailed recommendations for reporting crowd pilot studies.

3 METHOD

We conducted a systematic literature review [162] to investigate how pilot studies are being reported in the HCI and crowdsourcing literature (sections 3.1–3.4). The literature review is complemented with an online survey with requesters (see Section 3.5).

3.1 Scope of the Literature Review

Before we started our literature review, we needed to clearly define the research questions (see Section 1) and delineate the boundaries of the review [161]. Our work focuses specifically on pilot studies that are crowdsourced to a group of diverse and independent individuals online. Throughout

the remainder of this work, we refer to these types of pilot studies as **crowd pilot studies**. This concept has two components ('pilot study' and 'crowd') which we clarify and define in the following two sections.

3.1.1 Pilot study. From the onset of our literature review, we were particularly interested in small-scale formative pilot studies in the crowdsourcing domain. The way these studies are being reported in the scholarly literature is often opaque and we wanted to illuminate researchers' practices around reporting pilot studies. This is important because opaque reporting of pilot studies may obfuscate information and – from a systemic perspective – attenuate the spread of best practices. The first iteration of our literature review – as a scoping review [161] – found over 100 articles reporting small-scale pilot studies in a formative way. However, in this scoping review, we found that a considerable amount of pilot studies are also being conducted for summative purposes. We, therefore, extended the scope of our literature review to a broader definition of pilot studies that better captures the state of reporting on both formative and summative studies.

Many different user studies and experiments have been conducted on crowdsourcing platforms. In our work, a pilot study is any small-scale or larger-scale experiment or study that is being conducted to inform the design of a prototype, validate a proof of concept, or for other formative or summative reasons. The scope of our work is defined by the authors' use of the term crowdsourcing in combination with the 'pilot' keyword. For instance, user studies on crowdsourcing platforms are only included in our literature corpus if the studies were referred to by the authors as pilot studies.

3.1.2 Crowd. The pilot studies could be conducted internally by the researchers within a lab setting or with an external crowd [67]. In our work, we exclusively focus on pilot studies conducted with an external crowd. This external crowd needs to consist of people other than the researchers (otherwise it would be considered an "internal pilot study"). Crowdsourcing comes in many different forms (e.g., crowdfunding, contests, microtasking, among others). Our work follows a broad and integrative definition of the crowd. Our literature review includes studies conducted on traditional paid microtasking platforms, situated and mobile crowdsourcing [73, 92], or on other paid and unpaid crowdsourcing platforms with different types of participants (e.g., students and volunteers).

3.2 Creating a Corpus of Relevant Articles

We limited our search to articles published in the past ten years (2012 – 2021). The time frame of ten years was chosen to provide a representative window into current best practices that have emerged since the inception of crowdsourcing.

Our search was conducted in two bibliographic sources. First, we downloaded all articles (excluding posters) from the Proceedings of the Conference on Human Computation and Crowdsourcing (HCOMP), widely considered a primary venue for crowdsourcing-related research. This resulted in 215 articles (of which two articles from the oldest proceedings could not be retrieved). Second, we searched the Digital Library of the Association for Computing Machinery (ACM). The ACM-DL is the document storage for all articles published by the ACM and therefore covers a wide range of different conferences and journals (including the ACM CHI and ACM CSCW conferences where premier works on crowdsourcing research at the intersection of HCI are commonly published). The search in the ACM-DL used the following query string:

> *query: Fulltext:(pilot) AND Abstract:(crowdsourc*)*
> *filter: Article Type: Research Article,*
> *Publication Date (01/01/2012 to 12/31/2021)*

The choice of the 'pilot' keyword reflects the scope of our literature review. We found that limiting our search to occurrences of "crowdsourcing" (or derivations thereof) in the abstract was a good

compromise between identifying relevant literature and avoiding false positives (e.g., studies that mention crowdsourcing in the related work or the references).

Our aim was to arrive at a representative coverage of articles in the literature [219]. We focused on the proceedings of AAAI HCOMP and the ACM Digital Library due to their prominence and influence in the crowdsourcing research community. These venues are known for their rigorous peer-review processes and attract a wide range of high-quality submissions from researchers globally, making them representative sources for our analysis. While we acknowledge that relevant literature might also exist in the IEEE Xplore library and other repositories, our choice was driven by the desire to capture the core developments and trends in crowdsourcing from these leading communities. Our search resulted in a corpus of 513 articles (213 articles at HCOMP and 300 articles in the ACM-DL).

3.3 Article filtering and exclusion criteria

We filtered the corpus of 513 articles in five consecutive steps. The steps are depicted in the flowchart in Figure 1.

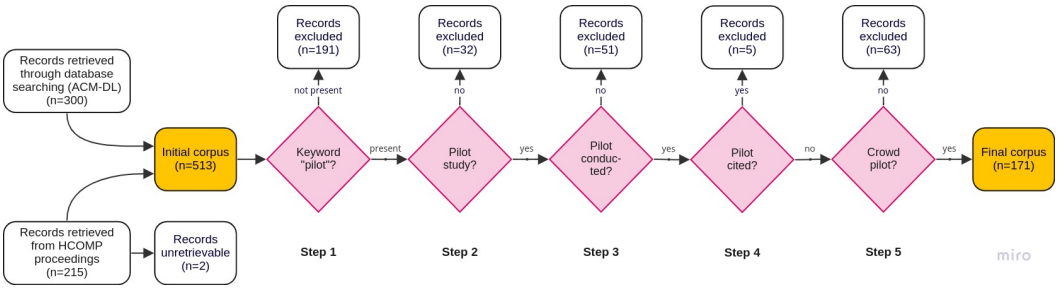


Fig. 1. Corpus screening procedure.

First, we identified whether the keyword "pilot" was present in the article. Articles from the HCOMP proceedings that did not include this term were excluded ($n = 191$). This step also excluded articles from the ACM-DL where the pilot keyword was only mentioned in the references. Second, we identified whether the term "pilot" refers to a pilot study or experiment, or whether it denotes something else (e.g., 'Palm Pilot' or 'airplane co-pilot'). This step excluded 32 articles. Third, we identified whether the pilot study was conducted by the authors in the article. This step excluded articles that mentioned pilot studies in the related work section or only provided recommendations for conducting pilot studies, without conducting one in the article. Articles that did not conduct a pilot study were excluded ($n = 51$). Next, we excluded articles in which authors referenced and discussed their pilot study in other articles. We decided not to conduct a backward reference search because the cited articles did not contain our search keyword and we wanted to focus on full articles. This step excluded five articles. Finally, we identified whether the pilot study involved crowd workers. As mentioned in Section 3.1.2, we apply a broad definition of crowd work that includes everything from situated crowdsourcing, citizen science (e.g., Zooniverse), and volunteering, to paid microwork. We specifically focused on pilot studies conducted with a crowd, excluding other participants (such as experts) in our analysis. If the pilot study did not include a pilot study conducted with a crowd, the article was excluded ($n = 63$). If it was not fully clear from the authors' statements whether the pilot study involved crowdsourcing, we included the article in our analysis ($n = 2$).

We validated the robustness of our 5-step literature screening procedure by conducting a sensitivity analysis. The sensitivity analysis involved slightly altering the criteria for the first step and

then applying the full 5-step filtering process on a subset of articles. More specifically, we expanded the first step to include "preliminary," "initial," and "formative" as search keywords, besides "pilot," and observed the impact on the sampled articles. As a subset for the sensitivity analysis, we selected two HCOMP proceedings (2017 and 2018) with 46 articles (8.97% of the initial corpus). Validation of our filtering methodology showed strong internal consistency. The sensitivity analysis revealed that modifications in the search keyword resulted in less than a 5% change in the subset of papers (2 out of 48; 4.35%), confirming the robustness of our filtering process. The sensitivity analysis also confirmed our initial suspicion that "pilot" is the most commonly used term to refer to a crowd pilot study.

Our final set of literature comprises 171 articles (23 articles from HCOMP and 148 articles from the ACM-DL). Throughout our work, direct quotes from articles are printed in *italics*. The literature corpus was analyzed as follows.

3.4 Analyzing the Corpus

We started our analysis by familiarizing ourselves with the articles. To this end, we manually extracted all verbatim statements that mentioned the pilot keyword from each article together with the surrounding context. Typically, there were only a few mentions of the pilot keyword in one paragraph or short sentences of the article. If there were pilot studies with multiple participant samples, we focused on the pilot studies conducted with the crowd in our analysis.

To answer our research questions, we followed an inductive approach based on grounded theory [70]. We iteratively revisited the verbatim statements to identify what could be reported about the articles (e.g., the number of pilot studies conducted in the article or whether payment to workers was reported). If our research questions could not be answered from the verbatim statements, we revisited the article for closer reading.

Coding was straightforward in cases when variables were binary (e.g., deciding whether the pilot study was reported in its own section) or when information had to be extracted (e.g., the year of publication). This straightforward coding required only one coder and no inter-rater agreement was calculated [135]. Other cases were more challenging. These cases were analyzed by two post-doctoral researchers. The coding was developed iteratively and from the bottom up in several coding passes. The first coding iteration stayed close to the information provided in the articles. This first iteration allowed us to form an understanding of categories in the data, which we then iteratively grouped into codes. The coding results were frequently discussed among all authors which resulted in codes being adjusted and articles iteratively being re-coded. The coding was done in an Excel sheet which was then used to produce graphs in R. All data and code relevant to this process will be shared publicly for the benefit of further research, in the spirit of Open Science.[2]

3.5 Survey Study

It is noteworthy that from the literature alone, we cannot discern anything about authors who decide not to report pilot studies. Therefore, our literature review can only provide incomplete insights into the prevalence of pilot studies and the requesters' motivations for running pilot studies. To limit if not alleviate this publication bias [160] of our literature review, we complemented our analysis with an online survey study with crowdsourcing researchers in academia and industry.

The survey was implemented on Qualtrics. Participants' consent was collected before starting the survey. Participation was incentivized with a raffle of 10 Amazon gift cards (each worth US$15). The survey included 25 items and was estimated to take between 10 and 15 minutes. Many questions were closed-ended (with an option to enter an open-ended response, if preferred or needed) to

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

be mindful of the researchers' time and not overburden the participants. The open-ended survey items focused on two key areas: 1) the motivation for running pilot studies and 2) the requester's practices around reporting pilot studies. The former included questions about the participants' motivation for conducting pilot studies, what they consider as a "good" pilot study, and criteria for running pilot studies. The latter asked what factors promote or obstruct the pilot study reporting and possible features of a crowdsourcing platform that could support requesters in running pilot studies. Participants with experience in collaborating with industry were asked about the differences between pilot studies in industry and academia.

We aimed to invite researchers and industry professionals with experience in crowdsourcing. For this reason, we followed a mix of snowball [74] and convenience sampling [86] to disseminate the survey study to experienced researchers in academia and industry. We also announced the study in communities dedicated to human computation and crowdsourcing, including the HCOMP Slack Community (with 396 members at the time of writing) and the Google Group on Crowdsourcing and Human Computation (with 579 members). The survey had valid responses from twelve participants, but we excluded one because she did not consent to the study. Participants included researchers from academia and industry with a background in computer science, human-computer interaction, and design. The sample includes five Ph.D. students, one postdoctoral researcher, and researchers at the professor level. Participants had between 2 and over 12 years of experience in crowdsourcing research.

4 THE STATE OF CROWD PILOT STUDY REPORTING

In this section, we first provide an overview of the literature corpus before we turn to answering our research questions in the subsequent sections.

4.1 Literature Corpus

The literature corpus includes 171 articles (see Table 1). We find the number of articles reporting crowd pilot studies has more than tripled in the past decade (see Figure 2).

Table 1. Research articles reporting crowd pilot studies per year.

Year	Articles
2012	[25], [57], [58], [111], [116], [189]
2013	[7], [13], [18], [24], [33], [49], [71], [81], [88], [96], [132], [134], [138], [149], [201]
2014	[56], [72], [83], [76], [84], [105], [109], [146], [216], [220], [224], [234], [237], [245]
2015	[8], [30], [82], [90], [120], [126], [127], [155], [167], [207], [233]
2016	[2], [3], [20], [23], [31], [100], [118], [119], [121], [128], [136], [139], [178], [190], [192], [205], [206], [231], [239], [241], [242]
2017	[28], [29], [37], [48], [63], [110], [97], [106], [114], [182], [187], [212], [226], [228], [232]
2018	[4], [5], [11], [15], [22], [27], [38], [41], [69], [91], [93], [102], [104], [143], [154], [157], [173], [184], [188], [197], [198], [213], [222], [230], [243]
2019	[1], [42], [51], [60], [89], [94], [115], [125], [140], [144], [150], [159], [163], [171], [175], [180], [194], [199], [200], [204], [209], [229], [246]
2020	[9], [14], [26], [87], [95], [107], [108], [124], [130], [145], [152], [164], [166], [172], [179], [183], [195], [214], [218], [223], [235]
2021	[10], [34], [44], [?], [52], [61], [79], [117], [122], [141], [142], [148], [158], [169], [176], [177], [181], [185], [225], [236]

The corpus includes articles published in academic conferences ($n = 144$), journals ($n = 24$), and workshops ($n = 3$). The articles have between 2 and 41 pages (including references and appendices). The distribution of articles over the different venues is long-tailed (see Figure 3). About half of the

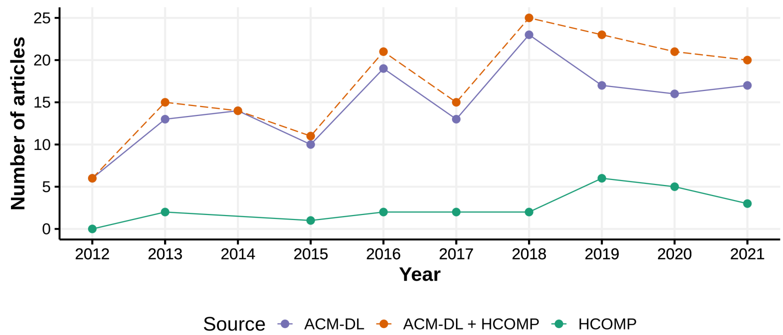


Fig. 2. Articles reporting crowd pilot studies in the ACM Digital Library and HCOMP proceedings.

articles ($n = 82$, 48.0%) are published in the three conferences on Human Factors in Computing Systems (CHI), Human Computation and Crowdsourcing (HCOMP), and Computer Supported Cooperative Work (CSCW), with other conference venues and journals reporting significantly fewer pilot studies, $\chi^2(70, N = 171) = 923.06$, $p < 0.000$.

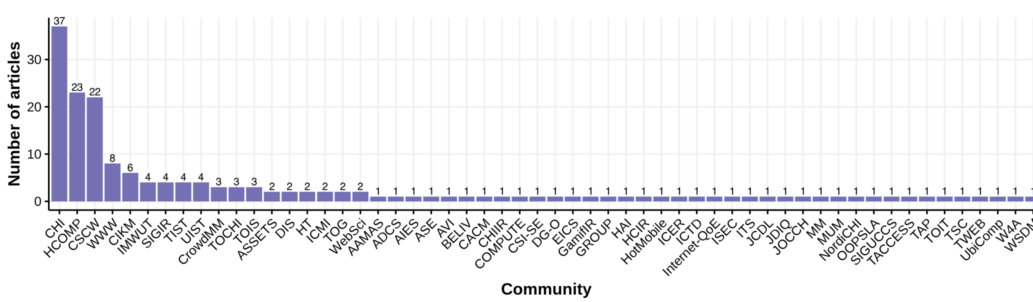


Fig. 3. Conference venues and journals in the corpus.

Analyzing the Computing Classification System (CCS) concepts provided by the authors, we find the bulk of the articles were classified as research on Human-Centered Computing ($n = 47$), HCI ($n = 45$), Information Systems ($n = 26$), Computing Methodologies ($n = 13$), and Applied Computing ($n = 7$). A similar human-centered pattern is notable in the author-provided keywords (see Figure 4) which revolve around human factors, design, user studies, and experimentation involving human subjects.

The corpus includes 770 affiliations (including double-affiliations) of authors from academia and industry. Authors are predominantly affiliated with universities ($n = 607$), with Carnegie Melon University (42 authors), Stanford University (26 authors), the University of Washington (24 authors), and the University of Waterloo (24 authors) having the most author affiliations in our corpus of literature. A smaller number of authors are from industry ($n = 89$) and other institutions (e.g., colleges and research institutes; 74 authors). Among the institutions from the industry are Microsoft (32 authors), Google (10 authors), and Disney Research (7 authors).

The corpus provides us with a source of documentation on the state of practice of crowd pilot study reporting in the HCI and crowdsourcing literature which we analyzed to answer research questions RQ1 – RQ4.

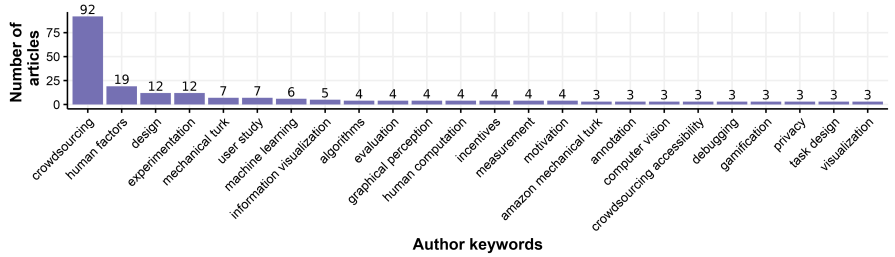


Fig. 4. Bar chart of author-provided keywords appearing in the analyzed papers, including only keywords mentioned at least three times.

4.2 RQ1: Why are Crowd Pilot Studies Typically Conducted?

Most crowd pilot studies in the literature are formative studies conducted during the design or development phase (n = 143, 83.6%). About 15% of the crowd pilot studies are summative studies (n = 26, 15.2%) conducted to evaluate or validate a prototype, proof of concept, or an idea. Three-quarters of the articles in our corpus report crowd pilot studies only in passing — in a few sentences, footnotes, or short paragraphs (128 articles, 74.9%).

We classified the articles based on the amount of space a crowd pilot study is given in the article and the type of crowdsourcing study (formative versus summative). In this classification, 'in passing' refers to articles that mention the pilot study only in a few sentences, footnotes, or short paragraphs. 'Detailed reporting' denotes articles that dedicate a larger amount of space (e.g., a full section) to the pilot study. Finally, 'main study' refers to articles in which the entire article is considered as being a pilot study. Based on this classification, we find there are five different types of crowd pilot studies with varying levels of prevalence in the literature (as depicted in Figure 5):

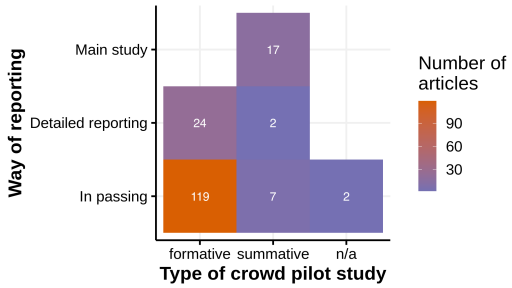


Fig. 5. Types of crowd pilot studies and ways of reporting crowd pilot studies in the literature.

