PhD Dissertation

International Doctorate School in Information and Communication Technologies

DISI - University of Trento

Strategies for addressing performance concerns and bias in designing, running, and reporting crowdsourcing experiments

Jorge Ramírez

Advisor: Prof. Fabio Casati

Co-Advisor: Prof. Luca Cernuzzi

November 2021

Abstract

Crowdsourcing involves releasing tasks on the internet for people with diverse backgrounds and skills to solve. Its adoption has come a long way, from scaling up problem solving to becoming an environment for running complex experiments. Designing tasks to obtain reliable results is not straightforward as it requires many design choices that grow with the complexity of crowdsourcing projects, often demanding multiple trial-and-error iterations to properly configure. These inherent characteristics of crowdsourcing, the complexity of the design space and heterogeneity of the crowd, set quality control as a major concern, making it an integral part of task design.

Despite all the progress and guidelines for developing effective tasks, crowdsourcing still is addressed as an "art" rather than an exact science, in part due to the challenges related to task design but also because crowdsourcing allows more complex use cases nowadays, where the support available has not yet caught up with this progress. This leaves researchers and practitioners at the forefront to often rely on intuitions instead of informed decisions. Running controlled experiments in crowdsourcing platforms is a prominent example. Despite their importance, experiments in these platforms are not yet first-class citizens, making researchers resort to building custom features to compensate for the lack of support, where pitfalls in this process may be detrimental to the experimental outcome.

In this thesis, therefore, our goal is to attend to the need of moving crowdsourcing from art to science from two perspectives that interplay with each other: providing guidance on task design through experimentation, and supporting the experimentation process itself. First, we select classification problems as a use case, given their importance and pervasive nature, and aim to bring awareness, empirical evidence, and guidance to previously unexplored task design choices to address performance concerns. And second, we also aim to make crowdsourcing accessible to researchers and practitioners from all backgrounds, reducing the requirement of in-depth knowledge of known biases in crowdsourcing platforms, experimental methods, as well as programming skills to overcome the limitations of crowdsourcing providers while running experiments.

We start by proposing task design strategies to address workers' performance, quality and time, in crowdsourced classification tasks. Then we distill the challenges associated with running controlled crowdsourcing experiments, propose coping strategies to address these challenges, and introduce solutions to help researchers report their crowdsourcing experiments, moving crowdsourcing forward to standardized reporting.

Keywords: crowdsourcing, classification, task design, crowdsourcing experiments

Acknowledgements

I am forever grateful to all the fantastic people who made it possible for me to overcome the challenges, grow, and successfully navigate this Ph.D. journey.

To Fabio, my advisor, thank you for giving me the opportunity to join your group, enabling me to explore interesting and challenging problems, and offering an amazing environment for doing research and guidance in how to approach this and succeed. Marcos, how much your guidance throughout the Ph.D. means to me can not be captured in words. You played a critical role in my growth and victories throughout this journey, and I am indebted for that. To Boualem and Luca, thanks for your collaboration and great feedback on the different projects I carried out in this Ph.D., which helped tremendously to shape our work. I also want to thank all my fellow Ph.D. and undergraduate students from the University of Trento for the collaboration, support, and long coffee chats that made this journey pleasant.

To my parents, brothers, and sister — this Ph.D. is for you too. Thank you for coping with the distance and always cheering for me, giving me the strength to keep going and growing. I would not be the person I am today without the values you all have taught me throughout my life. And to all my friends in Paraguay, thanks for your cross-continental support; it made this a fun journey.

To my father- and mother-in-law, and brother- and sister-in-law — this victory is for you as well. Thanks for rooting for me at the different milestones achieved during the Ph.D.; I am forever indebted to you for all the support. And to Coco, my dog, for being next to me in every single virtual conference and not barking his way through them.

Sofi, my wife, partner, and sun to my solar system — saying this Ph.D. is for you is an understatement. You have encouraged me to get out of my comfort zone and pursue challenges that I had thought were not mine to defeat. I treasure your support and advice, as well as our insightful conversations with our beautiful view of the Trentino mountains. Our journey has just started, and I am beyond excited for where destiny will take us. Thank you, my love.

Jorge

The work compiled in this thesis has been partially supported by funding from the EU Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 690962 and the Russian Science Foundation project No. 19-18-00282.

Contents

Chapter 1

Introduction

The adoption of crowdsourcing to solve problems has come a long way, from mainly serving as a tool to scale problem solving (e.g., label data for machine learning models [Snow et al., 2008; Liu et al., 2016]) to becoming a surrogate to complex experiments typically run in traditional laboratory settings [Paolacci et al., 2010; Mason and Watts, 2009; Schnoebelen and Kuperman, 2010; Crump et al., 2013] — however, crowdsourcing is still addressed as an "art" rather than an exact science.

Crowdsourcing involves releasing tasks on the internet for people with varying skills to solve, capitalizing on what is known as the "wisdom of crowds" — the collective performance outsmart the few individual ones. As such, crowdsourcing represents an attractive approach to practitioners and researchers as it allows to scale solutions requiring human intervention to levels where involving only experts is impractical. However, crowdsourcing is not straightforward to apply as designing proper tasks to obtain reliable results from the crowd is challenging and involves several design choices (e.g., instructions, compensation levels, allotted time to complete the task). Task design involves, besides defining the task interface, mechanisms to deploy, coordinate, collect, and curate the contributions from non-experts annotators. Despite all the progress and guidelines for developing effective tasks, getting the task right is still a trial-and-error process [Vaughan, 2017], amplified by the complexity of the design space and the heterogeneity of the crowd, people (known as workers) with diverse backgrounds, skills, and commitment levels [Gadiraju et al., 2015]. These inherent characteristics of crowdsourcing set quality control as a major concern. It is indeed an active area of research within the crowdsourcing community, where researchers propose techniques to control and assure quality, making quality control an integral part of task design (see Daniel et al. [2018] for a review).

Crowdsourcing is treated as an art not only because of the challenges associated with task design but because these challenges grow along with the complexity of crowdsourcing projects. These challenges result from the fact that crowdsourcing enables more complex use cases nowadays, but the support and guidance available has not yet caught up with this progress. This situation leaves researchers and practitioners at the forefront to often rely on their intuition instead of delivering informed decisions. Running controlled experiments in crowdsourcing platforms is a prominent example. Crowdsourcing researchers leverage experiments to, for example, improve their understanding of human behavior in crowdsourcing environments, propose new methods for obtaining quality results, and develop tools to support research and application of crowdsourcing. Despite their importance, experiments in crowdsourcing platforms are not yet first-class citizens, making researchers resort to building custom features to deploy their experiments [Mason and Suri, 2012]. These challenges may involve the selected experimental design (e.g., a between-subjects study), how this design is mapped to micro-tasks in a crowdsourcing platform, and the strategies used for sampling participants and running the experiments. Pitfalls in this process may result in the introduction of bias [Difallah et al., 2018; Qarout et al., 2019], wasted data due to invalid contributions [Kittur et al., 2008], or even deriving the wrong conclusions [Mason and Suri, 2012; Chandler et al., 2013].

The goal of this thesis is, therefore, to attend to the need of *"moving crowdsourcing from art to science"* from two perspectives that interplay with each other: providing guidance on task design through experimentation, and supporting the experimentation process itself. First, we aim to bring awareness, empirical evidence, and guidance to previously unaddressed task design choices in well-known problems. And second, we also aim to make crowdsourcing accessible to researchers and practitioners from all backgrounds, reducing the requirement of in-depth knowledge of known biases in crowdsourcing platforms, experimental methods, as well as programming skills to overcome the limitations of crowdsourcing providers while running experiments.

The approach we follow in this thesis thus unfolds into two main interconnected parts that take classification problems as a use case, aiming for results that apply to other contexts and problems. The choice of classification is given because it is a fairly popular task choice in major crowdsourcing platforms [Gadiraju et al., 2014], as well as within and beyond academic environments [Wallace et al., 2017; Wulczyn et al., 2017; Lan et al., 2017]. First, we focus on proposing previously unexplored task design choices to address performance concerns in crowdsourced classification tasks. Indeed, task design has many angles, and it is an active area of research, where explored dimensions (not only applicable to classification tasks) include, for example, worker compensation [Ho et al., 2015; Whiting et al., 2019], instructions [Wu and Quinn, 2017; Kittur et al., 2013], even worker environment [Gadiraju et al., 2017a]. As we need empirical evidence, through experimentation, to suggest what task design choices actually work, the second part of this thesis devotes its attention to addressing potential biases that could emerge from pitfalls in the design, deployment, and reporting of crowdsourcing experiments.

Providing guidance on structuring crowdsourcing tasks

As scientists, we frequently face text classification problems as part of our daily activities — a typical case being identifying documents that meet a set of conditions. In evidence-based medicine and other domains, a critical part of systematic literature reviews involve screening papers to select the subset of scientific articles that meet a set of inclusion criteria (e.g., studies involving older adults 65+ years of age living in nursing homes) [Wallace et al., 2017]. Places like Wikipedia Talk pages, where people collaborate (through text comments) to produce, discuss, and improve articles, often require administrators to moderate the discussion to avoid harmful and disrespectful behavior among contributors. Researchers typically brand this as a content moderation problem, with approaches often treating it as a classification task to identify negative comments and take some actions [Wulczyn et al., 2017]. In fact, its ubiquitous nature makes classification problems go beyond academic environments, as even simple tasks like choosing hotels matching a set of characteristics of interests can be cast as classification tasks [Lan et al., 2017]. Formally, these examples are instances of *multi-predicate classification* that aims to select items that meet a set of predicates, where the predicates correspond to desired characteristics.

A common method to solve multi-predicate classification problems is to leverage crowdsourcing [Parameswaran et al., 2012a; Park and Widom, 2013; Krivosheev et al., 2017, 2018]. In principle, the Condorcet's theorem from the 18th century states that if each member of a jury gives a correct judgment with probability higher than random ($p > 0.5$) and the votes are independent, then as the number of jurors grows the probability that the majority would make a correct decision increases (approaching 1). An interpretation of this theorem from a crowdsourcing perspective is that we can crowdsource any task and obtain quality results. This part of the thesis, therefore, focuses on multi-predicate classification problems to address task design challenges, considering the substantial share of tasks in crowdsourcing platforms that can be regarded as classification and the potential reusable knowledge that can be extracted from such problems. In fact, text classification problems are by far the most popular tasks published in major crowdsourcing platforms[1].

Specifically, this part investigates the following research question:

RQ1. How can we design the task shown to workers so as to support them and improve their individual and collective performance in crowdsourced classification tasks?

By answering this question, we aim to provide guidance on how to structure the task to obtain quality results, as well as how to render a complex question used to classify

[1] A recent survey shows that 45% of jobs in Appen, previously Figure Eight, can be regarded as classification tasks [Gadiraju et al., 2014].

items. The task shown to workers consists of reading a piece of text and answering a binary question (a predicate). Therefore, the complexity of the question comes from the fact that it is composite, involving multiple predicates that documents must satisfy.

Providing guidance on running and reporting crowdsourcing experiments

Addressing RQ1 required us to run controlled experiments in rather uncontrolled environments like crowdsourcing platforms, where gaps in the design and execution of an experiment may threaten the experimental outcome. As mentioned previously, this process involves overcoming challenges related to mapping and executing an experimental design onto a crowdsourcing platform. In this process, weaknesses in the task and experimental design may open room for inherent biases associated with crowdsourcing platforms that can ultimately threaten the results. For example, failing to properly put in place mechanisms to deal with deceiving workers could bias the population towards low-quality contributors and jeopardize the experiments as contributions may not meet quality criteria [Kittur et al., 2008]. From a deployment standpoint, the pool of active workers varies during the day. For instance, in Amazon Mechanical Turk, the majority comprises workers from India and the US [Difallah et al., 2018], available at distant time zones, and failing to account for such variability could introduce confounds that may swing the results or render the experimental treatments uncomparable [Qarout et al., 2019].

As crucial to delivering successful crowdsourcing experiments by overcoming challenges related to task and experimental design, as well as its operationalization, it is to ensure crowdsourcing experiments are reproducible. Reproducibility is the cornerstone of science [Wacharamanotham et al., 2020], and crowdsourcing research is no exception. The scrutiny of the different methodologies, by the research community, and the development of standardized protocols and methods for communicating results are critical players in the production of robust and repeatable experiments. Examples of this can be found in the medical domain where checklists are used to assess the rigor of systematic reviews [Shamseer et al., 2015], or in the machine learning community, we found checklists or datasheets to communicate in full the details underlying the production of datasets and performance of models [Bender and Friedman, 2018; Gebru et al., 2018; Mitchell et al., 2019; Arnold et al., 2019; Pineau et al., 2020]. Efforts in this context have been regarded as *repeatability*, *replicability*, or *reproducbility* to denote attempts at obtaining similar results, under certain conditions and error margins, by the same or different teams and experimental conditions [Plesser, 2018]. Setting terminology aside, it is critical that crowdsourcing experiments are communicated in acceptable levels of detail as it is fundamental to research and the scientific process.

The second part of this Ph.D. is devoted to addressing potential biases that could emerge from weaknesses in the design, deployment, and reporting of crowdsourcing experiments. Specifically, this part investigates the following research questions:

RQ2. How can we provide support to researchers in designing and running crowdsourcing experiments so as to address potential biases associated with this process?

RQ3. How can we aid researchers in reporting crowdsourcing experiments to ensure these are reproducible?

The research questions addressed by this thesis are complex, requiring more than a Ph.D. to be fully covered, and critical to advancing crowdsourcing towards science.

We aim to distill the explicit choices one has to make to build effective tasks and successful experiments, thus obtaining high-quality and reproducible results. These design and implementation choices often go unnoticed (or partially ignored) by people resorting to crowdsourcing, and this thesis also aims to address this issue by providing researchers and practitioners tools that guide them in making such choices.

Thesis structure

This thesis comprises the work carried out during the three years of the Ph.D. program at the University of Trento. The overall goal is to obtain high-quality contributions from crowd workers. First, from a task design perspective for classification tasks, devising strategies to support workers. And second, from the context of experiments, where quality also depends on the design and operationalization of the experiment, as well as how effectively researchers report the whole process to ensure reproducible research. The thesis starts by exploring task design strategies to improve the performance of workers. The focus then switches towards understanding and devising coping strategies to run controlled experiments in crowdsourcing platforms successfully. This thesis ends by proposing solutions to help researchers report their crowdsourcing experiments, thus directing crowdsourcing research towards standardized reporting.

Chapter 2 explores the impact of text highlighting on workers' performance in the context of text classification. The work is rooted in understanding to what extent highlighting relevant excerpts from the text could help workers solve the task and the potentially harmful effect that irrelevant or even deceiving parts of the text could have on workers' behavior and performance (RQ1).

Chapter 3 turns its attention to the actual question shown to workers as part of the task (RQ1). Here, we explore and provide guidance on a concrete but relevant aspect of task design for multi-predicate classification: how to ask "complex" questions to the crowd to classify a set of items.

Chapter 4 aims to understand the challenges associated with running controlled experiments in crowdsourcing platforms as well as propose coping mechanisms and tools to facilitate the job of researchers in such endeavor, addressing potential biases that could be detrimental to the validity of the experiments (RQ2).

Chapter 5 goal is to propose strategies to support researchers in reporting crowdsourcing experiments (RQ3). As a first step, this chapter aims to understand what constitutes crowdsourcing experiments. It then follows by identifying the current level of reporting in the crowdsourcing literature. These insights are leveraged to propose a checklist for crowdsourcing experiments to aid current reporting practices.

Publications

Research carried out during this Ph.D. resulted in the following publications.

Providing guidance on structuring crowdsourcing tasks

1. Jorge Ramírez, Marcos Baez, Fabio Casati, Luca Cernuzzi, Boualem Benatallah, Ekaterina A. Taran, and Veronika A. Malanina. *On the impact of predicate complexity in crowdsourced classification tasks*[2]. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM 2021).

2. Jorge Ramírez, Marcos Baez, Fabio Casati, and Boualem Benatallah. *Crowdsourced dataset to study the generation and impact of text highlighting in classification tasks.* In BMC Research Notes 12, 820 (2019).

3. Jorge Ramírez, Marcos Baez, Fabio Casati, and Boualem Benatallah. *Understanding the Impact of Text Highlighting in Crowdsourcing Tasks.* In Proceedings of the seventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2019).

4. Jorge Ramírez, Evgeny Krivosheev, Marcos Baez, Fabio Casati, and Boualem Benatallah. *CrowdRev: A platform for Crowd-based Screening of Literature Reviews.* In ACM Collective Intelligence Conference (CI 2018).

Providing guidance on running and reporting crowdsourcing experiments

5. Jorge Ramírez, Burcu Sayin, Marcos Baez, Fabio Casati, Luca Cernuzzi, Boualem Benatallah, and Gianluca Demartini. *On the State of Reporting in Crowdsourcing Experiments and a Checklist to Aid Current Practices*[3]. In Proceedings of the ACM on Human-Computer Interaction (PACM HCI), presented at the 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2021).

[2] This work was selected as an oral presentation at WSDM 2021.

[3] This work received the Methods Recognition award (intended to recognize significant methodological advances or prime examples of good methods implementation).

6. Jorge Ramírez, Marcos Baez, Fabio Casati, Luca Cernuzzi, and Boualem Benatallah. *DREC: towards a Datasheet for Reporting Experiments in Crowdsourcing*[4]. In CSCW'20 Companion: Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing (CSCW 2020).

7. Jorge Ramírez, Marcos Baez, Fabio Casati, Luca Cernuzzi, and Boualem Benatallah. *Challenges and strategies for running controlled crowdsourcing experiments.* In Proceedings of the XLVI Latin American Computing Conference (CLEI 2020).

8. Jorge Ramírez, Simone Degiacomi, Davide Zanella, Marcos Baez, Fabio Casati, and Boualem Benatallah. *CrowdHub: Extending crowdsourcing platforms for the controlled evaluation of tasks designs.* In Works-in-progress & Demonstrations track of the seventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2019).

[4] This work received the Outstanding Poster Recognition award.

Chapter 2

Understanding the impact of text highlighting in crowdsourcing tasks

Text classification is one of the most fundamental problems of machine learning (ML) projects [Aggarwal and Zhai, 2012], and also one of the most frequent human intelligence tasks in crowdsourcing platforms. It also occurs naturally in many activities we are faced in our work as scientists, such as identifying if a paper is relevant to a research topic [Wallace et al., 2017].

While ML has done impressive progress in some domains, it is still unable to accurately classify in many complex contexts. In the latter case we can resort to crowdsourcing, but this can be expensive especially when the problem is challenging or the text is long.

Recently, *hybrid* text classification algorithms, combining human computation and machine learning, have been proposed to improve accuracy and reduce costs. These techniques capitalize on the strength of humans and of machine classifiers to solve difficult tasks [Krivosheev et al., 2018; Gomes et al., 2011; Kamar et al., 2012; Cheng and Bernstein, 2015].

One way to capitalize on these complementary strengths is to have ML highlight or emphasize portions of text that it believes to be more relevant to the decision. Humans can then rely only on this text or read the entire text if the highlighted information is insufficient. Indeed, researchers in information management and psychology have shown that text highlighting can improve the reading time of humans [Wu and Yuan, 2003]. However, it can also be harmful when it is inappropriate or not relevant [Gier et al., 2009].

Previous research has explored the benefits of highlighting in: supporting workers in digitization tasks by highlighting target fields [Alagarai Sampath et al., 2014], recommending text excerpts to facilitate the job of text annotators [Wilson et al., 2016], requesting highlights as evidence to support judgments [Schaekermann et al., 2018], and as a tool to explain the output ML models [Nguyen, 2018].

In this chapter, under the first part of this thesis, we study if and under what conditions highlighting excerpts from the text can (or cannot) improve text classification cost and/or accuracy, and in general, how it affects the process and outcome of the human intelligence tasks. This is important both because highlighting can not only be a task in a two-step crowd classification procedure (highlight, then classify) but, perhaps most importantly, can also be used in hybrid classification processes where text summarization algorithms identify relevant portions of a text, thereby simplifying the subsequent (human) classification task. We do this through a series of crowdsourcing experiments running over different datasets with varying classification difficulty and document length, and with task designs imposing different cognitive demands.

Specifically, we make the following contributions:

- We are, to the best of our knowledge, the first to systematically study the effect of text highlighting in human computation, identifying the quality requirements that algorithms for text highlighting should possess to help with text classification and estimating the potential impact of good (and bad) highlighting.

- We uncover the potential of aggregating highlighting by multiple, independent annotators (or algorithms) showing that aggregation is practical and useful, somewhat analogously to what happens in a crowdsourced classification where we aggregate multiple votes on items.

- We discuss interesting and perhaps unexpected effects of highlighting, important to make them effective, such as giving time to workers to get used to working with highlights.

- We contribute an annotated dataset for researchers who want to study the problem.

2.1 Related Work

Highlighting is a common tool used to mark relevant sections in text [Strobelt et al., 2016]. The act of identifying what is important to highlight in a text have been shown useful for learning [Craik and Lockhart, 1972]. Fowler and Barker [Fowler and Barker, 1974] have shown that students had better recall of highlighted passages in a document in comparison to non-highlighted portions, after reading a document with preexisting highlighting. However, when the preexisting highlighting is inappropriate (the highlighted portions are not relevant to the content of the document), Gier et al. showed [Gier et al., 2009] that this could impair the reading comprehension. Besides understanding, studies have shown that highlighting could reduce the cognitive load, that is, the reading time [Wu and Yuan, 2003].

In crowdsourcing, researchers have used highlighting to facilitate the job of workers. Alagarai et al. [Alagarai Sampath et al., 2014] explored different variation of a form digitation task, showing that highlighting of the target fields improved the accuracy of workers. Wilson and colleagues [Wilson et al., 2016] studied the feasibility of crowdsourcing for annotating privacy policies and how automatic highlighting of relevant paragraphs can support annotators. They showed that highlighting reduces task completion time without hurting nor improving the accuracy of the annotators. Besides helping workers, highlighting have also been used to ask the crowd for evidence that support their judgement [Schaekermann et al., 2018; McDonnell et al., 2016].

In the context of interpretability of machine learning models, text highlighting have been used to present machine-generated explanations (relevant words) to humans for evaluation. In these settings, Nguyen [Nguyen, 2018] asked workers on AMT to guess the output of the model based on the text and the highlighted explanation, to determine how automatic evaluation compares to the human-level evaluation of explanations.

Researchers have shown the feasibility of non-expert annotations for NLP tasks [Snow et al., 2008], and the above works have shed some lights on the potential of highlighting as a tool for assisting workers. However, no study discusses the effects of highlighting in crowd classification, considering the quality and quantity of the highlighted text, and the behavior of workers on documents with varying difficulty. This is central to any study as it indicates how "good" highlighting needs to be to provide value to crowd classification.

2.2 Research Questions

The problem of hybrid classification via text highlighting has two sides: i) obtaining the highlighting and ii) using it in crowdsourcing tasks. In this work we focus on the latter problem, that of classifying using highlighted text support. The first problem is relevant only in terms of obtaining a rich and diverse dataset of highlighted text, that as we will see presents challenges in itself.

We set to study the impact of highlighting under different metrics, all important to crowdsourcing: the classification **accuracy** and the **decision time** (time spent on the task which, if we set the pay rate based on time, is directly related to costs). In analyzing the impact of highlighting, we focus particularly on the following research questions:

- **RQ1**. Does highlighting increase worker accuracy? Specifically, we consider three dimensions of the problem when assessing impact of highlighting: i) The *quality of highlighting*, meaning, whether the emphasised texts actually facilitates the classification tasks, is neutral, or possibly even hurts it, ii) the *difficulty* of the classification task, and iii) the *length* of the document and the proportion of highlighted text.

- **RQ2**. Does highlighting reduce decision time? And again, how is decision time impacted based on quality, difficulty and length?

Considering these factors is important because it helps us understand *how good* highlighting needs to be in order to be useful, thereby setting the bar for human computation or ML algorithms obtaining such highlighting. It also tells us for which kinds of tasks the impact may be more or less significant.

2.3 Crowdsourcing and Generating Highlights

As basis for our study of highlighting effectiveness, we obtained and assessed highlights for three datasets with different properties in terms of document length and classification accuracy, used in prior art [Krivosheev et al., 2018]. We obtain highlights from both humans and algorithms. Crowdsourced highlights allow us to obtain a wide set of highlights and highlighting patterns (e.g., individual words, full sentences) and of highlighting quality for the same text. Machine highlights, obtained via state of the art algorithms, help us assess the effectiveness of the hybrid "highlight then classify" approach to text classification that can be achieved today, as well as enabling us to assess machine highlighting quality with respect to the downstream task of efficient and accurate classification. Therefore, our focus here is not to improve ML algorithms but to assess how they perform.
Systematic Literature Review (SLR). This dataset contains a list of 900+ abstracts annotated by experts according to their relevance to an SLR. The dataset defines two relevance questions (filters): *1. SLR-OA*: *Does the paper describe a study that involves older adults (60+)?*[1], and *2. SLR-Tech*: *Does the paper describe a study that involves technology for online social interactions?* We considered each filter separately and created two datasets of 135 and 150 papers, respectively. The papers were randomly selected but controlled for the abstract length. We first excluded a long tail of outlier abstracts of length over 4000 characters, divided the remainder in three buckets of equal number of abstracts (the dividing points turned out to be 1050 and 2150 characters) and sampled an equal number of abstracts from each bucket.

For *SLR-OA* we also balanced the number of papers that described the population age explicitly vs those that refer to "older adults" or synonyms. We do so as we suspect (as it turned out) that this can impact worker behavior and performance. The distribution of the ground truth labels for *SLR-OA* is 41.5% no, 54.1% yes, 4.4% maybe; and for *SLR-Tech* is 56% no, 40% yes, 4% maybe.

[1] 60 is a commonly used age limit in scientific studies

Amazon Reviews. The dataset contains reviews about products sold on Amazon. It includes 100k items annotated with ground truth on two relevance questions, including *Is this review written on a book?*. We selected 400 reviews randomly (50% about books and 50% about other products), focusing only on short (200 reviews with < 1050 characters, but as long as at least the shortest SLR abstract that has 625 characters to make a fair comparison) and long reviews (200 reviews with > 2159 characters).

Crowd-generated highlights. To test the effects of highlighting of different quality in a controlled fashion, we ran a series of crowdsourcing tasks that requested users to classify items and highlight the reasons supporting their judgment. We do not discuss this task further as obtaining highlighting is not the focus of this work, but the interested reader can see the task description and results in the supplementary material[2]. We collected 3-7 highlighted excerpts per document and filter, totaling 2722 highlights (610 for *SLR-OA*, 616 for *SLR-Tech*, and 1496 for *Amazon*).

Two researchers assessed the quality of the highlighting provided by workers according to the following coding scheme: *bad*: the rationale could potentially lead a worker to make a wrong decision; *neutral*: it does not provide information to make a decision; *suboptimal*: it could potentially help but there are other fragments that are more suitable; *good*, it holds enough information that could help a worker in making the right decision.

The procedure for coding the quality of highlights involved both coders going over 20% of each dataset for tuning specific criteria, followed by independent coding on random splits of each dataset. Disagreements were down to a minimum and within the same usefulness class (mixing bad/neutral or good/suboptimal highlighting), the resulting Cohen's Kappa was 0.87 for SLR-OA, 0.72 for SLR-Tech and 0.66 for Amazon.

Machine-generated highlights. We generate highlights based on two approaches: state of the art algorithms for *extractive summarization* [Liu, 2019; Narayan et al., 2018], which are independent of the specific question being asked, and *question-specific* highlighting [Devlin et al., 2018]. For the first approach, we selected *BertSum* [Liu, 2019] and *Refresh* [Narayan et al., 2018]. These are recent algorithms for extractive summarization, where *BertSum* produces state-of-the-art results on a commonly used dataset. For *BertSum* we followed the training procedure with the indicated dataset, and for *Refresh* we used the available pre-trained model.

Leveraging a "generic" extractive summarization algorithm might give useful summaries but would not however be a fair comparison with crowd highlighting and probably not efficient for question-specific classification. We, therefore, chose to generate question-specific highlights by borrowing Q&A algorithms that provide answers as a subset of a text (e.g., the answer is a sentence or paragraph from a Wikipedia page that the algorithm

[2] Material available at *https: // tinyurl. com/ hcomp19-hl*

believes to contain the information necessary to answer the question). In other words, we use Q&A ML to obtain the *rationale* that can support an answer [Reddy et al., 2018], but not the answer per se which is left to the crowd, who may or may not make use of the rationale as a guide or as a way to help determine the answer more quickly.

Specifically, we leverage a fine-tuned version of BERT [Devlin et al., 2018] for question answering, or *Bert-QA* for brevity. For *Bert-QA*, we used the BERT-Base (uncased) pre-trained model and followed the fine-tuning procedure on the SQuAD dataset as indicated in the BERT paper.

Notice that the kinds of tasks we aim at covering include challenging tasks (such as SLR screening) requiring very high accuracy (SLR experts achieve over 0.96 accuracy, and the same is required – and can be obtained – by the crowd in this domain [Nguyen et al., 2015b; Mortensen et al., 2016; Krivosheev et al., 2018]). Today these are outside what machines can achieve when giving direct answers[3].

2.4 Experiment Design

We now have datasets with items of varying length and difficulty and with highlighting of different quality, corresponding to different control conditions. The basic task design to assess impact is inspired by basic screening task designs [Ramírez et al., 2018; Krivosheev et al., 2017], that have been modified to incorporate highlighting. The task is shown in Fig 2.1 for SLR-Tech and is analogous for the other datasets. Workers are presented with the text to classify, with some parts highlighted, and we mentioned that highlighting *might* (with emphasis) facilitate the classification.

The tasks were designed and run in Figure Eight (F8)[4]. This platform organises the items in a task in *pages*, where the first page acts as a test page (contains gold items only). Subsequent pages include a hidden test question to control for workers' accuracy.

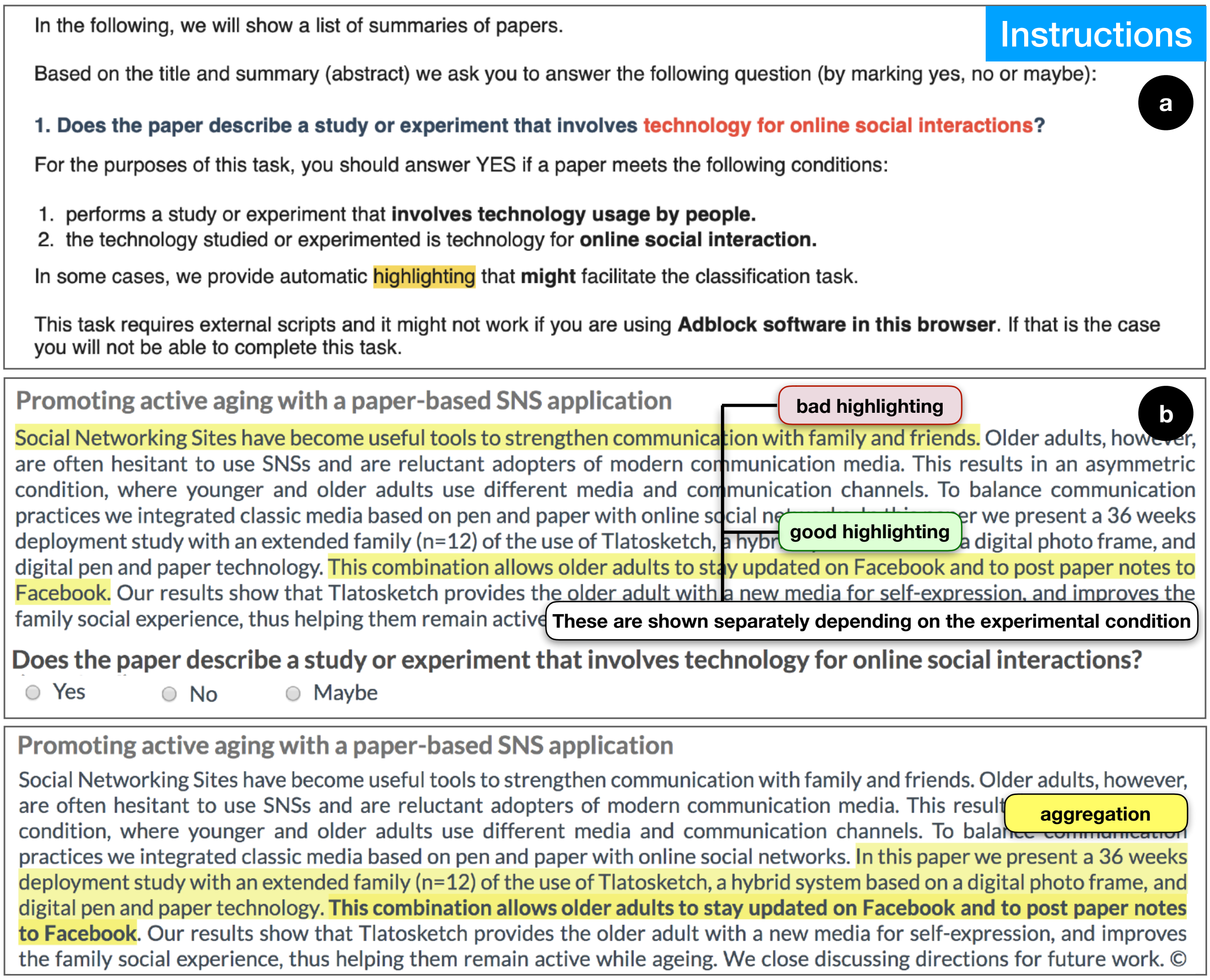
In the study we aim at observing the workers' *accuracy*, the *time to decision*, and the *retention* (how many pages a worker processes before deciding to quit) as key metrics. Notice that retention is important as dropouts make the task slower and more expensive (if a worker completes the initial tests, we are charged for the cost of test items as well, which means that we waste money if the worker abandons shortly after)[Han et al., 2019].

Given this setting, we run three main rounds of experiments with the following configurations.

Experiment 1, on the effects of highlighting of varying quality. Specifically, here we generate *six* conditions: four of them contain different proportion of abstracts or reviews

[3] see, e.g., https://rajpurkar.github.io/SQuAD-explorer/ and https://stanfordnlp.github.io/coqa/

[4] Figure Eight https://figure-eight.com



In the following, we will show a list of summaries of papers.

Based on the title and summary (abstract) we ask you to answer the following question (by marking yes, no or maybe):

1. Does the paper describe a study or experiment that involves technology for online social interactions?

For the purposes of this task, you should answer YES if a paper meets the following conditions:

1. performs a study or experiment that **involves technology usage by people.**
2. the technology studied or experimented is technology for **online social interaction.**

In some cases, we provide automatic highlighting that **might** facilitate the classification task.

This task requires external scripts and it might not work if you are using **Adblock software in this browser**. If that is the case you will not be able to complete this task.

Promoting active aging with a paper-based SNS application

Social Networking Sites have become useful tools to strengthen communication with family and friends. Older adults, however, are often hesitant to use SNSs and are reluctant adopters of modern communication media. This results in an asymmetric condition, where younger and older adults use different media and communication channels. To balance communication practices we integrated classic media based on pen and paper with online social ne er we present a 36 weeks deployment study with an extended family (n=12) of the use of Tlatosketch, a hybr a digital photo frame, and digital pen and paper technology. This combination allows older adults to stay updated on Facebook and to post paper notes to Facebook. Our results show that Tlatosketch provides the older adult with a new media for self-expression, and improves the family social experience, thus helping them remain active

Does the paper describe a study or experiment that involves technology for online social interactions?

○ Yes ○ No ○ Maybe

Promoting active aging with a paper-based SNS application

Social Networking Sites have become useful tools to strengthen communication with family and friends. Older adults, however, are often hesitant to use SNSs and are reluctant adopters of modern communication media. This result condition, where younger and older adults use different media and communication channels. To bala practices we integrated classic media based on pen and paper with online social networks. In this paper we present a 36 weeks deployment study with an extended family (n=12) of the use of Tlatosketch, a hybrid system based on a digital photo frame, and digital pen and paper technology. **This combination allows older adults to stay updated on Facebook and to post paper notes to Facebook**. Our results show that Tlatosketch provides the older adult with a new media for self-expression, and improves the family social experience, thus helping them remain active while ageing. We close discussing directions for future work. ©

Instructions
a
b
bad highlighting
good highlighting
These are shown separately depending on the experimental condition
aggregation

Figure 2.1: Task design, and example highlighting.

with "useful" highlighting (the *good* and *suboptimal* highlightings were considered as *useful*, while the *neutral* and *bad* as *not useful*). We create four conditions with 0%, *33%*, *66%*, and *100%* useful highlighting. Notice that the percentage refers to documents: for example, in the *66%* condition two out of three documents have only useful highlighting, while the third has non-useful ones. The purpose of this experimental design is to assess behaviors in situations where crowd worker could consistently trust or mistrust the highlights, as well as cases where the quality is mixed. During the qualitative assessment, the researchers generated the missing highlighting of the items (papers or reviews) with unbalanced highlights, those having only useful or not useful highlights.

In addition, we create an *aggregation condition* that fuses, for each item, all the highlighting obtained on that item. The aggregation strategy computes a score for each word in a text as the total number of highlighting that cover the word divided by the number of workers that produced these highlighting. With this score, the aggregation condition places more emphasis on words highlighted more often. If the score of a word is greater or equal than 0.33, then the word is highlighted, and the opacity value is equal to the score. If the score is at least one standard deviation away from the mean, then the word is boldfaced[5]. An example of how aggregated highlighting looks like can be seen in Fig 2.1. Finally, we add a baseline condition where items have no highlighting.

We followed a between-subject design to assign workers to one of the 0%, 33%, 66%, 100%, aggregation, and baseline condition. We defined that a worker could give a maximum of 18 judgments divided into 6 pages of 3 items each (6×3 design), and we set the accuracy threshold to be 76% and 100% for the SLR and Amazon datasets respectively. We set the payment to \$0.05 for the SLR datasets and \$0.02 for Amazon, aiming at a rate of 10USD/hour. We repeatedly ran the tasks over five weeks where each lasted between 2 to 5 days, collecting votes from F8's middle tier contributors.

Experiment 2 focused on the impact of highlighting on difficult and demanding tasks. We followed the same experimental setup as in Experiment 1, but modified the tasks to impose higher cognitive demands on the worker. We focused only on long documents and on the dataset with lower accuracy (*SLR-Tech*), and implemented two task designs: *Tech6×6*, featuring longer pages with 6 documents instead of 3, while maintaining the same number of pages; and *Tech12×3*, featuring a longer task with 12 pages, but keeping page size. We tested a condition with 83% quality highlighting (based on the promising range identified in the first experiment) against the baseline. We paid \$0.05 per item.

Experiment 3 focused on determining if the same relationships between quality of highlights and performance are observed in scenarios that rely on automatic highlighting. This experiment relies on six experimental conditions: three corresponding to the automatic highlight generation with *BertSum*, *Refresh* and *Bert-QA*, an aggregation of the output of three algorithms (*Aggregation-ML*), a condition with only high-quality highlighting from the automatic approaches *(100%ML)*, and a *baseline* without highlighting – the last two to provide a reference point for comparison. The highlight of the extractive summarization algorithms (Bertsum and Refresh) is produced by taking the top ranked sentence from the resulting summary. An additional pilot is also run considering the top three sentences as the resulting highlight, so as to assess the impact of longer highlighted text.