- *Formative crowd pilot study, mentioned in passing* (n = 119, 69.6%): Over two-thirds of the articles in our corpus contain formative crowd pilot studies that are mentioned in passing. These articles are by far the largest group in the literature. The articles in this group are often opaque in how the crowd pilot study is being reported and details about the crowd pilot study are often not provided. As this is the largest group of articles, we analyze these articles in more detail in Section 4.3.1.
- *Formative crowd pilot study, detailed reporting* (n = 24, 14.0%): This type of article devoted more than just a few sentences to the crowd pilot study. For instance, Winther et al. [233] report on a series of three formative pilot studies in which the authors iterated on parameters

related to the design of the task and campaign (such as the task design, task instructions, task reliability, task accuracy, task difficulty, and worker behavior) to inform a crowdsourcing campaign.

- *Summative crowd pilot study, pilot is the main study* (n = 17, 9.9%): Several articles presented a summative crowd pilot study as the main study of the article. This type of study is being conducted to test the feasibility, provide a proof of concept, or evaluate and validate a system. A representative article is the work by Dow et al. [49] who established the feasibility of using crowds for design feedback in the classroom. Other examples are the technical evaluation of the VidQuiz system by Davis et al. [41] and the work by Ramchurn et al. [167] who mention pilot studies with the task allocation system of a disaster response system.
- *Summative crowd pilot study, detailed reporting* (n = 2, 1.2%): Only very few articles contained summative pilot studies reported in detail (i.e., in a separate section of the article). These studies were being conducted to validate the design and functionality of systems or to show the generalizability of the system. Qiu et al. [164] conducted a pilot study with the prototype of a spatial crowdsourcing system. This pilot study was reported after the main results section (titled *"performance evaluation"*). Eickhoff et al. [57] used a crowd pilot study to demonstrate that their game generalizes to other domains.
- *Summative crowd pilot study, mentioned in passing* (n = 7, 4.1%): A few articles mentioned a summative pilot study in passing. Some articles in this group report the crowd pilot study in the context of evaluating a system, demonstrating a proof of concept, or validating the feasibility. Vaish et al. [212], for instance, report that a small team of participants conducted a pilot experiment. Winkler et al. [232] mention that a *"set of pilot runs"* was executed *"to ensure the feasibility of the study design"* (p. 34). Oppenlaender et al. [152] pilot tested their CrowdUI system to ensure the system's functionality before the main study. Similarly, Yang et al. [237] launched a pilot study with the intention of verifying the quality of results before deploying the main study. Other articles validate that a task produces the intended results, such as Noy et al. [149] who compared the results of a pilot study with the main results of their article, finding no differences.
- *Other articles:* Two articles (1.2%) could not be assigned to being formative or summative due to insufficient information in the article. These two instances are an article in which a pilot study is mentioned in the acknowledgments section [122] and the article by Kim and Follmer [110] in which the authors state that crowd workers were excluded if they had *"previously participated in any of [the authors'] pilot studies"* (p. 10).

Digging deeper into the details reported about the crowd pilot studies, we find that task design and crowdsourcing campaigns are the most commonly reported reasons for undertaking a crowd pilot study in the literature. We split our further analysis into three parts related to the motivation for running the pilot study: the design of the task, the crowdsourcing campaign, and other reasons.

4.2.1 Task design related reasons for conducting crowd pilot studies (n = 101, 59.1%). The design of crowdsourced tasks is a central component of crowdsourcing and includes factors affecting the performance of tasks and the quality of results, such as the task design, the usability of the user interface, and the clarity of the task instructions. Authors in our corpus often mentioned iterative experimentation aiming to improve the task design, but without providing details. Timmermans et al., for instance, mention that *"[p]ilots were run for optimizing the microtasks settings in terms of cost, amount of judgments, and task design"* (p. 2). Singh et al. [197] report that they *"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)"* (p. 4).

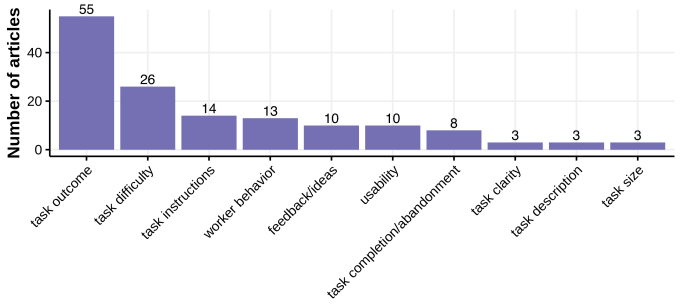


Fig. 6. Task design related reasons for conducting crowd pilot studies reported in the literature.

If details about the task design are mentioned in the article, we find it is most common for authors to report on the outcome of the crowd pilot study (see Figure 6). Under this category, we subsume any information reported by authors about the crowd pilot study's results, performance, accuracy, validity, and quality. Hu et al. [95], for instance, determined a similarity threshold *"[t]hrough a pilot study"* from the accuracy of results. Kim et al. [109] *"learned from pilot runs that longer video segments lead to lower annotation accuracy [...] and slower responses on Mechanical Turk"* (p. 4021). The task outcome was the most often reported factor related to task design ($n = 55$, 32.2%), used both in formative and summative crowd pilot studies.

Crowd pilot studies were also often conducted to assess the difficulty of the task during the formative design stage ($n = 26$, 15.2%). For instance, Swinger et al. [204] reported that *"in pilot experiments,"* workers were unfamiliar with items presented to them. The solution by the authors was to use custom qualifications. A qualification is "a set of questions [...] that the worker must answer to qualify and therefore work on the assignments [6]. Similarly, Kiesel et al. [108] determined *"[i]n pilot experiments"* that the *"task does not require expert workers, so we just required workers to have at least 100 previously approved HITs"* (p. 3051). Other authors carried out a more rigorous process to avoid systematic biases from seeping into the collected data [98]. For instance, Vogogias et al. [218] systematically experimented with different difficulty levels in a pilot study "*to identify the correct level of difficulty to avoid floor and ceiling effects*" (p. 5).

Another reason for conducting the crowd pilot study is the iterative design of task instructions ($n = 14$, 8.2%), including to improve a task's intelligibility (e.g., Wang et al. [223]) and clarity (e.g., Simoiu et al. [194]). However, many authors were not specific on how the task instructions were iterated on. For instance, Vertanen and Kristensson [216] simply reported that the "*exact instructions we gave workers evolved over the course of several pilot experiments*" (p. 18).

Analyzing worker behavior and preferences was another – although with 13 articles (7.6%) less common – reason for conducting crowd pilot studies. Qiu et al. [166], for instance, measured parameters of a worker model *"according to a pilot task on Figure Eight"* (p. 225). Roy et al. [180] reported on pilot experiments that *"revealed people had a strong preference to use manual control"* (p. 4). Acer et al. [1] investigated the workers' behavior during the crowd pilot study, including the response rate, and reported that workers already adopted the tasks as a habit during the crowd pilot study.

Some authors elicited open-ended feedback in the crowd pilot study and ideas for improving the study during the formative design phase ($n = 10$, 5.8%). For instance, Chen et al. [31] wrote that *"[t]o better inform the interface, we conducted a pilot study with 5 non-expert workers and asked them to rate the appearance of the marks after they finished their tasks"* (p. 9). Siangliulue et al. [192] reported that according to feedback from the pilot studies, their *"approach was intuitive and matched*

users' expectations well" (p. 613). Lykourentzou et al. [128] used the exit survey on CrowdFlower during the crowd pilot study to monitor the workers' satisfaction with the payment to *"ensure fair worker treatment"* (p. 265). However, CrowdFlower's exit survey was rarely used to specifically support the aims of crowd pilot studies.

Usability (n = 10, 5.8%) and the task abandonment [77, 78] or completion rate (n = 8, 4.7%) were also mentioned in articles. Wilson et al. [231], for instance, reported that the *"iterative design resulted in substantial usability improvements"* (p. 135) and Kandappu et al. [106] observed the task completion rate of 900 workers in *"a pilot study [with] over 900 workers in Sept 2015. From that study, [the authors] observed that 15% of the accepted tasks are not completed by the crowd workers"* (p. 907). The task description (n = 3, 1.8%) and the task clarity (n = 3, 1.8%) were only explicitly mentioned in a few articles, although these two items are implicitly part of the design of the task and its instructions. Another equally less frequently reported reason for conducting the crowd pilot study includes empirically determining the optimal size of a task (n = 3, 1.8%). For instance, Goncalves et al. [72] *"tested a variety of gameplay settings"* to determine the optimal number of items included in a task to *"not cause fatigue"* (p. 708).

4.2.2 Campaign design related reasons for conducting crowd pilot studies (n = 58, 33.9%). The design of a crowdsourcing campaign includes parameters necessary for launching the campaign, such as the number of tasks assigned to workers or batch sizes, the average task completion time, and the task pricing. These factors have been shown to influence task outcomes [32, 46, 47].

Related to the campaign design, crowd pilot studies are often conducted to empirically determine the price of the crowdsourced task (n = 35, 20.5%). As mentioned in the introduction, the task price is typically estimated from the workers' average task completion times in a crowd pilot study. We see evidence for this practice in our literature corpus. Typical ways of reporting this information include, for instance, Hara et al. [81] who reported that workers were *"paid $0.75 per HIT ($0.047–0.054 per labeling task); which was decided based on the task completion time in pilot studies (e.g., approximately $0.10 per minute)"* (p. 6). Another way was to provide information on a target hourly rate, such as Correll et al. [38] who mentioned that *"[b]ased on piloting, we paid participants $2 for participation, for a target rate of $8/hour"* (p. 5). Similarly, Han et al. [79] mention in a footnote that *"[b]ased on [a] pilot experiment"* the hourly pay was *"equivalent to US$13.5 per hour on average"* (p. 3). A slightly more extensive report was given by Roitero et al. [179] who *"performed several small pilots of the task, and after measuring the time and effort taken to successfully complete it, we set the HIT reward to $1.5. This was computed based on the expected time to complete it and targeting to pay at least the US federal minimum wage of $7.25 per hour"* (p. 441).

Besides the very common combination of task prices estimated from average task completion times, some authors also empirically determined other campaign-related parameters in crowd pilot studies. This includes determining a time limit for the task [5, 22, 72, 114, 119, 138, 181, 245] and determining the optimal sample size [30, 102, 115, 117, 126, 163, 171, 172] for the main study. Only a few crowd pilot studies involved qualifications for a task. Swinger et al. [204] used a *"qualification exam"* to identify well-performing workers based worker accuracy. Kiesel et al. [108] used results from a crowd pilot study to determine the minimum number of previously approved HITs for workers. Ramírez et al. [171] analyzed the geographic location of workers in the crowd pilot study to identify countries for the main study. Aigrain et al. [3] used a quiz on CrowdFlower to filter workers. Feyisetan and Simperl [60] ruled out the use of qualifying questions through crowd pilot studies to avoid an increase in attrition rate.

A common way of controlling the quality in crowdsourcing studies is gold questions (i.e., questions for which the answer is known) [40]. One way of developing and verifying gold standard questions would be through crowd pilot studies. However, only a few articles mentioned gold

standard questions in the context of crowd pilot studies. McDonnell et al. [136] found an *"inexplicable problem"* with the gold judgments and subsequently abandoned the use of the gold dataset in favor of another dataset. Chang et al. [30] used a crowd pilot study with seven MTurk workers to verify that the quality of work done is comparable to trained experts, concluding that *"judgments from 20 workers on Mturk can serve as the gold standard data set"* (p. 403). Nguyen et al. [146] mention that *"[s]mall pilot experiments were carried out while developing the design of the HIT"* which included *"gold labels"* (p. 323). This gold standard was taken from existing corpora and not verified. Last, Winther et al. [233] went one step further and presented experimentation on gold standards in one of their crowd pilot studies. The authors found that *"the gold standard proved to be too restrictive"* and *"gold tests and majority voting produced approximately the same acceptance results"* (p. 29).

4.2.3 Other reasons for conducting a crowd pilot study (n = 72, 42.1%). In the above, the majority of crowd pilot studies are conducted for formative reasons with the aim of iteratively designing a crowdsourced task in a rapid fashion with a small set of participants. We found that another formative reason for conducting a crowd pilot study is collecting data for the main study (20 articles, 11.7%). This category of articles can be split into articles that conducted the crowd pilot study with the sole purpose of collecting data (9 articles) and articles that conducted the crowd pilot study also for other purposes (11 articles). Amir et al. [13], for instance, collected *"pre-generated solutions represented common wrong solutions that were submitted by participants (as determined in a pilot study)"* (p. 5). Yu et al. [242] conducted a pilot study to collect constraints for a design task. In these articles, the collected data was then used in the main experiment or study of the article. For instance, Agarwal et al. [2] crowdsourced a dataset of tagged tweets in a crowd pilot study which was then used as input for a machine-based classifier, *"thereby making the classifier emotionally intelligent"* (p. 3).

Another reason for conducting the crowd pilot study was the iterative design of a study or an experiment. This purpose of a crowd pilot study was explicit in 11 articles (6.4%), although we believe this motive is implicitly present in many articles, such as the ones reporting on iteratively designing a task in the context of an experiment or the design of a system. As a single outstanding instance, d'Eon et al. [42] used a crowd pilot study to qualify and recruit participants for the main study.

As observed through our findings, crowd pilot studies are carried out for a broad range of compelling reasons – reasons that others who partake in carrying out crowdsourcing studies are very likely to face. It is, therefore, important to understand how and at what level of detail crowd pilot studies are reported in the literature.

4.3 RQ2: How are Crowd Pilot Studies Typically Reported?

In the previous section, we already provided some examples of how authors reported results of crowd pilot studies in the scholarly literature. In this section, we go in-depth and investigate how authors report crowd pilot studies. We analyze which terms and phrasing authors use to refer to pilot studies and how consistent they are in their wording (Section 4.3.1) as well as in which section of the article crowd pilot studies are being reported (Section 4.3.2).

4.3.1 How do authors refer to pilot studies? Different terms are being used in the scholarly literature to denote the provisional nature of pilot studies (see Figure 7). By far the most common term used by authors was 'pilot study' (90 articles, 52.6%), followed by 'pilot' (n = 44, 25.7%), 'pilot experiment' (n = 28, 16.4%) and 'pilot test' (n = 20, 11.7%). Most of the terms chosen by authors are nouns, although few authors also use 'pilot' as a verb (e.g., *"piloted"*), present participle (*"piloting"*), and gerund (*"pilot testing"*).

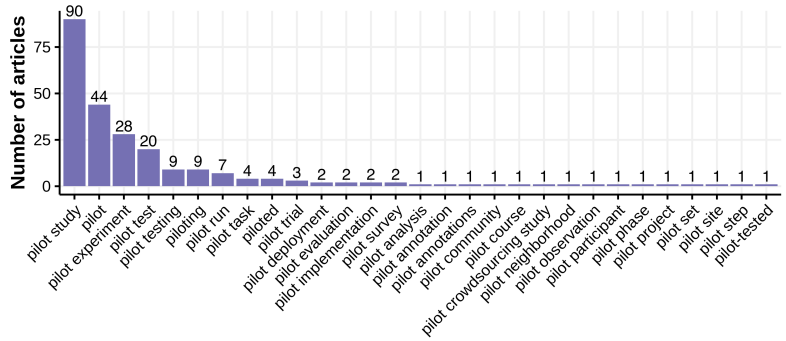


Fig. 7. Different terms used to refer to pilot studies in the articles. Note that some authors used multiple terms to refer to crowd pilot studies within their article.

Given that the vast majority of articles reported on pilot studies only in passing, a surprisingly low number of articles (n = 5, 2.9%) referred to the crowd pilot study as an *"informal"* study [7, 83, 178, 190, 213]. This informal study provided authors an *"informal sense"* of worker behavior in the context of an *"open-ended exploration"* [7] as well as support in the design of tasks to *"understand how different interfaces affected crowd performance"* [83].

We find that about two-thirds of the articles (n = 115, 67.3%) are internally consistent in how they refer to the pilot study within the article. In these articles, authors used only one single term to refer to the pilot study. In the other third of the articles, some authors used up to four different terms to refer to the pilot study (43 articles used two different terms, 12 articles used three terms, and one article used four terms). The most common combinations among the articles using multiple terms are 'pilot study' and 'pilot' (n=12), 'pilot test' and 'pilot study' (n=78), 'pilot study' and 'pilot experiment' (n=3), 'pilot experiment' and 'pilot' (n=3), and 'pilot study' and 'pilot run' (n=3).

In formative crowd pilot studies mentioned in passing, the term 'pilot study' is often used as a blanket statement to justify design decisions without providing details about the crowd pilot study. For this reason, we analyzed the phrasing used by authors of these crowd pilot studies in more detail (see Figure 8).

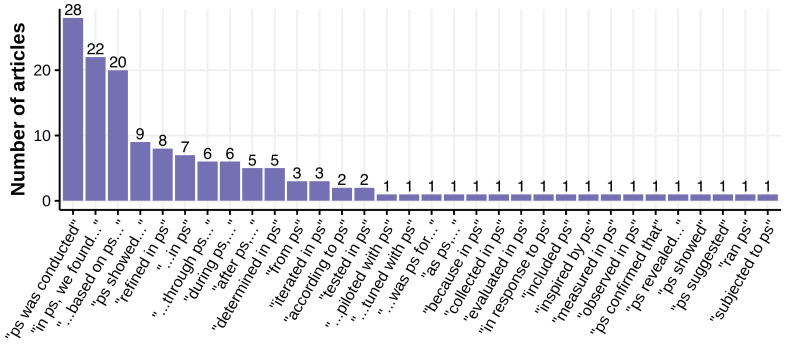


Fig. 8. Phrasing used to report on the results of pilot studies (abbreviated 'ps' in this figure) among articles that mention formative crowd pilot studies.

The most common way in which authors mention formative crowd pilot studies is by stating that a crowd pilot study was conducted, followed by selected details about the outcome of the pilot study. Nguyen et al. [146], for instance, mention that *"[s]mall pilot experiments were carried out while developing the design of the HIT,"* followed by details about the HIT, the assignment of the HIT to workers, and the price of the HIT. The phrase 'in a pilot study, we found' (or close derivations thereof) is also common. For instance, Hara et al. [81] report that *"[i]n early pilot studies, we found that users would get disoriented"* (p. 6). It is also very common for authors to derive design decisions 'based on a pilot study.' This phrasing was often used to refer to the estimation of the task price from average (or in some cases median) task completion times. For instance, Diakopoulos et al. [44] report that workers were offered $0.50 per rating *"based on the median time taken on a pilot task"* (p. 10). Similarly, Li et al. [125] estimated *"the time needed for each microtask based on pilot studies"* (p. 7). Some other phrases used among authors include that crowd pilot studies 'showed,' 'revealed,' or 'demonstrated' some specific results and that parameters were iteratively 'refined in pilot studies.'

Next, to draw further insights about pilot studies based on the context in which they are described, we explored sections of articles in which they are reported.

4.3.2 In which section are pilot studies reported? We analyzed in which section the authors report on the crowd pilot study, using a closed-coding approach. Our initial coding scheme reflected the standard structure of academic articles (i.e., Introduction, Related Work, Method, Results, Discussion, Conclusion, and Appendix). However, the codes were slightly modified after one iteration of coding to better accommodate differences in the methodological approaches used in the articles. The result of the coding is depicted in Figure 9.

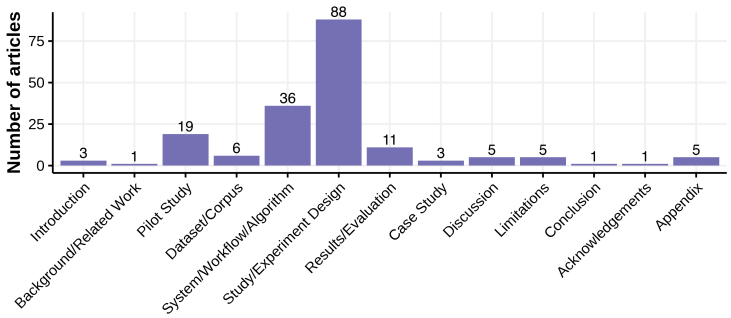


Fig. 9. Sections in which authors report the results of their pilot studies.

We find the majority of articles (n = 149, 87.1%) report on the crowd pilot studies in sections related to the methodology. These sections include the study design or experiment design (n = 88, 51.5%), the system design (or related sections; n = 36, 21.1%), dataset creation (n = 6, 3.5%), as well as separate sections dedicated to the pilot study, as found in about 10% of the articles (n = 19). The choice of section depends on the methodological approach taken in the article. For instance, articles that develop a novel system often mention the results of the crowd pilot study in the section on the system's design.

Besides the general trend described above, some outstanding instances of articles took a different approach to reporting on the crowd pilot study. Of the outstanding instances of articles that report on the crowd pilot study in the limitations section, we were expecting that the authors would discuss the weaknesses and limitations of the crowd pilot study. Instead, the crowd pilot study was, in some cases, used to validate the results of the main study. For instance, Sabou et al. [184] mention

in the limitations section that *"a set of pilot runs"* was executed *"to ensure the feasibility of the study design"* in an application domain to *"address external threats to validity"* (p. 171). Robertson et al. [177] discuss differences between their main study and an (independent) pilot study, reporting that the results *"were fully consistent with those from a pilot version of this study that we conducted in July 2019"* and that *"results are robust to pseudoreplication"* (p. 12).

Rekatsinas et al. [175] conducted a crowd pilot study in the introduction section to motivate their article. Rodríguez et al. [178] mentioned an *"independent"* crowd pilot study which was used to estimate the optimal price of the task and Robertson et al. [177] also mentioned an independent crowd pilot study. However, besides these three articles, crowd pilot studies were typically not conducted as independent studies, but as integral parts of the article. On the other hand, some authors used the extended space of the appendix to report on the crowd pilot study in detail. For instance, Fogliato et al. [61] report differences between the main experiment and the crowd pilot study in a separate appendix.

In the following section, we investigate in more detail what is known about the pilot studies from the reporting in the literature.

4.4 RQ3: What do Crowd Pilot Studies Report?

This section reports findings on which crowdsourcing platform is being used (Section 4.4.1), how many crowd pilot studies are being conducted in each article (Section 4.4.2), and which other key details are being reported about the crowd pilot study (Section 4.4.3).

4.4.1 Which crowdsourcing platform is being used? We analyzed which crowdsourcing platform is being used in crowd pilot studies. This information was often not explicitly stated and had to be inferred from context.

Amazon Mechanical Turk (MTurk) is by far the most common crowdsourcing platform (n = 102, 59.6%) in the literature corpus. Other platforms include, for instance, CrowdFlower/Figure Eight (now Appen) (n = 23, 13.5%), Prolific [117, 141, 150], Microworkers [220, 233], LabInTheWild [192], ZBJ [223], Clickworker [177], and the Yahoo! crowdsourcing platform [236], among others. In about 12% of the articles (n = 20, 11.7%), the crowd pilot study was conducted with other participant samples, such as students [104, 106, 182, 183], citizens [11, 58, 72, 100, 143, 201], and volunteers [25, 88, 89, 105, 134, 213, 214, 239]. In-house or custom crowdsourcing platforms were only reported in five studies (2.9%) [31, 93, 157, 164, 225]. These articles include a pilot study conducted on an *"indigenous crowdsourcing platform"* [157] and an article by authors from Google who *"ran numerous pilots to tune task hyper-parameters [...] sourced from contracted operators through an in-house crowdsourcing platform"* [225], a study with the prototype of a spatial crowdsourcing system [164], a study with a web-based platform for collecting ratings [93], and a study conducted with a crowdsourcing system designed for analyzing industrial tomographic images [31]. In ten articles (5.8%), neither the type of participant sample nor the crowdsourcing platform was mentioned.

4.4.2 How many pilot studies are being conducted in the article? As in the previous section, the information on how many pilot studies were conducted in the article had to, in many instances, be inferred from the wording used by the authors. In many cases, this wording was opaque and the exact number of crowd pilot studies could not be determined (see Figure 10). For instance, some authors mentioned conducting 'pilot studies' (e.g., Dimara et al. [48], Hara et al. [81]), 'pilots' (e.g., Feyisetan and Simperl [60], Luther et al. [127]), or 'pilot experiments' (e.g., Swinger et al. [204], Timmermans et al. [207]). From these terms, we can only infer that there was more than one crowd pilot study conducted. Due to the use of non-descriptive terms such as 'pilot testing' [15, 52, 116] and 'piloting' [38, 76, 84], the number of crowd pilot studies could not be determined in six articles.

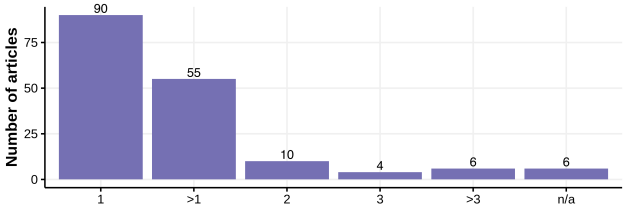


Fig. 10. Number of pilot studies conducted in the article.

About half of the articles (n = 90, 52.6%) report conducting one pilot study (see Figure 10). A third of the articles (n = 55, 32.2%) report conducting more than one pilot study. A high number of pilot studies within an article was rare (n = 6, 3.5%). Simoiu et al. [194], for instance, conducted *"six small pilot tests"* to *"ensure that the questions were clearly phrased, and of appropriate difficulty"* (p. 174). The highest number of pilot studies was reported by Inel et al. [102] who conducted extensive experimentation in eight crowd pilot studies.

4.4.3 Which key attributes are typically reported about crowd pilot studies? We analyzed what authors choose to report about crowd pilot studies. We specifically looked at three key attributes of crowd pilot studies: the number of workers participating in the crowd pilot study, the number of tasks (or any other information given in the article that would allow to determine the number of assignments to workers), and the rewards to workers.

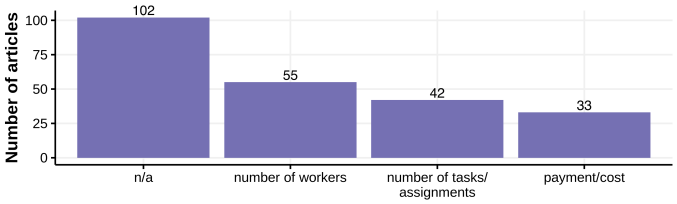


Fig. 11. Key statistics reported about crowd pilot studies.

About 60% of the articles did not provide any information on the three key statistics (n = 102, 59.6%; see Figure 11). We find authors who report one key statistic also often report other key statistics about the crowd pilot study. Twenty-four articles (14.0%) report on all three key statistics. These articles often dedicated a full section or the entire article to the crowd pilot study. About 40% of the articles (n = 68, 39.8%) report at least one of the three key statistics.

About a third of the articles mentioned the number of workers participating in the crowd pilot study (55 articles, 32.2%). In these articles, the number of workers ranged from three (e.g., Oppenlaender and Hosio [150]) to over 2,000 [88]. Some authors were imprecise about the number of participating workers, such as Ambrosino et al. [11] who reported *"almost 40"* participants (p. 5). Among the 50 articles in which we could identify or calculate the exact number of participants in the crowd pilot study, the average number of participants was 111.7 (SD = 169.9).

The number of tasks or assignments was mentioned in 42 articles (24.6%). This information was more difficult to analyze because some authors mentioned the number of assignments, others the number of tasks. Further complicating the analysis was the fundamental difference between the studies – some of which collected tags or annotations, others conducted situated crowdsourcing studies. In the articles in which we could infer the number of tasks, the number ranged from 10

tasks [224] to 55,000 [225] (assuming one rating per task). Because of the difficulty of determining the exact number, we do not report the mean and standard deviation of the number of tasks in these articles.

Monetary rewards were reported in 33 articles (19.3%) but often without explicitly mentioning MTurk fees. Of the 34 articles, eight involved unpaid volunteers [1, 11, 33, 88, 89, 105, 214, 239], two articles simply stated that participants were paid minimum wage [100, 139], and one article involved a raffle for an iPad [134]. Three articles reported the monetary rewards in the crowd pilot study as an average hourly [4, 132] or per minute wage [81]. Among the remaining 20 articles, the monetary pay for participating in the crowd pilot study ranged from $0.01 [149] to $10 [61] per task. Welbourne et al. [224] paid *"a maximum of $30 US"* (p. 3), but in this case, workers were recruited on Elance and ODesk (now Upwork) and the actual bids may have been lower.

Only a few articles (n = 3, 1.8%) reported experimenting with different price points. Kim et al. [111] experimented with paying workers in increments *"from $0.00 to $8.00 [...] up to the federal minimum wage in the United States ($7.25/hour as of April 2, 2011)"* (p. 4). Similarly, Borish and Lok [20] posted tasks *"in $.05 increments, starting at $.15 and going up through $.50"* (p. 10). Rodríguez et al. [178] investigated the robustness of results against varying levels of reward (from US$0.05 to US$0.10). Bonuses to workers in the context of crowd pilot studies, in general, were only mentioned in a few articles. Vonikakis et al. [220], for instance, mention experimenting with different incentive schemes that involve a bonus to well-performing workers and Huang and Fu [96] conducted a crowd pilot study to determine a bonus based on the workers' average accuracy.

4.5 Differences in how Crowd Pilot Studies are reported

4.5.1 Are there differences in crowd pilot study reporting between research communities? As mentioned in Section 4.1 and depicted in Figure 3, the bulk of crowd pilot studies were reported in three research communities: CHI, CSCW, and HCOMP. The former two are closely related human-centered venues and researchers often submit to both venues. The latter is a venue specialized in advancing the state of the art of human computation and crowdsourcing, but also in applying it practically. We explored differences between the three venues in how crowd pilot studies are being reported in human-centered conferences (CHI and CSCW) as compared to crowdsourcing research (HCOMP). Our initial hypothesis is that there will be a difference between the communities since best practices will likely emerge from within the community of practice [227] in the crowdsourcing-focused domain at HCOMP.

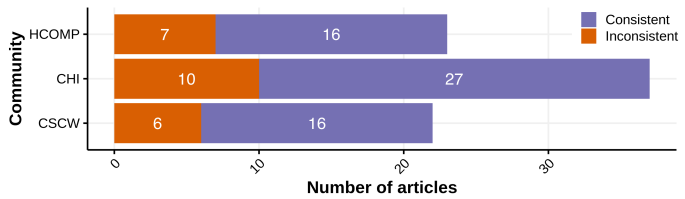


Fig. 12. Consistency of wording within articles in different research communities.

We first investigate how consistent crowd pilot studies are being reported in the three communities. We define 'consistency' in the reporting of crowd pilot studies as the uniformity in the use of terminologies, methodologies, and presentation of results across the surveyed articles. Specifically, an article is deemed 'internally consistent' if its descriptions, methodologies, and terminologies related to pilot studies remain coherent and unambiguous throughout the article's text. In contrast, articles with varied references to pilot studies are deemed 'inconsistent.' Looking at the ratio of

internally consistent articles in the three venues (cf. Figure 12), we find CHI and CSCW are about comparable in consistency (73.0% and 72.7%, respectively). The ratio of internally consistent articles published at HCOMP is slightly lower (69.6%), but this difference is not significant (pairwise t-tests, each with $p > 0.05$). We find there is no agreement between articles in the three venues on which term is used to denote pilot studies, even among the articles that use only one term. A wide range of different terms are being used in the three communities, with 'pilot study' being the most common, followed by 'pilot' and 'pilot experiment' (see Figure 13).

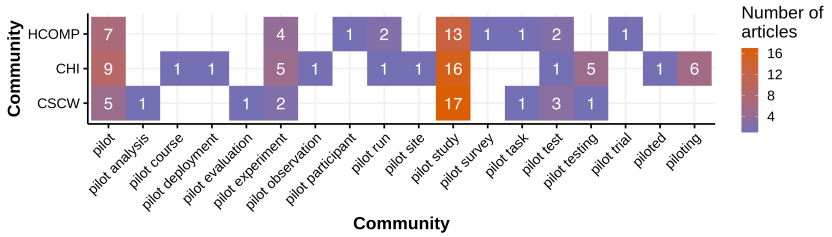


Fig. 13. Wording used to refer to crowd pilot studies within articles in different research communities.

Looking at the placement of pilot studies within articles (see Figure 14), we find that crowd pilot studies are often reported in sections relating to the methodology (e.g., study design or experiment design). There is no significant difference between the three venues when it comes to the section in which the pilot study is being reported, $\chi^2(22, N = 90) = 19.37$, $p = 0.6224$. Four HCOMP articles (17.4% of the articles in this venue) reported the pilot study in a separate section, which highlights the importance of pilot studies in the field of crowdsourcing, as compared to articles in CHI (5.4%) and CSCW (9.1%).

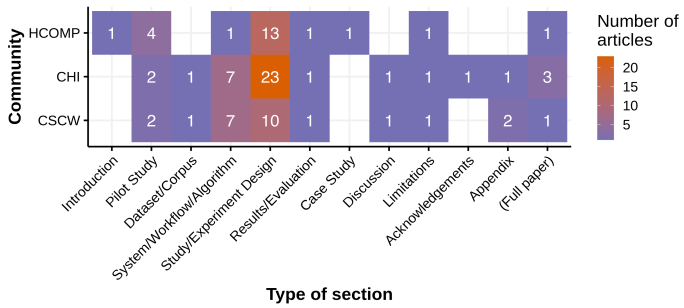


Fig. 14. Type of section in which pilot studies are being reported in different research communities.

The number of pilot studies conducted within an article is similar in all three venues, with one single crowd pilot study being the most common (see Figure 15). At CHI and HCOMP, it is also common for articles to report more than one crowd pilot study. The difference between the three venues is, however, not statistically significant ($\chi^2(10, N = 82) = 8.3317$, $p = 0.5965$).

When it comes to reporting key statistics about the crowd pilot studies, we find that in all three venues, the most common way of reporting a crowd pilot study is in passing without providing any details about the number of participating workers, the number of tasks, or the exact monetary rewards provided to workers. A large percentage of articles (between 65% to over 95% of the articles

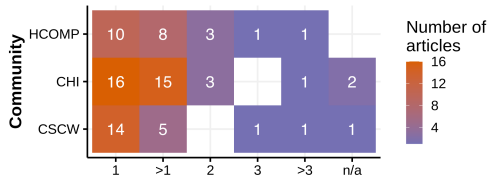


Fig. 15. Number of crowd pilot studies conducted in the articles in different research communities.

reporting crowd pilot studies in each conference venue) do not report these three key statistics (see Figure 16).

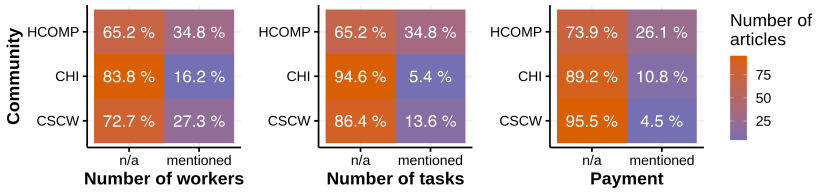


Fig. 16. Key statistics reported in different research communities.

There are, however, differences between HCOMP and CHI/CSCW when it comes to reporting details about the crowd pilot study. Authors in HCOMP are more likely to report key statistics about the crowdsourcing campaign as compared to CHI and CSCW (see Figure 16). HCOMP articles report the number of workers almost twice as often as CHI articles. Similarly, the number of tasks assigned to workers is more likely to be reported in HCOMP articles (34.8%) as compared to CHI (5.4%) and CSCW articles (13.6%). This difference is even more profound when it comes to reporting payments to workers. HCOMP articles reported payments to workers participating in crowd pilot studies in about a quarter of the HCOMP articles as compared to CHI (8.1% of the CHI articles) and CSCW (4.5% of the CSCW articles). One possible reason for this is that authors at HCOMP may be more sensitive to issues surrounding crowdwork due to the fairness of crowd work being a long-standing research topic in human computation and crowdsourcing. The differences between the three conference venues were, however, only statistically significant for the number of tasks, $\chi^2(2, N = 82) = 9.2864$, $p = 0.0096$.

In summary, we found no major differences between the HCOMP and CHI/CSCW communities in terms of the number of crowd pilot studies being conducted and the wording used within articles. The consistency of wording within articles was comparable between the three venues, with many different terms being used to denote the crowd pilot study (some more common than others). Authors in all three venues prefer to report the results of crowd pilot studies in a section relating to methods, with the study design section being the top choice of authors.