We followed a between-subject design to assign workers to one of the six experimental classification support conditions, and relied on the same task design (6×3), datasets (*SLR-OA*, *SLR-Tech*, *Amazon*), budget constraints and process as in Experiment 1.

Figure 2.2 shows the experimental conditions. Each task is a factorial combination

[5] During our highlighting collection experiments, we developed a visual tool to evaluate the aggregation strategy and determine the values that we end up using for opacity and boldface.

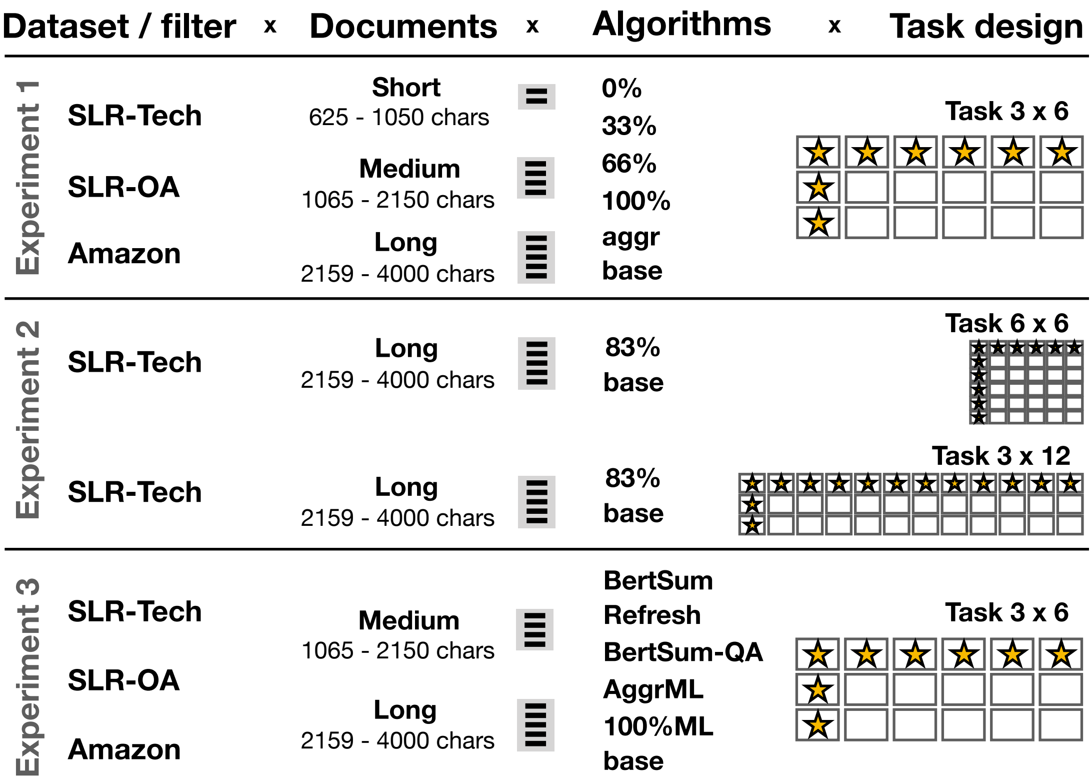


Figure 2.2: Experimental conditions

of dataset, document length and design. An external service controlled the random assignment of workers to the conditions. This allowed us to run all the conditions and baseline in parallel and reduce potential noise in the results, due to the same worker taking part in multiple tasks (something we experienced in the many preparation experiments we did, and caused us to waste some of our budget). Specifically, we implemented an external server that keeps track of the number of workers in each condition and uses this information when a new worker arrives to perform a random assignment among the conditions with the fewer number of workers (for balancing the assignment). F8 allows adding custom JavaScript code to the task interface that runs on every page load on the workers' browsers. We added code to call our external server to i) determine the condition for the worker (or retrieve a previous assignment), ii) obtain what parts to highlight for each item in the current page (unless the condition is baseline), iii) compute the decision time metric. To calculate decision time, we captured each time a worker clicks on one of the possible answers and compute the difference between the first and last values stored for each of the items[6].

To avoid workers judging items of different sizes (e.g., mixing short and long abstracts in a page) we split items in the dataset along this dimension and ran separate jobs for each size bucket respectively.

During our early pilot studies, we found that most workers came from a handful of countries. So to avoid this potential bias, we defined three geographical buckets where the head member of each bucket was one of the top three countries identified in our pilots. We ran our experiments at three different time slots (morning, afternoon and night) to orchestrate the assignment of geographical buckets to size buckets so that at any given time slot one group of countries work on one size bucket. We swapped this assignment of countries to size buckets at each time slot to make sure that one size bucket (short abstracts, for example) gets contributions from all of our target countries. This plan for running the jobs in F8 allowed us to block workers, during a particular time frame, from jumping between jobs after they finish, that is, workers that complete judging short abstracts and then continue with the long abstracts bucket (which would bias and introduce a correlation in the results).

2.5 Results

2.5.1 Experiment 1: Impact of highlighting quality

We collected a total of 14085 judgments from 1337 workers. Table 2.1 shows the distribution of these values considering the datasets. The number of workers was balanced among the experimental conditions.

Dataset	**#judgments**	**#workers**
SLR-OA	3327	424
SLR-Tech	4014	464
Amazon	6744	449

Table 2.1: Distribution of workers and judgements per dataset

[6] We captured the page load time and used it as the starting point for computing decision time of the first item of the page

Worker accuracy

The median accuracy of the workers in the baseline conditions was 0.67 for *SLR-OA* and *SLR-Tech*, and, as expected, much higher for *Amazon* (0.94). When comparing to the conditions with highlighting (see Fig. 2.3), we can see that the workers in the *100%* condition featured the same or better median accuracy (*SLR-OA*: 0.78, *SLR-Tech*: 0.67, *Amazon*: 0.94) than all the other conditions.

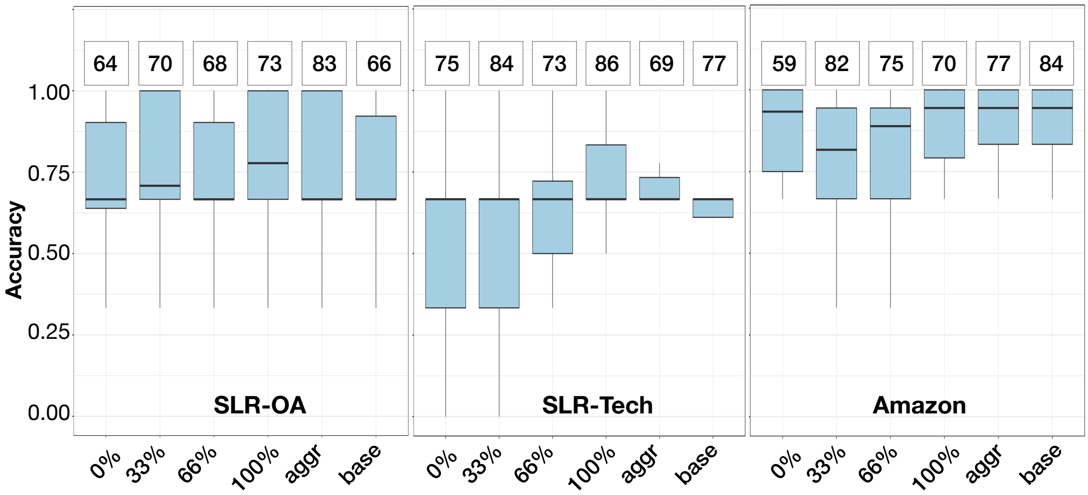


Figure 2.3: Worker accuracy boxplot (the top row shows the number of items in the condition).

A Kruskal-Wallis rank sum tests showed no significant difference for *SLR-OA* ($H(5) = 4.30$, $p = .51$), despite the trend in favor of the conditions with higher quality highlighting. In contrast, the results for *SLR-Tech* did show a statistically significant difference between the conditions ($H(5) = 12.74$, $p = .03$), but with the test of multiple comparisons [Dunn, 1964] (using Benjamini-Hochberg adjustment) indicating a significant difference only between the extremes *100%* and *0%* in favor of the former. In the *Amazon* dataset we also observed a statistically significant difference ($H(5) = 21.76$, $p < .001$), with the test of multiple comparisons showing the difference to be significant between *33%* and all the others conditions, and between *66%* and *100%* – these differences in detriment of the conditions with lower quality.

The above tell us that despite the trend in favor of the conditions with higher quality highlighting, and in particular the *100%*, *the highlighting support did not improve over the baseline. Instead, we have seen the opposite effect: bad highlighting can hurt accuracy.*

Decision time

The median decision time in the baseline conditions was 12.75s for *SLR-OA*, 32.52s for *SLR-Tech* and 15.62s for *Amazon*. Deciding whether an abstract is related to older adults required less effort than for *SLR-Tech*, we believe because the nature of former was more suitable for screening for keywords and age (e.g., "older adults", "aged 60 and older"). Surprisingly, workers took more time in screening *Amazon* reviews - a fairly easy task - than screening abstracts with the *SLR-OA* dataset.

In comparison, the best performance for the highlighting conditions improved on the baseline in all the filters (*aggr*=12.41s for *SLR-OA*; *0%*=18.51s for *SLR-tech*; *100%*=9.45s for *Amazon*). The general trend, as shown in Fig. 2.4, is that of conditions with higher-quality highlighting resulting in lower decision time, except for the curious case of *0%*, where workers achieved a performance not only better or at par with the baseline, but also with the conditions with mixed quality highlighting when considering all filters. We attribute this behavior to workers learning of the highlighting support not being useful (or being deceitful), which might have led to them dismissing the highlighted text and redirecting their attention to other parts of the document — thus having a similar effect as in the *100%* condition.

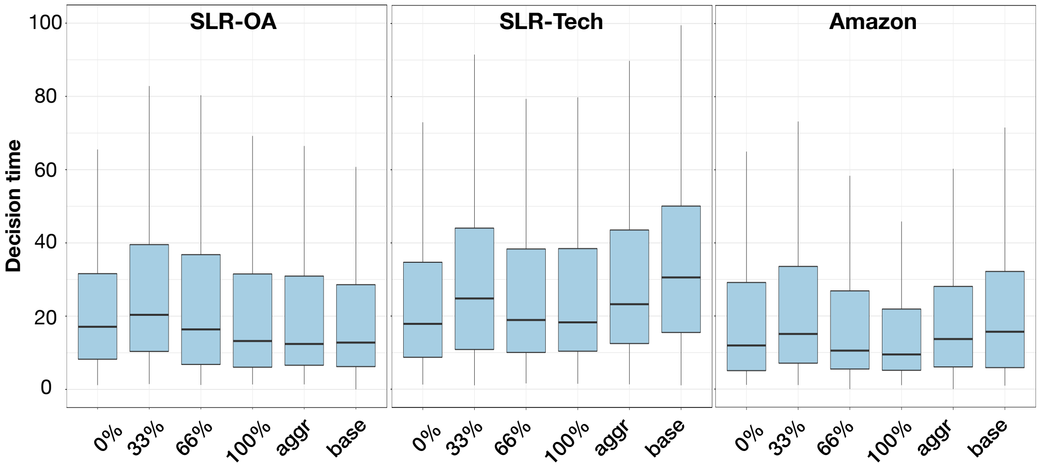


Figure 2.4: Decision time per condition.

Kruskal-Wallis rank sum tests show statistically significant difference between the conditions for all datasets ($H(5) = 43.78$, $p < .001$ for *SLR-OA*, $H(5) = 40.31$, $p < .001$ for *SLR-Tech*, and $H(5) = 50.52$, $p < .001$ for *Amazon*). Multiple comparison tests show that the *100%* condition has significantly faster decision times with respect to the baseline for *SLR-Tech* and *Amazon* and it significantly outperforms all other highlighting conditions (except for *SLR-Tech* where nearly every condition significantly outperforms the baseline). The test also confirms the curious effect of the *0%* condition outperforming the *33%* one. The detailed test results are available in the supplementary material.

Aside from the *SLR-OA* dataset, the above results indicate that good-quality highlights give an advantage to workers, reducing the time to judge. The benefit is pronounced in the 66% to 100% range, while the worst performance can be expected when mixing good highlighting with a majority of bad highlighting. This situation has proven to harm decision time more than having all documents with low quality highlighting.

2.5.2 Experiment 2: Impact in demanding tasks

In this experiment, we focused on understanding the impact of highlighting in situations of higher cognitive demand.

We collected 2481 judgements in total from 255 workers. Of these, 864 judgements from 76 workers in *Tech6×6*, and 1617 judgements from 179 workers in *Tech12×3*.

Accuracy. The median worker accuracy resulted in 0.67 for all of the designs and conditions, even though the distribution was elongated above the median for the conditions with highlighting. The Kruskal-Wallis test showed no statistically significant difference between the conditions for *Tech6×6* ($H(1) = 0.17$, $p = .68$) and *Tech12×3* ($H(1) = 0.20$, $p = .65$). Thus, accuracy did not improve in situations of higher cognitive demand.

In the case of longer pages (*Tech6×6*), what we did observe was a huge percentage of task abandonment in the first page. The majority of workers selected the job and perhaps even tried to complete it but ultimately did not submit their contributions. This happened significantly more in the baseline, where only a 23% percent of the workers assigned to the condition decided to take the task, compared to a 41% in the condition with highlighting. In our experiment, this difference in task abandonment means that highlighting can lower the perceived effort and attract more contributors, but at the same time, it also introduced a potential bias in our comparison of accuracy in attracting more committed workers in the baseline.

Decision time. The median decision time for longer pages (*Tech6×6*) resulted in 23.43s for the baseline, and 13.09s in the highlighting condition. Highlighting reduced decision time by 44% compared to no highlighting. A Kruskal-Wallis test showed the difference between the conditions to be statistically significant ($H(1) = 36.22, p < .001$). For longer tasks (*Tech12x3*) the median decision time was of 31.45s in the baseline, and 20.65s in the highlighting condition ($H(1) = 22.80$, $p < .001$). In this case, highlighting reduced decision time by 34% compared to the baseline.

2.5.3 Additional analyses

Classification performance

We computed the F_1 scores by aggregating the judgements by condition and highlighting quality as shown in Table 2.2, so as to assess and compare the output of the classification more in detail.

	Bad	**Neutral**	**Subopt**	**Good**	**All**	**None**
0%	.333	.700	-	-	-	-
33%	.178	.685	.589	.731	-	-
66%	.551	.681	.556	.725	-	-
100%	-	-	.71	.744	-	-
aggr	-	-	-	-	.727	-
base	-	-	-	-	-	.717

Table 2.2: Aggregated F_1 scores by condition and highlighting quality for *SLR-Tech*

The results show that F_1 scores for highlighting aggregation (F_1: *SLR-OA*=.845, *SLR-Tech*=.727, *Amazon*=.945) and "good" highlighting in the 66% (F_1: *SLR-OA*=.859, *SLR-Tech*=.725, *Amazon*=.953) and 100% conditions (F_1: *SLR-OA*=.845, *SLR-Tech*=.744, *Amazon*=.960) to be superior to that of the baseline (F_1: *SLR-OA*=.830, *SLR-Tech*=.717, *Amazon*=.937) for all datasets.

This suggests that by focusing on the highlighting of highest quality, the resulting classification can be superior to that of the baseline. Interestingly, *aggregating* the highlights can also result in superior classification performance, which opens up opportunities for bypassing quality annotation steps in the case of crowdsourced highlights, or using ensembles in the case of machine-generated ones. Part of the reason here is that useful highlighting as generated with the method described earlier (that is, by multiple independent annotators) outnumber non-useful ones, and aggregation enables to filter out the "noise" generated by low-quality, but more rare highlights.

Factors contributing to decision time and accuracy

We performed additional analyses to investigate how the key factors of our dimensions, such as quality and length of the highlighting, document size, worker experience (meaning number of “pages” contributed by the worker at the moment of providing the judgment), modified the impact of highlighting. We performed i) logistic regression analyses to predict *correct judgment* (true / false) and ii) multiple regression analyses to predict *decision time*, and compared the results for the baseline condition and the highlighting conditions. We include the regression analyses tables as supplementary materials. Below we summarise the main findings:

Experience with the task increases the benefits of highlighting. Experience (progression through the pages of the task) was a significant predictor of decision time, contributing to lower decision time in all three datasets in the highlighting conditions. For the baseline condition, it was significant for *SLR-OA* and *Amazon* datasets. However, in the highlighting conditions, experience also translated in workers being less likely to make mistakes, as it was a significant predictor of correct judgment, but not in the baseline. This insight suggests that experience, and possibly the amount of work given to the worker, increases the benefits of highlighting.

Workers adapt their behavior in longer documents. The size of the document was a significant predictor of decision time and correct judgements for all datasets in the highlighting models and baseline. The general insight is that workers are more likely to spend more time deciding on longer documents as well as more prone to make mistakes. However, we observed that, first, this is not the case in *SLR-Tech* where *“long” documents predict less time to judge compared to “short” documents in the highlighting conditions.* Second, judging a “long” document, despite being significant, predicts only from 1-5 seconds more in decision time than short documents, even when the length of the document is more than three times longer. Finally, deciding on “medium” documents but not on “long” documents increase the likelihood of incorrect judgements.

These results, and the *length of the highlighting* as a significant predictor, suggest that people adapt their behavior in longer documents, possibly relying more on the highlighted text, and therefore modifying the effect of highlighting.

2.5.4 Experiment 3: Impact of machine-generated highlighting

We collected a total of 8129 judgements from 1035 workers. The quality distribution of the highlights generated by the automated approaches - according to the qualitative assessment - is shown in Table 2.3 to put the results into context.

Worker accuracy

The conditions with highlight support did not significantly improve on accuracy over the baseline for any of the datasets. We should note, however, that the overall trend

	BertSum	Refresh	Bert-QA
SLR-OA	.38	.24	.49
SLR-Tech	.43	.18	.40
Amazon	.56	.62	.59

Table 2.3: Proportion of useful highlights generated.

correspond to the quality distribution of each condition, i.e., lower quality translates into a lower median accuracy or elongated tail, stressing our observation that bad highlighting affects accuracy (see Figure 2.5a).

Decision time

Highlighting support did not improve over the baseline for SLR-OA, which is in lines with our previous results for this dataset (Experiment 1), where highlighting support did not improve decision time regardless of the quality. In contrast, highlighting support did improve over the baseline for all conditions in SLR-Tech, as it was the case again in the first experiment. In the Amazon dataset, only the 100ML condition with high-quality highlighting improved over the baseline, but not the other conditions which are not in the promising zone (66%-100%) identified in the first experiment. The results are summarised in Figure 2.5b.

Classification performance

We computed the F_1 scores for the aggregated performance of each condition and dataset, as shown in Table 2.4. The low quality of the highlighting resulted in the automated approaches performing below the baseline. Improvement, in this context, was only achieved through aggregation or selecting the best highlights among the ones available. Notice that the quality of the underlying algorithms, and the space for improvement, limits the benefit of aggregation.

2.6 Discussion

The quality assessment of the machine-generated highlights provided us with insights into the nature and potential limitations of automated approaches.

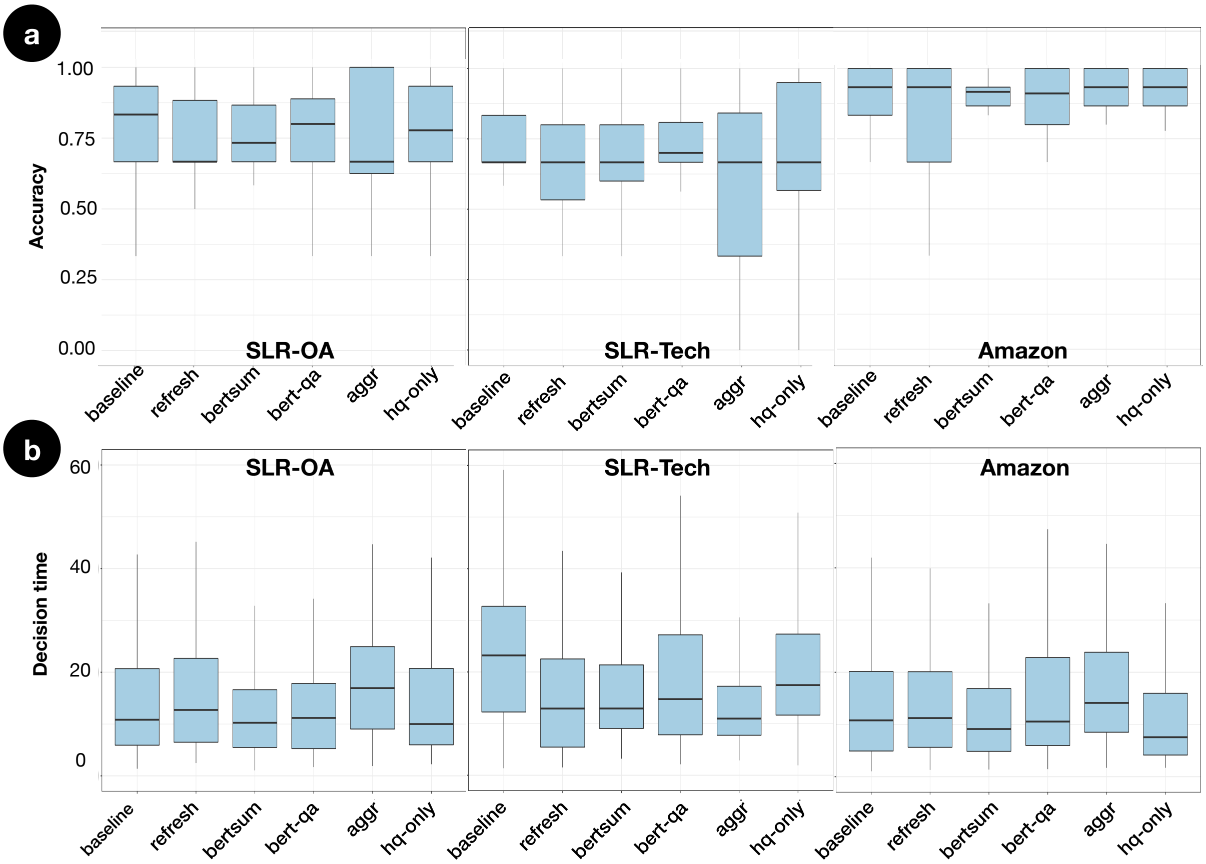


Figure 2.5: Worker accuracy per condition.

Extractive summarization approaches are not trained for a specific filter and therefore are prone to generate less useful highlights. *BertSum*, the algorithm of this class with the overall better performance, was particularly bad targeting "participants" (SLR-OA), but its performance improved when targeting the "objective" of the paper (SLR-Tech).

The Q&A-based approach, instead, generated shorter highlights specific for each dataset and resulted in overall higher quality. However, it was sensitive to how the questions were formulated, varying in the output with each attempt. *Bert-QA* also attempted to retrieve evidence for a question even when there was none. For example, if the paper is not about technology for social interaction, Bert-QA will still look for excerpts associated with these concepts, which can sometimes lead to deceiving (bad) highlights. Instead, in these cases, a counter-argument (e.g., highlighting a different focus) is desirable, or even indicating that the question is "unanswerable" (e.g., no highlighting at all).

The impression we got working with Bert-QA is that by training it specifically on the class of problems of interest (e.g., on SLRs in general), it could be possible to achieve a

	BertSum	Refresh	Bert-QA	AggrML	100ML	Base
SLR-OA	.831	.817	.842	**.860**	**.863**	.858
SLR-Tech	.684	.677	.678	.685	.712	.733
Amazon	.891	.907	.911	.918	.924	.938

Table 2.4: Aggregated F_1 scores by condition. Improvements over the baseline are highlighted.

high-quality result. Attempting this is in our work pipeline.

Besides these considerations on ML-generated highlights, the investigation into the impact of highlighting quality provided us the following main insights:

– **Bad highlighting support can hurt accuracy, while high quality offers no significant benefits**. High quality highlighting showed a positive trend in worker accuracy, improved over conditions of lower quality, but ultimately did not significantly improve over the baseline. Even when posing workers with tasks of higher cognitive demand, worker accuracy was not significantly better when providing good quality highlighting. The opposite however was consistent across all datasets: bad highlighting can hurt accuracy.

– **Higher quality highlighting can reduce decision time to almost a half**. We observed that highlighting quality in the 66% to 100% range offered significant improvements in decision time over the baseline in two of the three datasets analysed. In high demand scenarios, highlighting support can reduce the decision time by 44% compared to no highlighting, while maintaining the same level of accuracy. In a different domain, [Gaur et al., 2016] showed a similar insight, where automatic speech recognition (ASR) could facilitate workers at transcription tasks, but only when the ASR support was good enough.

– **Aggregating highlighting can increase overall classification performance**. The additional analyses also uncovered the potential of aggregating highlighting by independent annotators (or algorithms), which provided benefits analogous to that of aggregating votes in crowdsourced classification: while it did not improve on individual worker accuracy, the aggregated classification performance was superior to that of the baseline. Compared to other conditions with similar accuracy, this suggests that errors in aggregated highlighting might be more independent, an interesting effect that requires further exploration.

– **Highlighting can further decrease the decision time and perceived effort in**

high demand scenarios. The regression analyses also suggested that in higher demand scenarios (e.g., longer documents and increasing the number of contributions requested from workers) highlighting could increase its benefits. We confirmed the added benefits in terms of decision time, a reduction going from 16% up to 44% compared to the baseline for *SLR-Tech*, as well as perceived effort (lower task abandonment), but not in terms of accuracy. The difference in abandonment that we observed is in line with [Han et al., 2019], where the results on relevance judgements experiments show a similar ratio of submission to abandonment; and most of the workers tend to quit early, after a quick assessment of the effort for the tasks. [Wu and Quinn, 2017] observed a similar situation, where tasks with longer instructions showed a higher abandonment rate than more compact tasks.

- **Task difficulty does not affect the impact of highlighting**. According to our results, the impact of text highlighting on decision time was not modified by task difficulty (measured as accuracy at the baseline). The relative improvement of highlighting support in the two significant cases *SLR-Tech* (accuracy: .67) and *Amazon* (accuracy: .94) was of 60% with respect to their baselines, while for *SLR-OA* (accuracy: .67) was not significant. In the case of improvements in accuracy with respect to the baseline, the results were not significant regardless of task difficulty.

The takeaway message is that highlighting is a promising direction for text classification support, better suitable for situations where workers are faced with longer documents or are expected to provide a large number of contributions. Highlighting approaches should however consider the negative impact of bad highlighting, and use approaches that either i) limit the recommendations of highlighting to those with high level confidence (quality), or ii) aggregate the highlighting provided by independent annotators or algorithms – provided that the distribution of quality favors good highlighting or is at least balanced.

Experiments also show that highlighting support of good quality can significantly reduce the decision time by 44% while maintaining (but not necessarily increasing) worker accuracy. These benefits are elevated in situations of high cognitive demand, where workers not only see an effective decrease in decision time but also experience a lower barrier to participation. We identified the promising quality range for highlighting support, as well as the negative effects of bad highlighting, providing alternative approaches based on highlighting aggregation and quality (or confidence) level filtering. The former is a promising direction, as it can reduce the efforts in quality annotation and allow for combining the output of ensembles of algorithms. We provide the datasets used in this work in the supplementary material.

Chapter 3

On the impact of predicate complexity in crowdsourced classification tasks

Micro-task crowdsourcing today is still an art. Indeed, it is not surprising that companies charge hefty consulting fees to help businesses set up and run crowdsourcing tasks. Successful projects involve designing and harmonizing several aspects, from designing the user experience to task design, training and test settings, and to seemingly easy problems such as how to ask questions and elicit truthful, accurate answers [Daniel et al., 2018] — all while meeting budget constraints and treating your workforce fairly and with respect.

For example, longer instructions affect the task uptake by workers by three times, while showing concrete solution examples improve accuracy up to ten times [Wu and Quinn, 2017], depending on the type of task. Mechanisms to combat task spammers are often essential, since without them a task can easily get half of the answer as invalid even on simple tasks, although it raises the contributors' efforts (and cost) [Kittur et al., 2008]. Budget is also a limiting factor, and reward strategies and optimization can also affect the results [Cheng and Bernstein, 2015; Callaghan et al., 2018; Wallace et al., 2017; Krivosheev et al., 2018]. The list is almost endless, so much that it is motivating crowdsourcing researchers to prepare design and reporting guidelines for crowd experiments [Ramírez et al., 2020b].

In this chapter, we follow up on the first part of the thesis devoted to developing task design strategies contributing to improving worker individual and collective performance. In the previous chapter, we have learned that text highlights could potentially help perform the task faster without hurting quality. While this is good for settings where the focus is on speed, this chapter instead concentrates on scenarios where the focus is on quality. This chapter explores and provides guidance on a specific but important aspect of crowdsourcing task design: how to ask “complex” questions to the crowd to classify items. Classification in general is by far the most popular type of crowdsourcing tasks[1]. We study classification in the context of information retrieval and multi-predicate classification problems, that is, tasks where the crowd has to select items that meet a set of conditions. The “complexity” of the question comes therefore from the fact that it is *composite*, and we want our crowd worker to state if items satisfy our set of conditions (predicates). This is a very common task we do implicitly or explicitly countless of times in our daily life and that often appears in crowd tasks as well (from selecting hotels that have certain characteristics of interest [Lan et al., 2017] to screening papers for systematic literature reviews [Wallace et al., 2017]). Indeed, any conjunctive query is an instance of such problem and abundant prior research on crowd query processing studied how to efficiently retrieve items from a potentially large set [Franklin et al., 2011; Parameswaran et al., 2012a; Park and Widom, 2013].

We tackle this problem because it is common enough to be of widespread interest and nuanced enough (as we show in this chapter) to require a detailed investigation, and it can be framed so that it can result in reusable knowledge for task designers. In particular, we set to study the following research question:

How does the way we ask a composite question impacts the individual and aggregate performance of crowd workers?

We investigate the question both in the context of crowd-only classification and in *hybrid classification*, an increasingly common approach where humans and machines work together to solve a classification problem. We analyze the problem based on both characteristics of the question and of the task, such as task ”length” (e.g., length of the document to read for text classification tasks), task domain, task difficulty, and class balance.

Surprisingly, the crowdsourcing literature somewhat overlooked predicate complexity in classification tasks. First, complex predicates may require longer task instructions, which is known to correlate positively with the perceived complexity (as seen by workers [Yang et al., 2016]), impact task intake (most workers tend to quit after inspecting the instructions [Han et al., 2019]), and, therefore, the latency. Second, increased task complexity naturally demands more effort from workers, challenging accurate and fair compensation [Whiting et al., 2019]. Last, task complexity plays an important role in the quality of the results obtained [Cheng et al., 2015; Krause and Kizilcec, 2015].

The main contributions of this work are as follows. We introduce the problem of

[1] A relatively recent worker survey on Appen, previously Figure Eight, shows that 45% of jobs are classification tasks [Gadiraju et al., 2014]. Also, 60% of the builtin templates offered by Amazon Mechanical Turk constitute classification tasks, and 40% in Yandex Toloka.

predicate formulation for crowd classification tasks as a relevant design dimension (to enhancing worker performance and enabling Human-AI collaboration). We study complexity in classification problems on a broad landscape of tasks considering categorization and classification, verification, content moderation, and sentiment analysis tasks (see [Gadiraju et al., 2014] for a taxonomy of task types in crowdsourcing). Our experiments, therefore, cover multiple domains and leverages human and machine classifiers. We provide empirical evidence on the impact of predicate formulation on classification outcomes, suggesting performance gains when querying complex predicates as multiple simpler questions. We also provide insights into the expected performance of different formulation strategies under different i) problem settings such as predicate selectivity and class distribution, and ii) task design choices such as querying predicates on the same or separate tasks. The experiments also offer preliminary evidence on the potential of predicate formulation in the context of hybrid classification, suggesting performance gains even in its simplest collaborative approach, by assigning crowd and machines parts of a complex predicate they are more suited to classify. Last but not least, we contribute datasets derived from our experiments[2].

3.1 Related Work

Task design in crowdsourcing

Task design is a multi-dimensional problem with a rich body of work in the crowdsourcing literature [Jain et al., 2017]. "Design" does not only mean the actual task interface, but also the mechanisms to deploy, coordinate, and assign tasks to workers, the tools to assure high-quality contributions, and budget management [Daniel et al., 2018]. The lessons learned from this literature spawn on best practices for designing effective tasks (given the impact task design has on the resulting performance), and methods for performing crowdsourcing studies.

Crowdsourcing results are sensitive to subtle changes in task design. Poor instructions may lead workers to misinterpret the task and produce subpar responses [Wu and Quinn, 2017]. The clarity of the task [Gadiraju et al., 2017b] and how it is framed (whether meaningfully or not) [Chandler and Kapelner, 2013] may also swing workers' performance. The prevalence of malicious workers in platforms asks for design decisions that account for this and guard quality (e.g., equip tasks with mechanisms to combat spammers [Kittur et al., 2008]). Similarly, task design could aid worker performance, in the form of assistance to workers [Wilson et al., 2016; Ramírez et al., 2019a], proper compensation for effort-intensive tasks [Ho et al., 2015], or by rigorous training protocols [Liu et al., 2016] and

[2] `https://github.com/TrentoCrowdAI/simpler-predicates`

feedback loops [Dow et al., 2012]. Latency also matters and can be affected by ineffective instructions causing task abandonment [Han et al., 2019] or generating mistrust in task requesters [Kittur et al., 2013]. However, fair compensation can help to speed up task intake and how much workers contribute [Ho et al., 2015].

These lessons provided valuable insight into properly designing and running crowdsourcing studies. As design choices may swing the results obtained, it can also affect the validity of experimental outcomes [Kittur et al., 2008]. Choices in task design can amplify biases inherent to crowdsourcing environments. Task clarity influences how workers pick tasks and, therefore, introduce selection effects [Gadiraju et al., 2017b]. The active pool of workers varies as hours go by [Difallah et al., 2018], and with this, different decisions affecting when a crowdsourcing job runs could result in unanticipated performance differences and confounding factors [Qarout et al., 2019]. The lack of built-in support from crowdsourcing platforms makes it difficult to run controlled experiments, making simple between-subjects design a challenging endeavor [Kittur et al., 2008]. A common approach involves identifying workers via browser fingerprinting [Gadiraju and Kawase, 2017] and then using an external server to randomize participants to experimental conditions [Ramírez et al., 2019a]. These challenges motivated the research community towards developing guidelines for designing and reporting crowdsourcing experiments [Porter et al., 2020; Ramírez et al., 2020b].

Multi-predicate classification

We study predicate formulation in the context of problems regarded as *finite pool* classification [Nguyen et al., 2015a], where we have a finite set of items to classify according to a set of criteria (potentially) unique to the problem. Systematic literature reviews are one instance of this problem, and have been heavily-studied in the crowdsourcing literature [Mortensen et al., 2016; Krivosheev et al., 2017; Sun et al., 2016; Weiss, 2016]. Mortensen and colleagues [Mortensen et al., 2016] tested the feasibility of leveraging crowdsourcing, given the costs associated with producing SLRs [Wallace et al., 2017]. They found that task design plays a major role in the quality of the results, as well as this can vary from predicate to predicate. Krivoshev et al. [Krivosheev et al., 2017] proposed models and algorithms to crowdsource SLRs, offering quality and budget trade-offs to guide how to invest in the crowdsourcing tasks. Budget limits entire crowdsourced solutions, works have also focused on leveraging machine classifiers in tandem with crowd workers [Wallace et al., 2017; Krivosheev et al., 2018]. For example, leveraging strategies such as classifying "easy" items first with ML and crowd for the rest [Wallace et al., 2017] or modeling tasks and workers to determine promising predicates to filter out items.

Multi-predicate classification is also studied in the context of information retrieval. A common problem is to determine an optimal order of the predicates (to query the crowd for labels) to filter out tuples [Parameswaran et al., 2012a; Lan et al., 2017; Rekatsinas et al., 2019; Weng et al., 2017]. Similarly, work in crowd-powered databases studied how to leverage crowdsourcing to extend the capabilities of database systems to answer complex multipart queries over flexible (or on-demand) schemas [Franklin et al., 2011; Park and Widom, 2013].

Despite the vast body of work on task design and on information retrieval / multi-predicate classification, to the best of our knowledge, we are the first to study the impact on how the (complex) information retrieval question is formulated, a dimension that affects all of the prior art. Our experiments, over an ample range of tasks, emphasize the importance of the predicate formulation as a problem, and show its impact on classification outcomes.

3.2 Problem and Approach

We now define and scope the crowdsourced classification problem, and in Section 3.4, we introduce the crowd-machine variant.

The task we seek crowd help for is to identify all items in a set I that meet a complex predicate $\mathcal{P}$, defined as the conjunction of predicates $\{p_1, p_2, \ldots, p_n\}$. For example, taking a common problem from the literature (screening scientific papers [Krivosheev et al., 2017; Wallace et al., 2017]) , I could be a set of scientific articles returned by a keyword-based query on Scopus, and we may seek papers reporting experiments on older adults living in Africa ($\mathcal{P} = p_1 \wedge p_2$, where p_1: *"Is the study population 65+ years?"* and p_2: *"Is the population living in Africa?"*). To solve this problem, we have to our disposal a set of crowd workers W, a budget B, and a quality goal (or loss function) L to meet.

The predicate formulation problem seeks to determine how to ask the question in the context of multi-predicate classification. There are different ways to formulate a complex predicate, and, in this work, we study specifically three ways: i) ask the complex question (e.g., *"Is the study on 65+ years old adults living in Africa?"*), ii) break the composite question into component predicates, but ask them as part of the same task, and iii) make each predicate a task of its own (which also means that a crowd worker only sees one predicates and assesses many items for it).

We approach this problem systematically, considering both characteristics of the question and tasks. To give breadth to our analysis, we explore a broad landscape of tasks (categorization and classification, verification, content moderation, and sentiment analysis tasks) representing different domains and task difficulty levels. We focus our experiments on document retrieval (text classification) and consider documents of different lengths, given the associated effort incurred on workers to (understand and) assess text, and the potential influence of the documents' length on performance [Cheng et al., 2015; Ramírez et al., 2019a]. Our focus on text stems from the fact that it is a recurrent use case in the literature [Ho et al., 2015; Wilson et al., 2016; Krivosheev et al., 2017], and annotating images are deemed simpler in comparison to annotating texts [Krause and Kizilcec, 2015]. Finally, crowdsourcing tasks are prone to worker biases [Faltings et al., 2014], which could be caused by frequently assigning items to the same class. Therefore, we consider different class distribution scenarios to study the predicate formulation problem in crowdsourcing contexts.

3.3 Crowdsourcing Experiment

This experiment studies the impact of the task design alternatives on the performance of crowd workers. We focus on the simpler case where a complex predicate is composed of two simpler ones. We show the individual and collective performance gains related to predicate reformulation, and how the nature of the problem influences the resulting performance.

Datasets

We considered datasets with different characteristics in terms of domain, predicates, document length, and difficulty (classification accuracy), in line with prior art [Ramírez et al., 2019a; Krivosheev et al., 2018]. The datasets come from systematic literature reviews (SLRs), customer feedback analysis, content moderation and crowd verification, and are representative of multi-predicate screening problems from the literature [Krivosheev et al., 2018; Ramírez et al., 2019c; Wulczyn et al., 2017]. See the supplementary material[3] for details on the predicate composition for each of the reference datasets.