4.5.2 How do crowd pilot studies differ between academia and industry? In our survey study, we asked participants if they had industry experience or worked closely with the industry. In response to this, four participants (36.36%) responded they had industrial experience, while seven did not have experience (63.64%). Two out of four (50%) who had industrial experience did research in collaboration with an industrial partner, while one (25%) indicated that he is planning to conduct a pilot study with industry. When asked what differences the participants found between the academic and industrial crowd pilot studies, one indicated that *"industrial pilot studies cost more in*

salaries than academics" while another respondent stated that *"the pilot study was to create a digital asset for the company."*

We also asked about potential differences between crowdsourcing crowd pilot studies on in-house/internal platforms and other commercial platforms (e.g., MTurk). Participants came up with a variety of feedback. One participant indicated that *"the internal CS platform is more accurate than other commercial platforms"* and that *"internal systems are easier to use as managers would have no issue with them, external ones are more tricky due to privacy and security issues."*

4.6 RQ4: What makes a "good" crowd pilot study?

In response to this question, researchers in our survey (cf. Section 3.5) identified several qualities that define a good crowd pilot study. These qualities relate to the objectives for running a pilot study and may stand in tension with each other, as evident in the following sub-sections.

4.6.1 Mimicking the main experiment. Two researchers stated that a successful crowd pilot study is *"as similar to a formal experiment"* and *"one that only differs from the complete study by sample size."* This finding is also consistent with the recommendations of other researchers that a (crowd) pilot study should mirror all the processes of the main research and adhere to the identical protocol, including inclusion and exclusion criteria for participants, measuring tools, and training resources [101].

4.6.2 Exploration and experimentation. Others stated that the paramount quality of a good crowd pilot study is its exploratory nature, which provides them with different directions for their primary research questions or hypotheses. For instance, one noted that *"[a good pilot is the] one which gives a clear direction of which RQs/directions would be more promising to pursue in an actual study"* and one that *"should give researchers some useful inputs about their hypothesis or prototype."* This finding shows that researchers use pilot studies in the conception phase of their projects when they need supporting evidence to develop a research question and research plan [215].

4.6.3 Validating the feasibility of a study. Another quality mentioned by researchers is the ability of a crowd pilot study to assess the feasibility of an approach. This feasibility could also refer to the technical feasibility where researchers test rigorously through several trials that *"all functions are working, and log [that the] system can repeat what users have done."* Others viewed a good crowd pilot study as one that *"allows to validate the functioning of your task and it allows you to gather a sample of the final expected data"* because *"it costs high to redo a formal exp."* Other respondents defined assessing the task-related information as the main criteria that a crowd pilot study should incorporate, such as the *"task length, number of workers required, task complexity and task design"*. For instance, one researcher summarized this in the following words *"I think we need to always do a pilot study, to figure out if both the design and the technical problems are solved."*

4.6.4 Accurate estimation of campaign parameters. Another critical dimension that defines a good crowd pilot study is its ability to estimate sample size and power calculations. For instance, researchers reported that a good crowd pilot study could help to *"[correct] the sample size errors"* and *"help to calibrate the power calculation."* Similarly, one participant reported that a crowd pilot study should help to estimate the *"statistical data involved, e.g., mean/median/sd."* Thus, sample size and power calculations are another essential quality of a good crowd pilot study. These estimates are even crucial in crowdsourcing research when researchers need to hire hundreds of participants, thanks to the affordability and affordances of crowdsourcing platforms. However, the estimation of the sample size required for the main trial needs to be performed cautiously since a crowd pilot study only provides the estimated value of standardized effect size [123]. Moreover, one may also need to account for participant exclusion in such cases.

4.7 RQ5: What factors promote or obstruct the reporting of crowd pilot studies?

We approached this question by posing both closed- and open-ended questions. In a closed-ended question, we asked participants to provide potential reasons for characteristics that either encourage or restrict the reporting of crowd pilot studies. These reasons include page limit restrictions, funding availability, article types, and reviewer preferences (see Figure 17). Most participants indicated that page limitations were the most critical factor ($n = 6$), followed by availability of funding ($n = 5$). Only two respondents answered that the article type was the most essential factor, and one responded that he does crowd pilot studies because reviewers want them.

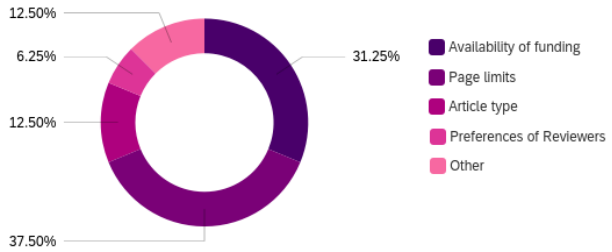


Fig. 17. Factors that promote or inhibit the reporting of crowd pilot studies.

4.7.1 Page limit restrictions. Regarding the *page limitations*, respondents felt that *"it limits the content length of the report"* and they would prefer actual experiments over crowd pilot studies *"because the results of the formal experiment are more interesting [than pilot studies]."* Another respondent believed that *"conf[erence] papers normally require a tight page limit which would squeeze the space for rather important content (e.g., results)."* Another respondent who worked in the area of crowd-powered applications responded that *"justifying some design decisions of a big crowd-powered system is probably not very critical. We will likely cut these justifications when we don't have space."* We also asked "if there is no page limit, will this make you more likely to report crowd pilot studies in your articles?"

Three out of eleven respondents believed that they would 'very likely' report pilot studies, two affirmed that they would undoubtedly report pilot studies, while one was neutral about this opinion. We also noted that no respondents selected 'unlikely' or 'highly unlikely', which shows that page limitation is a rather decisive factor. This trend is slowly shifting. For instance, conferences have been slowly transitioning to a revise and resubmit cycle along with more flexible manuscript lengths, which removes page restrictions and permits authors to expand the methodology and design sections, enabling the reporting of crowd pilot studies.

4.7.2 Availability of funding. Participants also felt that the *availability of funding* may encourage the scalability of an experiment and extensive testing of a product before it could be made available. A participant, for example, thought that *"funding is crucial in scaling the experiment, and the funding sources tend to encourage folks to include pilot results in the grant application."* Another participant who believed crowd pilot studies were important for iteration and testing stated this as follows:

> *A good project needs to be developed and tested for a long time before it can be released. Therefore, the initial investment is relatively large and stable sources of funds are needed.*

4.7.3 Article types. The *article type* also played a significant role in inhibiting the reporting of a pilot study. For example, one respondent responded as:

> *To report this we need to write a paragraph or at least some sentences describing it. If we need to cut down something due to exceeding the page limit, this would be an option. For conferences or journals, because the target audiences have different focuses. For some system-focused venues, we may shorten the description of data collection and experiment design by skipping this.*

4.7.4 Reviewer preferences. One respondent was of the opinion that *"reviewers perceive pilot studies as less impactful and therefore would not be willing to accept them for publication."*

4.8 RQ6: How can crowd pilot studies be facilitated with platform-specific features?

In our literature review, we found a handful of mentions of platform-specific technical features that were used for conducting and monitoring crowd pilot studies. In the remainder of this section, we reflect on the design of such features, based on our literature review, the results from our survey study, and our experience with different crowdsourcing platforms.

4.8.1 Exit surveys for facilitating crowd pilot studies. One possible feature for supporting crowd pilot studies is the 'exit survey.' An exit survey is a short questionnaire that workers fill out after completing tasks. Exit surveys were used by a few authors to measure or monitor workers' satisfaction with the payment during crowd pilot studies. For instance, Lykourentzou et al. [128] used the aggregated results of the exit survey (provided by CrowdFlower/Figure Eight) to validate and justify the choice of payment. The authors reported the results of the exit survey indicated *"that the chosen payment was considered acceptable by the workers"* and that *"the selected compensation was appropriate for the specific study setting"* (p. 265). Another use for an exit survey is collecting demographics, which is especially important on microtask platforms where tasks are typically too short to collect demographics. For instance, Wilson et al. [230] used a custom exit survey on Amazon Mechanical Turk to collect demographic information.

4.8.2 Reward calculation. Besides the above feature, participants in our survey mentioned a number of other features that could facilitate crowd pilot studies. Most often mentioned was a "reward calculator" which could calculate rewards based on estimated completion times. As found in our literature review, the calculation of rewards from average task completion times is one of the most common reasons for conducting crowd pilot studies. Prolific[3], a crowdsourcing platform for academic studies, already offers a recommendation for the price of a task, based on the estimated time. This is, however, only an incomplete solution because it is difficult for a requester to estimate the completion time — often, the very reason for conducting the crowd pilot study is finding this estimate. However, crowdsourcing platforms are host to many different types of tasks. Given the large variety and amount of tasks on the crowdsourcing platforms, it would be possible for platform operators to collect information on tasks and to devise machine learning-based platform features to support the estimation of task completion times and task rewards, based on empirical data collected on the crowdsourcing platform.

4.8.3 Better support for running qualification studies. Screening criteria were mentioned often by the survey participants. Crowdsourcing platforms differ in their capabilities to support screening and qualification studies. Custom qualifications can be created, but this requires running a study, collecting results, and then uploading a comma-separated values (CSV) file to Amazon Mechanical Turk to assign the custom qualifications to workers. Only then can the qualification be selected in future studies. Amazon Mechanical Turk offers only a limited set of qualification criteria for screening participants. Although Prolific offers a broader array of pre-defined qualification criteria,

[3] https://www.prolific.co

setting up custom qualification studies can be just as complex as in MTurk (when implemented via a survey study) or restricted to Prolific's in-built multiple-choice survey options. Better user interfaces for running qualification studies and setting (or deleting) qualifications are needed. The survey participants further perceived a need for an MTurk feature to extend running studies with more participants. On Amazon Mechanical Turk, no changes can be made to a running crowdsourcing campaign. This leads to disparate sets of survey results which then need to be manually integrated by the researcher. Last, the participants in our survey mentioned wanting better tools to communicate with workers, such as a chat or e-mail service. This speaks to the survey participants' need for a less dehuminizing communication with crowd workers [17]. Features to communicate with the crowd would allow requesters to better monitor ongoing studies and grow a base of trusting participants [191].

Clearly, there is an opportunity for the design of dedicated features on crowdsourcing platforms that could better support and facilitate running crowd pilot studies. Features, such as the above, could support best practices in crowdsourcing. We reflect on the importance of best practices and make recommendations for reporting crowd pilot studies in the following section.

5 MOVING FORWARD: BEST PRACTICES FOR REPORTING CROWD PILOT STUDIES

Crowd pilot studies are a common and required method in crowdsourcing research due to the empirical nature of the crowdsourcing paradigm. Unsurprisingly, many authors report having conducted crowd pilot studies in the scholarly literature. Yet, no scientific study spanning crowdsourcing has investigated this topic in depth. Our work aimed to fill this gap. In this section, we reflect on our findings and the current state of best practices on reporting crowd pilot studies.

5.1 Readdressing current practices for reporting crowd pilot studies in crowdsourcing research

Crowd pilot studies connect to two strains of research in the field of crowdsourcing that touch upon the very nature of crowdsourcing: fair and responsible crowdsourcing [193, 229] as well as reproducibility in empirical computer science [36]. These two issues have long been debated in the scholarly literature.

5.1.1 Best practices for fair and responsible crowdsourcing research. Crowd pilot studies account for a significant amount of work that is unaccounted for to a large extent in the scholarly literature. Since the majority of authors use opaque language masking the extent of studies, little is known about the real extent of crowd pilot studies. Further, due to the empirical nature of crowdsourcing, it is likely that crowd pilot studies often underpay participants. Estimating the rewards for crowdsourced tasks is hard and one way of addressing this shortcoming is to raise the basic level of payment or assign bonuses to workers in a post-hoc manner to fairly compensate participants in crowd pilot studies [16, 85, 165]. However, it is likely that workers in crowd pilot studies are substantially and potentially systematically underpaid [45, 80, 133]. Recent work has unearthed different forms of invisible labor that crowd workers put in as they strive to earn their livelihood in various crowdsourcing marketplaces [75, 208]. Prior work has also revealed how crowd workers are often subject to unfair rejections following qualification studies [55, 64, 137]. It is likely that such practices transcend to ill-reported crowd pilot studies. Interestingly, extremely few articles in our literature review reported that bonuses were given to workers in or after crowd pilot studies. Based on results from our literature review and survey, we find it is more typical – though still not common – to pay bonuses to participants for performing well in the main study.

5.1.2 Reproducibility in crowdsourcing research. The crowdsourcing paradigm has many known limitations. For instance, results obtained from crowdsourcing studies may be difficult to reproduce

due to the anonymity of the workforce. The opaque reporting of crowd pilot studies, as evidenced in our literature review, adds one additional layer to the issue of reproducibility. The strong prevalence of reporting on study results in passing accentuates and entrenches bias in research and helps bad practices to endure. For instance, some authors used the crowd pilot study to substantiate claims. Crowd pilot studies are sometimes used as a magic linguistic device to materialize results that are later used as input for the main study of the article. In this sense, much of the reporting on crowd pilot studies uses 'hedging' language, a "rhetorical means of gaining acceptance of claims" [99].

We argue that researchers should do their due diligence on the research claims made and report transparently on the aims and results of crowd pilot studies. Readers need to know the details about crowd pilot studies. For instance, a reader needs to know the number of participants in a crowd pilot study "to know that the study was big enough to justify the claims made" [129]. Authors need to realize that opaque reporting on crowd pilot studies — especially if it is done as a summative evaluation (as was the case in a few articles we reviewed) — weakens the claims of the authors' research. Insufficient details can impede the progress of science in general. The current state of reporting on crowd pilot studies exacerbates and affirms this widespread practice of opaque reporting. More transparency on reporting crowd pilot studies is needed to nudge the current state of reporting in the field of crowdsourcing toward a code of practiced ethics that values transparent reporting of crowd pilot studies. However, crowdsourcing is still a relatively young field where good practices need building.

5.1.3 Treating the crowd workforce fairly. One of the pivotal realizations that has emerged through research and practice within the crowdsourcing community over the last few years is the need to treat crowd workers fairly and with dignity – whether it is in terms of the hourly wages paid or with respect to communication with workers [103, 193, 229]. It is now commonplace in most HCI communities to declare the hourly wage that participants are paid in reported main studies. By raising the bar for what is expected in the reporting of crowd pilot studies in scholarly literature, we can hope to instill the otherwise (potentially) dormant desire to pay workers fairly within crowd pilot studies. This will increase the overall accountability of researchers and other requesters, and help bridge a gap in the invisible labor prevalent in crowdsourcing marketplaces [75].

Beyond crowd pilot studies, the broader domain of data annotation stands as another significant area where fairness in the treatment and payment of crowd workers is paramount. Data annotators play a foundational role in shaping machine learning models and other AI systems by providing high-quality labeled data. Yet, there have been growing concerns about the remuneration, working conditions, and well-being of these data annotators, especially given the labor-intensive nature of their tasks [75, 103, 153, 202, 221, 244]. Inadequate compensation for data annotators not only poses ethical dilemmas but also risks compromising the quality of annotated datasets. By ensuring fair wages and conditions for these workers, we not only uphold the principles of ethical research and practice but also contribute to the production of more reliable and robust AI systems. It is crucial for the HCI and broader AI communities to address this concern head-on, establishing standards that reflect the true value of this indispensable labor.

Creating a widespread change in how crowd pilot studies are reported will require widespread and collective action. This is especially required since well-meaning authors are often subject to a trade-off while reporting crowd pilot studies, as discussed in the following section.

5.2 The trade-offs around reporting crowd pilot studies

Researchers are influenced – consciously or unconsciously – in how they report crowd pilot studies. In this section, we discuss confounding factors and biases that may affect the reporting of crowd pilot studies in academia and industry.

5.2.1 Page limitations. Traditionally, the page limit at conference venues such as CHI was 10 pages (in two-column format, not including references). Some venues continue to uphold such strict restrictions on the number of pages in articles. Authors, therefore, face a difficult trade-off between reporting on pilot studies in detail and reporting on the main study. Our literature review is evidence for this trade-off, with many crowd pilot studies being reported briefly and casually. In recent years, however, many venues in HCI (e.g., CHI and CSCW) have relaxed the page limitations which could, in theory, encourage authors to dedicate more space to crowd pilot studies. However, a number of other biases and trade-offs may still make authors consider otherwise, such as the academic publishing model.

5.2.2 Academic publishing model. Publication bias is "the failure to publish the results of a study 'on the basis of the direction or strength of the study findings'" [43]. Due to the publication bias in academia, authors may feel the need to report positive findings in order to get published. A formative crowd pilot study, in particular, may – in the mind of authors and/or reviewers – not add to this goal. Further, if the authors feel the pilot study does not contribute towards the acceptance of the article, the authors may decide to omit the pilot study or shorten the reporting. Another concern that authors may have is that if a formative pilot study is given too much space in the article, reviewers may view the article as a work-in-progress and recommend it for acceptance in a lesser capacity (e.g., as a poster). Therefore, researchers may decide not to report pilot studies because of a perceived need to produce writing that pleases reviewers. However, in the past years, some conference venues have opened up to the possibility of submitting works following the principles of open science. These venues encourage the submission of replications and articles with null or negative results which have been traditionally hard to publish. While these advances, so far, have been limited to special tracks – such as the Open Science track at the Conference on Intelligent User Interfaces (IUI) 2023 – they could lead to a slow systemic change toward an academic system in which reporting on pilot studies is being encouraged. In this regard, some referees may consider it favorable if crowd pilot studies are being reported transparently and in detail.

5.2.3 Funding. The availability of funding is another important factor that may influence the authors' decision to conduct or report pilot studies. For instance, the availability of funding may affect the extent of crowd pilot studies. If researchers are short on budget, they may skip or reduce the number of formative pilot studies. However, even if funding is available, authors may decide not to run crowd pilot studies to not "waste" the funding organization's money on formative studies with an anonymous crowd. For similar reasons, authors may decide not to report on crowd pilot studies. On the other hand, iterative experimentation is important, especially in the field of Human-Computer Interaction (HCI) where the emphasis is placed on iterative and participatory design to ensure optimal outcomes in a variety of contexts [19]. The very process of design requires iteration and formative experimentation to arrive at an acceptable solution.

5.2.4 Corporate or organizational culture. If not the academic system or external funding, then the internal culture of an authors' organization could discourage conducting pilot studies. For instance, universities in Finland recently started following stricter directives from the Tax Administration in a move towards a system where any rewards to participants – whether it is cinema vouchers, gift cards, or monetary payments – need to be declared to the tax office, regardless of the monetary value of the rewards. Because monetary compensations to participants are subject to withholding tax, this causes overhead to the university administration. Even more concerning is that researchers are asked to collect private information from their study participants (name, address, and social security number) if participants are to be rewarded. Therefore, researchers in Finland are strongly

discouraged from using paid participant samples in their research. This development is deeply worrisome as it discourages researchers in Finland from conducting ethical and fair science.

5.3 Guidelines for Reporting Pilot Studies

Our analysis of the HCI and crowdsourcing literature allows us to provide recommendations for reporting crowd pilot studies. In this section, we revisit the research questions RQ1–RQ3 and connect the findings of our literature review to recommendations reporting crowd pilot studies.

5.3.1 Why are crowd pilot studies typically conducted? (RQ1).

Be transparent on the reasons for conducting crowd pilot studies. Most articles in our literature review conducted crowd pilot studies for formative reasons. However, in articles that report crowd pilot studies in passing, it was sometimes not explicitly stated why a pilot study was conducted. Clearly motivating the crowd pilot study will provide clarity to the writing and increase the readers' understanding of why a pilot study was needed.

Inform crowd workers that they are participating in a crowd pilot study. Crowd workers are subject to a wide range of tasks posted on crowdsourcing platforms. Some tasks are more and some less lucrative for the workers. Crowd pilot studies may fall into the latter category, especially if the task price is not estimated accurately. While some workers are not motivated by extrinsic factors and may enjoy participating in crowd pilot studies [151], other workers may want to avoid them. Workers should be informed that they are participating in a small-scale study (that may potentially be under-priced).

5.3.2 How are crowd pilot studies typically reported? (RQ2).

Use consistent wording. Academic writing requires precise language. Using different terms within an article to refer to crowd pilot studies may add confusion to an uninitiated reader. We recommend using the term 'pilot study' to denote crowd pilot studies. This term was the most common term used in the literature (cf. Figure 7).

Report crowd pilot study findings in one section. We found many authors scatter findings from their pilot studies throughout their papers. To improve the clarity of pilot study reporting and to better showcase the results of pilot studies, we recommend bundling the reporting of crowd pilot studies in a single section of the article. This would improve both the understanding of the reader of the extent of pilot studies conducted and the reproducibility of the pilot study. Our analysis of the literature indicates that it is most common to report the results of crowd pilot studies in design-related sections (cf. Figure 9).

5.3.3 What do crowd pilot studies report? (RQ3).

Report the number and extent of crowd pilot studies. A considerable amount of articles in our literature corpus (62 articles, about 36%) did not provide information on the exact number of crowd pilot studies being conducted. In other cases, the number of studies being conducted had to be calculated from information scattered in the article. Authors should clearly state the number of pilot studies and their respective extent.

Clearly identify the participants. There should be no room for interpretation when it comes to who participated in the crowd pilot study. In particular, researchers should identify whether crowd workers participated in the pilot study or whether the pilot study was conducted with a different participant sample (e.g., students, experts, or internal participants). This also includes information

on the crowdsourcing platform used in the pilot study, if it cannot be reliably inferred from the context in the article. Internal pilot studies should be clearly denoted as such.

Report the key attributes of each crowd pilot study. If page restrictions limit authors from reporting in-depth on crowd pilot studies, we recommend including at least the following key information when reporting crowd pilot studies:

- number of participating crowd workers,
- number of tasks (or assignments to workers),
- payment per task,
- participation constraints enforced (including platform settings), and
- the type of crowd or crowdsourcing platform.

The latter could be omitted if it is clear from the context that only one crowdsourcing platform was used throughout the article. The selection of rewards for workers in the crowd pilot study should be justified. If there are major discrepancies between the rewards paid in the crowd pilot study and the main study, it should be explained how these discrepancies came into existence and what measures were taken to remedy the discrepancies.

Report a minimum set of information. Inspired by scientific reporting guidelines, such as guidelines by the American Psychological Association [12], and based on the above recommendations while also considering the trade-offs discussed in Section 5.2, we propose a condensed format for reporting formative crowd pilot studies:

> *… pilot study (MTurk; N=12; 1000 HITs; US$4.5 per HIT) …*

In combination with any of the preferred methods of reporting on pilot studies, such as *"in a pilot study (…), we found…"* or *"… based on a pilot study (…)"* (cf. Figure 8), this condensed format provides key statistics about the formative crowd pilot study (i.e., the crowdsourcing platform, number of participants, number of HITs, currency, and price per HIT) without taking up an undue amount of space in the article.

We hope that authors will adopt at least this condensed way of reporting crowd pilot studies to increase the transparency and reproducibility of their research. In the same vein, we hope that reviewers in conferences and journals publishing research with crowd pilot studies will, in the future, place increased emphasis on seeing transparent reporting of crowd pilot studies.

5.4 Practical Suggestions for Supporting Better Crowd Pilot Study Reporting

Creating a centralized repository for crowd pilot studies in crowdsourcing research could be a possible way to enhance the reporting and transparency of such investigations. The repository would serve as a dedicated platform for researchers to submit their crowd pilot studies following a standardized report format, as suggested in the previous section. This format should encapsulate critical elements including research questions, methods, results, and challenges encountered, thereby enabling a comprehensive understanding of the study without exceeding paper page limits. Furthermore, the structured reporting style within the repository should include specifics such as sample size, study duration, data cleaning methods, and outcomes. This standardized approach, coupled with a requirement for authors to detail challenges and potential improvements, could not only foster transparency but also provide insights for researchers undertaking similar studies.

To further encourage crowd pilot study reporting, a system of incentives could be introduced for authors who make use of the repository, ranging from formal acknowledgments within the academic community, and citation opportunities, to reduced publication fees in affiliated journals.

At the same time, scientific journals and conferences could set forth clear guidelines encouraging the citation of crowd pilot studies from the repository in their submissions. Creating this repository and encouraging its use would help cultivate a research culture that values transparently reporting crowd pilot studies, ultimately leading to more accurate, rigorous, and replicable crowdsourcing research.

5.5 Limitations

In our literature review, we made pragmatic choices to limit the set of literature to what we believe is a representative coverage of the literature, as mentioned in Section 3.2. Our screened corpus included all articles from HCOMP, the premier venue for crowdsourcing research. The ACM Digital Library contains articles from human-centered journals and conferences, such as CHI and CSCW. We do not claim that the selected corpus generalizes to all publications involving pilot studies. However, this corpus provided a good view into the prevailing practices in diverse research communities.

Another aspect in achieving representativeness is the choice of search keywords. Our literature review may have missed articles that do not contain the 'pilot' keyword and, instead, refer to the pilot study in other ways, such as "preliminary study" or "formative study." However, both our survey and scoping review of the literature found that 'pilot study' is the most common term to refer to crowd pilot studies. Further, there is a difference between pilot studies and preliminary studies. The latter are primarily conducted to identify user needs and to define requirements. In the context of crowdsourcing, however, pilot studies are being conducted for the specific purpose of determining and validating the parameters of a crowdsourcing campaign. We argue that 'pilot study' is a standing term that is being used to refer to small-scale formative studies in crowdsourcing-based research. It is this type of study that we investigated in our paper.

We acknowledge that there may be more reasons that prevent authors from reporting crowd pilot studies more elaborately, which did not surface in our investigation. We therefore cannot treat this work as an exhaustive account of why or how crowd pilot studies are being reported.

6 CONCLUSION

In this paper, we provided an extensive investigation into the state of pilot study reporting in crowdsourcing research. Our systematic screening of over 500 publications at the intersection of HCI and crowdsourcing literature resulted in a corpus of 171 articles which we analyzed in depth. Our analysis revealed that authors are often vague about the extent and content of their crowd pilot studies. Insufficient details pertaining to such pilot studies can hinder replication and reproducibility, and stall the progress of scientific research. We explored the various reasons that drive authors to carry out crowd pilot studies (RQ1), how they are typically reported (RQ2), and what such reports contain (RQ3). Through synthesizing related literature and via a survey study with crowdsourcing researchers in academia and industry, we explored the desirable attributes of a crowd pilot study (RQ4) and the factors that influence the reporting of crowd pilot studies (RQ5). Finally, we explored platform-specific features that can support and facilitate crowd pilot studies (RQ6). Based on our findings, we reflected on how detailed reporting of crowd pilot studies can further aid fair, responsible, and reproducible crowdsourcing research. We presented insights into the trade-offs that authors make while reporting crowd pilot studies and proposed guidelines for reporting them. Our proposed guidelines for reporting crowd pilot studies and the APA-inspired way of doing so — concisely but effectively — can have important implications on the proliferation of crowdsourcing research, crowdsourcing as a sound scientific method, and on the anonymous crowd workers who undoubtedly play the most pivotal role in sustaining the crowdsourcing paradigm.

ACKNOWLEDGMENTS

This work was partially supported by the TU Delft Design@Scale AI Lab.

REFERENCES

[1] Utku Günay Acer, Marc van den Broeck, Claudio Forlivesi, Florian Heller, and Fahim Kawsar. 2019. Scaling Crowdsourcing with Mobile Workforce: A Case Study with Belgian Postal Service. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2, Article 35 (2019), 32 pages. https://doi.org/10.1145/3328906

[2] Bhoomika Agarwal, Abhiram Ravikumar, and Snehanshu Saha. 2016. A Novel Approach to Big Data Veracity Using Crowdsourcing Techniques and Bayesian Predictors. In *Proceedings of the 9th Annual ACM India Conference (COMPUTE '16)*. ACM, New York, NY, USA, 153–160. https://doi.org/10.1145/2998476.2998498

[3] Jonathan Aigrain, Arnaud Dapogny, Kévin Bailly, Séverine Dubuisson, Marcin Detyniecki, and Mohamed Chetouani. 2016. On Leveraging Crowdsourced Data for Automatic Perceived Stress Detection. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI '16)*. ACM, New York, NY, USA, 113–120. https://doi.org/10.1145/2993148.2993200

[4] Alan Aipe and Ujwal Gadiraju. 2018. SimilarHITs: Revealing the Role of Task Similarity in Microtask Crowdsourcing. In *Proceedings of the 29th on Hypertext and Social Media (HT '18)*. ACM, New York, NY, USA, 115–122. https://doi.org/10.1145/3209542.3209558

[5] Fouad Alallah, Ali Neshati, Nima Sheibani, Yumiko Sakamoto, Andrea Bunt, Pourang Irani, and Khalad Hasan. 2018. Crowdsourcing vs Laboratory-Style Social Acceptability Studies? Examining the Social Acceptability of Spatial User Interactions for Head-Worn Displays. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 1–7. https://doi.org/10.1145/3173574.3173884

[6] Omar Alonso. 2009. Guidelines for Designing Crowdsourcing-based Relevance Experiments.

[7] Omar Alonso, Catherine C. Marshall, and Marc Najork. 2013. Are Some Tweets More Interesting Than Others? #HardQuestion. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval (HCIR '13)*. ACM, New York, NY, USA, Article 2, 10 pages. https://doi.org/10.1145/2528394.2528396

[8] Omar Alonso, Catherine C. Marshall, and Marc Najork. 2015. Debugging a Crowdsourced Task with Low Inter-Rater Agreement. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '15)*. ACM, New York, NY, USA, 101–110. https://doi.org/10.1145/2756406.2757741

[9] Abdullah Alshaibani, Sylvia Carrell, Li-Hsin Tseng, Jungmin Shin, and Alexander Quinn. 2020. Privacy-Preserving Face Redaction Using Crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8, 1 (2020), 13–22. https://doi.org/10.1609/hcomp.v8i1.7459

[10] Gabriel Amaral, Alessandro Piscopo, Lucie-aimée Kaffee, Odinaldo Rodrigues, and Elena Simperl. 2021. Assessing the Quality of Sources in Wikidata Across Languages: A Hybrid Approach. *J. Data and Information Quality* 13, 4, Article 23 (2021), 35 pages. https://doi.org/10.1145/3484828

[11] Maria Anna Ambrosino, Jerry Andriessen, Vanja Annunziata, Massimo De Santo, Carmela Luciano, Mirjam Pardijs, Donato Pirozzi, and Gianluca Santangelo. 2018. Protection and Preservation of Campania Cultural Heritage Engaging Local Communities via the Use of Open Data. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age (dg.o '18)*. ACM, New York, NY, USA, Article 50, 8 pages. https://doi.org/10.1145/3209281.3209347

[12] American Psychological Association (Ed.). 2020. *Publication Manual of the American Psychological Association. The Official Guide to APA Style* (7th ed.). American Psychological Association, Washington, D.C.

[13] Ofra Amir, Yuval Shahar, Ya'akov Gal, and Litan Ilani. 2013. On the Verification Complexity of Group Decision-Making Tasks. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 1, 1 (2013), 2–8. https://doi.org/10.1609/hcomp.v1i1.13072

[14] Samreen Anjum, Chi Lin, and Danna Gurari. 2021. CrowdMOT: Crowdsourcing Strategies for Tracking Multiple Objects in Videos. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 266 (2021), 25 pages. https://doi.org/10.1145/3434175

[15] Jaime Arguello, Bogeum Choi, and Robert Capra. 2018. Factors Influencing Users' Information Requests: Medium, Target, and Extra-Topical Dimension. *ACM Trans. Inf. Syst.* 36, 4, Article 41 (2018), 37 pages. https://doi.org/10.1145/3209624

[16] Agathe Balayn, Gaole He, Andrea Hu, Jie Yang, and Ujwal Gadiraju. 2022. Ready Player One! Eliciting Diverse Knowledge Using A Configurable Game. In *Proceedings of the ACM Web Conference 2022*. ACM, New York, NY, USA, 1709–1719. https://doi.org/10.1145/3485447.3512241

[17] Natã M. Barbosa and Monchu Chen. 2019. Rehumanized Crowdsourcing: A Labeling Framework Addressing Bias and Ethics in Machine Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300773

[18] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. 2013. OpenSurfaces: A Richly Annotated Catalog of Surface Appearance. *ACM Trans. Graph.* 32, 4, Article 111 (2013), 17 pages. https://doi.org/10.1145/2461912.2462002

[19] David Benyon. 2013. *Designing Interactive Systems: A Comprehensive Guide to HCI, UX and Interaction Design.* Trans-Atlantic Publications, Inc.

[20] Michael Borish and Benjamin Lok. 2016. Rapid Low-Cost Virtual Human Bootstrapping via the Crowd. *ACM Trans. Intell. Syst. Technol.* 7, 4, Article 47 (2016), 20 pages. https://doi.org/10.1145/2897366

[21] Ria Mae Borromeo, Thomas Laurent, and Motomichi Toyama. 2016. The Influence of Crowd Type and Task Complexity on Crowdsourced Work Quality. In *Proceedings of the 20th International Database Engineering & Applications Symposium (IDEAS '16).* ACM, New York, NY, USA, 70–76. https://doi.org/10.1145/2938503.2938511

[22] Jonathan Bragg, Mausam, and Daniel S. Weld. 2018. Sprout: Crowd-Powered Task Design for Crowdsourcing. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST '18).* ACM, New York, NY, USA, 165–176. https://doi.org/10.1145/3242587.3242598

[23] Robin Brewer, Meredith Ringel Morris, and Anne Marie Piper. 2016. "Why Would Anybody Do This?": Understanding Older Adults' Motivations and Challenges in Crowd Work. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16).* ACM, New York, NY, USA, 2246–2257. https://doi.org/10.1145/2858036.2858198

[24] Steven Burrows, Martin Potthast, and Benno Stein. 2013. Paraphrase Acquisition via Crowdsourcing and Machine Learning. *ACM Trans. Intell. Syst. Technol.* 4, 3, Article 43 (2013), 21 pages. https://doi.org/10.1145/2483669.2483676

[25] Michele A. Burton, Erin Brady, Robin Brewer, Callie Neylan, Jeffrey P. Bigham, and Amy Hurst. 2012. Crowdsourcing Subjective Fashion Advice Using VizWiz: Challenges and Opportunities. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '12).* ACM, New York, NY, USA, 135–142. https://doi.org/10.1145/2384916.2384941

[26] Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. 2020. *Choice of Voices: A Large-Scale Evaluation of Text-to-Speech Voice Quality for Long-Form Content.* ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376789

[27] Gülcan Can, Jean-Marc Odobez, and Daniel Gatica-Perez. 2018. How to Tell Ancient Signs Apart? Recognizing and Visualizing Maya Glyphs with CNNs. *J. Comput. Cult. Herit.* 11, 4, Article 20 (2018), 25 pages. https://doi.org/10.1145/3230670

[28] Mark Cartwright, Ayanna Seals, Justin Salamon, Alex Williams, Stefanie Mikloska, Duncan MacConnell, Edith Law, Juan P. Bello, and Oded Nov. 2017. Seeing Sound: Investigating the Effects of Visualizations and Complexity on Crowdsourced Audio Annotations. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 29 (2017), 21 pages. https://doi.org/10.1145/3134664

[29] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17).* ACM, New York, NY, USA, 2334–2346. https://doi.org/10.1145/3025453.3026044

[30] Shuo Chang, Peng Dai, Jilin Chen, and Ed H. Chi. 2015. Got Many Labels? Deriving Topic Labels from Multiple Sources for Social Media Posts Using Crowdsourcing and Ensemble Learning. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion).* ACM, New York, NY, USA, 397–406. https://doi.org/10.1145/2740908.2745401

[31] Chen Chen, Paweł W. Woźniak, Andrzej Romanowski, Mohammad Obaid, Tomasz Jaworski, Jacek Kucharski, Krzysztof Grudzień, Shengdong Zhao, and Morten Fjeld. 2016. Using Crowdsourcing for Scientific Analysis of Industrial Tomographic Images. *ACM Trans. Intell. Syst. Technol.* 7, 4, Article 52 (2016), 25 pages. https://doi.org/10.1145/2897370

[32] Justin Cheng, Jaime Teevan, and Michael S. Bernstein. 2015. Measuring Crowdsourcing Effort with Error-Time Curves. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15).* ACM, New York, NY, USA, 1365–1374. https://doi.org/10.1145/2702123.2702145

[33] Parmit K. Chilana, Amy J. Ko, Jacob O. Wobbrock, and Tovi Grossman. 2013. A Multi-Site Field Study of Crowdsourced Contextual Help: Usage and Perspectives of End Users and Software Teams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13).* ACM, New York, NY, USA, 217–226. https://doi.org/10.1145/2470654.2470685

[34] Charles L. A. Clarke, Alexandra Vtyurina, and Mark D. Smucker. 2021. Assessing Top-k Preferences. *ACM Trans. Inf. Syst.* 39, 3, Article 33 (2021), 21 pages. https://doi.org/10.1145/3451161

[35] Cihan Cobanoglu, Muhittin Cavusoglu, and Turktarhan Gozde. 2021. A beginner's guide and best practices for using crowdsourcing platforms for survey research: The case of Amazon Mechanical Turk (MTurk). *Journal of Global Business Insights* 6, 1 (2021), 92–97. https://doi.org/10.5038/2640-6489.6.1.1177

[36] Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. 2020. Threats of a Replication Crisis in Empirical Computer Science. *Commun. ACM* 63, 8 (2020), 70–79. https://doi.org/10.1145/3360311

[37] Michael Correll and Jeffrey Heer. 2017. Regression by Eye: Estimating Trends in Bivariate Visualizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17).* ACM, New York, NY, USA, 1387–1396.