Virtual reality exergames. This dataset was produced and annotated by the authors as part of their investigation into overlaps between SLRs. We identified a pool of $80K+$ scientific articles from multiple SLRs that share some predicates. From this pool of papers, we built the *Exergame-VR* dataset that consists of 500 articles from 4 SLRs with high overlap in terms of predicates and papers within their scope. Additionally, we split the documents into two buckets based on their length: *short* (150 items with length ≤ 230 words) and *long* (350 items with length > 230 words).

Amazon product reviews. This dataset contains 100k reviews of products that are sold in Amazon [Krivosheev et al., 2018]. It is labeled according to the following two predicates: *1. Book: "Is it a book review?"*, and *2. Negative: "Is it a negative review?"*. We randomly selected 236 reviews (118 *short*, and 118 *long*) to create the *AMZ-reviews*

[3] `https://tinyurl.com/simpler-predicates-supp`

dataset.

Wikipedia detox. This dataset from Wulczyn et al. [Wulczyn et al., 2017] contains 100k comments from “Talk pages” in Wikipedia, labeled by crowd workers on whether each of the comments contains a personal attack (or an attack of another kind). From this pool of 100k items, we built *Content-Moderation*, a dataset of 118 *long* documents (comments with $>$ 230 words) labeled on two predicates.

Verifying crowd contributions. In [Ramírez et al., 2019c], the authors contribute datasets where workers provided a binary label to a relevance question, and a highlighted excerpt to justify the labeling. We built the *Verification* dataset based on [Ramírez et al., 2019c], selecting 118 *long* documents labeled according to predicates that determine whether the judgment and highlighted passage are correct. These tasks are relevant to iterative workflows, where workers act as reviewers [Little et al., 2010].

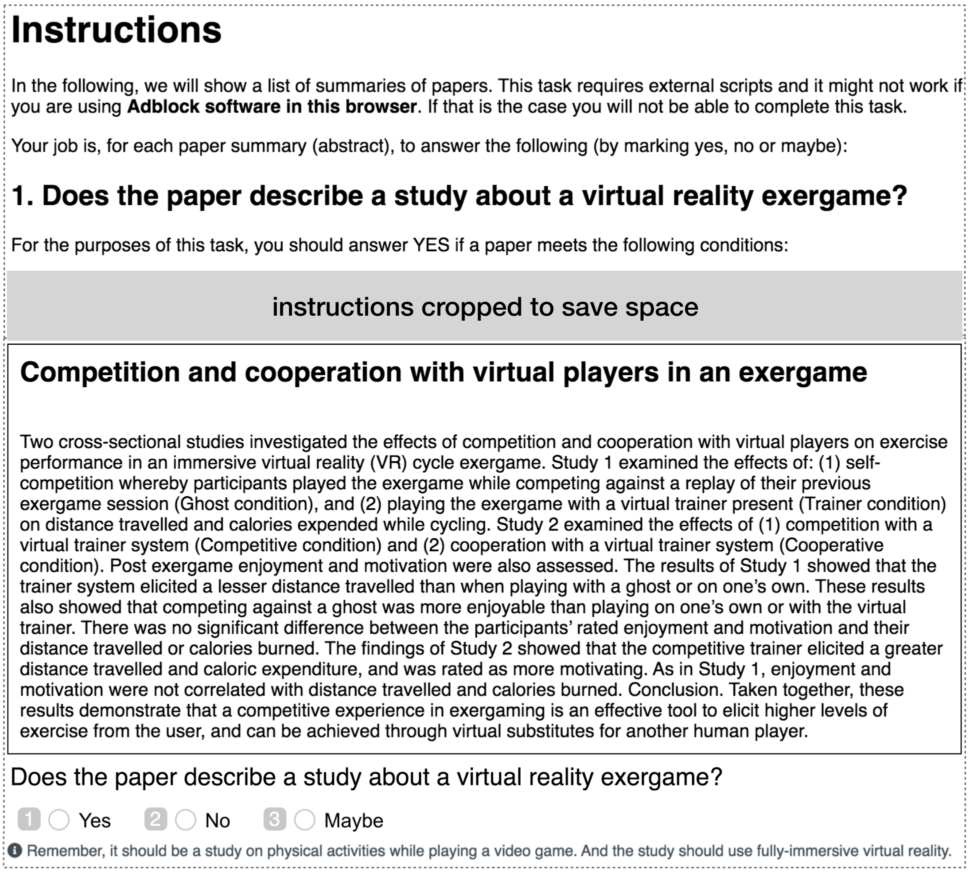
Economic inequalities in older adults. This dataset is part of an ongoing SLR on assessing the inequalities in older adults. It contains 2619 papers. From this pool of documents, we selected and labeled 151 items to build *Inequality-OA*, a dataset of *long* abstracts.

Design

The task performed by workers in our experiment consisted of reading a piece of text and answering one or two binary questions of different complexity levels depending on the task design. Figure 3.1 shows an example of a task (inspired by prior art [Krivosheev et al., 2017; Ramírez et al., 2018]).

We selected 118 items per dataset, reserving 18 for training workers (training items), and 100 for the actual task, where 34 of these items were used for quality control (control items). We consider two scenarios for the class distribution in these datasets: 60-40 and 80-20. In the *60-40* case, we selected items in each dataset according to a distribution of roughly 40% included (*IN*), 60% excluded (*OUT*). Included means that the documents satisfy all predicates $p_j \in \mathcal{P}$ for a given dataset (i.e., documents have a value of 1 for the predicates that constitute $\mathcal{P}$). Excluded documents are those that satisfy only one of the predicates or none of them. The excluded documents we distributed equally, whenever possible, between the three exclusion cases[4]. As the name suggests, the *80-20* case represents a setting with roughly 20% of items included and 80% excluded. This skewed setting tends to be problematic in crowdsourcing since it may bias workers towards the most frequent answer [Faltings et al., 2014]. For this reason, and quality control purposes,

[4] Representing the two predicates in each dataset as p_1 and p_2, the three exclusion cases are 1) $p_1 = 1$, $p_2 = 0$; 2) $p_1 = 0$, $p_2 = 1$; 3) $p_1 = 0$, $p_2 = 0$.



Instructions

In the following, we will show a list of summaries of papers. This task requires external scripts and it might not work if you are using **Adblock software in this browser**. If that is the case you will not be able to complete this task.

Your job is, for each paper summary (abstract), to answer the following (by marking yes, no or maybe):

1. Does the paper describe a study about a virtual reality exergame?

For the purposes of this task, you should answer YES if a paper meets the following conditions:

instructions cropped to save space

Competition and cooperation with virtual players in an exergame

Two cross-sectional studies investigated the effects of competition and cooperation with virtual players on exercise performance in an immersive virtual reality (VR) cycle exergame. Study 1 examined the effects of: (1) self-competition whereby participants played the exergame while competing against a replay of their previous exergame session (Ghost condition), and (2) playing the exergame with a virtual trainer present (Trainer condition) on distance travelled and calories expended while cycling. Study 2 examined the effects of (1) competition with a virtual trainer system (Competitive condition) and (2) cooperation with a virtual trainer system (Cooperative condition). Post exergame enjoyment and motivation were also assessed. The results of Study 1 showed that the trainer system elicited a lesser distance travelled than when playing with a ghost or on one's own. These results also showed that competing against a ghost was more enjoyable than playing on one's own or with the virtual trainer. There was no significant difference between the participants' rated enjoyment and motivation and their distance travelled or calories burned. The findings of Study 2 showed that the competitive trainer elicited a greater distance travelled and caloric expenditure, and was rated as more motivating. As in Study 1, enjoyment and motivation were not correlated with distance travelled and calories burned. Conclusion. Taken together, these results demonstrate that a competitive experience in exergaming is an effective tool to elicit higher levels of exercise from the user, and can be achieved through virtual substitutes for another human player.

Does the paper describe a study about a virtual reality exergame?

1 ○ Yes 2 ○ No 3 ○ Maybe

ℹ Remember, it should be a study on physical activities while playing a video game. And the study should use fully-immersive virtual reality.

Figure 3.1: The task interface used in the crowdsourcing experiments. The interface shows the complex predicate $\mathcal{P} = p_1 \wedge p_2$ for the *Exergame-VR* dataset, with p_1*: "Does the paper describe a study that uses an exergame?"*, and p_2*: "Does the paper describe a study that uses virtual reality for physical training?"*.

the training and control items follows a 30 *IN* and 70 *OUT* distribution, making sure that each page of work shows items from both classes.

We consider four experimental conditions for our crowdsourcing experiments, each condition represents a variation of the task interface shown in Figure 3.1. The *baseline* condition we use as control, and it asks workers a complex predicate $\mathcal{P}$ (a question that integrates both predicates in a dataset, as indicated in 3.1). The P_1-P_2 condition represents the task alternative that asks the constituents of $\mathcal{P}$ on the same task. The conditions P_1 and P_2 represent tasks that asks workers only one simpler predicate (predicates p_1 and p_2, respectively).

Initially, we considered both *short* and *long* documents. However, in a pilot study, we observed that the task alternatives did not improve over the *baseline* when considering *short* documents, suggesting that these may be more suitable for tasks where workers face longer documents [Wilson et al., 2016; Ramírez et al., 2019a]. Therefore, we consider only *long* documents in our study.

We followed a between-subject design, assigning workers to one of the four experimental conditions. Workers judged a maximum of 18 documents that we divided into 6 pages of 3 items per page (6 × 3 design), and we only allowed workers that understand English. We required workers to perform a training task (one page of work) before advancing to the main task, a quality control mechanism typically done in crowdsourcing research [Mitra et al., 2015]. Workers that scored 100% advanced to the main task, where we included control items as an additional quality assurance mechanism. We required workers to maintain an accuracy level of 100% for the *AMZ-reviews* dataset and 76% for the rest of the datasets[5]. We paid workers between \$0.09 and \$0.21 per page of work (depending on the condition and dataset), aiming at an hourly rate of 7.5 USD. We collected contributions from workers on the Yandex Toloka platform[6], asking 3 votes per item in the datasets. We defined a timeframe from 14:00 to 21:00 GMT+1 for running the experiment, running each dataset separately with a time gap between these. We executed each of the experimental conditions in parallel and balanced the contributions from each geographical bucket (˜33% per bucket within each condition).

We inspected the demographics of Toloka and noticed that roughly 90% of workers come from Russian-speaking countries, where Russia and Ukraine contribute the majority of the workers (∼79% and ∼10% respectively). Besides only allowing workers that understand English, we decided to create 3 geographic buckets: Russia, Ukraine, and the "Rest of the world", balancing the contributions from these buckets in our experiments to avoid any bias due to demographics.

We used an external server to assign workers to experimental conditions in a round-robin fashion, blocking workers from jumping between conditions (to avoid learning effect). We added a custom JavaScript code to the task interface to call the external server and render the experimental condition accordingly [Ramírez et al., 2019b].

3.3.1 Results

We collected a total of 8250 judgments from 1185 workers across the datasets we considered in this experiment. Here we describe our results to determine the impact of asking the complex predicate $\mathcal{P}$ vs. leveraging its simpler constituents on the classification performance of crowd workers.

[5] Prior art [Krivosheev et al., 2018] shows that the baseline performance was quite high for *AMZ-reviews*; therefore, we defined the 100% quality threshold for this dataset.

[6] `https://toloka.yandex.com/`

Worker accuracy

We use the ground-truth labels available in the datasets to determine the classification accuracy of workers in each of the experimental conditions. The median accuracy of workers in the *baseline* was 0.89 for AMZ-reviews and 0.67 for the rest (Exergame-VR, Content-moderation, Verification, and Inequality-OA), it would seem that workers found it easier to judge product reviews than documents from the rest of the domains.

We test the significance of the difference in worker accuracy using the *Kruskal-Wallis H test*[7]. The test indicates that there is a significant difference between the experimental conditions in 4 out of 5 datasets ($p < 0.05$ for Exergame-VR, and $p < 0.01$ for the rest). The results are depicted in Figure 3.2. We analyze all possible pairwise comparisons using the Dunn's Test of Multiple Comparisons [Dunn, 1964], using Benjamini-Hochberg correction to reduce the probability of Type I error. It can be noted that either P_1 or P_2 has a significant improvement over the *baseline* (3 out of 5 datasets). And that the P_1-P_2 condition significantly outperforms the *baseline* in the Content-moderation and Inequality-OA datasets.

In the skewed class distribution scenario, the *80-20* case, the median worker accuracy in the baseline condition was 0.67 for both Exergame-VR and Inequality-OA (figure omitted due to space limitations). The Kruskal-Wallis test shows no significant results between the experimental conditions (though, there is an interesting advantage of the P_1-P_2 condition where the median worker accuracy was 0.83 in both datasets).

For our predicate formulation problem, these results suggest that by asking simpler predicates instead of a complex question, we are likely to see an increase in worker accuracy in at least one of the simpler predicate. Furthermore, by asking more granular and simpler predicates we obtain valuable detailed information about crowd and task characteristics. For example, according to our results, workers were better at evaluating if a review was about a book than whether it was a negative review (simple verification vs. sentiment analysis tasks), suggesting different difficulty levels. This detailed information could equip crowd-machine algorithms to make better decisions about what to crowdsource and what to automate.

Classification performance

We analyze the results from a collective perspective and evaluate the impact of predicate formulation in the resulting classification.

The overall classification, for a given $\mathcal{P} = \{p_1, p_2\}$ in our datasets, is derived from the conjunction of the aggregated results from each of the simpler predicates (i.e., $p_1 \wedge p_2$ for each item $i \in I$). We use majority voting to aggregate the contributions from multiple workers and the F_1 score to assess the classification quality. The *baseline* condition already

[7] In our pilot study, we noticed that the observations do not follow a normal distribution

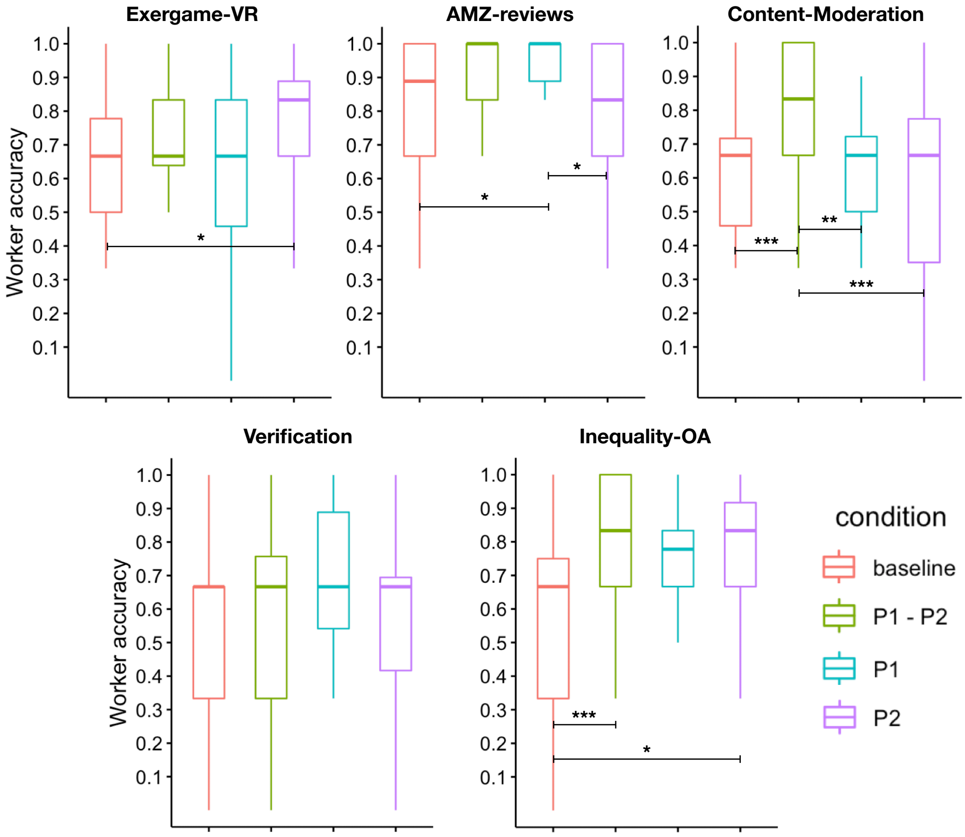


Figure 3.2: Worker accuracy by experimental condition for the *60-40* case. The lines indicate significant differences, coding p-values as *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$, ****: $p \leq 0.0001$.

provides classification on $\mathcal{P}$ since the simpler predicates are combined in a single question, and for the P_1-P_2 condition we simply take the conjunction of $\{p_1, p_2\}$. To make the results from P_1 and P_2 comparable to previous conditions, we introduce P_1 & P_2, which also takes the conjunction of each simpler predicate. To compute the F_1 score for these conditions we use as ground-truth label the conjunction of the simpler predicates.

Table 3.1 summarizes the classification performance for the experimental conditions across our five datasets for both *60-40* and *80-20* cases. The *baseline* performance ranges between 0.6 (Exergame-VR) and 0.909 (AMZ-reviews).

We compared the performance of asking $\mathcal{P}$ directly vs. asking the simpler predicates first and then combining the results (P_1-P_2 and P_1 & P_2). It can be observed that the conditions P_1-P_2 and P_1 & P_2, outperformed the baseline condition but not consistently across all datasets. P_1-P_2 outperformed the baseline in 3 out 5 datasets (Content-moderation, Verification, and Inequality-OA), with an increase in performance of up to

18%. P_1 & P_2 improved over the baseline in 2 out 5 datasets, with an increase of up to 9%. In the *80-20* case, the P_1-P_2 condition showed superior performance when compared to the *baseline*, with an increase of up to 27% (while the P_1 & P_2 fell behind the baseline).

We also compared the simpler predicates against the complex one. We observed superior classification results when formulating a composite predicate as multiple (more straightforward) questions leveraged on the same or separate tasks, even when votes are aggregated with simple majority voting. Asking two simple questions on the same task (the P_1-P_2 condition) resulted in performance gains ranging from 6% to 48%. And the conditions P_1 and P_2 that asked a simple question surpassed the *baseline* performance in all datasets, with an increase in F_1 score ranging from 2% and up to 47% In the *80-20* case the F_1 scores in the *baseline* were 0.571 for Exergame-VR, and 0.476 for Inequality-OA. In both datasets, task formulating simple predicates outperformed complex ones (when delivered separately or together on the same task), with an increase in classification performance of up to 97%.

A closer look into the performance on the complex predicates (*baseline* condition) across the two class distribution scenarios showed that overly skewed datasets may hurt the classification performance of the crowd — F_1 decreased 4% for Exergame-VR, and 31% for Inequality-OA. While by leveraging simple predicates, the classification performance could remain roughly the same, except for the unusual case of $P1$ for Inequality-OA, where the performance decreased 20%. We believe the selectivity of P_1 (see our supplementary material) played a role in this drop in performance since it is equal to 0.44 in the *60-40* version of Inequality-OA and 0.20 in the more skewed variant (also observed in the baseline).

In summary, there is evidence suggesting that complex multipart questions may benefit from disentanglement into simpler elements. As we observed, performance boosts can be obtained by formulating and presenting complex predicates as simple and more granular questions and combining back the results. However, there is no clear pattern for when each task design alternative (presenting simple predicates on the same- or separate tasks) will be the appropriate one to implement, an interesting direction for future work.

Worker effort

Although our main focus in this work is quality, we complement our analysis by looking at the impact on worker effort. We consider *decision time* as a proxy to estimate the effort incurred on workers.

The median decision time in the baseline condition was 22.66s for Exergame-VR, 33.88s for AMZ-reviews, 30.77s for Content-moderation, 23.38s for Verification, and 25.59s for Inequality-OA. In the *80-20* datasets, the median decision time in the baseline condition was 33.62s for Exergame-VR, and 33.05s for Inequality-OA.

Table 3.1: Classification performance (F_1 scores) by experimental condition from the crowdsourcing experiment.

Condition	**Exergame-VR**	**Inequality-OA**	**AMZ-reviews**	**Wiki-detox**	**Verification**
Distribution	*60-40 (80-20)*	*60-40 (80-20)*	*60-40*	*60-40*	*60-40*
Baseline	0.600 (0.571)	0.691 (0.476)	0.909	0.697	0.674
P1 - P2	0.583 (0.696)	**0.781** (0.606)	0.889	**0.825**	**0.707**
P1 & P2	**0.656** (0.629)	0.698 (0.435)	**0.947**	0.642	0.619
P1	0.819 (0.817)	0.744 (0.595)	0.981	0.838	0.889
P2	0.881 (0.853)	0.887 (0.942)	0.926	0.719	0.776

While formulating complex predicates as simpler multipart questions offer gains in quality, it results in slower task completion time. Workers in the P_1-P_2 condition spent significantly more time than the *baseline* in all datasets ($p < .01$), which intuitively makes sense since workers answered two questions rather than one (the decision time ranged between 36.77s and 53.56s). Likewise, the P_1 & P_2 condition was also significantly slower than the *baseline* ($p < .01$, with decision time between 36.96s and 45.08s)[8]. Also, there is no substantial evidence to suggest which task alternative (P_1-P_2 or P_1 & P_2) is better in terms of effort. The conditions P_1-P_2 and P_1 & P_2 had comparable results in 3 out of 5 datasets ($p > .05$), and P_1 & P_2 outperformed on the rest ($p < .05$).

Looking closer into the performance, we noticed that simpler predicates (when viewed in isolation) could potentially be faster than asking a complex predicate, but not always. When comparing P_1 and P_2 to the baseline, we noticed two competing observations. One of the simpler predicates was significantly faster than the baseline in some cases (40% faster for AMZ-reviews, 27% for Content-moderation, and 56% Verification) while significantly slower in some others (20% slower for AMZ-reviews, 55% for Verification, and 32% Inequality-OA). A similar result can also be observed in the *80-20* scenario. This suggests worker strategies such as short-circuit evaluation or focusing on simpler criteria when evaluating complex predicates, but the behavior requires further exploration.

To complement our analysis, we also explore how the predicate formulation may have influenced task intake. Overall, the percentage of workers who quit after a quick inspection of the task (during training) ranged between 18% and 73%. In particular, the task abandonment in the P_1-P_2 condition ranged between 50% and 73% (somewhat expected given that workers faced the same amount of instructions as in the baseline and had to

[8] To approximate the decision time for P_1 & P_2, we determine the median decision time (per document) for conditions P_1 and P_2 separately. Then for each document, we use the “slower” predicate as the decision time.

answer two questions rather than one). Across all datasets, either P_1 or P_2 obtained the highest task intake, aligning with the observation from the previous paragraph. To aid task intake, as task designers, we may seek to formulate a complex predicate as multiple (focused and simpler) questions and query them in isolation. Also, the instructions length should be kept in mind, in the P_1 and P_2 conditions our instructions were between 21% and 55% shorter than the baseline and P_1-P_2 . However, this suggestion demands further research, and we find it an interesting direction to explore.

Our results show that, as current literature suggests, there is a trade-off between quality and time. Besides, formulating complex predicates as multipart questions could also help identify which predicates may be more effort-intensive. Please refer to our supplementary material for a more detailed analysis.

3.3.2 Simulations

The high dimensionality of the problem makes it intractable to crowdsource for every possible parameter value. Here, we rely on simulations to assess how the performance of workers could vary under different parameterizations of the problem.

Conditions. The *baseline* task asks the complex predicate $\mathcal{P}$ directly, the *same-task* alternative queries the simpler predicates $\{p_1, \dots, p_n\}$ in one task (i.e., a worker answers n questions), and *separate-tasks* delivers these predicates on different tasks (i.e., a worker answers one of the p_j predicates). We use the terms conditions, cases and task alternatives, interchangeably.

Parameters & Metric. We parameterize the simulations based on 1) the number of simpler predicates n that constitute $\mathcal{P}$, 2) the selectivity s_j for predicates $p_j \in \mathcal{P}$, 3) the accuracy of workers drawn from a Beta distribution with mean μ and variance σ^2, 4) the budget b controlling the number of votes per item, and 5) the penalty γ that impacts the accuracy of the complex predicate $\mathcal{P}$. To assess different quality goals, we use F_β, for several values of β.

Worker accuracy. For a complex $\mathcal{P}$, the *separate-tasks* condition defines a beta distribution for each predicate with expected accuracy μ_j for $p_j \in \mathcal{P}$. The *same-task* condition defines a beta distribution with accuracy $\mu_s = \frac{1}{n}\sum u_j$. In contrast, the *baseline* defines a beta distribution with expected accuracy μ_b but adjusted based on the penalty γ.

We describe results for settings without penalty ($\gamma = 0$) and summarize the impact of γ at the end of this section, referring readers to our supplementary materials for further details on our parameterization and in-depth analysis.

Equal selectivity and accuracy. In this scenario, we define that predicates have equal selectivity s, and workers come from the same distribution. Figure 3.3 depicts the results for $n = 2$, $s = 0.5$, and for different expected accuracy values. It can be noticed than when precision and recall weight equally ($\beta = 1$), there is a difference between the task alternatives in favor of the baseline. However, the gap decreases as the accuracy of workers increases (until the conditions perform roughly the same). The same-task and separate-tasks alternatives outperform the baseline when precision is more relevant than recall ($\beta = .1$) and workers are better than random ($\mu \geq 0.6$). However, the conditions perform roughly the same when we consider higher selectivity ($s > 0.5$). The baseline outperforms the other alternatives when we value more recall ($\beta = 10$), and the difference holds as we increase the accuracy and selectivity (except for $s = 0.1$, a extremely low selectivity with high variance).

Increasing the budget (number of votes) does not affect the results in low-accuracy settings. But when accuracy is high ($\mu \geq 0.7$), the differences narrow until the conditions perform roughly the same. The number of predicates, however, harms the baseline performance, making the same- and separate-tasks superior choices for all settings.

These observations suggest that we may seek to formulate a complex predicate as a single question if we aim to optimize recall and the number of simpler predicates n is low, which intuitively makes sense and aligns with current guidelines for multi-class classification [Sabou et al., 2014]. While if we aim for precision or face scenarios with many predicates, we are better off by querying a complex predicate via its constituents. However, in the following, we assess more realistic settings (different selectivities and accuracies), and see how asking the simpler questions is preferable over $\mathcal{P}$.

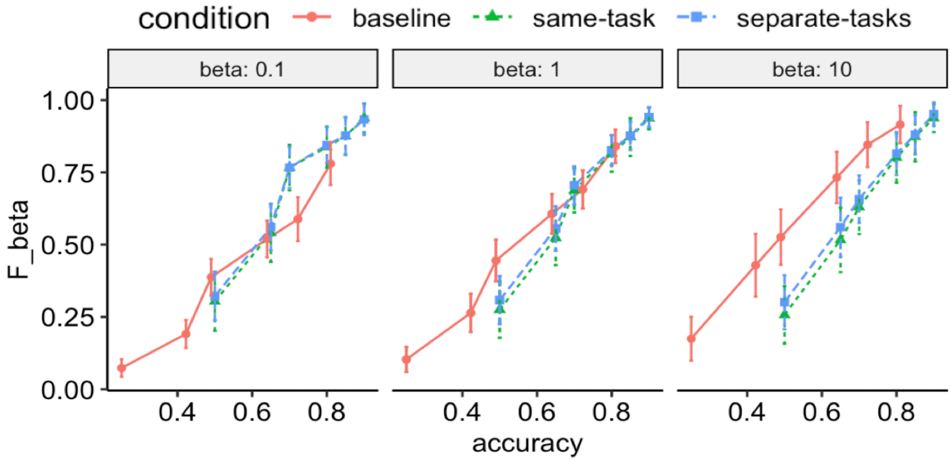


Figure 3.3: Classification performance for different accuracy values, number of predicates $n = 2$ and selectivity $s = 0.5$.

Different selectivity and same accuracy. We assign different selectivity values (either low or high) to the predicates, and we assume the same expected accuracy for the individual predicates p_j. We first considered two predicates p_1 and p_2 and two scenarios where the predicates have selectivities 1) $s_1 = 0.3$ and $s_2 = 0.7$; and 2) $s_1 = 0.7$ and $s_2 = 0.3$. We tested different accuracies $\mu \in [0.5, 0.9]$.

The results showed the same trend (figure omitted to save space) as in the simulations where we set predicates with equal selectivity and accuracy. Likewise, varying the number of predicates harmed the baseline performance, and increasing the budget narrowed the difference between the conditions when accuracy is high (until the task alternatives performed roughly equal).

Different selectivity and accuracy. Here we consider predicates with different selectivities and accuracies, a setting that aligns better with what we observed in the real crowdsourcing experiment. First, we simulated two predicates p_1 and p_2 with selectivity s_1 and s_2, and expected accuracy μ_1 and μ_2, respectively (where $s_1 \neq s_2$ and $\mu_1 \neq \mu_2$). We considered selectivity values $s_j \in \{0.3, 0.7\}$, a fixed accuracy $\mu_1 \in \{.6, .9\}$ and varied accuracy for μ_2 with $\mu_2 \in [0.6, 0.9]$. We tested all combinations combinations of selectivity and accuracy.

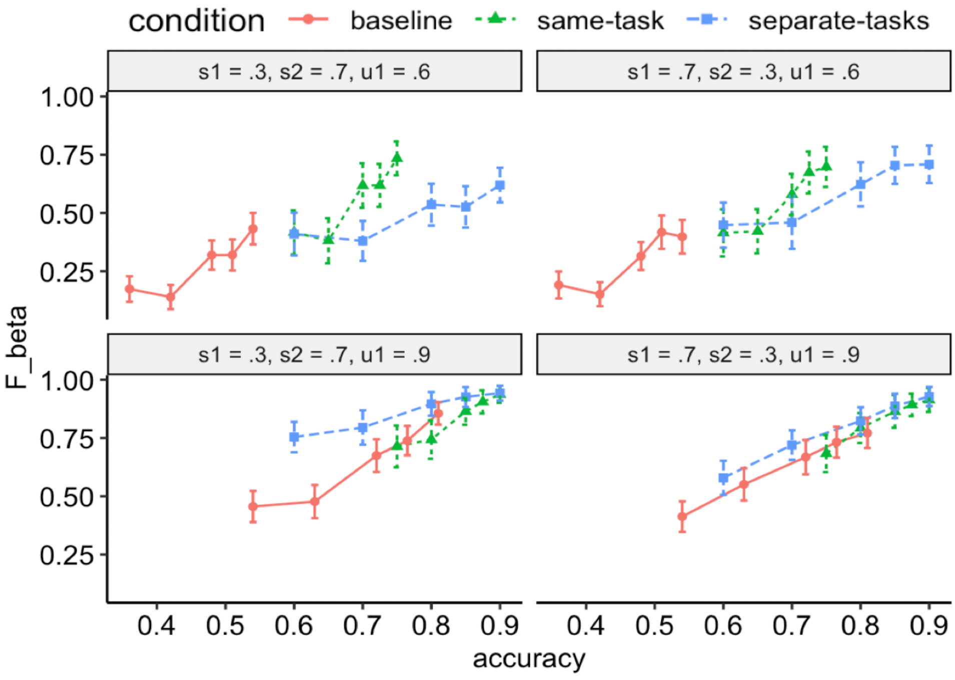


Figure 3.4: Classification performance for predicates with different selectivity and accuracy, $n = 2$ and $\beta = 1$.

Figure 3.4 shows the results for $\beta = 1$ (more details in the supplementary material). When we weight recall and precision equally, we noticed a difference in performance in favor of the same- and separate-tasks conditions (though less pronounced for $\mu_1 = .9$). Like in previous simulations, putting more weight to precision favors the same- and separate-tasks conditions. As for $\beta \geq 2$, the same- and separate-tasks condition also showed superior performance for the settings where the first predicate had an accuracy $u_1 = 0.6$. In contrast, for high-accuracy settings, $u_2 = 0.9$, the difference between the baseline and separate-task conditions narrowed until these performed roughly the same (both better than same-task).

We also considered the case of multiple predicates ($n = 4$) with different accuracies and selectivities. Like in previous simulations, a higher number of predicates hurts the baseline performance. In this setting, the same- and separate-tasks conditions outperformed the baseline across different values of budget b.

Summary. Our simulations without penalty showed how formulating a composite predicate as a single question is preferable for recall if we consider a small number of predicates with equal selectivity and accuracy. However, this is not always the case in real-world settings, where we have many predicates with different accuracy and selectivity. In these contexts, we noticed that formulating a complex predicate $\mathcal{P}$ as multiple simpler questions showed superior performance in general, which aligns with our real-world experiment. As we increase the penalty ($\gamma > 0$), the baseline tends towards 0.5 (random guessing), and naturally, the performance deteriorates, making the conditions that ask the individual predicates more suitable.

3.4 Hybrid Classification

3.4.1 Problem definition

We extend the crowdsourced classification by allowing to employ a set M of machine learning (ML) classifiers. We want to identify the items in I that meet the complex predicate $\mathcal{P}$, but we now can use ML classifiers alongside crowd workers.

To solve this problem, we now consider training ML classifiers as we cast votes from the crowd $\mathcal{W}$. The classifiers $\mathcal{M}$ can be trained for $\mathcal{P}$ directly, or for (some of) the simpler predicates $p_j \in \mathcal{P}$. Therefore, the solution space is naturally impacted by how well the ML classifiers can learn $\mathcal{P}$ or the individual constituents, and thus help in the crowdsourced classification problem.

Table 3.2: The cells correspond to F_1 scores for the best crowd performance (P_1-P_2 vs. P_1 & P_2), the best ML result on average (among single and ensemble of classifiers), and the best hybrid performance (Crowd-ML vs. ML-Crowd). The standard deviation for ML is ≤ 0.07.

Classifier	**Exergame-VR**	**Inequality-OA**	**AMZ-reviews**	**Content-moderation**
Distribution	*60-40 (80-20)*	*60-40 (80-20)*	*60-40*	*60-40*
Crowd	0.656 (0.696)	0.781 (0.606)	**0.947**	**0.825**
ML	**0.866** (0.821)	**0.853** (0.651)	0.753	0.183
Hybrid	0.775 (0.762)	0.800 (0.588)	0.931	0.485

3.4.2 Experiment

The crowdsourcing experiment showed how performance gains are obtained by querying a complex predicate as multiple (simpler) questions and then combining back the results. Here we situate the predicate formulation problem in the context of hybrid classification and test our insight from the crowdsourcing experiment. The literature suggests that hybrid classification offers superior results. And our intuition is that formulating a complex predicate as multiple questions would allow us to capitalize on the strength of crowd and ML classifiers and, therefore, obtain superior performance.

Design. We consider four datasets from the crowdsourcing experiment (excluding *Verification*), with 118 items in each dataset. The crowd judgments from the crowdsourcing experiment are aggregated using majority voting, and we combine these with machine predictions in two (simplistic) ways: Crowd-ML and ML-Crowd, where *Crowd-ML* leverages the crowd for the first predicate (p_1) and machine for the second (p_2), while *ML-Crowd* does the opposite (computing $p_1 \wedge p_2$ to derive the complex $\mathcal{P}$).

We use classifiers and ensembles of classifiers in this experiment. The four machine learning classifiers correspond to Logistic Regression (LR), Support Vector Machine (SVM), BERT [Devlin et al., 2018] and DistilBERT [Sanh et al., 2019]. The aim of covering different ML techniques is to give our analysis breadth and not to compare the models, primarily since we are operating with small datasets. The ensemble methods use LR, SVM, and Multinomial Naive Bayes (MNB) as base estimators. We considered voting classifiers ("hard", using majority voting, and "soft", using the predicted probabilities), a bagging classifier (with SVM as its base estimator), and a stacking classifier.

The models were trained on the complex $\mathcal{P}$ and its constituents. We used 10-fold stratified cross-validation, repeating the experiment 10 times (with different seeds) and reporting averages. We fine-tuned the deep learning models for 4 epochs with a learning rate of 0.001 using the AdamW optimizer [Loshchilov and Hutter, 2019]. We used an over-sampling technique [Chawla et al., 2002] to aid the LR, SVM, and MNB classifiers (alone and within an ensemble) in dealing with imbalanced classes.

3.4.3 Results

Hybrid classification, Table 3.2, showed a superior (or comparable) performance when compared to crowd classification for most of the datasets we considered (see our supplementary material for a more in-depth analysis). For the *60-40* case, the hybrid classifier outperformed the crowd for the Exergame-VR and Inequality-OA datasets (16% and 2% difference in performance, respectively). For AMZ-reviews, the performance was comparable (both classifiers with $F_1 > 0.9$) while for Content-moderation the crowd showed superior classification with a score of $F_1 = 0.82$ in comparison to only 0.48 for the hybrid approach (this was a difficult dataset in general for both crowd and ML classifiers). The hybrid classification outperformed in the *80-20* variant of the Exergame-VR dataset (9% difference in F_1), while the crowd obtained a slightly better performance for the *80-20* version of Inequality-OA (3% difference).

Hybrid classification outperformed ML for AMZ-reviews (21% difference) and Content-moderation datasets, although the hybrid performance was almost random for Content-moderation. In contrast, ML performed better for Exergame-VR and Inequality-OA datasets (11% and 6% difference, respectively), including the imbalanced variants, where the difference was at most 10%.

From a task design perspective, these results suggest that framing a complex predicate as multiple simpler questions translates into performance gains and plays nicely with recommendations from hybrid classification research. Querying a complex predicate $\mathcal{P}$ via its constituents allows for a (potentially) better coupling of crowd and machine classifiers. Our experiment showed that even this simple Human-AI collaboration approach gives a solid and consistent performance across different datasets and domains.

3.5 Discussion

Performance gains could be obtained depending on how we formulate a composite question in the context of crowdsourced and hybrid classification. From a task designer perspective, leveraging focused more straightforward questions offers more detailed information about crowd workers, and *can inform the use of different approaches more adapted to the characteristics (difficulty, selectivity) of each simpler predicate*, instead of committing to a single strategy (e.g., hiring different workers based on task difficulty [Haas et al., 2015; Retelny et al., 2017]).

Querying simpler predicates could enable *more effective coupling of ML classifiers and favor long term reusability of already trained models*. We believe that there is potential for training highly-specialized models that couple effectively with the performance of workers (instead of learning models classify items based on complex predicates directly). Besides, answering simpler questions outputs reusable (and detailed) knowledge about the capabilities of crowd and machine classifiers. For example, if we were to work on an SLR about *exergame usage in older adults*, we could rely on the current knowledge that we have built by querying the simpler predicates from the Exergame-VR and Inequality-OA datasets. From the perspective of crowd workers, this means reapplying learned skills, and for machines, it involves classifying unseen papers (and filter out at least articles that are “obviously” not relevant).

We focused on a specific but relevant aspect for task designers: how to frame a composite question used to classify items. Our results showed that superior classification performance could be obtained by querying a complex predicate as multiple (simpler) questions instead of asking a single coarse predicate. In a scenario with low accuracy and selectivity, asking the constituents of $\mathcal{P}$ (i.e., n questions) may increase the chances of misclassifying items, as observed in our simulations. In this case, we may rely on framing the complex $\mathcal{P}$ as a single question (limited by the number of predicates it contains) or framing $\mathcal{P}$ as a mix of simpler and coarse questions. To some extent, our competing results from either asking predicates on the same task vs. on separate tasks is related to this point (i.e., the error rate of a single worker answering n questions vs. n workers answering a question each). Both task design choices offer superior results over the baseline, but there is not enough evidence to inform decisions based on given problem settings. We find this an interesting direction of future work, where we design algorithms that model workers, tasks, and predicates to automatically learn how to formulate complex predicates to meet quality goals while operating under a budget.