https://doi.org/10.1145/3025453.3025922

[38] Michael Correll, Dominik Moritz, and Jeffrey Heer. 2018. Value-Suppressing Uncertainty Palettes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3174216

[39] Crowdsourcing-code.com. 2017. Ground Rules for Paid Crowdsourcing/Crowdworking. Guideline for a prosperous and fair cooperation between crowdsourcing companies and crowdworkers. https://www.crowdsourcing-code.com/media/documents/Code_of_Conduct_EN.pdf

[40] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. *ACM Comput. Surv.* 51, 1, Article 7 (2018), 40 pages. https://doi.org/10.1145/3148148

[41] Dan Davis, Claudia Hauff, and Geert-Jan Houben. 2018. Evaluating Crowdworkers as a Proxy for Online Learners in Video-Based Learning Contexts. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 42 (2018), 16 pages. https://doi.org/10.1145/3274311

[42] Greg d'Eon, Joslin Goh, Kate Larson, and Edith Law. 2019. Paying Crowd Workers for Collaborative Work. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 125 (2019), 24 pages. https://doi.org/10.1145/3359227

[43] Nicholas J. DeVito and Ben Goldacre. 2019. Catalogue of Bias: Publication Bias. *BMJ Evidence-Based Medicine* 24, 2 (2019), 53–54. https://doi.org/10.1136/bmjebm-2018-111107

[44] Nicholas Diakopoulos, Daniel Trielli, and Grace Lee. 2021. Towards Understanding and Supporting Journalistic Practices Using Semi-Automated News Discovery Tools. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 406 (2021), 30 pages. https://doi.org/10.1145/3479550

[45] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and Dynamics of Mechanical Turk Workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. ACM, New York, NY, USA, 135–143. https://doi.org/10.1145/3159652.3159661

[46] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, and Philippe Cudré-Mauroux. 2014. Scaling-up the Crowd: Micro-task Pricing Schemes for Worker Retention and Latency Improvement. In *Second AAAI Conference on Human Computation and Crowdsourcing*. AAAI, Palo Alto, CA, USA. https://doi.org/10.1609/hcomp.v2i1.13154

[47] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux. 2015. The Dynamics of Micro-Task Crowdsourcing: The Case of Amazon MTurk. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 238–247. https://doi.org/10.1145/2736277.2741685

[48] Evanthia Dimara, Anastasia Bezerianos, and Pierre Dragicevic. 2017. Narratives in Crowdsourced Evaluation of Visualizations: A Double-Edged Sword?. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 5475–5484. https://doi.org/10.1145/3025453.3025870

[49] Steven Dow, Elizabeth Gerber, and Audris Wong. 2013. A Pilot Study of Using Crowds in the Classroom. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 227–236. https://doi.org/10.1145/2470654.2470686

[50] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A Checklist to Combat Cognitive Biases in Crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9. AAAI, Palo Alto, CA, USA, 48–59. https://doi.org/10.1609/hcomp.v9i1.18939

[51] John J. Dudley, Jason T. Jacques, and Per Ola Kristensson. 2019. Crowdsourcing Interface Feature Design with Bayesian Optimization. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300482

[52] John J. Dudley, Jason T. Jacques, and Per Ola Kristensson. 2021. Crowdsourcing Design Guidance for Contextual Adaptation of Text Content in Augmented Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Article 731, 14 pages. https://doi.org/10.1145/3411764.3445493

[53] Dynamo Contributors. 2014. Guidelines for Academic Requesters. Version 1.1 (10/2/2014). , 25 pages. https://irb.northwestern.edu/docs/guidelinesforacademicrequesters-1.pdf

[54] Florian Echtler and Maximilian Häußler. 2018. Open Source, Open Science, and the Replication Crisis in HCI. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–8. https://doi.org/10.1145/3170427.3188395

[55] Tom Edixhoven, Sihang Qiu, Lucie Kuiper, Olivier Dikken, Gwennan Smitskamp, and Ujwal Gadiraju. 2021. Improving Reactions to Rejection in Crowdsourcing Through Self-Reflection. In *Proceedings of the 13th ACM Web Science Conference 2021 (WebSci '21)*. ACM, New York, NY, USA, 74–83. https://doi.org/10.1145/3447535.3462482

[56] Carsten Eickhoff. 2014. Crowd-Powered Experts: Helping Surgeons Interpret Breast Cancer Images. In *Proceedings of the First International Workshop on Gamification for Information Retrieval (GamifIR '14)*. ACM, New York, NY, USA, 53–56. https://doi.org/10.1145/2594776.2594788

[57] Carsten Eickhoff, Christopher G. Harris, Arjen P. de Vries, and Padmini Srinivasan. 2012. Quality through Flow and Immersion: Gamifying Crowdsourced Relevance Assessments. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. ACM, New York, NY, USA, 871–880. https://doi.org/10.1145/2348283.2348400

[58] Irene Eleta and Jennifer Golbeck. 2012. A Study of Multilingual Social Tagging of Art Images: Cultural Bridges and Diversity. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 695–704. https://doi.org/10.1145/2145204.2145310

[59] Shaoyang Fan, Ujwal Gadiraju, Alessandro Checco, and Gianluca Demartini. 2020. CrowdCO-OP: Sharing Risks and Rewards in Crowdsourcing. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 132 (2020), 24 pages. https://doi.org/10.1145/3415203

[60] Oluwaseyi Feyisetan and Elena Simperl. 2019. Beyond Monetary Incentives: Experiments in Paid Microtask Contests. *Trans. Soc. Comput.* 2, 2, Article 6 (2019), 31 pages. https://doi.org/10.1145/3321700

[61] Riccardo Fogliato, Alexandra Chouldechova, and Zachary Lipton. 2021. The Impact of Algorithmic Risk Assessments on Human Predictions and Its Analysis via Crowdsourcing Studies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 428 (2021), 24 pages. https://doi.org/10.1145/3479572

[62] Erin D. Foster and Ariel Deardorff. 2017. Open Science Framework (OSF). *Journal of the Medical Library Association (JMLA)* 105, 2 (2017), 203. https://doi.org/10.5195/jmla.2017.88

[63] Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. 2017. Modus Operandi of Crowd Workers: The Invisible Role of Microtask Work Environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 49 (2017), 29 pages. https://doi.org/10.1145/3130914

[64] Ujwal Gadiraju and Gianluca Demartini. 2019. Understanding Worker Moods and Reactions to Rejection in Crowdsourcing. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media (HT '19)*. ACM, New York, NY, USA, 211–220. https://doi.org/10.1145/3342220.3343644

[65] Ujwal Gadiraju, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze. 2015. Human Beyond the Machine: Challenges and Opportunities of Microtask Crowdsourcing. *IEEE Intelligent Systems* 30, 4 (2015), 81–85. https://doi.org/10.1109/MIS.2015.66

[66] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1631–1640. https://doi.org/10.1145/2702123.2702443

[67] Ujwal Gadiraju, Sebastian Möller, Martin Nöllenburg, Dietmar Saupe, Sebastian Egger-Lampl, Daniel Archambault, and Brian Fisher. 2017. Crowdsourcing Versus the Laboratory: Towards Human-Centered Experiments Using the Crowd. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*, Daniel Archambault, Helen Purchase, and Tobias Hoßfeld (Eds.). Springer International Publishing, Cham, 6–26. https://doi.org/10.1007/978-3-319-66435-4_2

[68] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2017. Clarity is a Worthwhile Quality: On the Role of Task Clarity in Microtask Crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media (HT '17)*. ACM, New York, NY, USA, 5–14. https://doi.org/10.1145/3078714.3078715

[69] Ujwal Gadiraju, Ran Yu, Stefan Dietze, and Peter Holtz. 2018. Analyzing Knowledge Gain of Users in Informational Search Sessions on the Web. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. ACM, New York, NY, USA, 2–11. https://doi.org/10.1145/3176349.3176381

[70] Barney G. Glaser and Anselm L. Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine Transaction, Piscataway, New Jersey.

[71] Jorge Goncalves, Denzil Ferreira, Simo Hosio, Yong Liu, Jakob Rogstadius, Hannu Kukka, and Vassilis Kostakos. 2013. Crowdsourcing on the Spot: Altruistic Use of Public Displays, Feasibility, Performance, and Behaviours. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. ACM, New York, NY, USA, 753–762. https://doi.org/10.1145/2493432.2493481

[72] Jorge Goncalves, Simo Hosio, Denzil Ferreira, and Vassilis Kostakos. 2014. Game of Words: Tagging Places through Crowdsourcing on Public Displays. In *Proceedings of the 2014 Conference on Designing Interactive Systems (DIS '14)*. ACM, New York, NY, USA, 705–714. https://doi.org/10.1145/2598510.2598514

[73] Jorge Goncalves, Simo Hosio, Maja Vukovic, and Shin'ichi Konomi. 2017. Mobile and situated crowdsourcing. *International Journal of Human-Computer Studies* 102 (2017), 1–3. https://doi.org/10.1016/j.ijhcs.2016.12.001

[74] Leo A. Goodman. 1961. Snowball Sampling. *The Annals of Mathematical Statistics* 32, 1 (1961), 148–170. http://www.jstor.org/stable/2237615

[75] Mary L. Gray and Siddharth Suri. 2019. *Ghost Work. How to stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt, Boston and New York, N.Y.

[76] Daniela Grijincu, Miguel A. Nacenta, and Per Ola Kristensson. 2014. User-Defined Interface Gestures: Dataset and Analysis. In *Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces (ITS '14)*. ACM, New York, NY, USA, 25–34. https://doi.org/10.1145/2669485.2669511

[77] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. All Those Wasted Hours: On Task Abandonment in Crowdsourcing. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, New York, NY, USA, 321–329. https://doi.org/10.1145/3289600.3291035

[78] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. The Impact of Task Abandonment in Crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering* 33, 5 (2019), 2266–2279. https://doi.org/10.1109/TKDE.2019.2948168

[79] Lei Han, Rudra Sawant, Shaoyang Fan, Glenn Kefford, and Gianluca Demartini. 2021. An Analysis of the Australian Political Discourse in Sponsored Social Media Content. In *Proceedings of the 25th Australasian Document Computing Symposium (ADCS '21)*. ACM, New York, NY, USA, Article 1, 5 pages. https://doi.org/10.1145/3503516.3503533

[80] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174023

[81] Kotaro Hara, Shiri Azenkot, Megan Campbell, Cynthia L. Bennett, Vicki Le, Sean Pannella, Robert Moore, Kelly Minckler, Rochelle H. Ng, and Jon E. Froehlich. 2013. Improving Public Transit Accessibility for Blind Riders by Crowdsourcing Bus Stop Landmark Locations with Google Street View. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '13)*. ACM, New York, NY, USA, Article 16, 8 pages. https://doi.org/10.1145/2513383.2513448

[82] Kotaro Hara, Shiri Azenkot, Megan Campbell, Cynthia L. Bennett, Vicki Le, Sean Pannella, Robert Moore, Kelly Minckler, Rochelle H. Ng, and Jon E. Froehlich. 2015. Improving Public Transit Accessibility for Blind Riders by Crowdsourcing Bus Stop Landmark Locations with Google Street View: An Extended Analysis. *ACM Trans. Access. Comput.* 6, 2, Article 5 (2015), 23 pages. https://doi.org/10.1145/2717513

[83] Kotaro Hara, Jin Sun, Robert Moore, David Jacobs, and Jon Froehlich. 2014. Tohme: Detecting Curb Ramps in Google Street View Using Crowdsourcing, Computer Vision, and Machine Learning. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST '14)*. ACM, New York, NY, USA, 189–204. https://doi.org/10.1145/2642918.2647403

[84] Chris Harrison and Haakon Faste. 2014. Implications of Location and Touch for On-Body Projected Interfaces. In *Proceedings of the 2014 Conference on Designing Interactive Systems (DIS '14)*. ACM, New York, NY, USA, 543–552. https://doi.org/10.1145/2598510.2598587

[85] Gaole He, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. 2022. It is Like Finding a Polar Bear in the Savannah! Concept-level AI Explanations with Analogical Inference from Commonsense Knowledge. In *Proceedings of the Conference on Human Computation and Crowdsourcing (HCOMP '22, Vol. 10)*. AAAI, Palo Alto, CA, USA, 89–101. https://doi.org/10.1609/hcomp.v10i1.21990

[86] Gary T. Henry. 2002. *Practical Sampling*. Sage, Newbury Park.

[87] Danula Hettiachchi, Niels van Berkel, Vassilis Kostakos, and Jorge Goncalves. 2020. CrowdCog: A Cognitive Skill Based System for Heterogeneous Task Assignment and Recommendation in Crowdsourcing. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 110 (2020), 22 pages. https://doi.org/10.1145/3415181

[88] Michiel Hildebrand, Maarten Brinkerink, Riste Gligorov, Martijn van Steenbergen, Johan Huijkman, and Johan Oomen. 2013. Waisda? Video Labeling Game. In *Proceedings of the 21st ACM International Conference on Multimedia (MM '13)*. ACM, New York, NY, USA, 823–826. https://doi.org/10.1145/2502081.2502221

[89] Matthias Hirth, Kathrin Borchert, Fabian Allendorf, Florian Metzger, and Tobias Hoßfeld. 2019. Crowd-Based Study of Gameplay Impairments and Player Performance in DOTA 2. In *Proceedings of the 4th Internet-QoE Workshop on QoE-Based Analysis and Management of Data Communication Networks (Internet-QoE'19)*. ACM, New York, NY, USA, 19–24. https://doi.org/10.1145/3349611.3355545

[90] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing High Quality Crowdwork. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 419–429. https://doi.org/10.1145/2736277.2741102

[91] Jonggi Hong and Leah Findlater. 2018. Identifying Speech Input Errors Through Audio-Only Interaction. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174141

[92] Simo Hosio, Jorge Goncalves, Vili Lehdonvirta, Denzil Ferreira, and Vassilis Kostakos. 2014. Situated Crowdsourcing Using a Market Model. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST '14)*. ACM, New York, NY, USA, 55–64. https://doi.org/10.1145/2642918.2647362

[93] Simo Johannes Hosio, Jaro Karppinen, Esa-Pekka Takala, Jani Takatalo, Jorge Goncalves, Niels van Berkel, Shin'ichi Konomi, and Vassilis Kostakos. 2018. Crowdsourcing Treatments for Low Back Pain. In *Proceedings of the 2018 CHI*

Conference on Human Factors in Computing Systems (CHI '18). ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173850

[94] Kevin Hu, Snehalkumar 'Neil' S. Gaikwad, Madelon Hulsebos, Michiel A. Bakker, Emanuel Zgraggen, César Hidalgo, Tim Kraska, Guoliang Li, Arvind Satyanarayan, and Çağatay Demiralp. 2019. VizNet: Towards A Large-Scale Visualization Learning and Benchmarking Repository. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300892

[95] Xiao Hu, Haobo Wang, Anirudh Vegesana, Somesh Dube, Kaiwen Yu, Gore Kao, Shuo-Han Chen, Yung-Hsiang Lu, George K. Thiruvathukal, and Ming Yin. 2020. *Crowdsourcing Detection of Sampling Biases in Image Datasets*. ACM, New York, NY, USA, 2955–2961. https://doi.org/10.1145/3366423.3380063

[96] Shih-Wen Huang and Wai-Tat Fu. 2013. Don't Hide in the Crowd! Increasing Social Transparency between Peer Workers Improves Crowdsourcing Outcomes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 621–630. https://doi.org/10.1145/2470654.2470743

[97] Yi-Ching Huang, Jiunn-Chia Huang, Hao-Chuan Wang, and Jane Hsu. 2017. Supporting ESL Writing by Prompting Crowdsourced Structural Feedback. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 5, 1 (2017), 71–78. https://doi.org/10.1609/hcomp.v5i1.13313

[98] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300637

[99] Ken Hyland. 1996. Writing Without Conviction? Hedging in Science Research Articles. *Applied Linguistics* 17, 4 (1996), 433–454. https://doi.org/10.1093/applin/17.4.433

[100] Kazushi Ikeda and Michael S. Bernstein. 2016. Pay It Backward: Per-Task Payments on Crowdsourcing Platforms Reduce Productivity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4111–4121. https://doi.org/10.1145/2858036.2858327

[101] Junyong In. 2017. Introduction of a Pilot Study. *Korean Journal of Anesthesiology* 70, 6 (2017), 601–605. https://doi.org/10.4097/kjae.2017.70.6.601

[102] Oana Inel, Giannis Haralabopoulos, Dan Li, Christophe Van Gysel, Zoltán Szlávik, Elena Simperl, Evangelos Kanoulas, and Lora Aroyo. 2018. Studying Topical Relevance with Evidence-Based Crowdsourcing. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. ACM, New York, NY, USA, 1253–1262. https://doi.org/10.1145/3269206.3271779

[103] Lilly C. Irani and M. Six Silberman. 2013. Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 611–620. https://doi.org/10.1145/2470654.2470742

[104] Kasthuri Jayarajah and Archan Misra. 2018. Predicting Episodes of Non-Conformant Mobility in Indoor Environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 4, Article 172 (2018), 24 pages. https://doi.org/10.1145/3287050

[105] Hernisa Kacorri, Kaoru Shinkawa, and Shin Saito. 2014. Introducing Game Elements in Crowdsourced Video Captioning by Non-Experts. In *Proceedings of the 11th Web for All Conference (W4A '14)*. ACM, New York, NY, USA, Article 29, 4 pages. https://doi.org/10.1145/2596695.2596713

[106] Thivya Kandappu, Archan Misra, and Randy Tandriansyah. 2017. Collaboration Trumps Homophily in Urban Mobile Crowdsourcing. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 902–915. https://doi.org/10.1145/2998181.2998311

[107] Alireza Karduni, Ryan Wesslen, Isaac Cho, and Wenwen Dou. 2020. Du Bois Wrapped Bar Chart: Visualizing Categorical Data with Disproportionate Values. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376365

[108] Johannes Kiesel, Florian Kneist, Lars Meyer, Kristof Komlossy, Benno Stein, and Martin Potthast. 2020. Web Page Segmentation Revisited: Evaluation Framework and Dataset. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)*. ACM, New York, NY, USA, 3047–3054. https://doi.org/10.1145/3340531.3412782

[109] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J. Guo, Robert C. Miller, and Krzysztof Z. Gajos. 2014. Crowdsourcing Step-by-Step Information Extraction to Enhance Existing How-to Videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 4017–4026. https://doi.org/10.1145/2556288.2556986

[110] Lawrence H. Kim and Sean Follmer. 2017. UbiSwarm: Ubiquitous Robotic Interfaces and Investigation of Abstract Motion as a Display. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 66 (2017), 20 pages. https://doi.org/10.1145/3130931