Chapter 4

Challenges and strategies for running controlled crowdsourcing experiments

This chapter introduces the second part of this thesis, where we focus on providing support to run and report experiments in crowdsourcing platforms. This shift in focus is motivated by the lessons we have learned in developing task design strategies for addressing performance concerns in crowdsourced classification tasks, which required us to run controlled experiments to provide empirical evidence on the approaches we proposed in Chapters 2 and 3. Moreover, we aim to make crowdsourcing more accessible to researchers and practitioners, offloading the need for in-depth knowledge of the inherent characteristics of crowdsourcing platforms and programming skills to make controlled experiments possible.

A crucial aspect in running a successful crowdsourcing project is identifying an appropriate task design [Jain et al., 2017], typically consisting of trial-and-error cycles. Task design goes beyond defining the actual task interface, involving the deployment, collection, and mechanisms for assuring the contributions meet quality objectives [Daniel et al., 2018].

The design of a task represents a multi-dimensional challenge. The instructions are vital for communicating the needs of requesters since poorly defined guidelines could affect the quality of the contributions [Wu and Quinn, 2017; Kittur et al., 2013; Gadiraju et al., 2017b; Liu et al., 2016], as well as task acceptance [Schulze et al., 2011]. Moreover, enriched interfaces could also help workers in performing tasks faster, and with results of potentially higher-quality [Sampath et al., 2014; Wilson et al., 2016; Ramírez et al., 2019a]. Accurate task pricing is also a relevant aspect since workers are paid for their contributions [Whiting et al., 2019], representing an incentive mechanism that impacts the

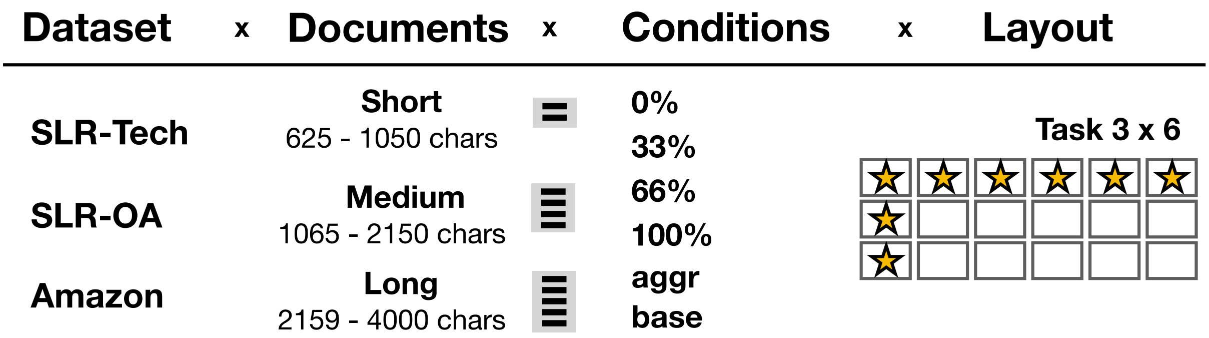


Figure 4.1: A summary of the experimental design used to discuss the challenges of running controlled crowdsourcing experiments. We use this experiment as our running example, and its goal is to study the impact of text highlighting in crowdsourcing tasks. In this case, the experiment uses a between-subjects design and considers datasets from multiple domains with documents of varying sizes, six experimental conditions, and tasks organized in pages. The figure is adapted from [Ramírez et al., 2019a].

number of worker contributions [Mason and Watts, 2009], as well as the quality, especially for demanding tasks [Ho et al., 2015]. The time workers are allowed to spend on a task can also affect the quality of the contributions [Maddalena et al., 2016; Krishna et al., 2016], and even characteristics of the crowd marketplace and work environment [Gadiraju et al., 2017a]. These insights constitute guidelines for articulating effective task designs and highlight the feasibility of running controlled experiments in crowdsourcing platforms.

Over the years, a vast body of work grew the scope of crowdsourcing [Paolacci et al., 2010; Buhrmester et al., 2011; Mason and Watts, 2009; Schnoebelen and Kuperman, 2010; Sun and Stolee, 2016; Crump et al., 2013], expanding its application beyond serving as a tool to create machine learning datasets [Snow et al., 2008; Liu et al., 2016]. Paid crowd work thus establishes as a mechanism for running user studies [Kittur et al., 2008; Buhrmester et al., 2011; Sun and Stolee, 2016], complex work that initially does not fit in the microtask market [Kittur et al., 2011; Ahmad et al., 2011; Kulkarni et al., 2012], and experiments beyond task design evaluation [Crump et al., 2013; Paolacci et al., 2010; Mason and Watts, 2009; Schnoebelen and Kuperman, 2010]. Naturally, the set of challenges also increases along with the scope and ambition of crowdsourcing projects, especially for crowdsourcing experiments.

Figure 4.1 depicts the study we use as our running example throughout this chapter to describe the challenges and strategies for running controlled experiments in crowdsourcing platforms. The goal of this project was to understand *if*, and *under what conditions*, highlighting text excerpts relevant to a given relevance question would improve worker performance [Ramírez et al., 2019a; Ramírez et al., 2019c]. This required testing different highlighting conditions (of varying quality) against a baseline without highlighting, given different document sizes and datasets of different characteristics. The resulting experimental design featured a combination of *dataset* (3) x *document size* (3) x *highlighting conditions* (6) — a total of 54 configurations.

The potential size of the design space, along with the individual and environmental biases [Barbosa and Chen, 2019; Balahur et al., 2010; Cheng and Cosley, 2013; Eickhoff, 2018; Nguyen et al., 2014; Sen et al., 2015], and the limitations of crowdsourcing platforms [Qarout et al., 2019; Paritosh, 2012], makes it difficult to run controlled crowdsourcing experiments. This means that researchers need to deal with the challenging task of mapping their study designs as simple tasks, managing the recruitment and verification of subjects, controlling for the assignment of subjects to tasks, the dependency between tasks, and controlling for the different inherent biases in experimental research. This requires in-depth knowledge of experimental methods, known biases in crowdsourcing platforms, and programming using the extension mechanisms provided by crowdsourcing platforms.

Current systems that extend crowdsourcing platforms focus on specific domains [Bernstein et al., 2010; Ramírez et al., 2018; Correia et al., 2018; Franklin et al., 2011] or kinds of problems that split into interconnected components [Kittur et al., 2011; Ahmad et al., 2011; Kulkarni et al., 2012]. The general purpose tooling available to task requesters comes in the form of extensions to (or frameworks build on top of) programming languages [Little et al., 2010; Minder and Bernstein, 2012; Barowy et al., 2012], which could potentially demand considerable work or lock the requester to a specific crowdsourcing platform.

Contributions. First, this chapter describes the challenges and strategies for running controlled crowdsourcing experiments, as a result of the lessons we have learned while running experiments in crowdsourcing platforms. And second, we introduce CrowdHub[1], a web-based platform for running controlled crowdsourcing projects. CrowdHub blends the flexibility from programming with requester productivity, offering a diagramming interface to design and run crowdsourcing projects. It offers features for systematically evaluating task design to aid researchers and practitioners during the design and deployment of crowdsourcing projects across multiple platforms, as well as features for researchers to run controlled experiments.

4.1 Related Work

Crowdsourcing platforms such as Amazon Mechanical Turk, Figure Eight, or Yandex Toloka, expose low-level APIs for common features associated with publishing a task to a pool of online workers. These features involve creating a task with a given template,

[1] `https://github.com/TrentoCrowdAI/crowdhub-web`

uploading data units, submitting and keeping track of the progress, rewarding workers, defining quality control mechanisms, among others. Naturally, exposing APIs open the room for additional extensions to the feature space, and we describe the related literature on technologies that extend the capabilities of crowdsourcing platforms. We identify that these technologies can be roughly categorized in domain-specific tooling and general-purpose platforms.

Domain-specific tools. Soylent [Bernstein et al., 2010] is a word processor that offers three core functionalities that allow requesters to ask crowdsourcing workers to edit, proofread, or perform an arbitrary task related to editing text documents. Soylent articulated and implemented the idea that crowdsourcing could be embedded in interactive interfaces and support requesters in solving complex tasks. Tools to support researchers in performing systematic literature reviews (SLRs) have also been developed. CrowdRev [Ramírez et al., 2018] is a platform that allows users (practitioners and human computation researchers) to crowdsource the screening step of systematic reviews. Practitioners can leverage crowd workers via an easy-to-use web interface, leaving CrowdRev in charge of dealing with the intricacies of crowdsourcing. For human computation researchers, however, CrowdRev offers the flexibility to tune (and customize) the algorithms involved in the crowdsourcing process (i.e., querying and aggregation strategies). SciCrowd [Correia et al., 2018] is a system that embeds crowdsourcing workers in a continuous information extraction pipeline where human and machine learning models collaborates to extract relevant information from scientific documents.

Crowdsourcing databases extend the capabilities of database systems to allow answering queries via crowd workers to aid data cleansing pipelines. CrowdDB [Franklin et al., 2011] extends the standard SQL by introducing crowd-specific operators and extensions to the data definition language (DDL) that the query engine can interpret and spawn crowdsourcing tasks accordingly. CrowdDB manages the details related to publishing crowdsourcing tasks, automatically generating task interfaces based on the metadata specified by developers using the extended DDL. Several other declarative approaches were proposed to embed crowdsourcing capabilities into query processing systems [Parameswaran et al., 2012b; Demartini et al., 2013; Marcus et al., 2011; Morishima et al., 2012].

General-purpose platforms. Several approaches have been proposed to manage complex problems that partition into interdependent tasks. Inspired by the MapReduce programming paradigm [Dean and Ghemawat, 2004], CrowdForge [Kittur et al., 2011] offers a framework to allow solving complex problems via a combination of partition, map and reduce tasks. Jabberwocky [Ahmad et al., 2011] is a social computing stack that offers three core components to tackle complex (and potentially interdependent) tasks. Dormouse represents the foundations of Jabberwocky, and it acts as the runtime that can

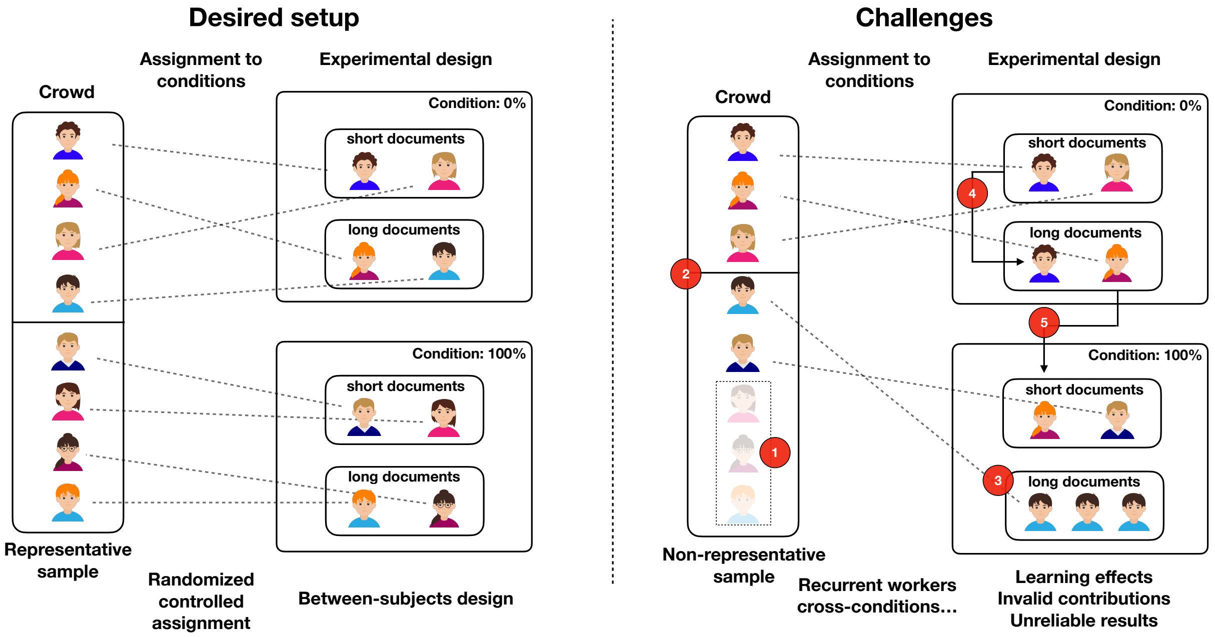


Figure 4.2: The challenges that arise while running controlled crowdsourcing experiments.

process human (and hybrid) computation, offering for cross-platform capabilities (e.g., going beyond Amazon Mechanical Turk). The ManReduce layer is similar to CrowdForge but implemented as a framework for the Ruby programming language, allowing map and reduce steps to be executed by either crowd or machines. Finally, Jabberwocky offers a high-level procedural language called Dog that compiles down to ManReduce programs. Turkomatic [Kulkarni et al., 2012], unlike previous approaches, allows the crowd to play an active role in decomposing the problem into the set of interdependent components. Turkomatic operates using a divide-and-conquer loop where workers actively refine the input problem (supervised by requesters) into subtasks that run on AMT, where a set of generic task templates are instantiated accordingly.

Turkit [Little et al., 2010] is a JavaScript programming toolkit for implementing human computation algorithms that run on Amazon Mechanical Turk. It offers functions that interface with AMT and introduce the crash-and-rerun programming model for building robust crowdsourcing scripts. CrowdLang [Minder and Bernstein, 2012] is a framework and programming language for human and machine computation, relying on three core components to allow requesters to implement complex crowdsourcing workflows. First, a programming library that encapsulates operators and reusable computation workflows (e.g., find-fix-verify [Bernstein et al., 2010], iterative improvements [Little et al., 2010]). Then, the engine component orchestrates and run human computation algorithms (and deals with the underlying challenges). And last, the integration layer bridges CrowdLang with distinct crowdsourcing platforms for cross-platform support. Similar to Turkit, AUTOMAN [Barowy et al., 2012] embeds crowdsourcing capabilities into a programming language (Scala in this case). However, it also offers features to automatically manage pricing, quality control, and scheduling of crowdsourcing tasks.

Substantial efforts have been devoted to offering solutions that sit on top of crowdsourcing platforms. However, these tools still demand considerable work from requesters (e.g., programming the complete workflows or experiments). For running controlled experiments, we find the TurkServer platform [Mao et al., 2012] to be closely related to our work. TurkServer is based on a JavaScript web framework and offers builtin features that enable researchers to run experiments on AMT. Even though programming frameworks give full flexibility to researchers, CrowdHub aims to offer a paradigm that blends flexibility with productivity. Concretely, we propose an easy-to-use platform that allows task requesters to focus on designing crowdsourcing workflows via a diagramming interface, reducing the programming efforts that current tooling would require.

4.2 Challenges in Evaluating Task Designs

In this section, we describe the challenges that arise when running crowdsourcing experiments. These challenges are derived from our own experiences in evaluating task designs while studying the impact of highlighting support in text classification [Ramírez et al., 2019a; Ramírez et al., 2019c].

We return to the study we use as our running example (see Figure 4.1) to describe the challenges while running controlled crowdsourcing experiments. The datasets used in this experiment come from two domains: systematic literature reviews (SLR) and Amazon product reviews. Using the number of characters as a proxy for document length, we categorized the documents as short, medium, or long. The workers that participated in the experiment performed the task in pages (up to six), where each page showed three documents with one item used for quality control[Daniel et al., 2018], except the first page which was used entirely for quality control. Each document in the datasets was associated with a list of text excerpts of varying quality. The highlighting conditions 0%-100% indicate the proportion of items in a given page that will highlight text excerpts of good quality (0% means non-useful highlights). The baseline was used as the control condition (no highlights), and the *aggr* condition first aggregated the available excerpts and then highlighted the result.

The left half of Figure 4.2 depicts the desired experimental setup, a between-subjects design. A representative sample of workers from the selected crowdsourcing platform is randomly assigned to one of the experimental conditions, assuring a well-balanced distribution of participants. Also, within each of the conditions, workers are restricted to assess only documents of a specific size. However, running this experiment on crowdsourcing platforms is far from being a straightforward task. Researchers need to put a lot of effort into overcoming the limitations of crowdsourcing platforms and successfully executing the desired experimental design. And in this process, challenges emerge that could hurt the outcome and validity of the experiment, as depicted in the right half of Figure 4.2.

In order to identify the challenges and quantify potential experimental biases in running an *uncontrolled* evaluation of task designs, we created individual tasks in Figure Eight for a subset (1 dataset) of the experimental conditions. We ran the tasks one after another, collecting a total of 6993 votes from 631 workers (16 tasks). In the following, we lay out the challenges that we encountered during the process (Figure 4.2).

Platforms lack native support for experiments. Crowdsourcing platforms such as Figure Eight (F8), Amazon Mechanical Turk (AMT), and Toloka offer the building blocks to design and run crowdsourcing tasks. In F8, for example, this implies defining i) data units to classify, ii) gold data to use for quality control, iii) task design, including instructions, data to collect, assignment of units to workers, iv) the target population (country, channels, trust), and v) the cost per worker contribution. F8 then manages the assignment of data units, the data collection and the computation of basic completion metrics. These features are suitable for running individual tasks, but less so when experimenting with different task designs with a limited pool of workers, where special care must be taken to run even simple between-subject designs [Kittur et al., 2008].

This lack of support left researchers with the laborious job of actually implementing the necessary mechanisms to deploy a controlled experiment. For our running example, this means that researchers need to create the tasks for each of the experimental conditions and document sizes (6 conditions and 3 document sizes, a total of 18 tasks). Eligibility control mechanisms are crucial to identify workers and randomly assign them to one of the tasks, controlling that workers only participate once. Besides, during deployment, researchers need to constantly monitor the progress of the crowdsourcing task to avoid potential demographic biases presence in the crowdsourcing platforms, assuring that a well-balanced and representative sample of the population participates in the experiment. Ultimately, this means that researchers need deep knowledge in both programming and experimental methods to implement the necessary mechanisms and controlling inherent biases in experimental research and crowdsourcing platforms.

Timezones ❶. A wide range of countries constitutes the population of workers in crowdsourcing platforms. However, the majority of workers tend to come from a handful of countries instead. The pool of workers that can participate in a task varies at different times of the day since workers come from a diverse set of countries with different timezones. For example, the population of active US workers could be at its pick while workers from

India are just starting the day, as shown by Difallah, Filatova, and Ipeirotis for the AMT platform [Difallah et al., 2018].

Running experiments without considering the mismatch of worker availability during the day could introduce confounding factors that hurt the experiment's results. In the case of our running example, this means that the experimental conditions may not be comparable. For instance, we noticed the worker performance in independent runs of our study varied by different factors even between runs of the same condition (e.g., from 24s to 14s in decision time between a first and a second run considering only new workers).

Collecting reliable and comparable results thus requires multiple systematic runs over an extended period of time. For our study, this means executing the experiment over chunks of time during the day and spread it over weeks, balancing across the experimental conditions.

Population demographics ❷ ❸. The demographics of a crowdsourcing platform (e.g., gender, age, country) defines the pool of active workers, along with the time an experiment runs. Researchers tend to resort to crowdsourcing since it represents a mechanism to access a large pool of participants. However, demographic variables tend to follow a heavy-tailed distribution. For example, workers from the US and India constitute the majority of the available workforce in Amazon Mechanical Turk [Difallah et al., 2018]. Many kinds of experiments could potentially be sensitive to the underlying population demographics. Thus, ensuring a diverse set of workers is a crucial endeavor that researchers must undertake to perform methodologically sound experiments.

Uncontrolled worker demographics could result in an imbalanced sample of the population ❷ and subgroups of workers dominating the task ❸, potentially amplifying human biases and produce undesired results [Barbosa and Chen, 2019]. In running uncontrolled tasks, we observed a participation dominated by certain countries, which prevented more diverse population characteristics. For example, the top contributing countries provided 48.1% of the total judgements (Venezuela: 28.5%, Egypt: 11.8%, Ukraine: 7.8%).

Crowdsourcing platforms offer basic demographic variables that can be tuned to control the population of workers participating in tasks. For our experiment, we identified the top three contributing countries and created buckets with each of the top countries as head of the groups. We then manually assigned the country buckets to the experimental conditions, distributing uniformly the tasks that these could perform and swapping buckets accordingly. This tedious but effective mechanism allowed us to overcome the heavy-tail distribution of the population demographics and give equal opportunity to the top countries that constitute the crowdsourcing platform.

Recurrent workers may impact the results ❹. While returning workers are desirable in any crowdsourcing project, they represent a potential source of bias in the context of crowdsourcing experiments, and special care must be taken to obtain independent contributions within and across similar experiments [Paolacci et al., 2010].

In our study, we published the combination of experimental conditions and document sizes as independent tasks in the crowdsourcing platform. Therefore, since tasks run in parallel, nothing prevents workers from proceeding with another of our tasks upon finishing the task where they first landed, which is the scenario depicted by point ❹ in Figure 4.2. This situation is potentially problematic since workers that return to complete more tasks might perform better due to the *learning effect*. As shown in Figure 4.3, we observed 38% of returning workers, who featured a lower completion time (i.e., workers were faster) but not higher accuracy[2].

Researchers must implement custom eligibility control mechanisms to deal with recurrent workers, and in general, to manage the population of workers that participate in the experimental conditions. Fortunately, crowdsourcing platforms provide the necessary means to extend its set of features. The task interface shown to workers is usually a combination of HTML, JavaScript, and CSS that researchers need to code. As part of this interface, special logic could be embedded to control worker participation. For our study, we identified workers by levering browser fingerprinting [Gadiraju and Kawase, 2017] and sending this information to an external server that performed control and random assignment of workers to conditions. Our task interface included JavaScript code that upon page load requested the server information about the worker, resulting in a "block" or "proceed" action that prevented or allowed the worker to continue with the task (in the case of the former the page showed a message with the reasons of the block).

Recurrent workers may cross conditions ❺. Closely related to the previous issue is the fact that returning workers can also land in a different experimental condition, as depicted by point ❺ in Figure 4.2. This scenario could make the experimental conditions difficult to compare, threatening the validity of the experiment, since returning workers that cross conditions could modify their behavior and resulting performance.

In our study, by comparing the 30% workers who crossed the experimental conditions with the "new workers" (those that never performed the task), we observed that switching between highlighting support and not support resulted in lower decision time ("highlighting to base" and "base to highlighting" in Figure 4.3). However, those workers that came from the "bad highlighting" condition and arrived at the condition with good highlighting support showed a higher decision time, possibly due to the lack of trust in the support. Workers switching from support to no support also featured higher accuracy than the new workers and those returning to the same condition.

[2] We noticed, however, that accuracy remained mostly unaffected by conditions and other factors across all our experiments, and it might have been less susceptible to the learning effect.

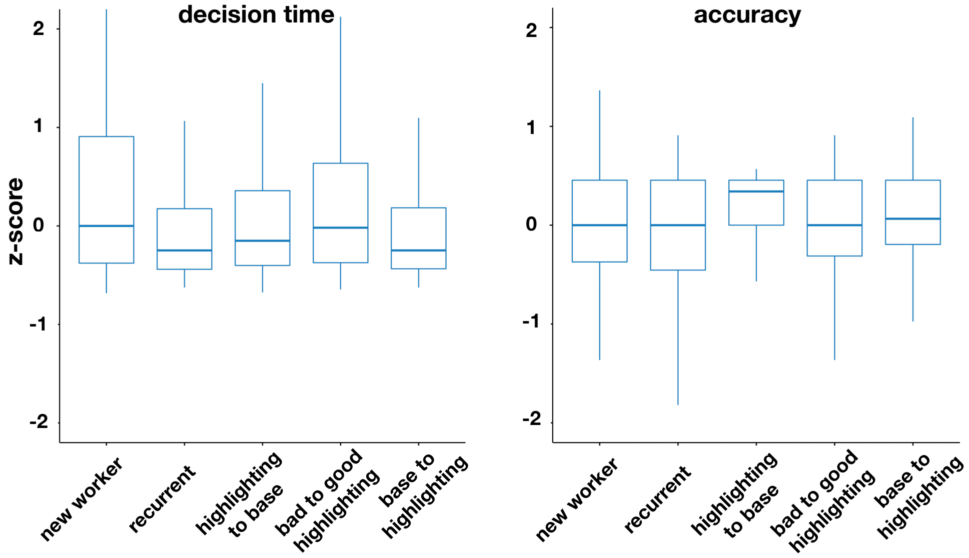


Figure 4.3: Decision time and accuracy for recurrent workers in the highlighting support experiment. For comparison purposes, the performance values from all conditions are aggregated using a normalized z-score that considers the median from the valid contributions in its computation. The distribution of values from non-valid contributions, organized under the different sources of bias, thus depict the deviation in performance from the normal population (i.e., new workers).

The same eligibility control mechanism via browser fingerprinting takes care of recurring workers that land in different conditions since it allows controlling for recurrent workers in the first place.

The above challenges emphasize that running a systematic comparison of task designs using the native building blocks of a crowdsourcing platform is thus a complex activity, susceptible to different types of experimental biases, which are costly to clean up (e.g., discarding 38% of the contributions). While our example may represent an extreme case, and it focuses on task design evaluation, it is still indicative of many of the challenges that task designers and researchers face in general when running controlled experiments in crowdsourcing platforms.

4.3 CrowdHub Platform

The above challenges motivated us to design and build a system that extends the capabilities of crowdsourcing platforms and allow researchers and practitioners to run controlled crowdsourcing projects. By extending major crowdsourcing platforms, CrowdHub offers cross-platform capabilities and aims to provide the building blocks to design and run crowdsourcing workflows, and for researchers, in particular, the features to run controlled crowdsourcing experiments.

4.3.1 Design goals

We now describe the design goals behind CrowdHub:

- **Offer cross-platform support**. CrowdHub extends crowdsourcing platforms and integrates the differences between these into features that allow task requesters to design tasks that can run across multiple platforms. This means that we can design a task (e.g., one that asks workers to classify images) and then run it on multiple platforms without dealing with the underlying details that set the platforms apart. This goal is particularly relevant for crowdsourcing experiments since there is evidence suggesting that results could vary across crowdsourcing platforms [Qarout et al., 2019].
- **Blend easy-of-use with flexibility**. While offering crowdsourcing capabilities by extending a programming language gives complete control to task requesters, it also demands more effort since requesters would need to code their solutions. With CrowdHub, we aim to mix the flexibility of programming with productivity, and thus instead offer a diagramming interface that does not require task requesters to code every piece of the crowdsourcing puzzle.
- **Support interweaving human and machine computation**. CrowdHub is designed to allow researchers and practitioners to extend its set of features and incorporate machine computation alongside human workers. This means, following on the image annotation example, that researchers could incorporate a machine learning model for classifying "easy" images and then derive "hard ones" to crowd workers.

4.3.2 Architecture

Figure 4.4 shows the internal architecture of the CrowdHub platform. We used a client-server architecture to implement CrowdHub, where both the backend and the frontend are implemented using the JavaScript programming language. CrowdHub offers a diagramming interface where *visual blocks* are the foundation to design crowdsourcing workflows. A workflow is essentially a graph representing a crowdsourcing project (e.g., an experiment) that allows data units to flow through the nodes, where the nodes represent tasks or executable code to transform data units.

The ***Workflow Manager*** exposes features that allow requesters to create, update, delete, and execute workflows. It uses the crash-and-rerun model [Little et al., 2010] to offer a robust mechanism for executing workflows. A node in the graph defining the workflow is called *block*. These blocks can be seen as "functions" that the workflow manager can run. The current implementation of CrowdHub offers two blocks *Do* and *Lambda*. The

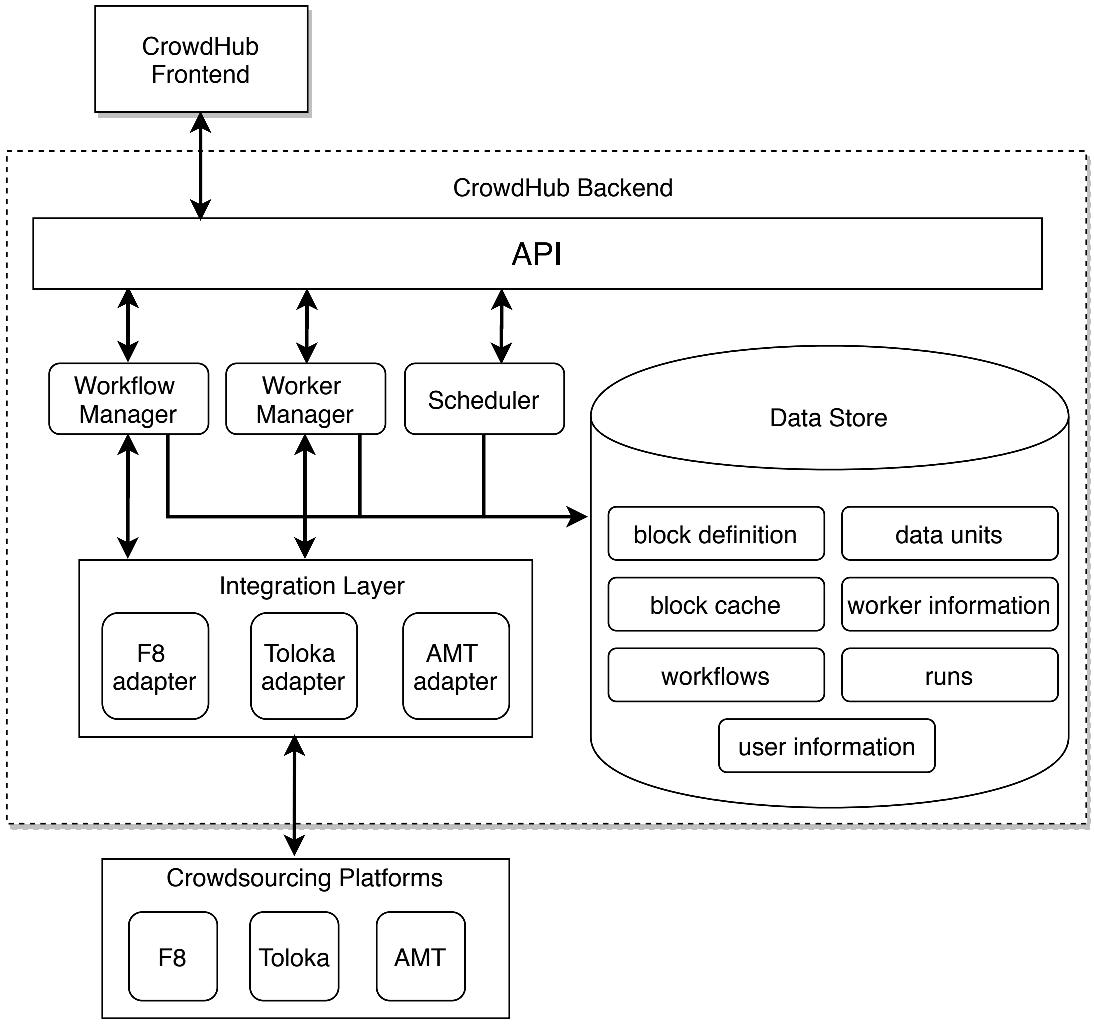
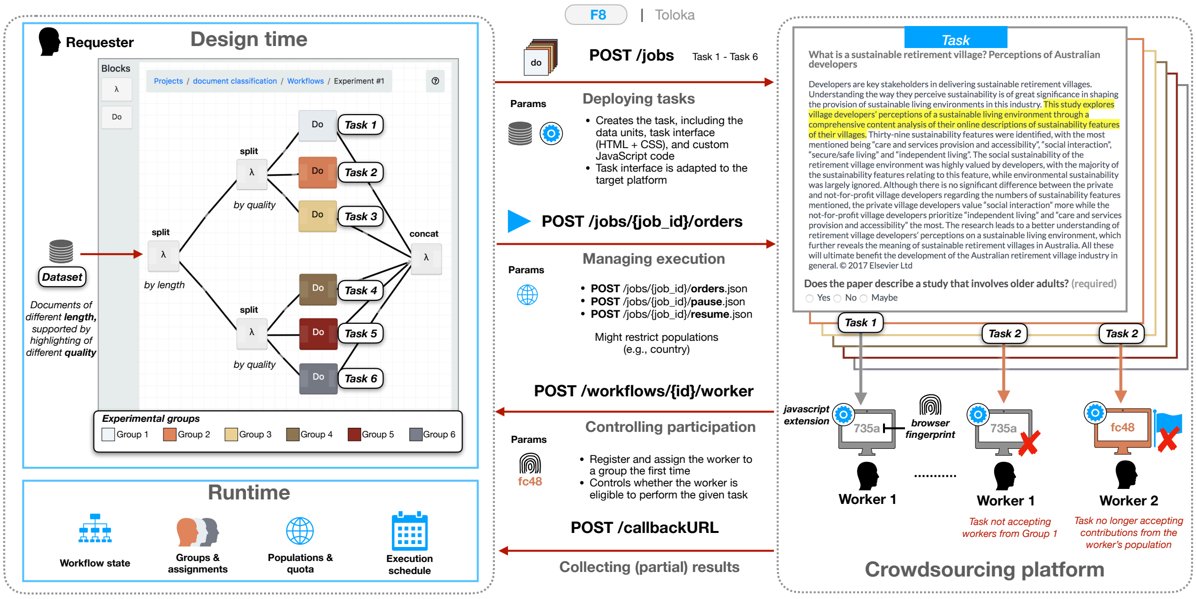


Figure 4.4: The architecture of the CrowdHub platform.

Do block represents a task that is published on a crowdsourcing platform. Using a *Do* block, requesters can configure the task interface using a builtin set of *UI elements* that safes requesters from coding the interface (which typically consists of HTML, CSS, and JavaScript). In addition to configuring the interface, other crowdsourcing task parameters can be specified (e.g., number of votes, monetary rewards). The *Lambda* block, accepts JavaScript code, and it represents an arbitrary function that receives and returns data (useful for data aggregation and partitioning, for example).

The ***Worker Manager*** offer features for eligibility control and population management. The *eligibility control* gives researcher the functionality to define the policy regarding returning workers and condition crossovers associated with the experimental design (between- or within-groups design). Through *population management* requesters can control for subgroups of workers dominating a dataset by assigning a specific quota.



- Creates the task, including the data units, task interface (HTML + CSS), and custom JavaScript code
- Task interface is adapted to the target platform

- **POST** /jobs/{job_id}/**orders**.json
- **POST** /jobs/{job_id}/**pause**.json
- **POST** /jobs/{job_id}/**resume**.json

Might restrict populations (e.g., country)

- Register and assign the worker to a group the first time
- Controls whether the worker is eligible to perform the given task

What is a sustainable retirement village? Perceptions of Australian developers

Developers are key stakeholders in delivering sustainable retirement villages. Understanding the way they perceive sustainability is of great significance in shaping the provision of sustainable living environments in this industry. This study explores village developers' perceptions of a sustainable living environment through a comprehensive content analysis of their online descriptions of sustainability features of their villages. Thirty-nine sustainability features were identified, with the most mentioned being "care and services provision and accessibility", "social interaction", "secure/safe living" and "independent living". The social sustainability of the retirement village environment was highly valued by developers, with the majority of the sustainability features relating to this feature, while environmental sustainability was largely ignored. Although there is no significant difference between the private and not-for-profit village developers regarding the numbers of sustainability features mentioned, the private village developers value "social interaction" more while the not-for-profit village developers prioritize "independent living" and "care and services provision and accessibility" the most. The research leads to a better understanding of retirement village developers' perceptions on a sustainable living environment, which further reveals the meaning of sustainable retirement villages in Australia. All these will ultimate benefit the development of the Australian retirement village industry in general. © 2017 Elsevier Ltd

Does the paper describe a study that involves older adults? (required)

○ Yes ○ No ○ Maybe

Task not accepting workers from Group 1

Task no longer accepting contributions from the worker's population

Figure 4.5: Example workflow for a between-subjects design using CrowdHub.

Altogether, these features allow requesters to be in control of their crowdsourcing project.

The ***Scheduler*** enables to control for confounding factors by scheduling task execution over a period of time. Task requesters specify the time and progress intervals at which the workflow should run, and the scheduler notifies the *Workflow Manager* accordingly to pause and resume the execution.

The ***Integration Layer*** implements the cross-platform capabilities that CrowdHub provides. This layer offers a set of functions that handle the differences between the crowdsourcing platforms, thus allowing requesters to publish tasks across multiple platforms. The workflow manager module uses the Integration Layer to publish, pause, and resume jobs in the crowdsourcing platforms supported by CrowdHub. When publishing a task, the Integration Layer first translates the UI components that constitute the task interface into the actual interface that will be shown to workers on the selected crowdsourcing platform. The worker manager module relies on the Integration Layer to handle worker demographics in a platform-agnostic manner, as well as implementing the eligibility control policy. CrowdHub manages the interactions with the crowdsourcing platform through their public APIs, and JavaScript extensions incorporated into the tasks interface allows for additional features such as worker control and identification (browser fingerprinting) [Gadiraju and Kawase, 2017]. The current implementation of CrowdHub supports Figure Eight and Toloka, with Amazon Mechanical Turk as a work in progress.

The ***Data Store*** is a SQL database that contains user information, workflow definition and runs (which allow for running a workflow multiple times), block definition and cache (to store the results from running the blocks), worker information and the data units.

4.3.3 Deployment

Figure 4.5 shows an example workflow using CrowdHub, where we define a crowdsourcing experiment following a between-subjects design to systematically evaluate task interface alternatives.

CrowdHub enables the entire task evaluation process, as shown in Figure 4.5. At *design time*, requesters use the workflow editor to define the experimental design, which includes the tasks (*Do* boxes) and the data flow (indicated by the arrows and the *lambda* functions describing data aggregation and partitioning). Experimental groups can also be defined and associated to one or more tasks, denoted in the diagram using different colors. When deploying the experiment, the *Workflow Manager* parses the workflow definition and creates the individual tasks in the target crowdsourcing platform with the associated data units and task design, relying on the Integration Layer to handle the selected platform. At *run time*, the requester can specify the population management strategy and time sampling, if any, and the platform will leverage the Worker Manager and Scheduler modules to launch, pause and resume the tasks, and manage workers participation.

4.4 Discussion

Running controlled experiments in crowdsourcing environments is a challenging endeavor. Researchers must put special care in formulating the task and effectively communicating workers what they should perform [Kittur et al., 2008]. Unintended worker behavior could be observed if researchers fail at delivering the task; for example, poor instructions could lead workers to produce low-quality work or discourage them from participating in the first place [Wu and Quinn, 2017]. The inherent biases associated with experimental research and those present in the crowdsourcing platforms hinder the job of researchers. However, most of these biases are unknown to researchers approaching crowdsourcing platforms and could harm the experimental results, reducing the acceptance of crowdsourcing as an experimental method [Crump et al., 2013; Paolacci et al., 2010]. For example, the underlying population demographics characterizes the workers that can take part in an experiment. If run uncontrolled, the sampled population may not be representative and include biases that hurt the experimental outcome [Barbosa and Chen, 2019].

We showed specific instances of how running crowdsourcing experiments without coping strategies can impact the experimental design, assignment, and workers participating in the experiments. Using task design evaluation, we distilled the challenges and quantified how it could change the outcomes of experiments. However, these challenges are not only tied to task design evaluation, and in general, they play a role in the success of crowdsourcing experiments. Qarout et al. [Qarout et al., 2019] identified that worker performance could vary significantly across platforms, showing that workers in AMT performed the same task significantly faster than those on F8. This difference highlights the importance and impact of the underlying population demographics [Difallah et al., 2018]. Recurring workers naturally affects longitudinal studies. Over prolonged periods, the population of workers could refresh [Difallah et al., 2018]; however, this still depends on the actual platform. Therefore, it is common for researchers to resort to employing custom mechanisms [Gadiraju and Kawase, 2017] or leveraging the actual worker identifiers to limit participation [Qarout et al., 2019]. This forces researchers to fill the gaps by programming extensions to crowdsourcing platforms and make it possible to run controlled experiments, one of the core challenges highlighted by behavioral researchers [Crump et al., 2013].

The lack of native support from crowdsourcing platforms to deliver experimental protocols could potentially affect the validity and generalization of experimental results. Researchers are forced to master advanced platform-dependent features or even extend their capabilities to implement coping strategies and bring control to crowdsourcing experiments. This is because crowdsourcing platforms were built with micro-tasking and data collection tasks in mind where results are important. However, the potential downside of manually extending crowdsourcing platforms is the learning curve that it incurs on researchers. This could lock researchers to specific platforms and discourage running experiments across multiple crowdsourcing vendors — potentially threatening how well results generalize to other environments [Qarout et al., 2019].

The reliability of results, associated with choosing the right sample size, is a critical challenge in experimental design. The inherent cost of recruiting participants forces researchers to trade-off sample size, time, and budget in laboratory settings [Kittur et al., 2008]. For crowdsourcing environments, this is not necessarily the case, since it naturally represents easy access to a large pool of participants [Crump et al., 2013]. In laboratory settings, it is possible to resort to magic numbers and formulas to derive the sample size, but these rely on established recruitment criteria that can ensure a certain level of homogeneity — in general, a homogenous population would require a small sample size. In crowdsourcing, however, it is not easy to screen participants, and sometimes researchers just accept whoever is willing to participate. This can bring a lot of variability into the results. One way to address this issue is to rely on techniques used in adaptive or responsive survey design, i.e., stopping when we estimate that another round of crowdsourcing (e.g., data collection) will have a low probability of changing our current estimates [Rao et al., 2008]. Another approach would be to run simulations inspired by k-fold cross validation

[James et al., 2013], which relies on specific data distribution (e.g., unimodal).

As crucial as controlling different aspects of task design, experimental protocol, and coping strategies to deal with the crowdsourcing environment is the fact that researchers must communicate these clearly to aid repeatable and reproducible crowdsourcing experiments [Qarout et al., 2019; Paritosh, 2012]. In the context of systematic literature reviews, PRISMA [Shamseer et al., 2015] defines a thorough checklist that aids researchers in the preparation and reporting of robust systematic reviews. For crowdsourcing researchers and practitioners, there is currently no concise guideline on what should be reported to facilitate reproducing results from crowdsourcing experiments, beyond those guidelines addressing concrete crowdsourcing processes [Liu et al., 2016; Barbosa and Chen, 2019; Sabou et al., 2014; Blanco et al., 2011]. We find this an exciting direction for future work, and we are currently initiating our project on developing guidelines for reporting crowdsourcing experiments and encouraging repeatable and reproducible research.

We created CrowdHub to provide features that enable requesters to build crowdsourcing workflows, such as creating datasets for training machine learning models or designing and executing complex crowdsourcing experiments.

CrowdHub is not a monolithic system but rather a collection of components that interact with each other. Requesters can register new adapters with the *Integration Layer* to add support for new crowdsourcing platforms, therefore growing the list of vendors that CrowdHub provides out-of-the-box. This could be particularly useful for integrating private in-house platforms that suites the specific needs of task requesters. The *Workflow Manager* enables researchers and practitioners to add new *blocks* to the set of available nodes for creating crowdsourcing workflows. With this feature, requesters can grow the scope of crowdsourcing workflows that can be created using CrowdHub. Therefore, new forms of computation could be added that run alongside crowd workers. An example is adding an “ML block” that uses data units to train a machine learning classifier.

To create task interfaces, requesters can resort to the built-in set of UI elements that encapsulates the necessary code for rendering the interface on the underlying crowdsourcing platforms. CrowdHub’s frontend application exposes these UI elements as visual boxes that requesters can drag and drop to arrange and configure the interface accordingly. The current set consists of elements for rendering text, images, form text inputs, and inputs for multiple- and single-choice selection. We also incorporated the possibility of highlighting text and image elements. This is useful, for example, to generate datasets for natural language processing and computer vision (e.g., question-answering and object detection datasets, respectively), or studying the impact of text highlighting in classification tasks [Ramírez et al., 2019a]. We plan to add support for actually coding the task interface, allowing requesters to use HTML, JavaScript, and CSS instead of using the current editor that offers draggable visual elements. This modality will give full flexibility to experienced requesters for designing task interfaces, and in this context, the current UI components will be available as special “HTML tags”.

We presented a demo of CrowdHub [Ramírez et al., 2019b] and received positive and constructive feedback from researchers in the human computation community. These discussions allowed us to arrive at the current design goals and set of features that constitute CrowdHub. The current implementation offers all the features we described for the Workflow manager, and a subset of the functionality associated with the Worker Manager that enables eligibility control, allowing researchers to map experimental designs and control worker participation. The system also supports collaboration between requesters and generating URLs for read-only access to workflows — a feature that aims to foster repeatability and reproducibility of results in crowdsourcing experiments. CrowdHub currently supports two crowdsourcing platforms: Figure Eight and Yandex Toloka. Support for Amazon Mechanical Turk is also in the roadmap.

Chapter 5

On the state of reporting in crowdsourcing experiments and a checklist to aid current practices

Crowdsourcing platforms are being widely adopted as an environment to run experiments with human subjects [Kittur et al., 2008; Mason and Watts, 2009; Paolacci et al., 2010; Crump et al., 2013]. Researchers are leveraging crowdsourcing to test hypotheses, comparing different study methods, designs or populations, as well as to run studies aiming at observing user behavior. For example, crowdsourcing is helping researchers evaluate the impact of different interface designs on user performance, comprehension and understanding [Steichen and Freund, 2015; Dimara et al., 2017; Ramírez et al., 2019a], assess the difference in performance between users with different expertise, background and even mood levels [Wu and Bailey, 2016; Hube et al., 2019; Xu et al., 2019]. These platforms (e.g., Amazon MTurk) give researchers easy access to a large and diverse population of participants, allowing them to scale experiments previously curbed to constrained laboratory settings.

Researchers need to articulate many elements to successfully map and run an experiment in a crowdsourcing platform, as depicted in Figure 5.1. The relevance of this is rooted in the need of incorporating more control and safeguards in an otherwise uncontrolled environment [Gadiraju et al., 2015]. For example, an experiment testing the quality of two alternative approaches to text summarization would require researchers to define, among others, i) how to implement the two text summarization conditions as micro-tasks in the crowdsourcing platform (e.g., both conditions in the same task or in different tasks), ii) how to sample and allocate crowd workers to have representative, diverse (e.g., in culture, education or mother-tongue) and comparable groups assigned to both experimental conditions,[1] iii) quality control measures to avoid malicious or low-quality contributions from workers, and iv) task design and configuration, including the user interface, training examples, number of text summaries to show to each contributor, and the compensation for their participation. In doing so, researchers also need to ensure that their entire setup meets ethical standards for research with human subjects.

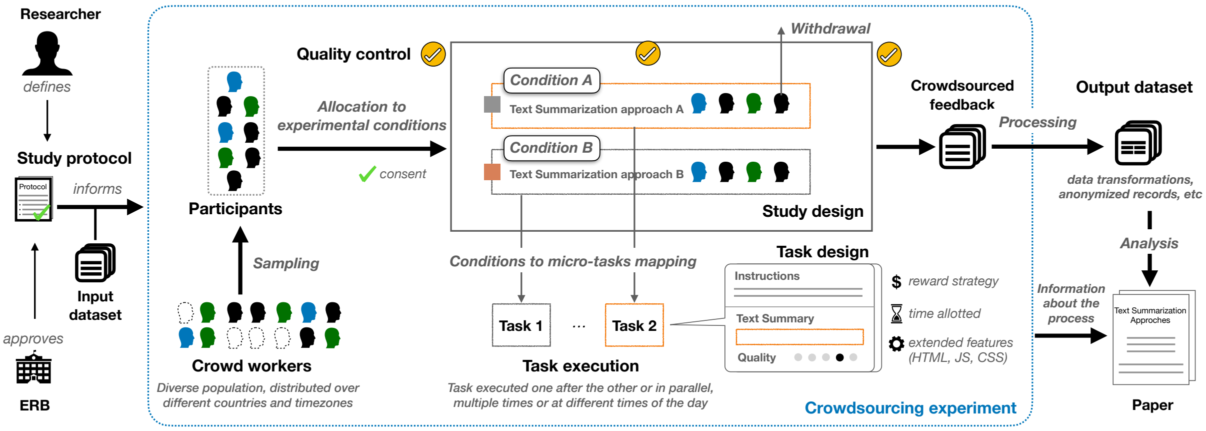

Researcher
defines
Study protocol
Protocol
informs
Input dataset
approves
ERB
Quality control
Participants
Allocation to experimental conditions
consent
Sampling
Crowd workers
Diverse population, distributed over different countries and timezones
Withdrawal
Condition A
Text Summarization approach A
Condition B
Text Summarization approach B
Study design
Conditions to micro-tasks mapping
Task 1
...
Task 2
Task execution
Task executed one after the other or in parallel, multiple times or at different times of the day
Task design
Instructions
Text Summary
Quality
$ reward strategy
time allotted
extended features (HTML, JS, CSS)
Crowdsourcing experiment
Crowdsourced feedback
Processing
Output dataset
data transformations, anonymized records, etc
Analysis
Information about the process
Text Summarization Approches
Paper

Figure 5.1: Mapping an experimental design to a crowdsourcing platform involves articulating many elements (none of which have a unique implementation). These elements constitute sources of variability if not properly reported.

While guidelines and best practices have emerged to help researchers navigate the implementation choices and inherent challenges of running experiments in a crowdsourcing environment [Rogstadius et al., 2011; Mason and Suri, 2012; Chandler et al., 2013], little attention has been paid to how to *report* on crowdsourcing experiments to facilitate assessment and reproducibility of the research. If not properly reported, the elements mentioned above constitute sources of variability that can introduce confounds affecting the repeatability and reproducibility of the experiments, and preventing a fair assessment of the strength of the empirical evidence provided.

Reproducibility of experiments is essential in science [Wacharamanotham et al., 2020]. The scrutiny of the research methods, by the academic community, and the development of standardized protocols and methods for communicating results are critical in the production of robust and repeatable experiments. Examples of this can be found in evidence-based medicine where systematic reviews follow a strict elaboration protocol [Shamseer et al., 2015], or, more recently, in the machine learning community where authors are encouraged to follow pre-established checklists or datasheets to communicate their

[1] While a perfect sampling and allocation strategy (of workers to conditions) may represent an ideal scenario detached from reality, we as researchers should ensure a reasonable balance in the underlying population to avoid confounds affecting the experimental outcome.

models and datasets effectively [Bender and Friedman, 2018; Gebru et al., 2018; Mitchell et al., 2019; Arnold et al., 2019; Pineau et al., 2020]. Efforts in this regard have been branded under the definitions of *repeatability*, *replicability* and *reproducibility*, to denote attempts at obtaining similar results, by the same or different teams and experimental conditions, under acceptable margins of error [Plesser, 2018]. Terminology notwithstanding, the importance of reporting experiments in sufficient detail has long been acknowledged as a fundamental aspect of research and of the scientific process.

The reporting of crowdsourcing experiments should be held to the same standards. Studies, under the umbrella of reproducibility and experimental bias, report on multiple aspects that affect the outcome of crowdsourcing experiments. These aspects include task design (e.g., instructions [Wu and Quinn, 2017], compensation [Ho et al., 2015], task interface [Sampath et al., 2014], and time [Maddalena et al., 2016] among others), external factors such as the workers' environment [Gadiraju et al., 2017a], platforms [Qarout et al., 2019], and population [Difallah et al., 2018] — all of which serve as the foundation for running user studies. In addition to the standard reporting for experimental research, it is paramount to identify and report the critical aspects and all possible knobs that take part in crowdsourcing experiments. In this regard, existing literature has barely grappled with proposing guidelines that aid the reporting of crowdsourcing experiments, with existing guidelines limited to crowdsourced data collection in the social sciences [Porter et al., 2020].

In this chapter we aim to fill this gap. We do so by identifying salient issues associated with the reporting of crowdsourcing experiments and propose solutions that can help address these issues. Specifically, we derive a *taxonomy of attributes* associated with crowdsourcing experiments and turn this into a checklist for reporting. The checklist aims at facilitating the reporting of crowdsourcing experiments so that they can be repeated and so that a reader can assess if the experiment design matches the desired intent. This chapter focuses on the reporting of *experiments* ran via crowdsourcing, that is, studies aiming to answer research questions/hypotheses by following an experimental design, mapping the design to a crowdsourcing task and the features of the platform that supports it, and recruiting crowd workers as subjects. This chapter therefore excludes qualitative crowdsourcing studies (e.g., surveys) from its scope and, in general, other kinds of tasks that are not experiments (e.g., data labeling tasks are often not framed as experiments). However, the challenges in reporting on experiments are likely to be a superset of those or generic crowdsourcing tasks.

Contributions. To aid the reporting of crowdsourcing experiments, we first need to understand the main factors that affect repeatability and that enable assessment of the quality of experiments by reviewers. We start by deriving the major design decisions of crowdsourcing experiments that play a role in crowdsourcing tasks and, therefore, on an experiment's outcomes. We do so by following an iterative approach — where we rely on literature reviews and interviews with experts — and derive a taxonomy of relevant attributes characterizing crowdsourcing studies. Using this taxonomy, we analyze the state of reporting in crowdsourcing literature and identify aspects that are frequently communicated and those that tend to go under-reported. We leverage these observations to discuss potential pitfalls and threats to validity of experiments. With feedback from experts, we then propose a checklist for crowdsourcing experiments to help researchers be more systematic in what they report.[2] This checklist seeks to help experimenters describe their setup in a standardized format and readers to understand the used methodology and how it was implemented, serving as a tool that complements existing experimental research guidelines

5.1 Related Work

Guidelines for reproducibility and reporting of scientific studies

Reporting and reproducibility are at the heart of science. Experiments allow researchers to manipulate a set of variables to test their influence into another group of variables of interest [Shadish et al., 2002]. Guidelines for reporting scientific studies emerge from the observed variance in the methodological rigor associated with the studies. Indeed, the output of experimental research is only meaningful as long as it is reproducible [Paritosh, 2012]. This property guides the adopted methodology, as well as how this methodology and the results are communicated.

For example, in the study and synthesis of scientific results obtained via systematic literature reviews (SLRs), papers adhere to precise reporting guidelines that describes the (systematic) approach to investigating a problem as discussed in the literature. The existence of study and study report protocols is what gives systematic reviews its methodological rigor, avoiding issues like a biased selection of clinical outcomes [Chan et al., 2004]. In this context, the PRISMA [Shamseer et al., 2015] guidelines propose a checklist to support the preparation and reporting of SLRs, making sure that such protocols exist and are reproducible.

In medicine, randomized control trials (RCTs) are the gold standard methodology to evaluate medical interventions. In this regard, guidelines like the CONSORT statement [Schulz et al., 2010] help authors properly report their RCTs and avoid potential issues resulting from the lack of methodological rigor (e.g., biased outcomes).

[2] The checklist can be found at `https://trentocrowdai.github.io/crowdsourcing-checklist/`

Guidelines in crowdsourcing contexts

Research has shown how crowdsourcing could be leveraged in a wide range of tasks and domains (e.g., from labeling data for ML [Snow et al., 2008; Sorokin and Forsyth, 2008] to serving as a platform for experimental research [Mason and Watts, 2009; Paolacci et al., 2010; Crump et al., 2013]). Existing guidelines and best practices focus on how to run experiments in crowdsourcing environments successfully, proposing strategies to overcome common pitfalls found in crowdsourcing platforms [Kittur et al., 2008; Rogstadius et al., 2011; Mason and Suri, 2012; Chandler et al., 2013]. However, how to properly report crowdsourcing experiments has, to the best of our knowledge, been somewhat overlooked. Our work complements experimental research in crowdsourcing by providing guidelines to aid researchers in reporting crowdsourcing studies.

Crowdsourcing acts as a surrogate to traditional participant samples, but with additional challenges that are not present in traditional experimental settings. The quality of the contributions provided by crowd workers is a major concern due to the diversity in worker skills and the level of commitment crowd workers put into the task [Gadiraju et al., 2015]. Many quality control mechanisms have been proposed (e.g., see [Daniel et al., 2018] for a review) as well as studies analyzing performance as a function of internal and external factors, showing that intrinsic motivation could help in increasing the performance of workers [Rogstadius et al., 2011]. Factors related to the design of the task could also contribute to obtaining subpar responses. Poorly written task instructions could misguide workers and result in low-quality work [Wu and Quinn, 2017; Kittur et al., 2013; Gadiraju et al., 2017b; Liu et al., 2016]. In contrast, enhanced interfaces may facilitate the job of crowd workers and aid these to improved performance [Sampath et al., 2014; Wilson et al., 2016; Ramírez et al., 2019a; Ramírez et al., 2019c], as well as adequately limiting the time to judge can accelerate task completion without compromising the quality of the results [Maddalena et al., 2016].

Another challenge relates to how we operationalize an experimental design in a crowdsourcing platform. As opposed to laboratory settings, there is an inherent lack of control over the participants of crowdsourcing (and online) experiments that represents a concern to researchers, amplified by the absence of built-in support from crowdsourcing platforms [Kittur et al., 2008]. Random assignment, although simple in principle, is not straightforward to implement. Using multiple tasks to map different experimental conditions is a typical approach [Ho et al., 2015]; however, self-selection effects could arise due to participants preferring a subset of the conditions over others. And ensuring that new workers arrive in the experiment is crucial to proper random assignment [Mason and Suri, 2012; Chandler et al., 2013], avoiding scenarios where workers participate multiple times in longitudinal studies or experiments that must run multiple times. The characteristics of the platform should also be kept in mind. The underlying demographics associated with the active workers [Difallah et al., 2018] play an important role since differences in the conditions could be attributed to differences in the population of workers rather than the conditions themselves [Qarout et al., 2019].

Recent guidelines also safeguard the ethics behind crowdsourcing experiments. Previous literature on humanizing crowd work pointed out relevant ethical issues present in common crowdsourcing practices (e.g., initially, crowdsourcing was seen as a "cheap labor market", as briefly reviewed in [Barbosa and Chen, 2019]). One aspect of this regards to accurate and fair task pricing [Whiting et al., 2019], which previous work has shown to impact the number of contributions produced by workers [Mason and Watts, 2009], the quality of their work [Ho et al., 2015], and being a relevant aspect from an ethical perspective. Another aspect concerns privacy, especially if some properties of the workers are being requested as part of the experiments [Mason and Suri, 2012].

As crucial as successfully conducting experiments in crowdsourcing environments, it is to make sure crowdsourcing experiments are reproducible [Paritosh, 2012; Qarout et al., 2019]. In this regard, proper reporting of the methodology, operationalization, and results of crowdsourcing experiments plays a relevant role. Surprisingly, guidelines for reporting crowdsourcing experiments have received little attention even though underreporting is a serious concern in science, and crowdsourcing is no exception [Buhrmester et al., 2018]. Existing literature has identified how current studies fail to adequately report the methodology behind their crowdsourcing experiments, showing that most of the papers tend to omit information about worker qualifications, task design, rejection or validation criteria [Porter et al., 2020; Ramírez et al., 2020b]. These efforts propose templates for reporting studies, capturing essential aspects of crowdsourcing experiments. However, these works are so far focused on specific use cases and platforms (data collection for social sciences in Amazon Mechanical Turk [Porter et al., 2020]), or are still work-in-progress [Ramírez et al., 2020b].

Existing guidelines cover very well how to effectively leverage crowdsourcing in different contexts, but it has barely grappled with how to report crowdsourcing studies properly. We aim to fill this gap by proposing guidelines for reporting the relevant aspects of crowdsourcing experiments. As a starting point, our work leverages on the preliminary taxonomy proposed in [Ramírez et al., 2020b]. The final taxonomy we propose differs from DREC's in terms of validation and scope of the attributes. The validation stems from a mixed approach involving a large-scale analysis of papers reporting crowdsourcing experiments, 171 articles from the literature, and feedback from experts in the field. This approach helped us to scope down and refine the taxonomy to only attributes relevant to crowdsourcing experiments, leaving off attributes that are well-understood and covered in guidelines for experimental research [3]. The final taxonomy also avoids narrow or too broad attributes (e.g., DREC's *synchronous* and *data analysis* attributes, respectively), as well as having attributes that can collapse into a single one (e.g., *reputation* and *environment* attributes are now part of the *target population*). In addition, our mixed approach allowed us to shape the final taxonomy into a checklist to aid how researchers report experiments in crowdsourcing.

5.2 Data & Methods

We aim to i) understand the status of reporting on crowdsourcing experiments to assess which aspects are covered or neglected and the extent to which reporting is consistent across the literature, and ii) provide guidelines for reporting to assist in achieving consistency in current practices and facilitate reproducibility and assessment.

To achieve these goals, we draw inspiration from standardizing efforts in evidence-based medicine and software engineering for the reporting of systematic literature reviews [Shamseer et al., 2015; Barbara and Charters, 2007]. We follow a mixed approach that involves (1) deriving a taxonomy of relevant attributes characterizing crowdsourcing experiments, (2) leveraging this taxonomy to analyze 171 papers published in major venues and get a picture of the current state of reporting, and (3) exploring potential alternatives of guidelines for reporting.

These steps were supported by literature reviews and interviews with experts in the field. Specifically, literature reviews informed the first two steps, with a specific review for each step, and we describe them in detail in Sections 5.3 and 5.4, respectively. The interviews with field experts provided a formative feedback along the entire process. They consisted of semi-structured interviews with researchers with ample experience i) performing experiments to study crowdsourcing (i.e., crowdsourcing was the main area of research), or ii) leveraging crowdsourcing platforms to run user studies and experiments on different domains. Participants were recruited from the extended network of the authors, considering as eligibility criteria a research track involving crowdsourcing experiments and publishing in SIGCHI conferences. Ten experts agreed to participate (2F, 8M) including 1 Ph.D. student, 6 senior researchers from academia and industry, and 3 professors.

The interviews took place over Skype and Zoom between August and September 2020. Before the start of the sessions, participants were informed of the goal and scope the interview and provided their consent to participate and for the session to be recorded. The

[3] In the final taxonomy, the *experimental design* dimension has 7 attributes vs. 13 in DREC. Here, the taxonomy omits attributes covered in guidelines for reporting study design and protocols (e.g., [Gergle and Tan, 2014]). Likewise, the *outcome* dimension, for example, omits the *data analysis* attribute in DREC, as this is addressed by guidelines for reporting statistics (e.g., [Association, 2010]).

interview (see the protocol in Appendix A.3) was organized into three parts that provided input to each of the three main steps in our methodology. The interviews were carried out independently by two researchers. All interviews were held in and transcribed to English by the interviewer. Only the transcripts were accessed for the analyses. The analyses performed and how they inform our entire process are described in its proper context in the following sections.

5.3 A Taxonomy of Relevant Attributes

This section introduces the taxonomy of relevant attributes that characterize different aspects of crowdsourcing experiments. The attributes are grouped around six main dimensions denoting: the task requester, the experimental design used, the participants of the experiments (i.e., the crowd), the task design and configuration, the quality control mechanisms used to guard the quality of the results, and the outcome of the experiment. The resulting taxonomy is summarized in Figure 5.2. In the following we describe the methods and described in detail the final taxonomy, highlighting the literature support and expert opinions.

5.3.1 Methods

We consider four sources to elicit the taxonomy: i) guidelines for research experiments in general and specific to crowdsourcing experiments, ii) features available in crowdsourcing platforms to support the deployment of experiments, iii) scientific papers describing crowdsourcing experiments, and iv) interviews with experts. With these sources we aim to convey in the taxonomy the landscape of elements taking part in crowdsourcing experiments: elements from experimental research, those inherent to crowdsourcing, and what features platforms offer.

We started by identifying relevant attributes from existing guidelines. For this, we took a small seed of well-known guidelines for experimental design and crowdsourcing [Gergle and Tan, 2014; Hosseini et al., 2015; Porter et al., 2020; Mason and Suri, 2012; Gadiraju et al., 2015] and expanded it through snowballing and keywords search using Google Scholar (*"crowdsourcing + guidelines"*, *"crowdsourcing + best practices"*, *"crowdsourcing + recommendations"*, and *"crowdsourcing + reporting + experiment"*) and screening the results based on title. This perspective was complemented with the analysis of practical task design attributes available in a example micro-task platform, Toloka [4], which provides all common features for managing microtask crowdsourcing. The analysis of platforms'

[4] https://toloka.yandex.com/

features allowed us to ground attributes to practical "knobs" that researchers need to consider to operationalize their experiments, and that often fall into assumptions (e.g., training steps, mapping of experimental design to concrete micro-tasks). Leveraging these sources, two researchers extracted an initial set of attributes (e.g., task interface, task instructions, compensation, crowd demographics, research design, random assignments, platform used, fair payments, among others) in a spreadsheet to form an emerging list. The two researchers then jointly discussed and organized these attributes into six dimensions as depicted by the top-level entries in Figure 5.2.

This organization of attributes required the two researchers to iteratively group the attributes identified in the seed of papers and selected crowdsourcing platform around common themes (e.g., dimensions like *pool of participants*, *workers*, and *study participants* were unified as the *crowd* dimension). This was followed by an analysis of the initial set of attributes extracted, merging (whenever possible) equivalent attributes from different sources (e.g., *task template* and *task UI* as *task interface*). Notice that during this process, the researchers defined the semantics and scope of each of the six dimensions, as well as that of the individual attributes. We should stress, however, that our aim is on the comprehensiveness in terms of the attributes we identified and not the way we organized these into dimensions, which is a specific way of viewing things. For example, we include the demographics attribute under Outcome, as we define this dimension as the one capturing the *results* of the different aspects of the crowdsourcing process (including the recruitment, application of quality control techniques, etc). Under different semantics, one might put demographics under the Crowd dimension, but we defined this dimension as the one capturing attributes and mechanisms to identify and sample participants from the Crowd.

This initial taxonomy was then refined by analyzing and piloting the extraction of relevant crowdsourcing experiment attributes from a list of research papers reporting crowdsourcing experiments. The process of identifying these papers is described in Section 5.4.1, as part of the analysis of the state of reporting. From this list, we took a random sample of 15 papers published in the last eight years, which we drafted incrementally until we reached saturation [Saunders et al., 2018]. In this piloting phase, researchers took note of the applicability of the attributes for certain types of experiments, new attributes not initially considered, as well as attributes to be discarded or merged. These observations were discussed and addressed jointly by the two researchers.

We then further refined and validated this taxonomy with the input from crowdsourcing experts. To this end, we used the input collected in Part 1 and (some bits of) Part 2 of the semi-structured interviews with experts. The goal was to tap into their experience to identify, through different trigger questions, relevant attributes that we might have missed.

We leveraged their input by inquiring about their experience running and designing crowdsourcing experiments (e.g., *“What design choices you found to be more critical, possibly affecting experiment outcomes?”*), reporting or trying to replicate experiments (e.g., *”What aspects you deem relevant and should be reported?”*), and their own experience reading or reviewing papers (e.g., *”What aspects do you find to be typically under-reported or poorly reported?”*). In doing so, the interviewer would bring up only the top level dimensions (e.g., quality control or task design) if they were not discussed by the participant. We avoided providing any details about specific attributes so as not to bias the participants towards our taxonomy. Then, Part 2 introduced a portion of the taxonomy (one or two dimensions) and asked participants to assess the relevance of the attributes and to suggest any missing one.

Transcripts were organized around the trigger questions in a document, where the interviewer highlighted excerpts touching on crowdsourcing attributes (e.g., *“[...] depending on the design of the interface you might get wide results.”*). These excerpts were then moved to a spreadsheet for analysis. The interviewer then coded the excerpts with the associated attribute if covered by our taxonomy or flagged them for discussion otherwise. The two researchers then discussed the coded excerpts and assessed whether i) the attribute was relevant to the scope of the taxonomy, ii) the scope and name of an attribute had be updated to cover a more general case, iii) the scope of the attribute had to be limited to account for scenarios where an attribute is not applicable.

The inclusion criteria we used to refine the taxonomy considered three main reasons. First, the taxonomy should focus on attributes directly associated with crowdsourcing, offloading general attributes to existing guidelines.[5] Second, the focus should also be on practical and essential attributes for the reproducibility of experiments in crowdsourcing (e.g., instead of just describing the experimental conditions, authors should explain how these were mapped to tasks in crowdsourcing platforms). And last, the attributes in the taxonomy should have a clear scope, avoiding “unique cases” or attributes that are too broad.[6] The resulting taxonomy is the starting point to develop a checklist for reporting crowdsourcing experiments.

[5] Initially, the taxonomy considered attributes such as hypotheses, independent, dependent, and control variables. In later iterations, we omitted these attributes since they are well understood and covered in guidelines for experimental research.

[6] For example, the taxonomy considered whether an experiment was “synchronous” (experiments with multiple phases where one phase's output serves as input to the next [Mason and Suri, 2012; Gadiraju et al., 2015]). We ultimately replaced this by introducing more specific attributes describing how the experimental design maps to crowdsourcing tasks and how they are executed.

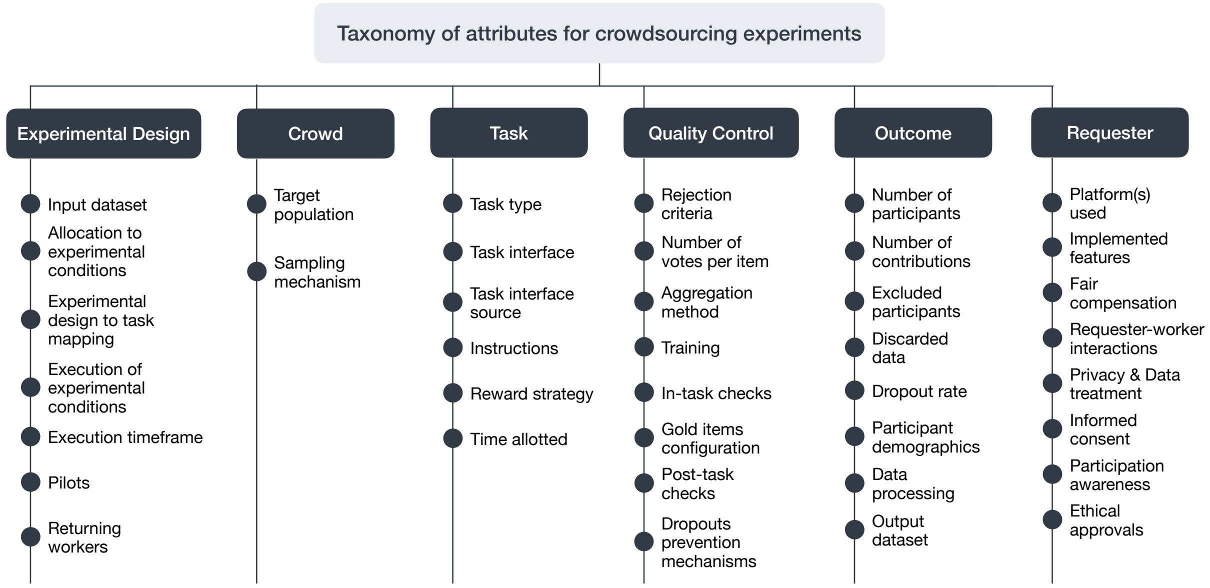


Figure 5.2: A taxonomy of relevant attributes characterizing experiments in crowdsourcing.

5.3.2 Experimental Design

The experimental design represents the building block for experiments of any kind, allowing to set the tone of the study and what level of conclusions can be derived from the experimental results. Proper experimental research involves several elements, from the research questions to the study design and analysis. These are well-understood aspects covered in experimental research guidelines and textbooks on research methods (e.g., [Gergle and Tan, 2014; Olson and Kellogg, 2014]).

The *experimental design* section of the taxonomy we propose focuses instead on those attributes that are more closely related to crowdsourcing, and that allow experiments to be run in such platforms. The limited support for running experiments in crowdsourcing platforms translates into many alternative strategies to map the experimental design to crowdsourcing tasks. Having multiple alternatives (and failing to accurately report these) can introduce confounds and ultimately affect the reproducibility of crowdsourcing experiments [Qarout et al., 2019]. The ● *experimental design to task mapping* attribute indicates how the experimental conditions (e.g., Condition A and Condition B) were mapped to crowdsourcing tasks. For example, as shown in Figure 5.1, a between-subjects design could map each experimental condition to a different micro-task (Condition A → Task 2, Condition B → Task 1) or randomize them within a single task (Conditions A,B → Task 1). Followed by this design choice, it is also relevant to define whether the experimental conditions were executed in parallel or sequentially (i.e., the ● *execution of experimental conditions* entry in the taxonomy) as different execution strategies may result in experimental conditions leveraging different sets of active crowd workers (e.g., population samples with different underlying characteristics).

Another factor comprises participant's ● *allocation to experimental conditions.* This attribute captures *if* and also *how* the randomization of the assignments was performed, considering the effect on the strength of the resulting experimental evidence. In crowdsourcing environments, where multiple micro-tasks may be running in parallel or sequentially to serve the overall experimental design, crowd workers may be able to engage in more than one micro-task, leading to what we refer to as ● *returning workers.* Depending on the experimental design and how it was handled, this situation may be desirable or introduce unintended biases affecting the integrity of the experimental conditions and, ultimately, the experiment results [Paolacci et al., 2010; Ramírez et al., 2019b]. Therefore, this entry in the taxonomy captures whether crowdsourcing experiments account for and communicate how returning workers were dealt with (or prevented in the first place).

The global scale of crowdsourcing platforms makes the active pool of workers to vary throughout the day [Difallah et al., 2018]. Thus, failing to account for (and therefore report on) this aspect may add confounding factors to the experiment that hurt its reproducibility [Qarout et al., 2019]. The taxonomy, therefore, needs to capture the ● *execution timeframe* of the experiment, by answering the simple question: over what timeframe was the experiment executed? (e.g., the experiment ran between January 1 and 10, every day at 2 PM). The ● *input dataset* fuels the task workers solve and is therefore crucial to indicate how this dataset was obtained and whether it is publicly available. Lastly, the crowdsourcing literature advises to fine-tune a crowdsourcing experiment through multiple pilots [Vaughan, 2017; Mason and Watts, 2009], and it is intuitively relevant that authors report whether ● *pilots* were performed before the main study.

Expert opinion

The interviews with experts organically touched on several of the experimental design attributes of crowdsourcing experiments, and while no new attributes emerged, their input allowed us to better scope and articulate the attributes. Their comments also provide a window into the challenges faced by experts in porting (and reporting on) experimental designs into crowdsourcing platforms.

Experts highlighted the importance of the experimental design in crowdsourcing experiments, indicating that "*the design is very critical*", and at the same time acknowledging the challenges posed by the uncontrolled environment provided by crowdsourcing platforms. One participant illustrated this aspect while bringing up the importance of the allocation of crowd workers to experimental conditions

> *"In the ideal case, if you [had] a full control of the crowdsourcing platform, you would actually be able to do [a proper] randomization of crowd workers [to experimental conditions]. And being a requestor in a crowdsourcing platform such as AMT or Crowdflower, you don' t have this opportunity (...) Therefore, by not having these opportunities you kind of [perform] certain workarounds."*

The importance of experimental design to task mapping, execution dates, execution of experimental conditions and returning workers were all covered by the participants during the interview. One expert conveyed through an illustrative example how all these aspect are interconnected:

> *Let's say you have two [conditions], like in an A/B test, and what differs in the two [associated] tasks is a design, literally a UI design. Then what you do is you run one task at 1pm today, and the other task tomorrow at 1pm as well. Well, time of the day is kind the same, but at the same time tomorrow [the platform would have a] slightly different population than today. It is very hard to control for some things. Maybe some participants that participate in your task today, will participate in your task tomorrow (recurrent workers), which definitely creates a certain bias and a "feel of work" effect.*

Accessing input dataset and other more general aspects of experimental design not covered by the taxonomy (e.g., hypotheses, research questions) were also mentioned. One participant went as far as to say that even in ideal case where all information is available, even running the experiment under a different requester name could introduce bias *"your requester name might be NASA (..) and people are just so excited to participate in studies run by NASA. It doesn' t mean that the effect is huge, but we know it [introduces] bias".*

5.3.3 Crowd

The active pool of workers in crowdsourcing platforms constitutes the population of human subjects who can participate in crowdsourcing experiments. The ● *target population* aims to explicitly capture the eligibility criteria used to screen crowd workers and determine potential participants of the experiment. Conversely, this entry captures if no specific criteria were applied, thus implicitly using the characteristics of workers in the selected crowdsourcing platform as eligibility criteria (i.e., the entire crowdsourcing population is considered eligible). For example, studies may consider using specific demographic attributes, concrete environments in which workers perform their work (OS, web browser, mobile phones), a threshold to the task acceptance rate, or the number of tasks completed (typically provided by the platforms).

Once the target population is defined, the ● *sampling mechanism* describes what strategies were used to recruit a diverse or representative set of participants from the target population. Crowdsourcing environments give researchers more affordable access to a large and diverse pool of subjects with a wide range of demographic attributes, as opposed to laboratory settings where scale (and diversity) is constrained by time and available funds [Rand, 2012; Gadiraju et al., 2015]. This diversity plays an essential part in the external validity of an experiment, similarly to how the soundness of the methodology contributes to its internal validity (indeed, crowdsourcing experiments can be both internally and externally valid as their laboratory counterparts [Horton et al., 2011]). However, sampling strategies in crowdsourcing environments face additional challenges not found in traditional settings. For example, how the underlying demographics are in flux based on the active workers [Difallah et al., 2018]; the lack of control over the subjects [Gadiraju et al., 2015]; and how easy it is for workers to quit, potentially causing non-random attrition and rendering the experimental conditions unbalanced [Rand, 2012].

Expert opinion

The interviews with the experts addressed and enriched the two attributes associated to this dimension. Participants also emphasized how crucial (and challenging) it is to find the right workers who suit the needs of the study and can participate in it, and make sure a reasonable and representative sample is obtained from this target population.

Accordingly, both the worker profile making up the target population and the mechanisms used to sample participants were deemed relevant by participants, giving examples such as demographics (*"I think is important to get to know some properties of the workers"*) and screening and qualification strategies applied to workers (*"First is your recruiting requirement that includes the screening process: how you selected your participants, you need to describe that"*).