[111] Sung-Hee Kim, Hyokun Yun, and Ji Soo Yi. 2012. How to Filter out Random Clickers in a Crowdsourcing-Based Study?. In *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors – Novel Evaluation Methods for Visualization (BELIV '12)*. ACM, New York, NY, USA, Article 15, 7 pages. https://doi.org/10.1145/2442576.2442591

[112] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 453–456. https://doi.org/10.1145/1357054.1357127

[113] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The Future of Crowd Work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, New York, NY, USA, 1301–1318. https://doi.org/10.1145/2441776.2441923

[114] Rachel Kohler, John Purviance, and Kurt Luther. 2017. Supporting Image Geolocation with Diagramming and Crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 5, 1 (2017), 98–107. https://doi.org/10.1609/hcomp.v5i1.13296

[115] Caitlin Kuhlman, Diana Doherty, Malika Nurbekova, Goutham Deva, Zarni Phyo, Paul-Henry Schoenhagen, MaryAnn VanValkenburg, Elke Rundensteiner, and Lane Harrison. 2019. Evaluating Preference Collection Methods for Interactive Ranking Analytics. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300742

[116] Anand Kulkarni, Matthew Can, and Björn Hartmann. 2012. Collaboratively Crowdsourcing Workflows with Turkomatic. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 1003–1012. https://doi.org/10.1145/2145204.2145354

[117] Xingyu Lan, Xinyue Xu, and Nan Cao. 2021. Understanding Narrative Linearity for Telling Expressive Time-Oriented Stories. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Article 604, 13 pages. https://doi.org/10.1145/3411764.3445344

[118] Edith Law, Ming Yin, Joslin Goh, Kevin Chen, Michael A. Terry, and Krzysztof Z. Gajos. 2016. Curiosity Killed the Cat, but Makes Crowdwork Better. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4098–4110. https://doi.org/10.1145/2858036.2858144

[119] Doris Jung-Lin Lee, Joanne Lo, Moonhyok Kim, and Eric Paulos. 2016. Crowdclass: Designing Classification-Based Citizen Science Learning Modules. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 4, 1 (2016), 109–118. https://doi.org/10.1609/hcomp.v4i1.13273

[120] Michael J. Lee and Amy J. Ko. 2015. Comparing the Effectiveness of Online Learning Approaches on CS1 Learning Outcomes. In *Proceedings of the Eleventh Annual International Conference on International Computing Education Research (ICER '15)*. ACM, New York, NY, USA, 237–246. https://doi.org/10.1145/2787622.2787709

[121] Iolanda Leite, André Pereira, Allison Funkhouser, Boyang Li, and Jill Fain Lehman. 2016. Semi-Situated Learning of Verbal and Nonverbal Content for Repeated Human-Robot Interaction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI '16)*. ACM, New York, NY, USA, 13–20. https://doi.org/10.1145/2993148.2993190

[122] Fritz Lekschas, Spyridon Ampanavos, Pao Siangliulue, Hanspeter Pfister, and Krzysztof Z. Gajos. 2021. Ask Me or Tell Me? Enhancing the Effectiveness of Crowdsourced Design Feedback. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Article 564, 12 pages. https://doi.org/10.1145/3411764.3445507

[123] Andrew C. Leon, Lori L. Davis, and Helena C. Kraemer. 2011. The Role and Interpretation of Pilot Studies in Clinical Research. *Journal of Psychiatric Research* 45, 5 (2011), 626–629. https://doi.org/10.1016/j.jpsychires.2010.10.008

[124] Blaine Lewis and Daniel Vogel. 2020. Longer Delays in Rehearsal-Based Interfaces Increase Expert Use. *ACM Trans. Comput.-Hum. Interact.* 27, 6, Article 45 (2020), 41 pages. https://doi.org/10.1145/3418196

[125] Tianyi Li, Chandler J. Manns, Chris North, and Kurt Luther. 2019. Dropping the Baton? Understanding Errors and Bottlenecks in a Crowdsourced Sensemaking Pipeline. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 136 (2019), 26 pages. https://doi.org/10.1145/3359238

[126] Zhaoliang Lun, Evangelos Kalogerakis, and Alla Sheffer. 2015. Elements of Style: Learning Perceptual Shape Style Similarity. *ACM Trans. Graph.* 34, 4, Article 84 (2015), 14 pages. https://doi.org/10.1145/2766929

[127] Kurt Luther, Nathan Hahn, Steven Dow, and Aniket Kittur. 2015. Crowdlines: Supporting Synthesis of Diverse Information Sources through Crowdsourced Outlines. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 3, 1 (2015), 110–119. https://doi.org/10.1609/hcomp.v3i1.13239

[128] Ioanna Lykourentzou, Angeliki Antoniou, Yannick Naudet, and Steven P. Dow. 2016. Personality Matters: Balancing for Personality Types Leads to Better Outcomes for Crowd Teams. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 260–273. https://doi.org/10.1145/2818048.2819979

[129] Malcolm MacLeod. 2021. An "Omics" Answer to the Replication Crisis. https://future.com/publomics-replication-crisis/

[130] Eddy Maddalena, Luis-Daniel Ibáñez, and Elena Simperl. 2020. Mapping Points of Interest Through Street View Imagery and Paid Crowdsourcing. *ACM Trans. Intell. Syst. Technol.* 11, 5, Article 63 (2020), 28 pages. https://doi.org/10.1145/3403931

[131] V.K. Chaithanya Manam, J. Thomas, and Alexander J. Quinn. 2022. TaskLint: Automated Detection of Ambiguities in Task Instructions. In *Proceedings of the Conference on Human Computation and Crowdsourcing (HCOMP '22)*. AAAI, Palo Alto, CA, USA. https://doi.org/10.1609/hcomp.v10i1.21996

[132] Andrew Mao, Ece Kamar, Yiling Chen, Eric Horvitz, Megan E Schwamb, Chris J Lintott, and Arfon M Smith. 2013. Volunteering Versus Work for Pay: Incentives and Tradeoffs In Crowdsourcing. In *First AAAI Conference on Human Computation and Crowdsourcing*. AAAI, Palo Alto, CA, USA. https://doi.org/10.1609/hcomp.v1i1.13075

[133] David Martin, Benjamin V. Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a Turker. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 224–235. https://doi.org/10.1145/2531602.2531663

[134] Thomas Mattauch. 2013. Innovate through Crowd Sourcing. In *Proceedings of the 41st Annual ACM SIGUCCS Conference on User Services (SIGUCCS '13)*. ACM, New York, NY, USA, 39–42. https://doi.org/10.1145/2504776.2504796

[135] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 72 (2019), 23 pages. https://doi.org/10.1145/3359174

[136] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 4, 1 (2016), 139–148. https://doi.org/10.1609/hcomp.v4i1.13287

[137] Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. 2016. Taking a HIT: Designing Around Rejection, Mistrust, Risk, and Workers' Experiences in Amazon Mechanical Turk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '2016)*. 2271–2282. https://doi.org/10.1145/2858036.2858539

[138] Andrew J. McMinn, Yashar Moshfeghi, and Joemon M. Jose. 2013. Building a Large-Scale Corpus for Evaluating Event Detection on Twitter. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM '13)*. ACM, New York, NY, USA, 409–418. https://doi.org/10.1145/2505515.2505695

[139] Róisín McNaney, Mohammad Othman, Dan Richardson, Paul Dunphy, Telmo Amaral, Nick Miller, Helen Stringer, Patrick Olivier, and John Vines. 2016. Speeching: Mobile Crowdsourced Speech Assessment to Support Self-Monitoring and Management for People with Parkinson's. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4464–4476. https://doi.org/10.1145/2858036.2858321

[140] Vikram Mohanty, Kareem Abdol-Hamid, Courtney Ebersohl, and Kurt Luther. 2019. Second Opinion: Supporting Last-Mile Person Identification with Crowdsourcing and Face Recognition. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (2019), 86–96. https://doi.org/10.1609/hcomp.v7i1.5272

[141] Luiz Morais, Yvonne Jansen, Nazareno Andrade, and Pierre Dragicevic. 2021. Can Anthropographics Promote Prosociality? A Review and Large-Sample Study. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Article 611, 18 pages. https://doi.org/10.1145/3411764.3445637

[142] Yashar Moshfeghi and Alvaro Francisco Huertas-Rosero. 2021. A Game Theory Approach for Estimating Reliability of Crowdsourced Relevance Assessments. *ACM Trans. Inf. Syst.* 40, 3, Article 60 (2021), 29 pages. https://doi.org/10.1145/3480965

[143] Daniel Mutembesa, Christopher Omongo, and Ernest Mwebaze. 2018. Crowdsourcing Real-Time Viral Disease and Pest Information: A Case of Nation-Wide Cassava Disease Surveillance in a Developing Country. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 6, 1 (2018), 117–125. https://doi.org/10.1609/hcomp.v6i1.13322

[144] Pranathi Mylavarapu, Adil Yalcin, Xan Gregg, and Niklas Elmqvist. 2019. Ranked-List Visualization: A Graphical Perception Study. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300422

[145] Anelise Newman, Barry McNamara, Camilo Fosco, Yun Bin Zhang, Pat Sukhum, Matthew Tancik, Nam Wook Kim, and Zoya Bylinskii. 2020. *TurkEyes: A Web-Based Toolbox for Crowdsourcing Attention Data.* ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376799

[146] Dong Nguyen, Dolf Trieschnigg, and Mariët Theune. 2014. Using Crowdsourcing to Investigate Perception of Narrative Similarity. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. ACM, New York, NY, USA, 321–330. https://doi.org/10.1145/2661829.2661918

[147] Carolina Nobre, Dylan Wootton, Zach Cutler, Lane Harrison, Hanspeter Pfister, and Alexander Lex. 2021. ReVISit: Looking Under the Hood of Interactive Visualization Studies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Article 25, 13 pages. https://doi.org/10.1145/3411764.3445382

[148] Zahra Nouri, Ujwal Gadiraju, Gregor Engels, and Henning Wachsmuth. 2021. What Is Unclear? Computational Assessment of Task Clarity in Crowdsourcing. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media (HT '21)*. ACM, New York, NY, USA, 165–175. https://doi.org/10.1145/3465336.3475109

[149] Natalya F. Noy, Jonathan Mortensen, Mark A. Musen, and Paul R. Alexander. 2013. Mechanical Turk as an Ontology Engineer? Using Microtasks as a Component of an Ontology-Engineering Workflow. In *Proceedings of the 5th Annual*

ACM Web Science Conference (WebSci '13). ACM, New York, NY, USA, 262–271. https://doi.org/10.1145/2464464.2464482

[150] Jonas Oppenlaender and Simo Hosio. 2019. Design Recommendations for Augmenting Creative Tasks with Computational Priming. In *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia (MUM '19)*. ACM, New York, NY, USA, Article 35, 13 pages. https://doi.org/10.1145/3365610.3365621

[151] Jonas Oppenlaender, Kristy Milland, Aku Visuri, Panos Ipeirotis, and Simo Hosio. 2020. Creativity on Paid Crowdsourcing Platforms. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, New York, NY, USA, Article 548, 14 pages. https://doi.org/10.1145/3313831.3376677

[152] Jonas Oppenlaender, Thanassis Tiropanis, and Simo Hosio. 2020. CrowdUI: Supporting Web Design with the Crowd. *Proc. ACM Hum.-Comput. Interact.* 4, EICS, Article 76 (2020), 28 pages. https://doi.org/10.1145/3394978

[153] Jonas Oppenlaender, Aku Visuri, Kristy Milland, Panos Ipeirotis, and Simo Hosio. 2020. What do crowd workers think about creative work?. In *Workshop on Worker-Centered Design: Expanding HCI Methods for Supporting Labor*. 4 pages. http://jultika.oulu.fi/files/nbnfi-fe2020052538841.pdf

[154] Maike Paetzel, James Kennedy, Ginevra Castellano, and Jill Fain Lehman. 2018. Incremental Acquisition and Reuse of Multimodal Affective Behaviors in a Conversational Agent. In *Proceedings of the 6th International Conference on Human-Agent Interaction (HAI '18)*. ACM, New York, NY, USA, 92–100. https://doi.org/10.1145/3284432.3284469

[155] Anshul Vikram Pandey, Katharina Rall, Margaret L. Satterthwaite, Oded Nov, and Enrico Bertini. 2015. How Deceptive Are Deceptive Visualizations? An Empirical Analysis of Common Distortion Techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1469–1478. https://doi.org/10.1145/2702123.2702608

[156] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (2010), 411–419. https://doi.org/10.1017/S1930297500002205

[157] Manasi Patwardhan, Abhishek Sainani, Richa Sharma, Shirish Karande, and Smita Ghaisas. 2018. *Towards Automating Disambiguation of Regulations: Using the Wisdom of Crowds*. ACM, New York, NY, USA, 850–855. https://doi.org/10.1145/3238147.3240727

[158] Weiping Pei, Zhiju Yang, Monchu Chen, and Chuan Yue. 2021. Quality Control in Crowdsourcing Based on Fine-Grained Behavioral Features. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 442 (2021), 28 pages. https://doi.org/10.1145/3479586

[159] Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar. 2019. What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (2019), 125–134. https://doi.org/10.1609/hcomp.v7i1.5281

[160] Mark Petticrew and Helen Roberts. 2006. *Exploring Heterogeneity and Publication Bias*. John Wiley & Sons, Ltd, Malden, MA, Chapter 7, 215–246. https://doi.org/10.1002/9780470754887.ch7

[161] Mark Petticrew and Helen Roberts. 2006. *Starting the Review: Refining the Question and Defining the Boundaries*. John Wiley & Sons, Ltd, Chapter 2, 27–56. https://doi.org/10.1002/9780470754887.ch2

[162] Mark Petticrew and Helen Roberts. 2006. *Systematic Reviews in the Social Sciences. A Practical Guide*. Blackwell Publishing, Malden, MA.

[163] Rehab Qarout, Alessandro Checco, Gianluca Demartini, and Kalina Bontcheva. 2019. Platform-Related Factors in Repeatability and Reproducibility of Crowdsourcing Tasks. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (2019), 135–143. https://doi.org/10.1609/hcomp.v7i1.5264

[164] Chenxi Qiu, Anna Squicciarini, Zhuozhao Li, Ce Pang, and Li Yan. 2020. Time-Efficient Geo-Obfuscation to Protect Worker Location Privacy over Road Networks in Spatial Crowdsourcing. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)*. ACM, New York, NY, USA, 1275–1284. https://doi.org/10.1145/3340531.3411863

[165] Sihang Qiu, Alessandro Bozzon, Max V Birk, and Ujwal Gadiraju. 2021. Using Worker Avatars to Improve Microtask Crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28. https://doi.org/10.1145/3476063

[166] Sihang Qiu, Alessandro Bozzon, and Geert-Jan Houben. 2020. *VirtualCrowd: A Simulation Platform for Microtask Crowdsourcing Campaigns*. ACM, New York, NY, USA, 222–225. https://doi.org/10.1145/3366424.3383546

[167] Sarvapali D. Ramchurn, Trung Dong Huynh, Yuki Ikuno, Jack Flann, Feng Wu, Luc Moreau, Nicholas R. Jennings, Joel E. Fischer, Wenchao Jiang, Tom Rodden, Edwin Simpson, Steven Reece, and Stephen J. Roberts. 2015. HAC-ER: A Disaster Response System Based on Human-Agent Collectives. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '15)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 533–541.

[168] Jorge Ramírez, Marcos Baez, Fabio Casati, Luca Cernuzzi, and Boualem Benatallah. 2020. *DREC: Towards a Datasheet for Reporting Experiments in Crowdsourcing*. ACM, New York, NY, USA, 377–382. https://doi.org/10.1145/3406865.3418318

[169] Jorge Ramírez, Marcos Baez, Fabio Casati, Luca Cernuzzi, Boualem Benatallah, Ekaterina A. Taran, and Veronika A. Malanina. 2021. On the Impact of Predicate Complexity in Crowdsourced Classification Tasks. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM '21)*. ACM, New York, NY, USA, 67–75. https://doi.org/10.1145/3437963.3441831

[170] Jorge Ramírez, Burcu Sayin, Marcos Baez, Fabio Casati, Luca Cernuzzi, Boualem Benatallah, and Gianluca Demartini. 2021. On the State of Reporting in Crowdsourcing Experiments and a Checklist to Aid Current Practices. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 387 (2021), 34 pages. https://doi.org/10.1145/3479531

[171] Jorge Ramírez, Marcos Baez, Fabio Casati, and Boualem Benatallah. 2019. Understanding the Impact of Text Highlighting in Crowdsourcing Tasks. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (2019), 144–152. https://doi.org/10.1609/hcomp.v7i1.5268

[172] Amy Rechkemmer and Ming Yin. 2020. Motivating Novice Crowd Workers through Goal Setting: An Investigation into the Effects on Complex Crowdsourcing Task Training. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8, 1 (2020), 122–131. https://doi.org/10.1609/hcomp.v8i1.7470

[173] Khairi Reda, Pratik Nalawade, and Kate Ansah-Koi. 2018. Graphical Perception of Continuous Quantitative Maps: The Effects of Spatial Frequency and Colormap Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173846

[174] Janice Redish and Sharon J. Laskowsk. 2009. *Guidelines for Writing Clear Instructions and Messages for Voters and Poll Workers*. Technical Report NISTIR 7596. National Institute of Standards and Technology. https://www.nist.gov/publications/guidelines-writing-clear-instructions-and-messages-voters-and-poll-workers

[175] Theodoros Rekatsinas, Amol Deshpande, and Aditya Parameswaran. 2019. CRUX: Adaptive Querying for Efficient Crowdsourced Data Extraction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19)*. ACM, New York, NY, USA, 841–850. https://doi.org/10.1145/3357384.3357976

[176] Samuel Rhys Cox, Yunlong Wang, Ashraf Abdul, Christian von der Weth, and Brian Y. Lim. 2021. Directed Diversity: Leveraging Language Embedding Distances for Collective Creativity in Crowd Ideation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Article 393, 35 pages. https://doi.org/10.1145/3411764.3445782

[177] Ronald E Robertson, Alexandra Olteanu, Fernando Diaz, Milad Shokouhi, and Peter Bailey. 2021. "I Can't Reply with That": Characterizing Problematic Email Reply Suggestions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Article 724, 18 pages. https://doi.org/10.1145/3411764.3445557

[178] Carlos Rodríguez, Florian Daniel, and Fabio Casati. 2016. Mining and Quality Assessment of Mashup Model Patterns with the Crowd: A Feasibility Study. *ACM Trans. Internet Technol.* 16, 3, Article 17 (2016), 27 pages. https://doi.org/10.1145/2903138