Some participants went as far as to suggest that the mechanisms for inferring the properties of the target population, such as demographics and qualifications, are important and should be reported. For example,

> *"Some tasks are designed in a so academic way. You are literally being asked 'what do you see in the picture' [prompting the worker to annotate the picture], and then you have 25 questions about your age, income, race and something like that. I am a bit skeptical about this task design because first of all, it creates a huge bias. You know it's an academic study, you might really enjoy faking it, (..) or produce irrelevant results for example. It might depend on [the] platform, in some you might get some data about the population."*

In relation to the above, another participant reflected on some uncertainties about using demographics, casting some light into the above behavior *"it's not always clear if the gender of the participants is expected to be collected and reported, since not always [it] is relevant but it might be required by the reviewers"*.

The participants also provided specific examples of sampling strategies used in their experience, including sampling over time and geographical regions.

5.3.4 Task

The actual crowdsourcing tasks solved by workers can be regarded as the actual instruments or materials presented to the participants. The group of attributes presented here characterizes the tasks delivered as part of crowdsourcing experiments, considering organizational and operational details known to affect the results.

The nature and goal of the experiment shape the kind of tasks that are sent down to workers in crowdsourcing platforms, these aspects have been considered by previous research to propose a categorization of micro-tasks [Gadiraju et al., 2014]; we included in the taxonomy the ● *task type* attribute to capture this information.

The ● *task interface* in tandem with the ● *instructions* concern the exact user interface and guidelines presented to workers. Poorly written instructions can misdirect workers and affect their performance in the experiments, resulting in subpar responses [Wu and Quinn, 2017]. Besides, it can also reduce task intake and negatively affect how long the experiment takes to finish [Han et al., 2019]. Variants of a task interface could unravel performance differences [Mortensen et al., 2016]. And similarly, current evidence suggests enriched interfaces may aid workers, allowing them to attain higher performance [Sampath et al., 2014; Wilson et al., 2016; Ramírez et al., 2019a]. The full disclosure of operational details such as task and instructions materials favors reproducible research. To this end, the taxonomy also captures the exact ● *task interface source* (typically a combination of HTML, CSS, and JavaScript) uploaded to the crowdsourcing platform or related system (e.g., TurkServer [Mao et al., 2012]).

The ● *time allotted* workers to perform the task, as well as the ● *reward strategy* used to motivate them, can also influence the progress of the experiment and resulting performance of workers. In general, extrinsic factors such as proper monetary rewards can impact how much workers contribute [Mason and Suri, 2012]. For effort-intensive tasks, in particular, adequate payments can motivate workers to produce results of higher quality [Ho et al., 2015]. Low payments, however, may curb task intake and affect how fast experiments progress [Rogstadius et al., 2011; Han et al., 2019]. Payment mechanisms are not limited to the amount being paid but also how. Different payment strategies could also lead to effects on how workers perform [Difallah et al., 2014]. Tasks can also include elements that target intrinsic motivational factors, for example, to aid how workers engage and commit [Gadiraju and Dietze, 2017] (although one could argue that intrinsic factors are effects we may want to reduce in an experimental setting unless it is the subject of study). The time allotted workers to spend on the task naturally limits the associated cost of the experiment, besides studies in cost-effective crowdsourcing shed light on the impact of time on worker behavior and performance [Maddalena et al., 2016; Krishna et al., 2016]. We draw on these reasons to include in the taxonomy attributes that capture what mechanisms were employed to reward and motivate workers and whether time constraints where imposed.

Expert opinion

Aspects of task design such as the clarity of the instructions, the task interface and compensation were all suggested as critical design choices for the success of crowdsourcing experiments. Intuitively, participants indicated that these aspects are relevant and must be reported.

Participants weighted in regarding different levels of information provided about the task interface, some mentioning that, for example, the reporting of the task interface should go beyond screenshots to include links to source files (*"people report task design like a screenshot [...] It is maybe relevant to report the actual task design, like HTML, JavaScript, CSS files, so that people can reproduce it"*). Other experts, provided more nuanced opinions, e.g., *"If there is a paper that talks about different treatments that are interface-related and there is no interface, it is a red flag. If there is no source code, it' s a yellow flag - for me at least."*

In relation to the above, instructions and task descriptions were deemed critical. One participant illustrated how this information could completely shape the outcomes of an experiment, in the specific case of open-ended input:

> *"The prompts and tips that we are giving before starting the task will definitely impact the final result. So, for example, if you give any useful examples of the feedback that crowd workers can generate, most probably, all of the workers will start imitating the same examples."*

Instructions, along with the payment and perceived effort (time) affect were reported as influencing not only the quality of the results but the decision to participate in the first place.

5.3.5 Quality Control

The varying skills, motivations, backgrounds, and behavior of online workers make quality control a major challenge in crowdsourcing [Daniel et al., 2018]. Therefore, quality control mechanisms are fundamental to any crowdsourcing task to safeguard the resulting quality of the contributions. An initial step concerning quality is to define what constitutes an invalid answer or contribution, and the ● *rejection criteria* aims to capture this information.

Quality mechanisms can be set at different points in the crowdsourcing task. ● *Training* protocols can be used to prepare workers for the job, which can have a positive impact on their resulting performance [Liu et al., 2016]. Another group of mechanisms can be categorized as ● *in-task checks* or strategies embedded in the task to safeguard quality as contributions are collected. Whenever the nature of the task allows it, a fairly common practice is to embed gold items or attention checks to monitor the collected answers against the ground truth regularly. This simple technique can help to deal with malicious workers and avoid wasting collected contributions [Kittur et al., 2008]. The ● *gold items configuration* attribute then captures how these items were selected (typically a subset of the input dataset), how frequently these gold items appear in the task, and what threshold was used to filter out workers underperforming on these items. Other strategies involve imposing quality by design, such as adding feedback loops (where workers self-assess or receive external feedback) [Dow et al., 2012], or making workers collaborate [Drapeau et al., 2016; Schaekermann et al., 2018; Chen et al., 2018].

Another set of approaches are ● *post-task checks* or mechanisms leveraged upon task completion, which can complement in-task checks and training protocols. Manual inspection (e.g., by an expert) can be used to review contributions from workers, limited by the scale of crowdsourcing and the cost associated with expert feedback [Dow et al., 2012]. The assessment can also be computation-drive, for example, removing contributions that do not fall above a minimum time or agreement threshold [Marshall and III, 2013; Hansen et al., 2013].

Depending on the type of task in the experiment, one may rely on redundancy, considering the ● *number of votes per item* to be more than one, and leverage the same task to multiple workers to compensate potential noise. The ● *aggregation method* comes in tandem with redundancy. A typical strategy is to use majority voting; however, more sophisticated and effective alternatives have been developed to derive the right answer even when the majority may be wrong (e.g., [Dawid and Skene, 1979; Whitehill et al., 2009; Dong et al., 2013]).

Participants may drop out of experiments for different reasons and at different rates depending on the experimental conditions, introducing potential selection and attrition bias. Different experimental treatments may be less appealing to workers and therefore result in non-random or selective attrition, introducing confounds in the experiment affecting the comparisons of the conditions [Horton et al., 2011; Rand, 2012]. Copying with attrition, especially in online experiments, is instrumental to the study's internal validity, making sure dropouts stay roughly the same across conditions. Therefore, the taxonomy captures what ● *dropouts prevention mechanisms* were used to deal with the prevalence of task abandonment in crowdsourcing [Han et al., 2019; **?**]. For example, on Amazon Mechanical Turk, one can resort to using (neutral) qualification tasks to create a pool of potential participants and then send invitations to the actual tasks based on the assigned conditions [Ho et al., 2015]. Or, if one treatment incurs different effort levels than another, one may manipulate both treatments to include an effort-intensive activity and align how the difficulty of both tasks are perceived, regularizing the attrition rates [Rand, 2012].

Expert opinion

Quality control was a hot topic during the interviews, cited as *"one of the biggest concerns in designing the tasks"*. Participants considered quality control mechanisms as critical design choices that impact the results of experiments.

The quality control mechanism suggested by the participants focused on the task flow, from strategies applied before (e.g., training or qualification assessment), during (e.g., gold items and attention checks) and after the completion of the task (e.g., filtering out contributions). In describing the control mechanisms applied to specific experimental settings, the participants highlighted that quality control mechanisms and when they are applied depend on the type of task. In some contexts, involving open-ended tasks such as content-creation, strategies such as gold items may not be applicable, requiring manual checks after task completion.

> *"In my experiments, the main problem - and that you would always have - is that there is no golden data for the answers that the crowd workers provide. When we are asking the crowd workers to provide open-ended text or feedback or anything, they can write anything they like. Here the quality control is very challenging and almost near to impossible to achieve."*

Indeed, a participant cited the exploration of automatic approaches as a *"major area of research"* in his/her domain.

5.3.6 Outcome

The outcome dimension concerns details of the experimental results, more closely related to crowdsourcing, to aid their understanding, verification, and reproducibility. The soundness of the data analysis process and how to articulate the findings aid to derive the right conclusions from experiments [Gergle and Tan, 2014]. However, capturing these aspects are well beyond the scope of the taxonomy and refer them to available guidelines (e.g., [Abelson, 1995; Association, 2010]).

The ● *number of participants* and the ● *number of contributions* collected (in total and across conditions) help to understand the scale of the experiment. Also, the ● *dropout rate* and ● *demographics* quantify the level of attrition and diversity present in the experimental treatments, respectively. Based on the rejection criteria, the experimenter may regard some of the contributions as invalid. The taxonomy identifies two related elements, the number of ● *excluded participants*, perhaps malicious (or underperforming) workers, rendering a whole batch of answers from them as invalid. And ● *discarded data*, i.e., specific contributions that were excluded before the data analysis.

While the previous attributes cover different quantities in the outcome, the ● *data processing* captures any data manipulation step performed on the collected data. Intuitively, this information foster reproducible research alongside providing the ● *output dataset* derived from the experiment (i.e., the raw or aggregated contributions from crowd workers).

Expert opinion

Interestingly, the comments from the participants would naturally flow more towards attributes that would allow researchers to repeat crowdsourcing experiments, and to a lesser extent the assessment of the outcomes. Among the few to touch on outcome attributes, two participants mentioned that the output dataset, as well as potential confounds like demographics, are relevant aspects of crowdsourcing experiments that should be reported (*"it is important to report those aspects you did not control, such as demographic information"*).

5.3.7 Requester

The task requester is the person (or group of people) responsible for the design and execution of the experiment. The organizational and operational details related to how the requester set up the experiment are also important. The selected crowdsourcing ● *platform(s)* constitutes the environment in which the experiment runs. It represents an essential element that guides how the requester operationalizes the study based on the features offered by the platform. However, the available functionality tend to be limited for the requirements imposed by the experimental design [Kittur et al., 2008], often requiring requesters to implement additional features to cover this lack of support. These ● *implemented features* represent additional experimental instruments that also affect how feasible it is to replicate an experiment.

This limited support also translates into the available tools to facilitate interactions with workers (e.g., chat rooms and emails). ● *Requester-worker interactions* could potentially impact how workers engage and perform in the tasks from the experiment [Dow et al., 2012], so it is important to capture in the taxonomy if and how these interactions happen. However, simple designs may not require complex interactions (or no interaction at all). For example, requesters may not need to interact with workers to study task design in the context of classification tasks, relying entirely on leveraging clear guidelines to articulate what they expect from workers [Ramírez et al., 2019a; Mortensen et al., 2016].

Crowdsourcing environments define new legal grounds where policies may not be sufficiently defined (see [Felstiner, 2011] for a review of the legal aspects around crowdsourcing). The community is increasingly voicing the need for ● *ethical approvals* for experiments run in crowdsourcing environments [Graber and Graber, 2013; Martin et al., 2014]. ● *Informed consent* tends to be often required for research with human subjects, as well as ● *privacy and data treatment* statements for experiments that need to collect and store sensitive information. ● *Fair compensation* is also a relevant aspect from an ethical perspective. Computing fair wages indeed represents an active research area, considering that providing a minimum wage is not necessarily fair [Whiting et al., 2019]. Initially, crowdsourcing platforms such as Amazon Mechanical Turk were deemed a "marked for lemons" [Ipeirotis, 2010 (accessed August 26, 2020], an environment restraining committed workers from earning at least a legal minimum wage due to the prevalence of less-committed or malicious workers. Proper compensation can become a more frequent practice due to current features (like qualifications or badges) in crowdsourcing environments, advances in techniques to safeguard quality, plus underpayment issues being a reiterating topic addressed by recent literature [Kittur et al., 2013; Hara et al., 2018; Barbosa and Chen, 2019; Whiting et al., 2019]. Additional context can be given to online workers, so they become aware they take part in an experiment. This ● *participation awareness* is a relevant aspect because it is known to play a role in participant behavior [McCambridge et al., 2012].

Expert opinion

The operational context navigated by the requester was prompted by the participants in terms of the treatment of crowd workers, data management, and the technical environment.

In terms of ethics and data management, participant stressed that workers should receive a fair compensation for their contributions and protected from potential harm, and that it should be clear that this is the case (*"from the ethical perspective, [it should be reported] how much [crowd workers] were paid, whether they were exposed to certain content they were not naturally expected to see, like watching porn, cut bodies, and so* *on"*). Reflecting on the role of the different stakeholders in ensuring these practices, and whether ethical approvals should be required, some participant turned to practical recommendations:

> *"I think [ethical] approval is a different thing. It is up to the institution where the researchers are from, but on the very generic case we just want to make sure that the crowd workers were told in advance [what they were exposed to]".*

The friction between observing the privacy of the crowd workers while making sure enough data was collected was also raised *"I think is important to get to know some properties of the workers. But of course, it has implications in privacy".*

As for the technical environment, participants expressed the importance of reporting from simple attributes, such as the actual crowdsourcing platform, to platform-specific features and complex configurations. One participant illustrated the latter for quality control configurations (*"the whole quality control mechanism has a set of parameters that you have to set up, complex parameters that can lead to complex settings"*). But even for platform, it was argued that the role should be properly described (*"You can say that F8/Append is kind of a hub for accessing other different platforms"*). Moreover, the case of F8 (now Append) serving as a hub to other platforms could be problematic if it is not reported properly, as indicated by the participant:

> *"Very few papers tell you, I ran my stuff on F8, but under the hood, F8 was forwarding the tasks to these 50 different platforms [...] These are like variables you add to your experiment and make the whole thing so complex".*

5.4 Analyzing the State of Reporting

To study the state of reporting we surveyed 171 crowdsourcing experiments from the literature, and assessed them on the basis of the relevant crowdsourcing experiment attributes from our taxonomy. We used the attributes to assess the *completeness* and *reporting style* used in communicating the attributes. By doing this, we aim to shed light on the level of reporting in current practices. In the following, we describe the methodology and results.

5.4.1 Methods

The assessment considered the i) *completeness*, referring to whether the attribute in question could be derived from the reported experiment, and ii) *reporting style*, i.e., how the attribute was reported in the paper. We detail the process below.

We started with a systematic search for scientific papers describing crowdsourcing experiments. We iteratively refined a query for Elsevier's Scopus database to cover a wide range of crowdsourcing experiments in computer science and ensure the taxonomy fits well with a broad range of experiments. The query consisted of keywords covering different usages or words associated with *crowdsourcing* (e.g., "crowd-sourcing", "micro-task", "human-computation") and *experiments* (e.g., "experimental design", "study", "evaluation", "analysis", "intervention"). We also limited the search space to papers published in major conferences (e.g., CSCW, CHI, IUI, UIST, HCOMP) between January 2013 and June 2020, excluding journal papers from the query to keep the search space focused. We complemented this search by downloading from DBLP the list of papers in the proceedings of relevant conferences not indexed by Scopus.

The search identified a total of 670 candidate papers. These papers were screened by two researchers to include papers describing experiments or user studies engaging crowd workers as subjects through a crowdsourcing platform. We excluded experiments leveraging on existing crowdsourced datasets, engaging workers in a role not related to the evaluation, experiments in laboratory settings, purely qualitative studies (e.g., surveys) and experiments where crowd workers did not come from an open call to a crowdsourcing platform but rather from a more restricted environment. The screening started with a random subset of 50 papers to calibrate the aforementioned eligibility criteria between the researchers, and then proceeded with the researchers screening independently the rest of the documents (the researchers agreed on 94% of the decisions in the set of 50 papers, resolving disagreements by consensus). This process identified 172 papers reporting crowdsourcing experiments. We refer readers to the Appendix A.2 for more details on the search and screening process, including the query and eligibility criteria.

The analysis then proceeded in three phases that aimed at assessing the reporting of experiments in terms of the taxonomy of relevant crowdsourcing experiment attributes. It started with a small sample of 15 papers, as explained in the above section, to iteratively refine the initial taxonomy and the assessment metrics. Second, to calibrate the interpretation of the taxonomy and the assessment metrics, a subset of 40 papers (~25% of the total number of included papers) was analyzed by three researchers, resulting in an average agreement of 90%. The researchers assessed all the taxonomy attributes for the same set of papers, considering as a match if they agreed on the completeness value (i.e., the presence of an attribute in a paper). Finally, the rest of the documents were distributed equally among the researchers and assessed independently. Only one experiment was analyzed per research article, and in the case of papers reporting on multiple experiments, one experiment was randomly selected. Any doubts emerging during the independent tagging were flagged by the researchers for discussion and resolved by consensus.

The analysis followed a meticulous procedure where the researchers leveraged on their experience (and the input from experts) to assess the completeness of the attributes of an experiment, evaluating only those applicable to i) the type of experiment reported in the paper, and ii) the objective of the study. This approach was followed as it was clear from our pilots, interviews with experts and our own experience that not all attributes were relevant for each experiment. With these criteria, the researchers ended up applying 26 of the 39 attributes in all cases, with the other 13 deemed not applicable (N/A) depending on the type and goal of the experiment. We found that, on average, four attributes were N/A and that 80% of the papers in our analysis had at most six N/A attributes. Notice that we adjusted the total number of attributes on a per-paper basis, excluding the N/A attributes. Similarly, the percentages we report for each attribute in Figure 5.4 are also adjusted by removing N/A cases. We refer readers to Appendix A.4 for further details.

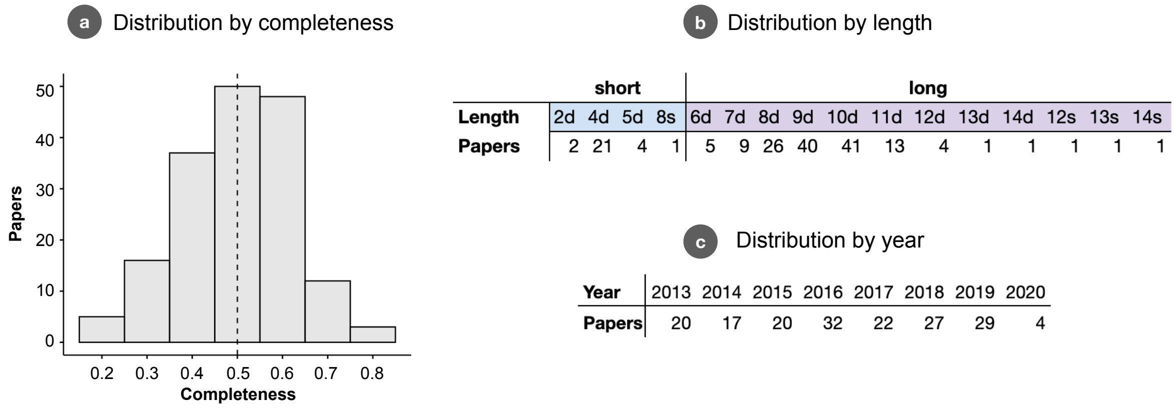
Taking the quality control attribute *In-task checks* as a concrete example, the researchers first checked whether the paper explicitly mentioned any quality checks performed during the task (e.g., gold items, attention checks), and to the best of their abilities assessed whether they were indeed applicable for the particular type of experiment (e.g., the experiment studied cheating behavior where in-task checks would not make sense). Researchers then continued by checking if the details relevant to this quality control check could be derived from the description, even if the papers did not explicitly describe it (and in this case, it was counted as "implicit reporting"). There are also some attributes with a "directional association". For example, we considered the interface and instructions as complete if the paper provided the source code (and, for example, omitted screenshots of these). Likewise, if *informed consent* was explicitly reported, we considered as complete the *participation awareness* attribute.

The assessment of the reporting covered not only the main content of the paper but also the appendix, supplementary materials and any source code or repository associated to the paper. To identify these additional sources, the researchers checked for links referenced or cited in the paper, as well as the official page of the paper in the publisher's website. This procedure was followed to assess the completeness of all the attributes in the taxonomy.

5.4.2 Results

We analyzed 171[7] papers published in major venues to understand how (and to what extend) crowdsourcing experiments are reported. As mentioned previously, we used the attributes in the taxonomy to guide the analysis, assessing the completeness (e.g., *is the attribute addressed?* And, *is it explicitly reported?*), and the reporting style used in communicating the attributes (e.g., using screenshots, a figure, etc.). By doing this, we

[7] One paper was excluded from the analysis as the full-text was not available to the researchers.



	short				long											
Length	2d	4d	5d	8s	6d	7d	8d	9d	10d	11d	12d	13d	14d	12s	13s	14s
Papers	2	21	4	1	5	9	26	40	41	13	4	1	1	1	1	1

Year	2013	2014	2015	2016	2017	2018	2019	2020
Papers	20	17	20	32	22	27	29	4

Figure 5.3: Some descriptive statistics of the 171 papers we analyzed. a) The distribution of papers by completeness, i.e., the proportion of attributes reported. The dashed line indicates the median. b) The distribution of papers based on their length (*"d"*stands for double-column and *"s"* for single-column format. c) The distribution of papers based on their year of publication.

seek to elucidate the level of reporting in current practices. We limited the analysis to 38 attributes, excluding the *implemented features* attribute.[8]

Figure 5.3 depicts some descriptive statistics of the papers we analyzed. The papers reported, on average, 49.9% of the attributes (19 attributes), with 89 of the 171 articles reporting at least 50% of the attributes. Of the papers analyzed, 89 of 171 were published between 2013 and 2016, and 82 of 171 were published between 2017 and June of 2020. Most of the documents (167 papers) were in double-column format and four in single-column format. The majority (143/171) are long articles (defined as those with #pages $\geq$ 6 double-column, with 3 papers of 12+ pages long in single-column format), and 28 are short (#pages $<$ 6 double-column, and one paper with 8 pages single-column).

We analyzed the level of reporting by year and length to contrast past and recent reporting efforts and differences based on paper length due to more pages available. We did not observe a clear pattern of recent papers addressing more attributes when compared to older ones, but we noticed an interesting trend of increasing reporting in longer documents. As for the attributes in our taxonomy, in 32/38, we noticed an increase in the percentage of papers reporting them, with relative differences of up to $3x$. However, the reporting was low, as the number of attributes covered by at least 50% of the papers was 14 for short and 16 for long. Overall, the median completeness for short papers was 42.4% and 51.5% for long. Also, we did not observe any trend in the usage of supplementary materials to cover the lack of space, as only 15/171 papers provided supplementary material[9]. Of these,

[8] It is important to report on the features implemented to operationalize an experiment, but we found it difficult to reliably determine whether this attribute was applicable and being reported when analyzing the papers.

[9] The supplementary material included screenshots of the task (7/15), task source code (5/15), input dataset

three were short papers. A detailed breakdown of the results by year and length can be found at `https://tinyurl.com/ReportingState`.

Figure 5.4 summarizes the analysis for each of the six dimensions in the taxonomy. In the following, we discuss the results for the attributes in each dimension, and when necessary, we will touch on interesting and relevant differences on the reporting level based on the year and length.

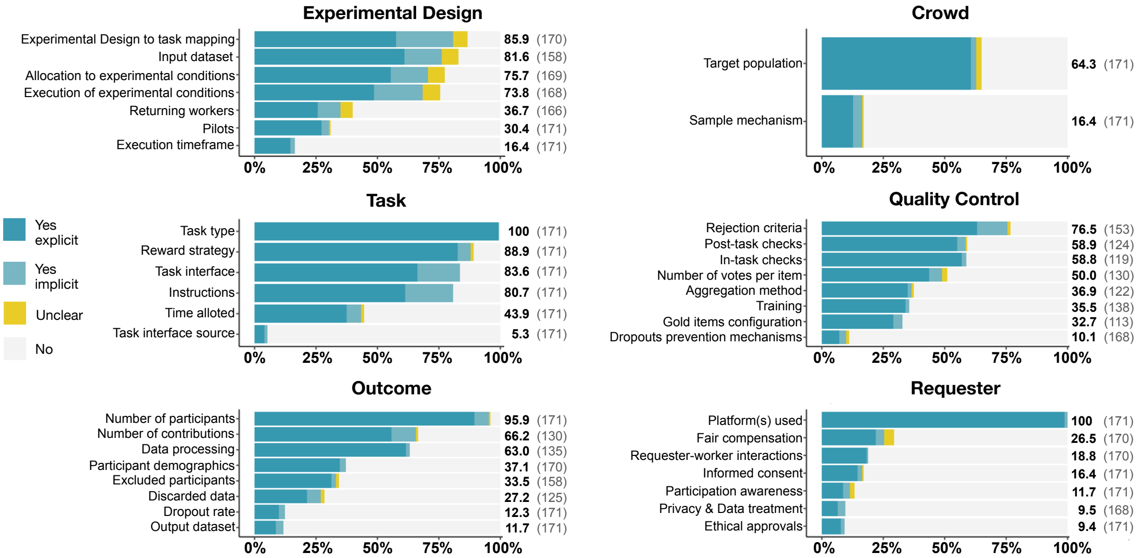


Figure 5.4: The state of reporting, based on the attributes in the taxonomy, for a set of 171 papers describing crowdsourcing experiments published in major venues. The number of papers to which the attribute applied is indicated in parenthesis, and in boldface, the percentages of papers reporting on it (explicitly and implicitly).

Experimental design

Despite being the cornerstone of crowdsourcing experiments, the explicit reporting of experimental design attributes is for the most part shallow and unclear, as indicated by the low levels of explicit as well as high levels of unclear reporting – the highest among the rest of the dimensions. Overall, 4/7 attributes in the experimental design dimension – those defining the design, mapping and execution of the experiment in the crowdsourcing platform – are reported explicitly by 52.4% to 65.2% of the papers, representing the highest reported attributes for this dimension. When the experimental design is relatively

(8/15), output dataset (8/15), and other details related to data analysis (e.g., scripts, notebooks). Most of this material was provided as external links to a code repository (9/15) or document (3/15), two were in the appendix, and one was found via the publisher's digital library (video presentation).

simple and straightforward to map to a crowdsourcing environment (e.g., a within-subjects design mapping to a single crowdsourcing task), these four attributes can be derived with sufficient confidence, as indicated by the 16.4% to 24.7% of implicit reporting. But as soon as the design is more complex (e.g., a mixed design with multiple conditions and no clear hint on how these map to tasks), this is no longer the case, as indicated by the 0.6% to 7.7% of shallow and unclear reporting.

In terms of reporting style, the input dataset allowed for the richer set of approaches, compared to the rest of the attributes that were reported mostly in text. A 65.2% of the papers reported the actual input dataset used in the tasks given to workers, providing references or a link to an external site. However, some papers only indicated the process that was followed to prepare the input dataset and that would allow a researcher to construct a dataset of similar characteristics (which we coded as implicit, 16.4% of papers). In two cases, the provided link was no longer reachable, which brings a different issue: reporting links to supplementary materials that do not survive the "test of time."

Under-reported details include whether pilots were performed, if and how returning workers were controlled, and the experiments' execution timeframe. Pilots help tune the design of the experiment; however, it was only reported by 27.5% of the papers. Returning workers, applicable depending on the design and how it is mapped, was reported by very few papers (only 27.1%). Similarly, only 14.6% of the papers reported the period in which the experiment run. For long papers, all of the attributes show an increase in explicit reporting, with relative differences of up to $2x$. But, only 4/7 attributes are reported by at least 50% of papers against 3/7 in short articles.

Crowd

In general, 62% of the papers explicitly reported properties characterizing the target population of workers suitable for the study, with 2.3% of the papers implicitly suggesting what constitutes the target population (e.g., based on the goal of the study, we inferred that the task was open to everybody in the selected platform). Ultimately, this information was not clear in 2.3% of the analyzed articles. The exact mechanism used to sample a diverse and representative set of participants was reported in only 12.9% of the papers (e.g., one possible strategy would be to sample systematically at different times of the day). In 3.5% of the papers, this information was not explicitly addressed, but we derived by inference, which we coded as implicit (e.g., by providing the name of an external tool, by an auxiliary task to reach potential participants first). The above comes at a surprise, given the highly diverse and changing nature of crowd workers, and the potential influence of participants' characteristics and associated environment in task performance. In terms of reporting styles, these attributes are embedded in the paper as text. The target population was reported by 64.3% of long papers vs. 50% for short. And while the sampling mechanism was largely under-reported, it also shows an increase (14% for long, and 7% for short).

Task

The task dimension is relatively well covered. The type of task, the reward strategy, and how the task looks like (including the instructions) are the attributes reported by most of the papers. Of the papers analyzed, 83.6% of them explicitly described the strategy used to reward workers for their contributions (almost all papers described monetary compensations as the reward strategy, except for three articles resorting to volunteering). The task interface was explicitly reported by 66.1% of the papers. Some articles (17.5%) did not explicitly include a screenshot of the task and instead described it partially in text (which we coded as implicit). When reported, the interface is mostly depicted using screenshots and just a handful of papers (4.1%) provided the actual source code of the interface, and in two cases the link provided was no longer available (coded as implicit). But even relevant information such as the instructions is reported explicitly by 61.4% of the papers. The instructions can be reconstructed from partial screenshots and textual descriptions in some cases (19.3% of the papers). Explicit time constraints are reported in only 38% of the papers. In longer documents, we noticed an improvement in explicit reporting for details regarding the task interface (72% for long, 35.7% for short), instructions (66.4% for long, 35.7% for short), and the reward strategy (88.1% for long, 60.7% for short). The other attributes also show an increase, but the level of reporting is under 50%.

Quality control

Quality control mechanisms can take place before, during, and after the task. Despite their importance, the papers did not fully report the mechanisms employed. Most reported attributes include the criteria used to reject contributions (rejection criteria, 64.1%), the mechanisms used after the task to safeguard quality (post-task checks, 55.6%), and in-task mechanisms for quality control (in-task checks, 57.1%). Though we found cases where these attributes were addressed rather implicitly (12.4% for rejection criteria, 3.3% for post-task checks, and 1.7% for in-task checks)[10]. Training sessions, when applicable, were reported by only 34.1% of the papers. And details regarding the redundancy employed (number of votes per item) and aggregation method were reported by 44.6% and 35.2%

[10] Examples of implicit reporting of rejection criteria would be when one could infer a paper actually accepted all contributions and filtered out on the data analysis part based on a metric like completion time, or a participant was not considered because they did not provide some demographic information.

of papers, respectively. In cases where gold items where used as quality checks, only 29.2% of papers reported explicitly the configuration (gold items configuration). Dropout prevention mechanisms, closely related to engagement, were reported by very few papers (7.1%). As for the reporting style, quality control attributes are described as text, with few cases relying on additional tables and figures to report information such as the number of votes per item (1 paper) and training sessions (2 papers). Although, for the most part, still inadequate, there is an increase in explicit reporting in 6/8 attributes for long papers, though only two of these are reported by at least 50% of papers (rejection criteria 64.6%, and in-task checks 57.6%).

Outcome

The number of participants is reported in most cases (by 90.1% of the papers), and by inference, the number of contributions, described explicitly in just 56.2% of papers. In some cases, 5.8% of papers for number of participants and 10% for number of contributions, these details were not explicitly addressed, and we coded as implicit (e.g., the number of participants could be inferred based on the total number of contributions and the contributions per annotator). Data processing steps on top of the contributions from the crowd are reported by 61.5% of the papers. For the rest of the attributes, the papers do not paint a clear picture. For example, only 21.6% to 31.6% of the papers reported explicitly discarded data and excluded participants (due to quality or rejection criteria), and just 9.9% described dropouts (workers leaving the experiment for different reasons). The output dataset and participant demographics were also poorly reported, with only 8.8% and 34.7% of the papers providing these details explicitly. Papers reported the attributes mostly as text, accompanied by figures and tables for attributes such as the number of participants (21 papers) and contributions (11 papers), and participant demographics (5 papers). For long documents, we noticed an increase in explicit reporting of 5/8 attributes, though, for the most part, reporting of these attributes is under 40%, except for data processing (63.5% for long, 50% for short) and the number of contributions (57% for long, 52.2% for short).

Requester

Overall, the attributes in this dimension were vastly under-reported. The papers reported the most basic information, the selected platform, but they poorly addressed the rest of the attributes (6/7 attributes reported explicitly by 6.5% to 22.9% of the papers). Most of the under-reported attributes relate to the ethics of the experiment. These attributes cover if compensation was fair (at least a minimum wage), whether workers gave their consent and were aware they took part in an experiment, if the study received ethical approvals and were in compliance with data privacy policies. Regarding the reporting style, the papers described these attributes as text. Interestingly, we noticed an improvement in recent years from an ethical perspective, when comparing papers from *2013-2016* and *2017-2020*. Reporting ethical approvals, fair compensation, and privacy and data treatment attributes increased. This insight aligns with current works addressing issues such as underpayments in crowdsourcing (see [Barbosa and Chen, 2019] for a brief overview). However, these improvements in reporting are still far from ideal.

5.4.3 Potential threats to validity

The strength of the experimental evidence, as well as the ability of researchers to repeat and reproduce experiments, relies on the underlying methodology and how well this methodology and results are described. We have observed a reporting gap for the different dimensions in our taxonomy, which raises questions about experimental validity of crowdsourcing experiments reported in the literature. This section briefly discusses issues associated with under-reported attributes and how these connect to known biases from experimental research [Pannucci and Wilkins, 2010] that affect the validity and integrity of scientific experiments.

Sampling bias

Researchers can introduce bias by inferring conclusions from a sample that is not representative of their target population. In crowdsourcing this can happen by not properly defining (or reporting on) a target population or because the mechanisms provided by the platform (or implemented by the researchers) fail to obtain a proper sample. As observed, the concrete target population and sampling mechanisms were not properly reported by 38% and 84.2% of the papers, respectively. The resulting demographics were omitted for 62.9% of the experiments – without weighing in the mechanisms to derive this information. Even returning workers, under-reported by 68.7% of the papers, may also bias the sample since these can hinder the goal of reaching a broader and more diverse set of workers. The lack of proper reporting of these attributes not only affects the proper assessment of population samples, but it may also introduce practical challenges to replicability given the diverse characteristics of crowd worker populations.

Selection bias

Bias can also occur due to the strategies chosen to assign participants to different conditions or cohorts. In crowdsourcing this could happen, for example, when the tasks associated to different conditions are executed under different operational settings (such as the ones mentioned in the following). Of the analyzed papers, 83.6% did not report over what timeframe the experiment run, which is known to determine the characteristics of the active pool of workers [Difallah et al., 2018]. In this context, under-reporting how the conditions were executed, as in 33.9% of the papers, can amplify this issue and render conditions with different sets of workers (e.g., conditions run sequentially with one capturing workers typically active during the morning and other conditions with night workers). Task design attributes can also influence what type of workers are attracted to the tasks in the experiment. We observed 19.3% of papers failing to report the instructions, 12.3% the compensation, and 57.3% the time constraints associated with the tasks. Reporting on these attributes is thus important, as they can reveal unintentional and intentional bias benefiting experimental conditions.

Observation and response bias

Also called the "The Hawthorne Effect" [Sedgwick and Greenwood, 2015], observation bias arises when participants, being aware that they are taking part to a study, modify their behavior or contributions. In a crowdsourcing experiment, this might be triggered when providing informed consents and acknowledging the scientific purpose of the task or, as made clear by the experts, in a more subtle way by the requester name, task design and even by the (fair) compensation. In relation to this, participants might feel compelled to orient their responses towards what they believe the expected findings are, in what is called response bias. Only a few papers reported on participation awareness, and 90.1% failed to report whether participants were informed of contributing to a study.

Design bias

Researchers may fail to account for inherent biases present in experiments leading to what is called design bias. In crowdsourcing, in addition to other forms of experimental bias, there is a large body of literature on different forms of bias introduced by all aspects of task design and execution [Wu and Quinn, 2017; Ho et al., 2015; Sampath et al., 2014; Maddalena et al., 2016; Gadiraju et al., 2017a; Qarout et al., 2019]. Our analysis shows that 20.6% of the papers under-report how the experimental design was mapped to the selected crowdsourcing platform. Moreover, papers omitted information related to task design such as interface (16.4%), instructions (19.3%), and compensation (12.3%). Even the implementation of desirable practices, such as running pilots to refine the experiments design, was reported explicitly in only 27.5% of the papers. This information gap can open the room to non-comparable conditions, especially when translating the experiment to other platforms with different features and workers.

Measurement bias

Bias can arise from measuring instruments of varying quality or errors in the data collection process. In crowdsourcing contexts, we can associate this bias to quality control. Overall, quality control attributes were under-reported, with 41.9% of papers failing to report post-tasks checks, 41.2% in-task checks, and 64.5% did not describe training sessions. Omitting these details may raise questions about the quality of the collected data, and allow for differences in quality when trying to rerun (or build upon) an experiment. For example, one could obtain better results by including training sessions.

Ethical integrity

We observed that the vast majority of the papers failed to report on attributes such as fair compensation (77.6%), informed consent (84.2%), privacy & data treatments (90.5%), and ethical approvals (90.6%). Not reporting on these attributes makes it difficult to assess whether experiments followed ethical guidelines and made sure workers were treated fairly and not exposed to any harm. This is clearly the ultimate goal. Understandably, processes for securing the ethics of experiments may vary from one institution to another, with some institutions enforcing ethical approvals on all studies while some put the burden on the researchers. What is clearly missing, and reflected on the interviews with experts and our own experience, are practical ethical guidelines that would allow researchers to make informed decisions about their crowdsourcing experiments, and proper support from crowdsourcing platforms to make sure these guidelines can be properly observed.

5.5 A Checklist for Reporting

The insights from the literature overview, expert interviews and the state or reporting, paint a daunting image for the reporting of crowdsourcing experiments, calling for the development of better guidelines and resources. In this section we take a small step in this direction, and describe the process that led to the development of a reporting checklist for crowdsourcing experiments. We stress that the goal of the type of support we explore here is to contribute to more a transparent reporting that can enable a better assessment of the validity of experiments, as well as to repeatability.

5.5.1 Methods

The checklist is the result of a process that involved three main steps: i) internally growing the taxonomy into a reporting sheet that was used in the internal assessment and pilots; ii) obtaining feedback from experts on the generic reporting sheet and views on alternatives strategies for reporting, and iii) developing and refining the checklist.

Starting from the taxonomy, two researchers defined each of the attributes and prepared questions that would require authors to specify if and how each attribute is reported in a paper. These definitions were then used as the initial template for piloting the reporting of three papers of the authors, leading to the definition of the first "sheet" for reporting. The sheet contained the attributes in our taxonomy as rows along with their definitions. We filled out the sheet with excerpts from the papers explicitly addressing the attributes in our taxonomy. This step gave us a concrete example that we could use to discuss among the team members. This discussion led to a reframing of attributes and deciding what to include as part of the reporting. The discussion also led to brainstorming alternatives for the presentation, where we considered datasheets Gebru et al. [2018]; Mitchell et al. [2019] or checklists Shamseer et al. [2015]; Pineau et al. [2020] as potential options for guidelines for reporting.

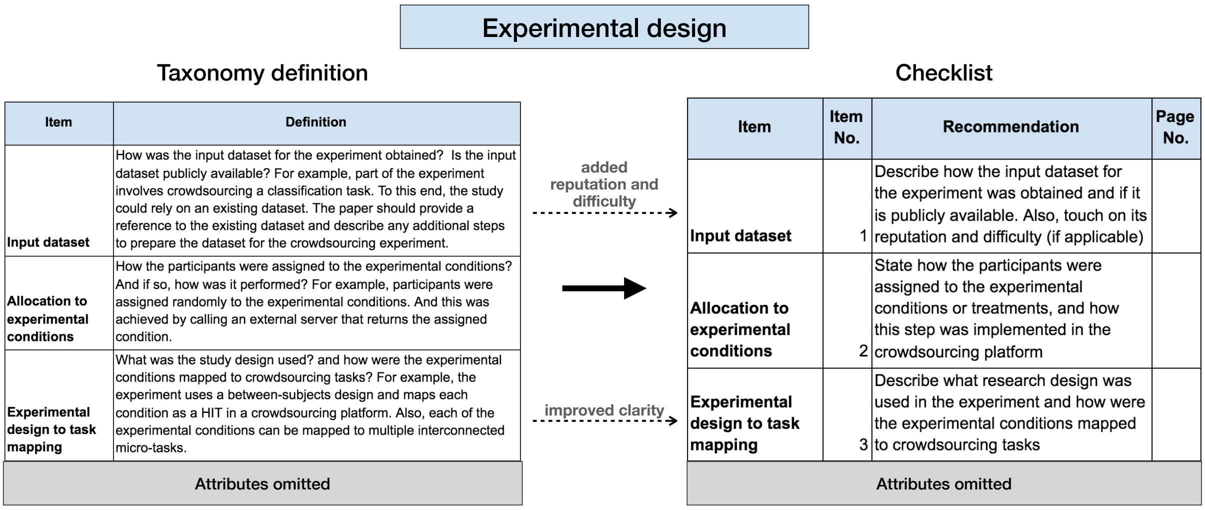
We focused Parts 2 and 3 of the interview with experts on assessing the definition of the attributes (as well as spotting missing ones) and collecting suggestions on how to present the guidelines for reporting. Part 2 of the interview introduced participants with a portion of the final taxonomy (one or two dimensions). Participants (besides giving feedback on the relevance of the attributes and spot any missing one) were asked to read aloud attributes description and to assess and provide feedback on the framing. Part 3 asked, *"How do you think these aspects of crowdsourcing experiments could be framed into a tool for reporting?"*, providing examples such as datasheets or checklists.

The answers collected from these parts of the interview were used, first, to improve the definitions of the attributes and, second, to derive the final form of the guidelines: a checklist for reporting crowdsourcing experiments.

5.5.2 Proposed checklist

The second and third parts of the interview with experts focused on 1) assessing the *clarity* and *completeness* of the attributes in the taxonomy, and 2) how we could exploit this taxonomy and turn it a tool for reporting. Based on the taxonomy and feedback we received in the interviews, we derived a checklist for reporting crowdsourcing experiments, detailed in Appendix A.1.

The *completeness* of the taxonomy makes sure the major ingredients of crowdsourcing experiments are well covered, while the *clarity* concerns the interpretation and understanding of the attributes in the taxonomy. We found the taxonomy to be quite complete: the participants who identified missing elements (6/10) suggested attributes that were already present, but in other parts of the taxonomy they did not assess. One of the participants, *P2*, suggested a "when to stop" attribute (not present in the taxonomy) for quality control, saying, *"[...] when we are trying to collect quality data, we do not want to stop the task until we reach a threshold. Also, we might finish the task because we have already collected very good quality data"*. While we find this to be an excellent point, we argue that it is more related to data collection practices, for instance, to obtain data for ML, than for experiments in crowdsourcing. Therefore, we ultimately did not incorporate this aspect as an attribute in the taxonomy (besides, it could be covered by the existing post-task checks attribute).


Experimental design
Taxonomy definition
Checklist
added reputation and difficulty
improved clarity

Item	Definition
Input dataset	How was the input dataset for the experiment obtained? Is the input dataset publicly available? For example, part of the experiment involves crowdsourcing a classification task. To this end, the study could rely on an existing dataset. The paper should provide a reference to the existing dataset and describe any additional steps to prepare the dataset for the crowdsourcing experiment.
Allocation to experimental conditions	How the participants were assigned to the experimental conditions? And if so, how was it performed? For example, participants were assigned randomly to the experimental conditions. And this was achieved by calling an external server that returns the assigned condition.
Experimental design to task mapping	What was the study design used? and how were the experimental conditions mapped to crowdsourcing tasks? For example, the experiment uses a between-subjects design and maps each condition as a HIT in a crowdsourcing platform. Also, each of the experimental conditions can be mapped to multiple interconnected micro-tasks.
Attributes omitted	

Item	Item No.	Recommendation	Page No.
Input dataset	1	Describe how the input dataset for the experiment was obtained and if it is publicly available. Also, touch on its reputation and difficulty (if applicable)	
Allocation to experimental conditions	2	State how the participants were assigned to the experimental conditions or treatments, and how this step was implemented in the crowdsourcing platform	
Experimental design to task mapping	3	Describe what research design was used in the experiment and how were the experimental conditions mapped to crowdsourcing tasks	
Attributes omitted			

Figure 5.5: Example of updates we introduce to the attribute definitions in the taxonomy based on the feedback from the interviews (Experimental design dimension in this case), including the final framing as a checklist.

In general, the attributes were clear and relevant to the participants. The suggestions of what attributes to add were used instead to improve the definition of the existing attributes, adding to the concrete feedback that focused on improving the definitions. Figure 5.5 depicts some of the changes to the experimental design dimension and the final framing as a checklist. Of the 39 attributes in the taxonomy, 31 were assessed by participants.[11] Participants provided feedback, on average, to 10 attributes (between 7 and 14), with an attribute being assessed by 2 to 4 participants. We noticed that 16

[11] The requester dimension was not discussed with any of the participants. But aspects such as platform, ethics, and privacy were considered by some participants in the initial part of the interview.

attributes were clear "as is", 9 were clear, but received suggestions to improve them, 3 were clear only after some clarification from the interviewer, and only 3 were not clear.[12]

As for how we could exploit the taxonomy, we asked participants, *"How do you think these aspects could be framed into a tool for reporting?"*. We clarified that the framing naturally depends on the intended usage and shared two examples. The first example was the PRISMA checklist [Shamseer et al., 2015] that assesses the methodological rigor of systematic reviews. And the other was datasheets found in the ML community, which are self-contained structured summaries of a dataset creation pipeline or model performance [Gebru et al., 2018; Mitchell et al., 2019]. The participants favored the checklist format (5/10) over the datasheet alternative (1/10), while the rest did not explicitly mention a checklist but suggesting it based on their answer.

5.5.3 Intended usage and adoption

The intended usage of the taxonomy, how it is framed as a tool for reporting, and the adoption by the research community go hand in hand.

A checklist format gives paper authors more flexibility to describe the different aspects of the experiments in the paper's main content and supplementary materials (weighing in typical page limits), indicating where an attribute is being described. We aim for the checklist to serve as a resource that guides researchers in what they report, helping them be thorough and systematic in communicating the details of their crowdsourcing experiments (i.e., serving as a reminder, *"when we are very deep into our data analysis part we forget the basic stuff that should be reported"*). Unlike a datasheet, a checklist is not self-contained, which was indicated by one of the participants, *"To me, it is easier to see it as a report, with all the information on the same page [...]"*. While a self-contained summary as a supplementary material could be more convenient to readers, it demands more effort from authors since the main ingredients of the experiment would still need to be described in the paper, at least at a high level. Guided by the feedback from the experts, we ultimately opted for the checklist format, considering that it helps authors report their experiments, while avoiding additional efforts, and readers to navigate the details of a crowdsourcing experiment.

During the interviews, the participants raised challenges associated with the adoption of a checklist — *"it all comes down to motivation"*. One of this challenges was associated with the research community and the current practices around crowdsourcing experiments. As explained by a participant: *"you can probably check 30% of those boxes and the*

[12] We coded an attribute as clear if it was clear to all the participants who assessed it. To code as "clear with suggestions", "clear after clarification", or "not clear", we only expected at least one feedback to fall in these categories, prioritizing the not clear option.

paper will be published [...] no major motivations in the academic world other than your paper being published". Adopting a checklist would go in tandem with evolving current community practices for assessing papers reporting on crowdsourcing experiments (as seen, for example, in the ML community, where reproducibility checklists are part of the submission/reviewing process[13]). However, we hope that bringing awareness about the potential issues of not standardizing practices can push the community in the right direction. In response to this challenge, the experts highlighted the importance of growing adoption organically by making people aware of the benefits and lowering the barriers to adoption. As commented by the participants: *"convince the people that these are the important things that you should follow"*, and *"make researchers' life easier"*.

We foresee promising avenues of future work that could address these challenges and facilitate adoption by the research community. In this chapter, however, we limited ourselves to starting the conversation towards reproducible crowdsourcing experiments. For instance, convincing people to adopt the checklist could be addressed by providing empirical evidence on how using the checklist aids reproducible results. Or making its usage so trivial that researchers just adopt it. For example, *"As a way to make their lives easier, you say, hey, tell me the platform, tell me the task ID, provide credentials, click ENTER, and a GitHub repository is created with all this information"*.

5.6 Discussion

The current state of reporting in crowdsourcing research still misses providing details beyond basic attributes associated with task design, quality control, requester, and experiment design and outcome. According to our analysis, at least 70% of papers report the selected platform, how the experimental design maps and executes, reward strategy, task interface, instructions, rejection criteria, and the number of participants. However, if we consider only explicit reporting, this list narrows to reward strategy, the number of participants, and the selected platform. While these attributes are relatively well-covered, either explicitly or implicitly, most tend to be under-reported by at least 50% of the papers. Among the six dimensions in our taxonomy, the requester — with attributes covering the ethics of experiments — was among the least reported by the papers. These issues open the room for potential threats to validity associated with missing details regarding the experimental design and its operationalization.

Under-reporting poses the interesting question of why the attributes are poorly reported in the first place and how we can overcome this situation. Our analysis and feedback from experts attribute this issue to the limited guidance and awareness on what needs to

[13] `https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist`

be reported. It is therefore of paramount importance to encourage further transparency in reporting, as to ensure that we as a community aim for higher standards of evidence and reproducibility of results. As our results shows, however, we are still lacking in this regard. By providing a checklist and depicting where the research community stands in terms of reporting practices, we expect our work to stimulate additional efforts to move the transparency agenda forward.

Chapter 6

Conclusions and Future work

We devoted our efforts in this thesis to move crowdsourcing forward and closer to being treated as a science rather than art. We approached this goal on two related fronts in crowdsourcing research: providing guidance on task design, and supporting the process of running and reporting experiments. The selection of multi-predicate classification problems as our use case was driven by its importance, emphasized by the share of relevant tasks in crowdsourcing marketplaces. The first part of this thesis proposed strategies to aid worker performance in classification tasks and highlighted relevant unexplored dimensions of task design. This required us to provide empirical evidence by running controlled experiments in rather uncontrolled environments like crowdsourcing platforms, which in turn motivated the second part of the thesis. This part aimed to aid in running and reporting crowdsourcing experiments to make crowdsourcing more accessible to researchers and practitioners, offloading the need for in-depth knowledge of the inherent characteristics of crowdsourcing platforms and programming skills to make controlled experiments possible.

This chapter introduces the lessons learned as a result of these three years of the Ph.D. program at the University of Trento. Here we reflect on these lessons and introduce a summary of our contributions, as well as discuss limitations and avenues of future work.

6.1 Lessons Learned

To structure crowdsourced classification tasks such as to provide support and obtain performance improvements **(RQ1).**

We systematically studied the effect of text highlighting in human computation in Chapter 2, identifying the quality requirements that automatic techniques for text highlighting should possess to help with text classification and estimating the potential impact of good (and bad) highlighting. We uncovered the potential of aggregating highlighting by multiple, independent annotators (or algorithms) showing that aggregation is practical and useful, somewhat analogously to what happens in a crowdsourced classification where we aggregate multiple votes on items. Then, we discussed interesting and perhaps unexpected effects of highlighting, important to make them effective, such as giving time to workers to get used to working with highlights. Our results ultimately show that text highlighting could support workers in solving the tasks faster without sacrificing quality.

In Chapter 3, we explored and provided guidance on a somewhat unexplored aspect of task design in multi-predicate classification: how to ask complex questions to classify items. A comprehensive crowdsourcing experiment involving multiple datasets from domains comprising systematic literature reviews, customer feedback analysis, content moderation, and crowd verification allowed us to provide empirical evidence on the impact of predicate formulation strategies on individual and collective worker performance. Our results showed that superior classification performance could be obtained by querying a complex predicate as multiple (simpler) questions instead of asking a single coarse predicate. The results also highlight that the predicate formulation strategies we explored could result in slower task completion time, representing an important trade-off for task designers.

The preliminary results from our experiment on hybrid classification in Chapter 3 emphasizes the importance of predicate formulation as a task design dimension. We showed that querying simpler predicates could enable more effective coupling of ML classifiers and favor long term reusability of already trained models. We believe that there is potential for training highly-specialized models that couple effectively with the performance of workers (instead of learning models classify items based on complex predicates directly). Besides, answering simpler questions outputs reusable (and detailed) knowledge about the capabilities of crowd and machine classifiers. From the perspective of crowd workers, this means reapplying learned skills, and for machines, it involves classifying unseen items (and filter out at least items that are "obviously" not relevant).

Providing support to researchers in designing and running crowdsourcing experiments so as to address potential biases associated with this process **(RQ2).**

We draw from our experience and distilled the challenges and coping strategies to run controlled experiments in crowdsourcing environments. We showed specific instances of how running crowdsourcing experiments without coping strategies can impact the experimental design, assignment, and workers participating in the experiments. Using task design evaluation, we distilled the challenges and quantified how it could change the outcomes of experiments. However, these challenges are not only tied to task design evaluation, and in general, they play a role in the success of crowdsourcing experiments.

Inspired by these lessons, and how frequently they occur in the literature, we designed and implemented CrowdHub, a system that extends crowdsourcing platforms and allows requesters to run controlled crowdsourcing projects. We presented a demo of CrowdHub

[Ramírez et al., 2019b] and received positive and constructive feedback from researchers in the human computation community. These discussions allowed us to arrive at the current design goals and set of features that constitute the system. The available features enable task requesters to build crowdsourcing workflows, such as creating datasets for training machine learning models or designing and executing complex crowdsourcing experiments. CrowdHub is an open-source project, and we made available on GitHub the source code of the frontend[1] and backend[2] layers.

Aiding researchers in reporting crowdsourcing experiments to ensure these are reproducible **(RQ3).**

Chapter 5 brought to light the gap in current reporting practices for the different dimensions in our taxonomy. The state of reporting in crowdsourcing research still misses providing details beyond basic attributes like the task interface, instructions, what was the rejection criteria, and the platform used. However, how the experiment was actually mapped and run on a crowdsourcing platform is, for the most part, shallow and unclear. We also noticed that the vast majority of the papers failed to report on attributes related to the ethics of the experiment. For example, attributes covering if compensation was fair and if the study received ethical approvals. The bottom line here is that "the bias is in the details", and we stress this in Section 5.4.3 by connecting the under-reported attributes to known biases from experimental research.

Our insights have many implications and aim to push the community towards standardized reporting of crowdsourcing experiments. However, current guidelines somewhat overlook advising researchers on what and how to report, and they instead focus primarily on effective task design and practical recommendations for running experiments. The research community should also seek to develop guidelines and best practices for reporting and reproducibility of crowdsourcing experiments. As a first step in this direction, we introduced a taxonomy of relevant ingredients characterizing crowdsourcing experiments and used this taxonomy to analyze the state of reporting of 171 articles published in top venues. This process allowed us to identify gaps in current reporting practices. To help address these issues, we leveraged the resulting taxonomy and feedback from experts to propose a checklist for reporting crowdsourcing studies.

It is clear from Chapter 5 that improved transparency in reporting is a shared responsibility among the different stakeholders in the crowdsourcing ecosystem. ***Researchers*** may be asked to agree to a code of conduct or follow guidelines, like our checklist, to improve the current level of reporting. As a ***research community***, we need to develop guidelines and best practices (e.g., on how we set up and report experiments) to increase

[1] CrowdHub frontend: `https://github.com/TrentoCrowdAI/crowdhub-web`

[2] CrowdHub backend: `https://github.com/TrentoCrowdAI/crowdhub-api`

transparency, strength, and reproducibility of crowdsourcing experiments. And such guidelines should also emphasize the ethics and fairness behind experiments. In turn, ***venues*** need to enforce and adopt higher reporting standards, mirroring the initiatives taking place in other communities.

Platforms can benefit from the insights and design recommendations emerging from this thesis and address operational barriers to running and reporting experiments. Platform providers may aim for solutions that are "experiment-aware" (e.g., by offering features to treat experiments as first-class citizens). Indeed, platforms can play a major role in both i) helping task requesters to design experiments that are consistent with accepted ethical guidelines (from informed consent to minimum wage) and ii) helping to generate reports that facilitate publication of relevant experiment information to aid reproducibility.

6.2 Contributions

Our contributions can be summarized as follows:

1. We provide evidence on the positive impact of text highlights as a tool for supporting workers in classification tasks, specifically for scenarios where the focus is on task completion speed. We show that text highlights can reduce task completion speed without losses in quality of the derived contributions. The experiments in Chapter 2 indicate that providing text highlights of good quality can significantly reduce the decision time by almost half while maintaining (but not necessarily increasing) the accuracy of workers [Ramírez et al., 2019a].

2. The quality assessment of the machine-generated highlights in Chapter 2 provided us with insights into the nature and potential limitations of automated approaches. Extractive summarization approaches are not trained for a specific predicate and therefore are prone to generate less useful highlights. Instead, a question-answering model generated shorter highlights specific for each predicate and dataset and resulted in overall higher quality (but sensitive to how the predicate was formulated) [Ramírez et al., 2019a].

3. We introduce the predicate formulation problem as an additional task design dimension to consider when structuring classification tasks and show how the resulting performance is affected by different predicate formulation strategies. Our comprehensive experiments in Chapter 3 provide empirical evidence on how the different strategies to pose a complex question to the crowd can offer gains in quality but at a slower task completion speed [Ramírez et al., 2021a].

4. The experiments in Chapter 3 also offer preliminary evidence on the potential of predicate formulation in the context of hybrid classification, suggesting performance gains even in its simplest collaborative approach, by assigning crowd and machine classifiers parts of a complex predicate they are more suited to classify [Ramírez et al., 2021a].

5. In Chapter 4, we discuss the challenges associated with running controlled experiments in crowdsourcing platforms and propose coping mechanisms to deal with these challenges, as well as highlight the potential impact of running experiments uncontrolled [Ramírez et al., 2020a].

6. The insights from Chapter 4, the lessons we learned from running crowdsourcing experiments, and the limitations of current systems that extend crowdsourcing platforms led us to design and implement a tool for running controlled crowdsourcing projects. It offers features for systematically evaluating task design to aid researchers and practitioners during the design and deployment of crowdsourcing projects across multiple platforms, as well as features for researchers to run controlled experiments [Ramírez et al., 2019b, 2020a].

7. Chapter 5 derives the major design decisions of crowdsourcing experiments that play a role in crowdsourcing tasks and, therefore, on an experiment's outcome. A bottom up approach rooted in the literature allowed us to derive a taxonomy of attributes characterizing experiments run in crowdsourcing platform [Ramírez et al., 2020b; Ramírez et al., 2021b]. We then leveraged this taxonomy to analyze the state of reporting in crowdsourcing literature and identify aspects that are frequently communicated and those that tend to go under-reported.

8. To address the gap in reporting practices, Chapter 5 then proposes a checklist to facilitate the job of understanding and replicating crowdsourcing experiments [Ramírez et al., 2021b]. The checklist seeks to help experimenters describe their setup in a standardized format and readers to understand the used methodology and how it was implemented, serving as a tool that complements existing experimental research guidelines.

9. There is a limited amount of datasets with individual crowd votes to study multi-predicate classification problems. Therefore, this thesis also contributes multiple datasets covering a broad landscape of tasks and expanding different domains and difficulty levels [Ramírez et al., 2019c, 2021a].

6.3 Limitations

Our work on structuring tasks to aim for better performance in multi-predicate classification has many limitations. The experiments on the impact of text highlighting on annotation speed and accuracy were focused on binary text classification tasks. We see the potential of highlighting as a tool to support workers in other kinds of tasks. However, we did not cover different use cases for text highlighting, such as supporting information retrieval and question answering tasks. This work also looked at additional factors influencing the impact of text highlighting, showing that, for example, experience with the task increases the benefits of highlighting, and workers adapt their behavior when documents are longer. However, we did not consider the environment in which tasks are being solved (e.g., working from a laptop or a phone, and other behavioral traces), which could also potentially play a role in the resulting performance of workers [Gadiraju et al., 2017a].

Chapter 3 is limited to studying predicate formulation for the simpler case: a complex predicate composed of two individual predicates. Indeed, a relevant question is whether or not similar results on predicate formulation can be achieved if the predicate complexity goes beyond two predicates. Focusing on the simpler case allowed us to study one aspect of the predicate formulation problem — task design — as the high dimensionality of this problem makes it intractable to crowdsource for every possible scenario in our experiment. Therefore, we did not consider settings with more than two individual predicates to make the crowdsourcing experiment manageable. Yet, the results from the simulations to cover settings (like going beyond two predicates) are encouraging, as they align with the insights we derived from our crowdsourcing experiment. However, further experiments are needed to support our insights in these settings. Another relevant aspect of this problem concerns developing algorithms for querying complex predicates (e.g., to know how to split a predicate and "route" predicates to the more suitable workers). Our work did not study this aspect, but we see some potential connections with existing work on multi-predicate classification [Krivosheev et al., 2021].

Chapter 4 summarized the challenges and coping strategies for running controlled crowdsourcing experiments, showed the impact of uncontrolled experiments, and proposed CrowdHub based on these learnings — but this chapter leaves unexplored how (and to what extent) CrowdHub would support researchers running controlled experiments. Answering this question would require user studies (for problems beyond multi-predicate classification) to assess the extent to which CrowdHub supports researchers and practitioners in designing and running crowdsourcing projects. While CrowdHub's features emerged from lessons we learned by running controlled experiments for a specific problem, we believe they are generic enough to serve as building blocks to a wide range of crowdsourcing studies.

Besides, researchers can easily extend CrowdHub to other kinds of tasks, despite its support limiting to multi-predicate classification. These points suggest that CrowdHub could support different kinds of experiments and problems, but further studies are needed to validate its support.

Our work on aiding current reporting practices in crowdsourcing research is limited to what is reported, which might be an incomplete picture of the design and operationalization of the experiments — but that is what is eventually available to the community to build upon and replicate. While a systematic process was followed to cover as much research landscape as possible, the search can not be considered an exhaustive account of crowdsourcing experiments in the literature. Yet, the 670 conference papers screened and the 171 analyzed in detail provide a representative sample of the current state of affairs. Also, it is worth noting that we did not manage to cover the four stakeholders we mentioned in Section 6.1, as we did not interview people (e.g., managers and crowd workers) from major crowdsourcing platforms. However, the participants we interviewed possess ample experience designing and executing crowdsourcing experiments in these platforms, with research output published in major SIGCHI conferences; they helped provide thoughtful design and operational angles to the attributes in the taxonomy and the final checklist. Finally, while we proposed a checklist to aid current reporting practices, we did not explore how to facilitate the adoption of such checklist.

6.4 Future Work

This thesis identifies many interesting avenues of future work.

Our work on text highlighting leveraged human- and machine-generated highlights, identifying the quality range for text highlights to become useful, as well as the impact of low-quality highlights — but we ultimately focused on the effects and not "the how". How to effectively and efficiently obtain text highlights is indeed a relevant research question. We see the efficient and effective generation of text highlights as a relevant problem for Human-AI approaches, and we expect the insights from Chapter 2 to stimulate and inform research in this direction. Here, we envision pipelines combining crowd and machines to produce useful text highlights. For example, the crowd may help in giving feedback to automatic highlighting techniques (e.g., by assessing the quality of machine-generated highlights or even providing concrete examples of good highlights) following an active learning setup.

In our experiments, workers spent more time on the task as the documents grew in size, but this was not the case for one of the datasets in our study on text highlights. We left this behavioral aspect unexplored, and we find it an interesting direction of future work because it shows how people adapt their behavior when documents are longer, which could change the effect of highlighting.

Our work on the predicate formulation problem considers two task design choices for presenting individual predicates to the crowd: either asking the predicates on the same task vs. on separate tasks. Both task design choices offer superior results over the baseline that asks the complex predicate directly, but there is not enough evidence to inform decisions based on given problem settings. We find this an interesting direction of future work, where we design algorithms that model workers, tasks, and predicates to automatically learn how to formulate complex predicates to meet quality goals while operating under a budget.

As part of our future work, we plan to run user studies to evaluate the extent to which CrowdHub supports researchers in running crowdsourcing experiments and practitioners in deploying crowdsourcing workflows.

Our work on aiding current reporting practices also identifies promising avenues of future work. We plan to evaluate if and how the proposed checklist aid reproducible research, explore methods to automatically derive crowdsourcing experiment reports from existing crowd platforms, extend the scope of the reporting to other crowdsourcing tasks such as data collection, and develop a system with crowdsourcing experiments as first-class citizens.

Bibliography

Abelson, Robert P. *Statistics As Principled Argument.* Psychology Press, 1995. ISBN 0805805281.

Aggarwal, Charu C. and Zhai, ChengXiang. A survey of text classification algorithms. In *Mining Text Data.* 2012. doi: 10.1007/978-1-4614-3223-4_6. URL https://doi.org/10.1007/978-1-4614-3223-4_6.

Ahmad, Salman; Battle, Alexis; Malkani, Zahan, and Kamvar, Sepandar D. The jabberwocky programming environment for structured social computing. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, Santa Barbara, CA, USA, October 16-19, 2011*, pages 53–64, 2011. doi: 10.1145/2047196.2047203. URL https://doi.org/10.1145/2047196.2047203.

Alagarai Sampath, Harini; Rajeshuni, Rajeev, and Indurkhya, Bipin. Cognitively inspired task design to improve user performance on crowdsourcing platforms. In *CHI 2014*. ACM, 2014.

Arnold, Matthew; Bellamy, Rachel K. E.; Hind, Michael; Houde, Stephanie; Mehta, Sameep; Mojsilovic, Aleksandra; Nair, Ravi; Ramamurthy, Karthikeyan Natesan; Olteanu, Alexandra; Piorkowski, David; Reimer, Darrell; Richards, John T.; Tsay, Jason, and Varshney, Kush R. Factsheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM J. Res. Dev.*, 63(4/5):6:1–6:13, 2019. doi: 10.1147/JRD.2019.2942288. URL https://doi.org/10.1147/JRD.2019.2942288.

Association, American Psychological. *Publication manual of the American Psychological Association.* American Psychological Association, sixth edition, 2010.

Balahur, Alexandra; Steinberger, Ralf; Kabadjov, Mijail A.; Zavarella, Vanni; der Goot, Erik Van; Halkia, Matina; Pouliquen, Bruno, and Belyaeva, Jenya. Sentiment analysis in the news. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*, 2010. URL http://www.lrec-conf.org/proceedings/lrec2010/summaries/909.html.

Barbara, Kitchenham and Charters, Stuart. Guidelines for performing systematic literature reviews in software engineering. 2, 01 2007.

Barbosa, Natã Miccael and Chen, Monchu. 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 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 543, 2019. doi: 10.1145/3290605.3300773. URL https://doi.org/10.1145/3290605.3300773.

Barowy, Daniel W.; Curtsinger, Charlie; Berger, Emery D., and McGregor, Andrew. Automan: a platform for integrating human-based and digital computation. In *Proceedings of the 27th Annual ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications, OOPSLA 2012, part of SPLASH 2012, Tucson, AZ, USA, October 21-25, 2012*, pages 639–654, 2012. doi: 10.1145/2384616.2384663. URL https://doi.org/10.1145/2384616.2384663.

Bender, Emily M. and Friedman, Batya. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6: 587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://www.aclweb.org/anthology/Q18-1041.

Bernstein, Michael S.; Little, Greg; Miller, Robert C.; Hartmann, Björn; Ackerman, Mark S.; Karger, David R.; Crowell, David, and Panovich, Katrina. Soylent: a word processor with a crowd inside. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology, New York, NY, USA, October 3-6, 2010*, pages 313–322, 2010. doi: 10.1145/1866029.1866078. URL https://doi.org/10.1145/1866029.1866078.

Blanco, Roi; Halpin, Harry; Herzig, Daniel M.; Mika, Peter; Pound, Jeffrey; Thompson, Henry S., and Tran, Duc Thanh. Repeatable and reliable search system evaluation using crowdsourcing. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 923–932, 2011. doi: 10.1145/2009916.2010039. URL https://doi.org/10.1145/2009916.2010039.

Buhrmester, Michael; Talaifar, Sanaz, and Gosling, Samuel. An evaluation of amazon' s mechanical turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, 13:149–154, 03 2018. doi: 10.1177/1745691617706516.

Buhrmester, Michael D.; Kwang, Tracy Nai, and Gosling, Samuel D. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science : a journal of the Association for Psychological Science*, 6 1:3–5, 2011.

Callaghan, William; Goh, Joslin; Mohareb, Michael; Lim, Andrew, and Law, Edith. Mechanicalheart: A human-machine framework for the classification of phonocardiograms. *PACMHCI*, 2(CSCW):28:1–28:17, 2018. doi: 10.1145/3274297. URL https://doi.org/10.1145/3274297.

Chan, An-Wen; Hrobjartsson, Asbjorn; Haahr, Mette T; Gotzsche, Peter C, and Altman, Douglas G. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA*, 291(20):2457–65, May 2004.

Chandler, Dana and Kapelner, Adam. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, 90:123 – 133, 2013. ISSN 0167-2681. doi: https://doi.org/10.1016/j.jebo.2013.03.003. URL http://www.sciencedirect.com/science/article/pii/S016726811300036X.

Chandler, Jesse; Mueller, Pam, and Paolacci, Gabriele. Nonnavetamong amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavior research methods*, 46, 07 2013. doi: 10.3758/s13428-013-0365-7.

Chawla, Nitesh V.; Bowyer, Kevin W.; Hall, Lawrence O., and Kegelmeyer, W. Philip. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16:321–357, 2002. doi: 10.1613/jair.953. URL https://doi.org/10.1613/jair.953.

Chen, Quanze; Bragg, Jonathan; Chilton, Lydia B., and Weld, Daniel S. Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing. *CoRR*, abs/1810.10733, 2018. URL http://arxiv.org/abs/1810.10733.

Cheng, Justin and Bernstein, Michael S. Flock: Hybrid crowd-machine learning classifiers. In *CSCW 2015*, 2015. doi: 10.1145/2675133.2675214. URL http://doi.acm.org/10.1145/2675133.2675214.

Cheng, Justin and Cosley, Dan. How annotation styles influence content and preferences. In *24th ACM Conference on Hypertext and Social Media (part of ECRC), HT '13, Paris, France - May 02 - 04, 2013*, pages 214–218, 2013. doi: 10.1145/2481492.2481519. URL `https://doi.org/10.1145/2481492.2481519`.

Cheng, Justin; Teevan, Jaime; Iqbal, Shamsi T., and Bernstein, Michael S. Break it down: A comparison of macro- and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, pages 4061–4064, 2015. doi: 10.1145/2702123.2702146. URL `https://doi.org/10.1145/2702123.2702146`.

Correia, António; Schneider, Daniel; Paredes, Hugo, and Fonseca, Benjamim. Scicrowd: Towards a hybrid, crowd-computing system for supporting research groups in academic settings. In *Collaboration and Technology - 24th International Conference, CRIWG 2018, Costa de Caparica, Portugal, September 5-7, 2018, Proceedings*, pages 34–41, 2018. doi: 10.1007/978-3-319-99504-5_4. URL `https://doi.org/10.1007/978-3-319-99504-5_4`.

Craik, Fergus I.M. and Lockhart, Robert S. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11(6), 1972. ISSN 0022-5371. doi: https://doi.org/10.1016/S0022-5371(72)80001-X. URL `http://www.sciencedirect.com/science/article/pii/S002253717280001X`.

Crump, Matthew J. C.; McDonnell, John V., and Gureckis, Todd M. Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PLOS ONE*, 8(3):1–18, 03 2013. doi: 10.1371/journal.pone.0057410. URL `https://doi.org/10.1371/journal.pone.0057410`.

Daniel, Florian; Kucherbaev, Pavel; Cappiello, Cinzia; Benatallah, Boualem, and Allahbakhsh, Mohammad. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Comput. Surv.*, 51(1):7:1–7:40, 2018. doi: 10.1145/3148148. URL `https://doi.org/10.1145/3148148`.

Dawid, A. P. and Skene, A. M. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C Applied Statistics*, 28(1), 1979.

Dean, Jeffrey and Ghemawat, Sanjay. Mapreduce: Simplified data processing on large clusters. In *6th Symposium on Operating System Design and Implementation (OSDI 2004), San Francisco, California, USA, December 6-8, 2004*, pages 137–150, 2004. URL `http://www.usenix.org/events/osdi04/tech/dean.html`.

Demartini, Gianluca; Trushkowsky, Beth; Kraska, Tim, and Franklin, Michael J. Crowdq: Crowdsourced query understanding. In *CIDR 2013, Sixth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 6-9, 2013, Online Proceedings*, 2013. URL `http://cidrdb.org/cidr2013/Papers/CIDR13_Paper137.pdf`.

Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton, and Toutanova, Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Difallah, Djellel Eddine; Catasta, Michele; Demartini, Gianluca, and Cudré-Mauroux, Philippe. Scaling-up the crowd: Micro-task pricing schemes for worker retention and latency improvement. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

Difallah, Djellel Eddine; Filatova, Elena, and Ipeirotis, Panos. Demographics and dynamics of mechanical turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 135–143, 2018. doi: 10.1145/3159652.3159661. URL `https://doi.org/10.1145/3159652.3159661`.

Dimara, Evanthia; Bezerianos, Anastasia, and Dragicevic, Pierre. Narratives in crowdsourced evaluation of visualizations: A double-edged sword? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5475–5484, 2017.

Dong, Xin Luna; Berti-Equille, Laure, and Srivastava, Divesh. Data fusion : Resolving conflicts from multiple sources. In *Procs of WAIM2013*. Springer, 2013. ISBN 9783642362576. doi: 10.1007/978-3-642-36257-6.

Dow, Steven; Kulkarni, Anand Pramod; Klemmer, Scott R., and Hartmann, Björn. Shepherding the crowd yields better work. In *CSCW '12 Computer Supported Cooperative Work, Seattle, WA, USA, February 11-15, 2012*, pages 1013–1022, 2012. doi: 10.1145/2145204.2145355. URL `https://doi.org/10.1145/2145204.2145355`.

Drapeau, Ryan; Chilton, Lydia B; Bragg, Jonathan, and Weld, Daniel S. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*, 2016.

Dunn, Olive Jean. Multiple comparisons using rank sums. *Technometrics*, 6(3), 1964. ISSN 00401706.

Eickhoff, Carsten. Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 2018.

Faltings, Boi; Jurca, Radu; Pu, Pearl, and Tran, Bao Duy. Incentives to counter bias in human computation. In *Proceedings of the Seconf AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2014, November 2-4, 2014, Pittsburgh, Pennsylvania, USA*, 2014. URL `http://www.aaai.org/ocs/index.php/HCOMP/HCOMP14/paper/view/8945`.

Felstiner, Alek. Working the crowd: employment and labor law in the crowdsourcing industry. *Berkeley J. Emp. & Lab. L.*, 32:143, 2011.

Fowler, Robert and Barker, Anne. Effectiveness of highlighting for retention of text material. *Journal of Applied Psychology*, 59, 06 1974. doi: 10.1037/h0036750.

Franklin, Michael J.; Kossmann, Donald; Kraska, Tim; Ramesh, Sukriti, and Xin, Reynold. Crowddb: answering queries with crowdsourcing. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2011, Athens, Greece, June 12-16, 2011*, pages 61–72, 2011. doi: 10.1145/1989323.1989331. URL `https://doi.org/10.1145/1989323.1989331`.

Gadiraju, Ujwal and Dietze, Stefan. Improving learning through achievement priming in crowdsourced information finding microtasks. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference, Vancouver, BC, Canada, March 13-17, 2017*, pages 105–114, 2017. URL `http://dl.acm.org/citation.cfm?id=3027402`.

Gadiraju, Ujwal and Kawase, Ricardo. Improving reliability of crowdsourced results by detecting crowd workers with multiple identities. In *International Conference on Web Engineering*, pages 190–205. Springer, 2017.

Gadiraju, Ujwal; Kawase, Ricardo, and Dietze, Stefan. A taxonomy of microtasks on the web. In *25th ACM Conference on Hypertext and Social Media, HT '14, Santiago, Chile, September 1-4, 2014*, pages 218–223, 2014. doi: 10.1145/2631775.2631819. URL `https://doi.org/10.1145/2631775.2631819`.

Gadiraju, Ujwal; Möller, Sebastian; Nöllenburg, Martin; Saupe, Dietmar; Egger-Lampl, Sebastian; Archambault, Daniel W., and Fisher, Brian. Crowdsourcing versus the laboratory: Towards human-centered experiments using the crowd. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments - Dagstuhl Seminar 15481, Dagstuhl Castle, Germany, November 22-27, 2015, Revised Contributions*, pages 6–26, 2015. doi: 10.1007/978-3-319-66435-4_2. URL `https://doi.org/10.1007/978-3-319-66435-4_2`.

Gadiraju, Ujwal; Checco, Alessandro; Gupta, Neha, and Demartini, Gianluca. Modus operandi of crowd workers: The invisible role of microtask work environments. *IMWUT*, 1(3):49:1–49:29, 2017a. doi: 10.1145/3130914. URL `https://doi.org/10.1145/3130914`.

Gadiraju, Ujwal; Yang, Jie, and Bozzon, Alessandro. 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 2017, Prague, Czech Republic, July 4-7, 2017*, pages 5–14, 2017b. doi: 10.1145/3078714.3078715. URL `https://doi.org/10.1145/3078714.3078715`.

Gaur, Yashesh; Lasecki, Walter S.; Metze, Florian, and Bigham, Jeffrey P. The effects of automatic speech recognition quality on human transcription latency. In *W4A 2016*, 2016. doi: 10.1145/2899475.2899478. URL `https://doi.org/10.1145/2899475.2899478`.

Gebru, Timnit; Morgenstern, Jamie; Vecchione, Briana; Vaughan, Jennifer Wortman; Wallach, Hanna M.; III, Hal Daumé, and Crawford, Kate. Datasheets for datasets. *CoRR*, abs/1803.09010, 2018. URL `http://arxiv.org/abs/1803.09010`.