[179] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. *Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background*. ACM, New York, NY, USA, 439–448. https://doi.org/10.1145/3397271.3401112

[180] Quentin Roy, Futian Zhang, and Daniel Vogel. 2019. Automation Accuracy Is Good, but High Controllability May Be Better. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 1–8. https://doi.org/10.1145/3290605.3300750

[181] Sabirat Rubya, Joseph Numainville, and Svetlana Yarosh. 2021. Comparing Generic and Community-Situated Crowdsourcing for Data Validation in the Context of Recovery from Substance Use Disorders. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Article 449, 17 pages. https://doi.org/10.1145/3411764.3445399

[182] Lalit Mohan S, Priya Raman, Venkatesh Choppella, and Y. R. Reddy. 2017. A Crowdsourcing Approach for Quality Enhancement of ELearning Systems. In *Proceedings of the 10th Innovations in Software Engineering Conference (ISEC '17)*. ACM, New York, NY, USA, 188–194. https://doi.org/10.1145/3021460.3021483

[183] Marta Sabou, Klemens Käsznar, Markus Zlabinger, Stefan Biffl, and Dietmar Winkler. 2020. Verifying Extended Entity Relationship Diagrams with Open Tasks. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8, 1 (2020), 132–140. https://doi.org/10.1609/hcomp.v8i1.7471

[184] Marta Sabou, Dietmar Winkler, Peter Penzerstadler, and Stefan Biffl. 2018. Verifying Conceptual Domain Models with Human Computation: A Case Study in Software Engineering. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 6, 1 (2018), 164–173. https://doi.org/10.1609/hcomp.v6i1.13325

[185] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *Commun. ACM* 64, 9 (2021), 99–106. https://doi.org/10.1145/3474381

[186] Niloufar Salehi, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. 2015. We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1621–1630.

https://doi.org/10.1145/2702123.2702508

[187] Niloufar Salehi, Jaime Teevan, Shamsi Iqbal, and Ece Kamar. 2017. Communicating Context to the Crowd for Complex Writing Tasks. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1890–1901. https://doi.org/10.1145/2998181.2998332

[188] Mike Schaekermann, Joslin Goh, Kate Larson, and Edith Law. 2018. Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 154 (2018), 19 pages. https://doi.org/10.1145/3274423

[189] Todd W. Schiller and Michael D. Ernst. 2012. Reducing the Barriers to Writing Verified Specifications. *SIGPLAN Not.* 47, 10 (2012), 95–112. https://doi.org/10.1145/2398857.2384624

[190] Oliver S. Schneider, Hasti Seifi, Salma Kashani, Matthew Chun, and Karon E. MacLean. 2016. HapTurk: Crowdsourcing Affective Ratings of Vibrotactile Icons. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3248–3260. https://doi.org/10.1145/2858036.2858279

[191] Sebastian Schäfer, David Antons, Dirk Lüttgens, Frank Piller, and Torsten Oliver Salge. 2017. Talk to Your Crowd. *Research-Technology Management* 60, 4 (2017), 33–42. https://doi.org/10.1080/08956308.2017.1325689

[192] Pao Siangliulue, Joel Chan, Steven P. Dow, and Krzysztof Z. Gajos. 2016. IdeaHound: Improving Large-Scale Collaborative Ideation with Crowd-Powered Real-Time Semantic Modeling. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*. ACM, New York, NY, USA, 609–624. https://doi.org/10.1145/2984511.2984578

[193] M. S. Silberman, B. Tomlinson, R. LaPlante, J. Ross, L. Irani, and A. Zaldivar. 2018. Responsible Research with Crowds: Pay Crowdworkers at Least Minimum Wage. *Commun. ACM* 61, 3 (2018), 39–41. https://doi.org/10.1145/3180492

[194] Camelia Simoiu, Chiraag Sumanth, Alok Mysore, and Sharad Goel. 2019. Studying the "Wisdom of Crowds" at Scale. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (2019), 171–179. https://doi.org/10.1609/hcomp.v7i1.5271

[195] Rachel N. Simons, Danna Gurari, and Kenneth R. Fleischmann. 2020. "I Hope This Is Helpful": Understanding Crowdworkers' Challenges and Motivations for an Image Description Task. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 105 (2020), 26 pages. https://doi.org/10.1145/3415176

[196] Elena Simperl. 2021. How to Use Crowdsourcing Effectively: Guidelines and Examples. *LIBER Quarterly: The Journal of the Association of European Research Libraries* 25, 1 (2021), 18–39. https://doi.org/10.18352/lq.9948

[197] Divit P. Singh, Lee Lisle, T. M. Murali, and Kurt Luther. 2018. CrowdLayout: Crowdsourced Design and Evaluation of Biological Network Visualizations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173806

[198] Kinga Skorupska, Manuel Nunez, Wieslaw Kopec, and Radoslaw Nielek. 2018. Older Adults and Crowdsourcing: Android TV App for Evaluating TEDx Subtitle Quality. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 159 (2018), 23 pages. https://doi.org/10.1145/3274428

[199] Stephen Smart and Danielle Albers Szafir. 2019. Measuring the Separability of Shape, Size, and Color in Scatterplots. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300899

[200] Marc Spicker, Franz Götz-Hahn, Thomas Lindemeier, Dietmar Saupe, and Oliver Deussen. 2019. Quantifying Visual Abstraction Quality for Computer-Generated Illustrations. *ACM Trans. Appl. Percept.* 16, 1, Article 5 (2019), 20 pages. https://doi.org/10.1145/3301414

[201] Colin Stanley, Heike Winschiers-Theophilus, Michel Onwordi, and Gereon K. Kapuire. 2013. Rural Communities Crowdsource Technology Development: A Namibian Expedition. In *Proceedings of the Sixth International Conference on Information and Communications Technologies and Development: Notes - Volume 2 (ICTD '13)*. ACM, New York, NY, USA, 155–158. https://doi.org/10.1145/2517899.2517930

[202] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Article 341, 14 pages. https://doi.org/10.1145/3411764.3445092

[203] James Surowiecki. 2005. *The Wisdom of Crowds*. Anchor, New York, NY, USA.

[204] Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What Are the Biases in My Word Embedding?. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. ACM, New York, NY, USA, 305–311. https://doi.org/10.1145/3306618.3314270

[205] John C. Tang, Gina Venolia, and Kori M. Inkpen. 2016. Meerkat and Periscope: I Stream, You Stream, Apps Stream for Live Streams. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4770–4780. https://doi.org/10.1145/2858036.2858374

[206] Brandon Taylor, Anind K. Dey, Daniel Siewiorek, and Asim Smailagic. 2016. Using Crowd Sourcing to Measure the Effects of System Response Delays on User Engagement. In *Proceedings of the 2016 CHI Conference on Human Factors*

in Computing Systems (CHI '16). ACM, New York, NY, USA, 4413–4422. https://doi.org/10.1145/2858036.2858572

[207] Benjamin Timmermans, Lora Aroyo, and Chris Welty. 2015. Crowdsourcing Ground Truth for Question Answering Using CrowdTruth. In *Proceedings of the ACM Web Science Conference (WebSci '15)*. ACM, New York, NY, USA, Article 61, 2 pages. https://doi.org/10.1145/2786451.2786492

[208] Carlos Toxtli, Siddharth Suri, and Saiph Savage. 2021. Quantifying the Invisible Labor in Crowd Work. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 319 (2021), 26 pages. https://doi.org/10.1145/3476060

[209] George Trimponias, Xiaojuan Ma, and Qiang Yang. 2019. Rating Worker Skills and Task Strains in Collaborative Crowd Computing: A Competitive Perspective. In *The World Wide Web Conference (WWW '19)*. ACM, New York, NY, USA, 1853–1863. https://doi.org/10.1145/3308558.3313569

[210] Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* (1973), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

[211] Stephen Uzor, Jason T. Jacques, John J Dudley, and Per Ola Kristensson. 2021. Investigating the Accessibility of Crowdwork Tasks on Mechanical Turk. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Article 381, 14 pages. https://doi.org/10.1145/3411764.3445291

[212] Rajan Vaish, Snehalkumar (Neil) S. Gaikwad, Geza Kovacs, Andreas Veit, Ranjay Krishna, Imanol Arrieta Ibarra, Camelia Simoiu, Michael Wilber, Serge Belongie, Sharad Goel, James Davis, and Michael S. Bernstein. 2017. Crowd Research: Open and Scalable University Laboratories. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST '17)*. ACM, New York, NY, USA, 829–843. https://doi.org/10.1145/3126594.3126648

[213] Rajan Vaish, Shirish Goyal, Amin Saberi, and Sharad Goel. 2018. Creating Crowdsourced Research Talks at Scale. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1–11. https://doi.org/10.1145/3178876.3186031

[214] Gerard van Alphen, Sihang Qiu, Alessandro Bozzon, and Geert-Jan Houben. 2020. Analyzing Workers Performance in Online Mapping Tasks Across Web, Mobile, and Virtual Reality Platforms. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8, 1 (Oct. 2020), 141–149. https://doi.org/10.1609/hcomp.v8i1.7472

[215] Edwin Van Teijlingen and Vanora Hundley. 2002. The Importance of Pilot Studies. *Nursing Standard* 16, 40 (2002), 33. https://doi.org/10.7748/ns2002.06.16.40.33.c3214

[216] Keith Vertanen and Per Ola Kristensson. 2014. Complementing Text Entry Evaluations with a Composition Task. *ACM Trans. Comput.-Hum. Interact.* 21, 2, Article 8 (2014), 33 pages. https://doi.org/10.1145/2555691

[217] Ruben Vicente-Saez and Clara Martinez-Fuentes. 2018. Open Science Now: A Systematic Literature Review for an Integrated Definition. *Journal of business research* 88 (2018), 428–436. https://doi.org/10.1016/j.jbusres.2017.12.043

[218] Athanasios Vogogias, Daniel Archambault, Benjamin Bach, and Jessie Kennedy. 2020. Visual Encodings for Networks with Multiple Edge Types. In *Proceedings of the International Conference on Advanced Visual Interfaces*. ACM, New York, NY, USA, Article 37, 9 pages. https://doi.org/10.1145/3399715.3399827

[219] Jan vom Brocke, Aalexander Simons, Kai Riemer, Bjoern Niehaves, Ralf Plattfaut, and Anne Cleven. 2015. Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research. *Communications of the Association for Information Systems* 37 (2015). https://doi.org/10.17705/1CAIS.03709

[220] Vassilios Vonikakis, Ramanathan Subramanian, Jonas Arnfred, and Stefan Winkler. 2014. Modeling Image Appeal Based on Crowd Preferences for Automated Person-Centric Collage Creation. In *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia (CrowdMM '14)*. ACM, New York, NY, USA, 9–15. https://doi.org/10.1145/2660114.2660126

[221] Ding Wang, Shantanu Prabhat, and Nithya Sambasivan. 2022. Whose AI Dream? In Search of the Aspiration in Data Annotation.. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New York, NY, USA, Article 582, 16 pages. https://doi.org/10.1145/3491102.3502121

[222] Nai-Ching Wang, David Hicks, and Kurt Luther. 2018. Exploring Trade-Offs Between Learning and Productivity in Crowdsourced History. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 178 (2018), 24 pages. https://doi.org/10.1145/3274447

[223] Yihong Wang, Konstantinos Papangelis, Ioanna Lykourentzou, Hai-Ning Liang, Irwyn Sadien, Evangelia Demerouti, and Vassilis-Javed Khan. 2020. In Their Shoes: A Structured Analysis of Job Demands, Resources, Work Experiences, and Platform Commitment of Crowdworkers in China. *Proc. ACM Hum.-Comput. Interact.* 4, GROUP, Article 07 (2020), 40 pages. https://doi.org/10.1145/3375187

[224] Evan Welbourne, Pang Wu, Xuan Bao, and Emmanuel Munguia-Tapia. 2014. Crowdsourced Mobile Data Collection: Lessons Learned from a New Study Methodology. In *Proceedings of the 15th Workshop on Mobile Computing Systems and Applications (HotMobile '14)*. ACM, New York, NY, USA, Article 2, 6 pages. https://doi.org/10.1145/2565585.2565608

[225] Chris Welty, Lora Aroyo, Flip Korn, Sara M. McCarthy, and Shubin Zhao. 2021. Rapid Instance-Level Knowledge Acquisition for Google Maps from Class-Level Common Sense. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9, 1 (2021), 143–154. https://doi.org/10.1609/hcomp.v9i1.18947

[226] Miaomiao Wen, Keith Maki, Steven Dow, James D. Herbsleb, and Carolyn Rose. 2017. Supporting Virtual Team Formation through Community-Wide Deliberation. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 109 (2017), 19 pages. https://doi.org/10.1145/3134744

[227] Etienne Wenger. 2011. Communities of Practice: A Brief Introduction. http://hdl.handle.net/1794/11736

[228] Mark E. Whiting, Dilrukshi Gamage, Snehalkumar (Neil) S. Gaikwad, Aaron Gilbee, Shirish Goyal, Alipta Ballav, Dinesh Majeti, Nalin Chhibber, Angela Richmond-Fuller, Freddie Vargus, Tejas Seshadri Sarma, Varshine Chandrakanthan, Teogenes Moura, Mohamed Hashim Salih, Gabriel Bayomi Tinoco Kalejaiye, Adam Ginzberg, Catherine A. Mullings, Yoni Dayan, Kristy Milland, Henrique Orefice, Jeff Regino, Sayna Parsi, Kunz Mainali, Vibhor Sehgal, Sekandar Matin, Akshansh Sinha, Rajan Vaish, and Michael S. Bernstein. 2017. Crowd Guilds: Worker-Led Reputation and Feedback on Crowdsourcing Platforms. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, New York, NY, USA, 1902–1913. https://doi.org/10.1145/2998181.2998234

[229] Mark E. Whiting, Grant Hugh, and Michael S. Bernstein. 2019. Fair Work: Crowd Work Minimum Wage with One Line of Code. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (2019), 197–206. https://doi.org/10.1609/hcomp.v7i1.5283

[230] Shomir Wilson, Florian Schaub, Frederick Liu, Kanthashree Mysore Sathyendra, Daniel Smullen, Sebastian Zimmeck, Rohan Ramanath, Peter Story, Fei Liu, Norman Sadeh, and Noah A. Smith. 2018. Analyzing Privacy Policies at Scale: From Crowdsourcing to Automated Annotations. *ACM Trans. Web* 13, 1, Article 1 (2018), 29 pages. https://doi.org/10.1145/3230665

[231] Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah A. Smith, and Frederick Liu. 2016. Crowdsourcing Annotations for Websites' Privacy Policies: Can It Really Work?. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 133–143. https://doi.org/10.1145/2872427.2883035

[232] Dietmar Winkler, Marta Sabou, Sanja Petrovic, Gisele Carneiro, Marcos Kalinowski, and Stefan Biffl. 2017. Improving Model Inspection with Crowdsourcing. In *Proceedings of the 4th International Workshop on CrowdSourcing in Software Engineering (CSI-SE '17)*. IEEE, 30–34. https://doi.org/10.1109/CSI-SE.2017.2

[233] Bård Winther, Michael Riegler, Lilian Calvet, Carsten Griwodz, and Pål Halvorsen. 2015. Why Design Matters: Crowdsourcing of Complex Tasks. In *Proceedings of the Fourth International Workshop on Crowdsourcing for Multimedia (CrowdMM '15)*. ACM, New York, NY, USA, 27–32. https://doi.org/10.1145/2810188.2810190

[234] Peng Xu and Martha Larson. 2014. Users Tagging Visual Moments: Timed Tags in Social Video. In *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia (CrowdMM '14)*. ACM, New York, NY, USA, 57–62. https://doi.org/10.1145/2660114.2660124

[235] Xiaotong Xu, Judith Fan, and Steven Dow. 2020. Schema and Metadata Guide the Collective Generation of Relevant and Diverse Work. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8, 1 (2020), 178–182. https://doi.org/10.1609/hcomp.v8i1.7479

[236] Shota Yamanaka. 2021. Utility of Crowdsourced User Experiments for Measuring the Central Tendency of User Performance to Evaluate Error-Rate Models on GUIs. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9, 1 (2021), 155–165. https://doi.org/10.1609/hcomp.v9i1.18948

[237] Huahai Yang, Yunyao Li, and Michelle X. Zhou. 2014. Understand Users' Comprehension and Preferences for Composing Information Visualizations. *ACM Trans. Comput.-Hum. Interact.* 21, 1, Article 6 (2014), 30 pages. https://doi.org/10.1145/2541288

[238] Jie Yang, Judith Redi, Gianluca Demartini, and Alessandro Bozzon. 2016. Modeling Task Complexity in Crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 4. AAAI, Palo Alto, CA, USA, 249–258. https://doi.org/10.1609/hcomp.v4i1.13283

[239] Pinar Yelmi, Hüseyin Kuşcu, and Asim Evren Yantaç. 2016. Towards a Sustainable Crowdsourced Sound Heritage Archive by Public Participation: The Soundsslike Project. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction (NordiCHI '16)*. ACM, New York, NY, USA, Article 71, 9 pages. https://doi.org/10.1145/2971485.2971492

[240] Ming Yin and Yiling Chen. 2015. Bonus or Not? Learn to Reward in Crowdsourcing. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*. AAAI, Palo Alto, CA, USA, 201–207. https://doi.org/10.5555/2832249.2832277

[241] Ming Yin, Mary L. Gray, Siddharth Suri, and Jennifer Wortman Vaughan. 2016. The Communication Network Within the Crowd. In *Proceedings of the 25th International Conference on World Wide Web (WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1293–1303. https://doi.org/10.1145/2872427.2883036

[242] Lixiu Yu, Robert E. Kraut, and Aniket Kittur. 2016. Distributed Analogical Idea Generation with Multiple Constraints. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 1236–1245. https://doi.org/10.1145/2818048.2835201

[243] Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. 2018. Predicting User Knowledge Gain in Informational Search Sessions. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. ACM, New York, NY, USA, 75–84. https://doi.org/10.1145/3209978.3210064

[244] Angie Zhang, Alexander Boltz, Chun Wei Wang, and Min Kyung Lee. 2022. Algorithmic Management Reimagined For Workers and By Workers: Centering Worker Well-Being in Gig Work. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New York, NY, USA, Article 14, 20 pages. https://doi.org/10.1145/3491102.3501866

[245] Yinglong Zhang, Jin Zhang, Matthew Lease, and Jacek Gwizdka. 2014. Multidimensional Relevance Modeling via Psychometrics and Crowdsourcing. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. ACM, New York, NY, USA, 435–444. https://doi.org/10.1145/2600428.2609577

[246] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. 2019. Dissonance Between Human and Machine Understanding. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 56 (2019), 23 pages. https://doi.org/10.1145/3359158

Received July 2023; revised October 2023; accepted November 2023