Gergle, Darren and Tan, Desney S. *Experimental Research in HCI*, pages 191–227. Springer New York, New York, NY, 2014. ISBN 978-1-4939-0378-8. doi: 10.1007/978-1-4939-0378-8_9. URL `https://doi.org/10.1007/978-1-4939-0378-8_9`.

Gier, Vicki Silvers; Kreiner, David S., and Natz-Gonzalez, Amelia. Harmful effects of preexisting inappropriate highlighting on reading comprehension and metacognitive accuracy. *The Journal of general psychology*, 136 3, 2009.

Gomes, Ryan; Welinder, Peter; Krause, Andreas, and Perona, Pietro. Crowdclustering. In *NeurIPS 2011*, 2011. URL `http://papers.nips.cc/paper/4187-crowdclustering`.

Graber, Mark A and Graber, Abraham. Internet-based crowdsourcing and research ethics: the case for irb review. *Journal of medical ethics*, 39(2):115–8, Feb 2013. doi: 10.1136/medethics-2012-100798.

Haas, Daniel; Ansel, Jason; Gu, Lydia, and Marcus, Adam. Argonaut: Macrotask crowdsourcing for complex data processing. *PVLDB*, 8(12):1642–1653, 2015. doi: 10.14778/2824032.2824062. URL `http://www.vldb.org/pvldb/vol8/p1642-haas.pdf`.

Han, Lei; Roitero, Kevin; Gadiraju, Ujwal; Sarasua, Cristina; Checco, Alessandro; Maddalena, Eddy, and Demartini, Gianluca. All those wasted hours: On task abandonment in crowdsourcing. In *WSDM 2019*, 2019. doi: 10.1145/3289600.3291035. URL `https://doi.org/10.1145/3289600.3291035`.

Hansen, Derek L.; Schone, Patrick John; Corey, Douglas; Reid, Matthew, and Gehring, Jake. Quality control mechanisms for crowdsourcing: peer review, arbitration, & expertise at familysearch indexing. In *CSCW 2013*, pages 649–660, 2013. doi: 10.1145/2441776.2441848. URL `https://doi.org/10.1145/2441776.2441848`.

Hara, Kotaro; Adams, Abigail; Milland, Kristy; Savage, Saiph; Callison-Burch, Chris, and Bigham, Jeffrey P. 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 2018, Montreal, QC, Canada, April 21-26, 2018*, page 449, 2018. doi: 10.1145/3173574.3174023. URL `https://doi.org/10.1145/3173574.3174023`.

Ho, Chien-Ju; Slivkins, Aleksandrs; Suri, Siddharth, and Vaughan, Jennifer Wortman. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pages 419–429, 2015. doi: 10.1145/2736277.2741102. URL `https://doi.org/10.1145/2736277.2741102`.

Horton, John J.; Rand, David G., and Zeckhauser, Richard J. The online laboratory: conducting experiments in a real labor market. *Experimental Economics*, 14(3):399–425, Sep 2011. ISSN 1573-6938. doi: 10.1007/s10683-011-9273-9. URL `https://doi.org/10.1007/s10683-011-9273-9`.

Hosseini, Mahmood; Shahri, Alimohammad; Phalp, Keith; Taylor, Jacqui, and Ali, Raian. Crowdsourcing: A taxonomy and systematic mapping study. *Computer Science Review*, 17, 05 2015. doi: 10.1016/j.cosrev.2015.05.001.

Hube, Christoph; Fetahu, Besnik, and Gadiraju, Ujwal. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. 2019.

Ipeirotis, Panos. *Mechanical Turk, Low Wages, and the Market for Lemons*, 2010 (accessed August 26, 2020). URL `https://www.behind-the-enemy-lines.com/2010/07/mechanical-turk-low-wages-and-market.html`.

Jain, Ayush; Sarma, Akash Das; Parameswaran, Aditya G., and Widom, Jennifer. Understanding workers, developing effective tasks, and enhancing marketplace dynamics: A study of a large crowdsourcing marketplace. *PVLDB*, 10(7):829–840, 2017. doi: 10.14778/3067421.3067431. URL `http://www.vldb.org/pvldb/vol10/p829-dassarma.pdf`.

James, Gareth; Witten, Daniela; Hastie, Trevor, and Tibshirani, Robert. *An Introduction to Statistical Learning: With Applications in R*, pages 181–186. Springer Publishing Company, Incorporated, 2013. ISBN 1461471370.

Kamar, Ece; Hacker, Severin, and Horvitz, Eric. Combining human and machine intelligence in large-scale crowdsourcing. In *AAMAS 2012*, 2012. URL `http://dl.acm.org/citation.cfm?id=2343643`.

Kittur, Aniket; Chi, Ed H., and Suh, Bongwon. Crowdsourcing user studies with mechanical turk. In *Proceedings of the 2008 Conference on Human Factors in Computing Systems, CHI 2008, 2008, Florence, Italy, April 5-10, 2008*, pages 453–456, 2008. doi: 10.1145/1357054.1357127. URL `https://doi.org/10.1145/1357054.1357127`.

Kittur, Aniket; Smus, Boris; Khamkar, Susheel, and Kraut, Robert E. Crowdforge: crowdsourcing complex work. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, Santa Barbara, CA, USA, October 16-19, 2011*, pages 43–52, 2011. doi: 10.1145/2047196.2047202. URL `https://doi.org/10.1145/2047196.2047202`.

Kittur, Aniket; Nickerson, Jeffrey V.; Bernstein, Michael S.; Gerber, Elizabeth; Shaw, Aaron D.; Zimmerman, John; Lease, Matt, and Horton, John J. The future of crowd work. In *Computer Supported Cooperative Work, CSCW 2013, San Antonio, TX, USA, February 23-27, 2013*, pages 1301–1318, 2013. doi: 10.1145/2441776.2441923. URL `https://doi.org/10.1145/2441776.2441923`.

Krause, Markus and Kizilcec, René F. To play or not to play: Interactions between response quality and task complexity in games and paid crowdsourcing. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2015, November 8-11, 2015, San Diego, California, USA.*, pages 102–109, 2015. URL `http://www.aaai.org/ocs/index.php/HCOMP/HCOMP15/paper/view/11575`.

Krishna, Ranjay A; Hata, Kenji; Chen, Stephanie; Kravitz, Joshua; Shamma, David A; Fei-Fei, Li, and Bernstein, Michael S. Embracing error to enable rapid crowdsourcing. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 3167–3179. ACM, 2016.

Krivosheev, Evgeny; Casati, Fabio; Caforio, Valentina, and Benatallah, Boualem. Crowdsourcing paper screening in systematic literature reviews. In *HCOMP 2017*, 2017.

Krivosheev, Evgeny; Casati, Fabio; Baez, Marcos, and Benatallah, Boualem. Combining crowd and machines for multi-predicate item screening. *PACMHCI*, 2(CSCW), November 2018. ISSN 2573-0142. doi: 10.1145/3274366. URL http://doi.acm.org/10.1145/3274366.

Krivosheev, Evgeny; Casati, Fabio, and Bozzon, Alessandro. Active hybrid classification. *CoRR*, abs/2101.08854, 2021. URL https://arxiv.org/abs/2101.08854.

Kulkarni, Anand Pramod; Can, Matthew, and Hartmann, Björn. Collaboratively crowdsourcing workflows with turkomatic. In *CSCW '12 Computer Supported Cooperative Work, Seattle, WA, USA, February 11-15, 2012*, pages 1003–1012, 2012. doi: 10.1145/2145204.2145354. URL https://doi.org/10.1145/2145204.2145354.

Lan, Doren; Reed, Katherine; Shin, Austin, and Trushkowsky, Beth. Dynamic filter: Adaptive query processing with the crowd. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2017, 23-26 October 2017, Québec City, Québec, Canada.*, pages 118–127, 2017. URL https://aaai.org/ocs/index.php/HCOMP/HCOMP17/paper/view/15932.

Little, Greg; Chilton, Lydia B.; Goldman, Max, and Miller, Robert C. Turkit: human computation algorithms on mechanical turk. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology, New York, NY, USA, October 3-6, 2010*, pages 57–66, 2010. doi: 10.1145/1866029.1866040. URL https://doi.org/10.1145/1866029.1866040.

Liu, Angli; Soderland, Stephen; Bragg, Jonathan; Lin, Christopher H.; Ling, Xiao, and Weld, Daniel S. Effective crowd annotation for relation extraction. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 897–906, 2016. URL http://aclweb.org/anthology/N/N16/N16-1104.pdf.

Liu, Yang. Fine-tune BERT for extractive summarization. *arXiv preprint arXiv:1903.10318*, 2019. URL http://arxiv.org/abs/1903.10318.

Loshchilov, Ilya and Hutter, Frank. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Maddalena, Eddy; Basaldella, Marco; De Nart, Dario; Degl'Innocenti, Dante; Mizzaro, Stefano, and Demartini, Gianluca. Crowdsourcing relevance assessments: The unexpected benefits of limiting the time to judge. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*, 2016.

Mao, Andrew; Chen, Yiling; Gajos, Krzysztof Z.; Parkes, David C.; Procaccia, Ariel D., and Zhang, Haoqi. Turkserver: Enabling synchronous and longitudinal online experiments. In *The 4th Human Computation Workshop, HCOMP@AAAI 2012, Toronto, Ontario, Canada, July 23, 2012*, 2012. URL http://www.aaai.org/ocs/index.php/WS/AAAIW12/paper/view/5315.

Marcus, Adam; Wu, Eugene; Karger, David R.; Madden, Samuel, and Miller, Robert C. Human-powered sorts and joins. *PVLDB*, 5(1):13–24, 2011. doi: 10.14778/2047485.2047487. URL http://www.vldb.org/pvldb/vol5/p013_adammarcus_vldb2012.pdf.

Marshall, Catherine C. and III, Frank M. Shipman. Experiences surveying the crowd: reflections on methods, participation, and reliability. In *Web Science 2013 (co-located with ECRC), WebSci '13, Paris, France, May 2-4, 2013*, pages 234–243, 2013. doi: 10.1145/2464464.2464485. URL https://doi.org/10.1145/2464464.2464485.

Martin, David B.; Hanrahan, Benjamin V.; O'Neill, Jacki, and Gupta, Neha. Being a turker. In *CSCW 2014*, pages 224–235, 2014. doi: 10.1145/2531602.2531663. URL https://doi.org/10.1145/2531602.2531663.

Mason, Winter and Suri, Siddharth. Conducting behavioral research on amazon's mechanical turk. *Behavior Research Methods*, 44(1):1–23, Mar 2012. ISSN 1554-3528. doi: 10.3758/s13428-011-0124-6. URL `https://doi.org/10.3758/s13428-011-0124-6`.

Mason, Winter A. and Watts, Duncan J. Financial incentives and the "performance of crowds". *SIGKDD Explorations*, 11(2):100–108, 2009. doi: 10.1145/1809400.1809422. URL `https://doi.org/10.1145/1809400.1809422`.

McCambridge, Jim; de Bruin, Marijn, and Witton, John. The effects of demand characteristics on research participant behaviours in non-laboratory settings: a systematic review. *PloS one*, 7(6):e39116, 2012. doi: 10.1371/journal.pone.0039116.

McDonnell, Tyler; Lease, Matthew; Kutlu, Mucahid, and Elsayed, Tamer. Why is that relevant? collecting annotator rationales for relevance judgments. In *HCOMP 2016*, 2016. URL `http://aaai.org/ocs/index.php/HCOMP/HCOMP16/paper/view/14043`.

Minder, Patrick and Bernstein, Abraham. Crowdlang: A programming language for the systematic exploration of human computation systems. In *Social Informatics - 4th International Conference, SocInfo 2012, Lausanne, Switzerland, December 5-7, 2012. Proceedings*, pages 124–137, 2012. doi: 10.1007/978-3-642-35386-4_10. URL `https://doi.org/10.1007/978-3-642-35386-4_10`.

Mitchell, Margaret; Wu, Simone; Zaldivar, Andrew; Barnes, Parker; Vasserman, Lucy; Hutchinson, Ben; Spitzer, Elena; Raji, Inioluwa Deborah, and Gebru, Timnit. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 220–229, 2019. doi: 10.1145/3287560.3287596. URL `https://doi.org/10.1145/3287560.3287596`.

Mitra, Tanushree; Hutto, Clayton J., and Gilbert, Eric. Comparing person- and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, pages 1345–1354, 2015. doi: 10.1145/2702123.2702553. URL `https://doi.org/10.1145/2702123.2702553`.

Morishima, Atsuyuki; Shinagawa, Norihide; Mitsuishi, Tomomi; Aoki, Hideto, and Fukusumi, Shun. Cylog/crowd4u: A declarative platform for complex data-centric crowdsourcing. *Proc. VLDB Endow.*, 5(12):1918–1921, 2012. doi: 10.14778/2367502.2367537. URL `http://vldb.org/pvldb/vol5/p1918_atsuyukimorishima_vldb2012.pdf`.

Mortensen, Michael L.; Adam, Gaelen P.; Trikalinos, Thomas A.; Kraska, Tim, and Wallace, Byron C. An exploration of crowdsourcing citation screening for systematic reviews. *Research Synthesis Methods*, 2016. ISSN 1759-2887.

Narayan, Shashi; Cohen, Shay B., and Lapata, Mirella. Ranking sentences for extractive summarization with reinforcement learning. In *NAACL 2018*, 2018.

Nguyen, An Thanh; Wallace, Byron C., and Lease, Matthew. Combining crowd and expert labels using decision theoretic active learning. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2015, November 8-11, 2015, San Diego, California, USA.*, pages 120–129, 2015a. URL `http://www.aaai.org/ocs/index.php/HCOMP/HCOMP15/paper/view/11567`.

Nguyen, An Thanh; Wallace, Byron C., and Lease, Matthew. Combining crowd and expert labels using decision theoretic active learning. In *HCOMP 2015*, 2015b. URL `http://www.aaai.org/ocs/index.php/HCOMP/HCOMP15/paper/view/11567`.

Nguyen, Dong. Comparing automatic and human evaluation of local explanations for text classification. In *NAACL-HLT 2018*, 2018. URL `https://aclanthology.info/papers/N18-1097/n18-1097`.

Nguyen, Dong; Trieschnigg, Dolf; Dogruöz, A. Seza; Gravel, Rilana; Theune, Mariët; Meder, Theo, and de Jong, Franciska. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 1950–1961, 2014. URL `https://www.aclweb.org/anthology/C14-1184/`.

Olson, Judith S. and Kellogg, Wendy A. *Ways of Knowing in HCI*. Springer Publishing Company, Incorporated, 2014. ISBN 1493903772.

Pannucci, Christopher J and Wilkins, Edwin G. Identifying and avoiding bias in research. *Plastic and reconstructive surgery*, 126(2):619, 2010.

Paolacci, Gabriele; Chandler, Jesse, and Ipeirotis, Panagiotis G. Running experiments on amazon mechanical turk. 2010.

Parameswaran, Aditya G.; Garcia-Molina, Hector; Park, Hyunjung; Polyzotis, Neoklis; Ramesh, Aditya, and Widom, Jennifer. Crowdscreen: algorithms for filtering data with humans. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012*, pages 361–372, 2012a. doi: 10.1145/2213836.2213878. URL `https://doi.org/10.1145/2213836.2213878`.

Parameswaran, Aditya G.; Park, Hyunjung; Garcia-Molina, Hector; Polyzotis, Neoklis, and Widom, Jennifer. Deco: declarative crowdsourcing. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, pages 1203–1212, 2012b. doi: 10.1145/2396761.2398421. URL `https://doi.org/10.1145/2396761.2398421`.

Paritosh, Praveen. Human computation must be reproducible. In *Proceedings of the First International Workshop on Crowdsourcing Web Search, Lyon, France, April 17, 2012*, pages 20–25, 2012. URL `http://ceur-ws.org/Vol-842/crowdsearch-paritosh.pdf`.

Park, Hyunjung and Widom, Jennifer. Query optimization over crowdsourced data. *PVLDB*, 6(10):781–792, 2013. doi: 10.14778/2536206.2536207. URL `http://www.vldb.org/pvldb/vol6/p781-park.pdf`.

Pineau, Joelle; Vincent-Lamarre, Philippe; Sinha, Koustuv; Larivière, Vincent; Beygelzimer, Alina; d'Alché-Buc, Florence; Fox, Emily B., and Larochelle, Hugo. Improving reproducibility in machine learning research (A report from the neurips 2019 reproducibility program). *arXiv preprint arXiv:2003.12206*, 2020. URL `https://arxiv.org/abs/2003.12206`.

Plesser, Hans E. Reproducibility vs. replicability: a brief history of a confused terminology. *Frontiers in neuroinformatics*, 11:76, 2018.

Porter, Nathaniel D.; Verdery, Ashton M., and Gaddis, S. Michael. Enhancing big data in the social sciences with crowdsourcing: Data augmentation practices, techniques, and opportunities. *PLOS ONE*, 15(6):1–21, 06 2020. doi: 10.1371/journal.pone.0233154. URL `https://doi.org/10.1371/journal.pone.0233154`.

Qarout, Rehab Kamal; Checco, Alessandro; Demartini, Gianluca, and Bontcheva, Kalina. Platform-related factors in repeatability and reproducibility of crowdsourcing tasks. In *HCOMP 2019*, 2019.

Ramírez, Jorge; Krivosheev, Evgeny; Báez, Marcos; Casati, Fabio, and Benatallah, Boualem. Crowdrev: A platform for crowd-based screening of literature reviews. In *Collective Intelligence, CI 2018*, 2018.

Ramírez, Jorge; Baez, Marcos; Casati, Fabio, and Benatallah, Boualem. Understanding the impact of text highlighting in crowdsourcing tasks. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2019*, volume 7, pages 144–152. AAAI, October 2019a.

Ramírez, Jorge; Degiacomi, Simone; Zanella, Davide; Báez, Marcos; Casati, Fabio, and Benatallah, Boualem. Crowdhub: Extending crowdsourcing platforms for the controlled evaluation of tasks designs. *CoRR*, abs/1909.02800, 2019b. URL `http://arxiv.org/abs/1909.02800`.

Ramírez, Jorge; Baez, Marcos; Casati, Fabio, and Benatallah, Boualem. Crowdsourced dataset to study the generation and impact of text highlighting in classification tasks. *BMC Research Notes*, 12(1):820, 2019c. ISSN 1756-0500. doi: 10.1186/s13104-019-4858-z. URL `https://doi.org/10.1186/s13104-019-4858-z`.

Ramírez, Jorge; Báez, Marcos; Casati, Fabio; Cernuzzi, Luca, and Benatallah, Boualem. Challenges and strategies for running controlled crowdsourcing experiments. In *Proceedings of the XLVI Latin American Computing Conference (CLEI 2020)*, 2020a.

Ramírez, Jorge; Baez, Marcos; Casati, Fabio; Cernuzzi, Luca, and Benatallah, Boualem. Drec: towards a datasheet for reporting experiments in crowdsourcing. In *CSCW 2020*, 2020b.

Ramírez, Jorge; Baez, Marcos; Casati, Fabio; Cernuzzi, Luca; Benatallah, Boualem; Taran, Ekaterina A., and Malanina, Veronika A. On the impact of predicate complexity in crowdsourced classification tasks. In *WSDM 2021*, 2021a.

Ramírez, Jorge; Sayin, Burcu; Baez, Marcos; Casati, Fabio; Cernuzzi, Luca; Benatallah, Boualem, and Demartini, Gianluca. On the state of reporting in crowdsourcing experiments and a checklist to aid current practices. In *Proceedings of the ACM on Human-Computer Interaction (PACM HCI), presented at the 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2021). October 2021*, 2021b.

Rand, David G. The promise of mechanical turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, 299:172 – 179, 2012. ISSN 0022-5193. doi: https://doi.org/10.1016/j.jtbi.2011.03.004. URL `http://www.sciencedirect.com/science/article/pii/S0022519311001330`. Evolution of Cooperation.

Rao, R. Sowmya; Glickman, Mark E., and Glynn, Robert J. Stopping rules for surveys with multiple waves of nonrespondent follow-up. *Statistics in Medicine*, 27(12):2196–2213, 2008. doi: 10.1002/sim.3063. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3063`.

Reddy, Siva; Chen, Danqi, and Manning, Christopher D. Coqa: A conversational question answering challenge. *arXiv preprint arXiv:1808.07042*, 2018. URL `http://arxiv.org/abs/1808.07042`.

Rekatsinas, Theodoros; Deshpande, Amol, and Parameswaran, Aditya G. CRUX: adaptive querying for efficient crowdsourced data extraction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 841–850, 2019. doi: 10.1145/3357384.3357976. URL `https://doi.org/10.1145/3357384.3357976`.

Retelny, Daniela; Bernstein, Michael S., and Valentine, Melissa A. No workflow can ever be enough: How crowdsourcing workflows constrain complex work. *PACMHCI*, 1(CSCW):89:1–89:23, 2017. doi: 10.1145/3134724. URL `https://doi.org/10.1145/3134724`.

Rogstadius, Jakob; Kostakos, Vassilis; Kittur, Aniket; Smus, Boris; Laredo, Jim, and Vukovic, Maja. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *Proceedings of the Fifth*

International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011, 2011. URL `http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2778`.

Sabou, Marta; Bontcheva, Kalina; Derczynski, Leon, and Scharl, Arno. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 859–866, 2014. URL `http://www.lrec-conf.org/proceedings/lrec2014/summaries/497.html`.

Sampath, Harini Alagarai; Rajeshuni, Rajeev, and Indurkhya, Bipin. Cognitively inspired task design to improve user performance on crowdsourcing platforms. In *CHI Conference on Human Factors in Computing Systems, CHI'14, Toronto, ON, Canada - April 26 - May 01, 2014*, pages 3665–3674, 2014. doi: 10.1145/2556288.2557155. URL `https://doi.org/10.1145/2556288.2557155`.

Sanh, Victor; Debut, Lysandre; Chaumond, Julien, and Wolf, Thomas. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL `http://arxiv.org/abs/1910.01108`.

Saunders, Benjamin; Sim, Julius; Kingstone, Tom; Baker, Shula; Waterfield, Jackie; Bartlam, Bernadette; Burroughs, Heather, and Jinks, Clare. Saturation in qualitative research: Exploring its conceptualization and operationalization. *Quality & quantity*, 52(4):1893–1907, 2018.

Schaekermann, Mike; Goh, Joslin; Larson, Kate, and Law, Edith. Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. *CSCW 2018*, 2018.

Schnoebelen, Tyler and Kuperman, Victor. Using amazon mechanical turk for linguistic research. *Psihologija*, 43: 441–464, 2010.

Schulz, K.F.; Altman, D.G.; Moher, D., and others, . Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC medicine*, 8(1):18, 2010.

Schulze, Thimo; Seedorf, Stefan; Geiger, David; Kaufmann, Nicolas, and Schader, Martin. Exploring task properties in crowdsourcing - an empirical study on mechanical turk. In *19th European Conference on Information Systems, ECIS 2011, Helsinki, Finland, June 9-11, 2011*, page 122, 2011. URL `http://aisel.aisnet.org/ecis2011/122`.

Sedgwick, Philip and Greenwood, Nan. Understanding the hawthorne effect. *Bmj*, 351:h4672, 2015.

Sen, Shilad; Giesel, Margaret E.; Gold, Rebecca; Hillmann, Benjamin; Lesicko, Matt; Naden, Samuel; Russell, Jesse; Wang, Zixiao (Ken), and Hecht, Brent J. Turkers, scholars, "arafat" and "peace": Cultural communities and algorithmic gold standards. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14 - 18, 2015*, pages 826–838, 2015. doi: 10.1145/2675133.2675285. URL `https://doi.org/10.1145/2675133.2675285`.

Shadish, W.R.; Cook, T.D., and Campbell, D.T. *Experimental and quasi-experimental designs for generalized causal inference.* Houghton, Mifflin and Company, 2002.

Shamseer, Larissa; Moher, David; Clarke, Mike; Ghersi, Davina; Liberati, Alessandro; Petticrew, Mark; Shekelle, Paul, and Stewart, Lesley A. Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015: elaboration and explanation. *BMJ*, 349, 2015. doi: 10.1136/bmj.g7647. URL `https://www.bmj.com/content/349/bmj.g7647`.

Snow, Rion; O'Connor, Brendan; Jurafsky, Daniel, and Ng, Andrew Y. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu,* *Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 254–263, 2008. URL http://www.aclweb.org/anthology/D08-1027.

Sorokin, Alexander and Forsyth, David A. Utility data annotation with amazon mechanical turk. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2008, Anchorage, AK, USA, 23-28 June, 2008*, pages 1–8, 2008. doi: 10.1109/CVPRW.2008.4562953. URL https://doi.org/10.1109/CVPRW.2008.4562953.

Steichen, Ben and Freund, Luanne. Supporting the modern polyglot: A comparison of multilingual search interfaces. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3483–3492, 2015.

Strobelt, Hendrik; Oelke, Daniela; Kwon, Bum Chul; Schreck, Tobias, and Pfister, Hanspeter. Guidelines for effective usage of text highlighting techniques. *IEEE Trans. Vis. Comput. Graph.*, 22(1), 2016. doi: 10.1109/TVCG.2015.2467759. URL https://doi.org/10.1109/TVCG.2015.2467759.

Sun, Peng and Stolee, Kathryn T. Exploring crowd consistency in a mechanical turk survey. In *Proceedings of the 3rd International Workshop on CrowdSourcing in Software Engineering, CSI-SE@ICSE 2016, Austin, Texas, USA, May 16, 2016*, pages 8–14, 2016. doi: 10.1145/2897659.2897662. URL https://doi.org/10.1145/2897659.2897662.

Sun, Yalin; Cheng, Pengxiang; Wang, Shengwei; Lyu, Hao; Lease, Matthew; Marshall, Iain James, and Wallace, Byron C. Crowdsourcing information extraction for biomedical systematic reviews. *CoRR*, abs/1609.01017, 2016. URL http://arxiv.org/abs/1609.01017.

Vaughan, Jennifer Wortman. Making better use of the crowd: How crowdsourcing can advance machine learning research. *Journal of Machine Learning Research*, 18, 2017. URL http://jmlr.org/papers/v18/17-234.html.

Wacharamanotham, Chat; Eisenring, Lukas; Haroz, Steve, and Echtler, Florian. Transparency of CHI research artifacts: Results of a self-reported survey. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–14, 2020. doi: 10.1145/3313831.3376448. URL https://doi.org/10.1145/3313831.3376448.

Wallace, Byron C.; Noel-Storr, Anna; Marshall, Iain James; Cohen, Aaron M.; Smalheiser, Neil R., and Thomas, James. Identifying reports of randomized controlled trials (rcts) via a hybrid machine learning and crowdsourcing approach. *JAMIA*, 24(6), 2017. doi: 10.1093/jamia/ocx053. URL https://doi.org/10.1093/jamia/ocx053.

Weiss, Michael. Crowdsourcing literature reviews in new domains. 2016.

Weng, Xueping; Li, Guoliang; Hu, Huiqi, and Feng, Jianhua. Crowdsourced selection on multi-attribute data. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 307–316, 2017. doi: 10.1145/3132847.3132891. URL https://doi.org/10.1145/3132847.3132891.

Whitehill, Jacob; Wu, Ting-fan; Bergsma, Jacob; Movellan, Javier R, and Ruvolo, Paul L. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. pages 2035–2043, 2009.

Whiting, Mark E.; Hugh, Grant, and Bernstein, Michael S. Fair work: Crowd work minimum wage with one line of code. In *HCOMP 2019*, 2019.

Wilson, Shomir; Schaub, Florian; Ramanath, Rohan; Sadeh, Norman; Liu, Fei; Smith, Noah A., and Liu, Frederick. Crowdsourcing annotations for websites' privacy policies: Can it really work? In *WWW 2018*, 2016. ISBN 978-1-4503-4143-1. doi: 10.1145/2872427.2883035. URL https://doi.org/10.1145/2872427.2883035.

Wu, Jen-Her and Yuan, Yufei. Improving searching and reading performance: the effect of highlighting and text color coding. *Information & Management*, 40(7), 2003. doi: 10.1016/S0378-7206(02)00091-5. URL https://doi.org/10.1016/S0378-7206(02)00091-5.

Wu, Meng-Han and Quinn, Alexander J. Confusing the crowd: Task instruction quality on amazon mechanical turk. In *HCOMP 2017*, 2017. URL https://aaai.org/ocs/index.php/HCOMP/HCOMP17/paper/view/15943.

Wu, Y Wayne and Bailey, Brian P. Novices who focused or experts who didn't? In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4086–4097, 2016.

Wulczyn, Ellery; Thain, Nithum, and Dixon, Lucas. Wikipedia Talk Labels: Personal Attacks. 2 2017. doi: 10.6084/m9.figshare.4054689.v6. URL https://figshare.com/articles/Wikipedia_Talk_Labels_Personal_Attacks/4054689.

Xu, Luyan; Zhou, Xuan, and Gadiraju, Ujwal. Revealing the role of user moods in struggling search tasks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1249–1252, 2019.

Yang, Jie; Redi, Judith; Demartini, Gianluca, and Bozzon, Alessandro. Modeling task complexity in crowdsourcing. In *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2016, 30 October - 3 November, 2016, Austin, Texas, USA.*, pages 249–258, 2016. URL http://aaai.org/ocs/index.php/HCOMP/HCOMP16/paper/view/14039.

Appendix A

Supporting material for Chapter 5

A.1 Checklist for Crowdsourcing Experiments

In this section, we introduce the checklist, depicted in Table A.1. The checklist should be filled out per experiment, in case the paper reports on multiple studies involving the crowd as subjects. Besides, suppose an experiment uses different (potentially interconnected) micro-tasks. In that case, the Task and Quality control sections should be reported per task (or at least the Task section in case the quality control mechanisms are the same for all tasks).

Table A.1: Checklist for reporting crowdsourcing experiments

Item	Item N.	Recommendation	Page N.
Experimental design			
Input dataset	1	Describe how the input dataset for the experiment was obtained and if it is publicly available. Also, touch on its reputation and difficulty (if applicable)	
Allocation to experimental conditions	2	State how the participants were assigned to the experimental conditions or treatments, and how this step was implemented in the crowdsourcing platform	

Experimental design to task mapping	3	Describe what research design was used in the experiment and how were the experimental conditions mapped to crowdsourcing tasks	
Execution of experimental conditions	4	Report how the crowdsourcing tasks, representing the experimental conditions, were executed (e.g., in parallel, sequentially, or mixed)	
Execution timeframe	5	State over what timeframe the experiment was executed	
Pilots	6	Describe if pilot studies were performed before the main experiment	
Returning workers	7	Report the strategies used to prevent returning workers, i.e., workers who finish the experiment and then reenter it later because the study was still running	
Crowd			
Target population	8	Describe the criteria used to determine the workers who are allowed to participate (e.g., acceptance rate, tasks completed, demographics, working environment). And also indicate the strategy used to identify such workers	
Sampling mechanism	9	Report what strategies were used to recruit a diverse or representative set of workers from the target population	
Task			
Task interface	10	Report and show the task interface as seen by workers	
Task interface source	11	Provide a link to an online repository with the source code of the task interface (typically a combination of HTML, CSS, and JavaScript)	
Instructions	12	Describe and show the instructions of the task as seen by workers	

Reward strategy	13	State the mechanisms used to reward and motivate workers (e.g., payments)	
Time allotted	14	Report if a time constraint was defined for workers to complete the task (if so, describe also how much)	
Quality control			
Rejection criteria	15	State the criteria used to accept or reject a contribution from a worker (e.g., workers can be allowed to submit the task and reject it afterward, submissions can be blocked based on prior rejections or on time spent on the task)	
Number of votes per item	16	Describe, if applicable, how many workers solved the same item or data unit	
Aggregation method	17	Report, if applicable, how the contributions from workers were aggregated (e.g., majority voting)	
Training	18	State if workers performed a training session or pre-task qualification test. If so, describe 1) the training, 2) the items used as the training set, and 3) if it was performed before or as part of the task	
In-task checks	19	Report the mechanisms embedded in the task to guard the quality of the results. Also, state if and how workers were allowed to revise their answers. In case gold items or attention checks were used, describe how these items were selected, how frequently they appear, and the threshold used to filter out workers underperforming on these items.	
Post-task checks	20	Report the steps performed upon task completion to safeguard the quality of the results (e.g., post hoc analysis)	

Dropouts prevention mechanisms	21	Indicate the strategies used to deal with worker dropouts (i.e., workers who leave the task unfinished)	
Outcome			
Number of participants	22	Indicate how many workers participated in the experiment (in total and per condition)	
Number of contributions	23	Report the number of contributions (e.g., votes) in total and per condition	
Excluded participants	24	Indicate the number of participants Nt considered for the data analysis, including the reason for exclusion.	
Discarded data	25	State the number of contributions excluded before the data analysis	
Dropout rate	26	Describe the dropout rate of the participants in the experimental conditions. If applicable, also show breakdowns per milestone of progress within the task (e.g., after 2, 3, and 5 questions).	
Participant demographics	27	Report the demographics of the participants (e.g., age, country, language)	
Data processing	28	Report any data transformation, augmentation, and/or filtering step performed on the raw dataset obtained from the crowdsourcing platform.	
Output dataset	29	Provide a link to the dataset resulting from the experiment. Also indicate if the dataset contains the aggregated or individual contributions from workers	
Requester			
Platform(s) used	30	Indicate the crowdsourcing platform(s) selected for the experiment	

Implemented features	31	Report any additional feature implemented to support the experiment, covering missing functionality from the selected platform(s)	
Fair compensation	32	State whether workers were compensated fairly and according to legal minimum wage	
Requester-Worker interactions	33	Describe concrete requester-worker interactions taking place as part of the experiment	
Privacy & Data Treatment	34	Report any relevant privacy regulations and methods used to comply, especially if the output is put online (e.g., the data could be aNnymized to meet privacy policies).	
Informed consent	35	Indicate if an informed consent was used	
Participation awareness	36	State if workers were informed they took part in an experiment	
Ethical approvals	37	Report if the study received ethical approval from the corresponding institutional authority	

A.2 Identifying Papers Reporting Crowdsourcing Experiments

This section introduces the query that was used to retrieve papers (potentially) reporting crowdsourcing experiments and the exact eligibility criteria used to filter out retrieved articles.

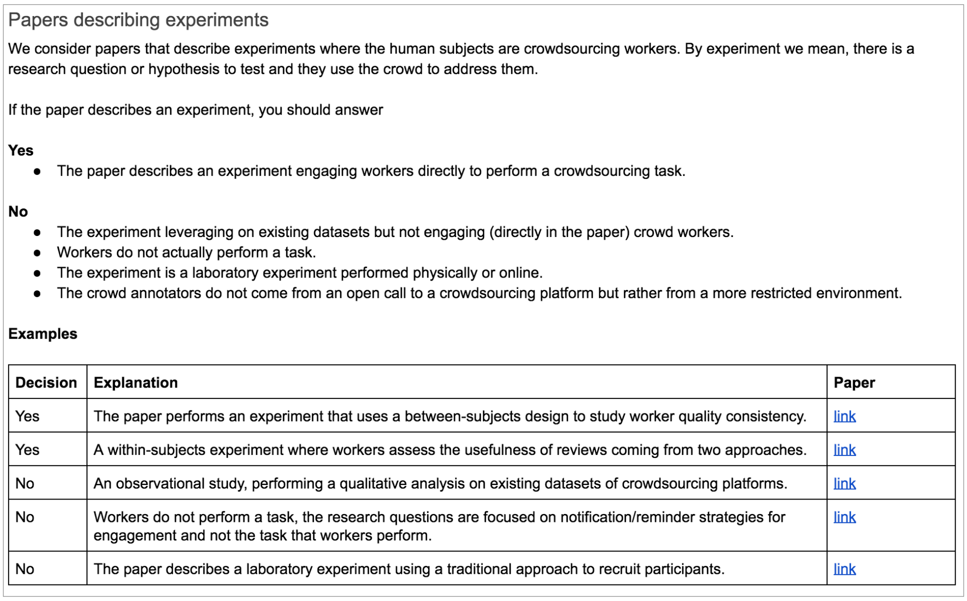
A.2.1 Scopus query

```
TITLE-ABS-KEY(crowdsource OR crowd-source OR crowd-sourcing OR
              crowdsourcing OR "human computation" OR
              crowdsourc* OR crowd-sourc* OR m*cro-task OR m*crotask)
AND
TITLE-ABS-KEY(experiment OR "experimental design" OR stud* OR evaluation OR
              intervention OR analysis)
AND
TITLE-ABS-KEY(user* OR behavio* OR worker*)
AND
```

```
(CONF(CHI) OR CONF(HCI) OR CONF(CSCW) OR CONF(WWW) OR CONF(HCOMP) OR
CONF(WSDM) OR CONF(CIKM) OR CONF(SIGIR) OR CONF(ICWE) OR CONF(IUI) OR
CONF(UIST) OR CONF(ICWSM) OR CONF("Human Factors"))
AND
(LIMIT-TO (DOCTYPE ,  "cp"))
```

A.2.2 Screening instructions

Figure A.1 depicts the instructions used by the researchers to identify papers reporting crowdsourcing experiments.



Papers describing experiments

We consider papers that describe experiments where the human subjects are crowdsourcing workers. By experiment we mean, there is a research question or hypothesis to test and they use the crowd to address them.

If the paper describes an experiment, you should answer

Yes

- The paper describes an experiment engaging workers directly to perform a crowdsourcing task.

No

- The experiment leveraging on existing datasets but not engaging (directly in the paper) crowd workers.
- Workers do not actually perform a task.
- The experiment is a laboratory experiment performed physically or online.
- The crowd annotators do not come from an open call to a crowdsourcing platform but rather from a more restricted environment.

Examples

Decision	Explanation	Paper
Yes	The paper performs an experiment that uses a between-subjects design to study worker quality consistency.	link
Yes	A within-subjects experiment where workers assess the usefulness of reviews coming from two approaches.	link
No	An observational study, performing a qualitative analysis on existing datasets of crowdsourcing platforms.	link
No	Workers do not perform a task, the research questions are focused on notification/reminder strategies for engagement and not the task that workers perform.	link
No	The paper describes a laboratory experiment using a traditional approach to recruit participants.	link

Figure A.1: The screening instructions for identifying papers reporting crowdsourcing experiments.

A.3 Interview Protocol

The interview protocol can be found at `https://tinyurl.com/ReportingInterviewProtocol`.

A.4 Applicability of Attributes

As we mentioned in Section 5.4.1, we considered 13 of the 39 attributes as potentially not applicable (N/A) based on the experiment's goal. Here, we detail these attributes.

The *input dataset* was N/A if the paper does not necessarily use an input dataset for the crowdsourcing tasks. For example, in creative tasks, workers are just given instructions, and they provide input. The *returning workers* attribute was N/A if the paper studied mechanisms to deal with workers that return to the experiment, or the study needed returning workers as part of their setup. The attributes in the quality control dimension were considered N/A if the paper actually studied quality control in crowdsourcing, including also strategies to deal with workers dropouts. The *aggregation method* was N/A if the aggregation of contributions was not suitable for the experiment, and likewise, the *gold items configuration* was N/A if the experiment did not use gold items. The *number of contributions* and *discarded data* were N/A if they were just equal to the number of participants and excluded participants, respectively. An example of this is a study on worker behavior, which could collect a single contribution from each participant. Finally, the *data processing* was N/A if contributions were used as-is.