



The  
University  
Of  
Sheffield.

# Novel Methods for Designing Tasks in Crowdsourcing

By:

Rehab K. QAROUT

*A thesis submitted in partial fulfilment of the requirements  
for the degree of Doctor of Philosophy  
in the*

Faculty of Engineering  
Department of Computer Science

July 2019



## Declaration of Authorship

I, Rehab K. QAROUT, declare that this thesis titled, “Novel Methods for Designing Tasks in Crowdsourcing” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

---

Date:

---



THE UNIVERSITY OF SHEFFIELD

## *Abstract*

Faculty of Engineering

Department of Computer Science

Doctor of Philosophy

### **Novel Methods for Designing Tasks in Crowdsourcing**

by Rehab K. QAROUT

Crowdsourcing is becoming more popular as a means for scalable data processing that requires human intelligence. The involvement of groups of people to accomplish tasks could be an effective success factor for data-driven businesses. Unlike in other technical systems, the quality of the results depends on human factors and how well crowd workers understand the requirements of the task, in order to produce high quality results. Looking at previous studies in this area, we found that one of the main factors that affect workers' performance is the design of the crowdsourcing tasks. Previous studies of crowdsourcing task design covered a limited set of factors. The main contribution of this research is the focus on some of the less-studied technical factors, such as examining the effect of task ordering and class balance and measuring the consistency of the same task design over time and on different crowdsourcing platforms. Furthermore, this study ambitiously extends work towards understanding workers' point of view in terms of the quality of the task and the payment aspect by performing a qualitative study with crowd workers and shedding light on some of the ethical issues around payments for crowdsourcing tasks. To achieve our goal, we performed several crowdsourcing experiments on specific platforms and measured the factors that influenced the quality of the overall result.



## Acknowledgements

I am grateful to God for the excellent health and willpower that were necessary to complete this thesis.

A special thanks to my supervisors Kalina Bontcheva and Alessandro Checco for their support and guidance throughout my PhD. I am grateful to Gianluca Demartini - my previous supervisor- for introducing me to the field of crowdsourcing and inspire me to work on this research.

My family, words cannot express how grateful I am, to my brothers and sisters for their support that push foreword.

I would like to thank all my friends who supported me in my journey and motivate me to reach my goal. Amal and Areej your unlimited support was the only reasons that make me keep working and passed all the difficulties through these 3 years. Banan, Heba, and Hawazen you are the first friends I had in this journey and the best memories of the true friendship. To all other friends that became like sisters and family, you are all precious.

Last but not least, I would like to express my appreciation to my beloved husband, Aiman. Thank you for all of the sacrifices that you have made to be at my side, I would not finish my PhD without you.

Finally, to the sunshine in my life, my little daughters Zinah and Ruba, thank you for your never-ending smiles. Thank you Zinah, for being a wonderful big sister and the most responsible one. Ruba, my little star, thank you for always being a happy child and your kindness. I am really sorry for all the time that I spent away from you both.





# Contents

<b>Declaration of Authorship</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>vii</b>
<b>Contents</b>	<b>ix</b>
<b>List of Figures</b>	<b>xiii</b>
<b>List of Tables</b>	<b>xv</b>
<b>List of Abbreviations</b>	<b>xvii</b>
<b>1 Introduction</b>	<b>3</b>
1.1 Problem definition . . . . .	4
1.2 Aim and objectives . . . . .	4
1.3 Research questions . . . . .	5
1.4 Research contributions . . . . .	6
1.5 Research design . . . . .	7
1.6 Selection criteria for literature review . . . . .	10
1.7 Structure of the thesis . . . . .	10
1.8 Previously published material . . . . .	11
<b>2 Literature review</b>	<b>13</b>
2.1 Introduction . . . . .	13
2.2 Human Computation and Crowdsourcing . . . . .	14
2.3 The rise of Crowdsourcing: definitions and the origins of the concept . . . . .	16
2.3.1 Definitions . . . . .	16
2.3.2 The Origin of Crowdsourcing . . . . .	17
2.3.3 The value of using crowdsourcing . . . . .	18
2.4 Crowdsourcing factors . . . . .	19
2.4.1 Human Factor . . . . .	19
2.4.2 Crowdsourcing process . . . . .	22
2.4.3 Types of crowdsourcing tasks . . . . .	25
2.4.4 Other factors influencing the design of a crowdsourcing task . . . . .	27
- Graphical User Interface (GUI) of the task . . . . .	28

	- Training questions . . . . .	30
	- Length of the task . . . . .	31
	- Ordering of the data in the task . . . . .	32
2.5	Classification of crowdsourcing systems . . . . .	33
2.5.1	Based on the nature of the job . . . . .	33
2.5.2	Based on the motivation of the crowd . . . . .	35
2.5.3	Based on the identity of the crowd . . . . .	36
2.5.4	Based on the nature of the platforms . . . . .	36
2.6	Crowdsourcing platforms . . . . .	37
2.6.1	Overview of current crowdsourcing platforms . . . . .	37
2.6.2	The evaluation of crowdsourcing platforms . . . . .	42
2.6.3	The consistency and reliability of platform results . . . . .	43
2.6.4	Crowdsourcing approaches . . . . .	43
2.7	The Assignment and aggregation mechanisms of crowdsourcing tasks . . . . .	45
2.8	Chapter Summary . . . . .	47
<b>3</b>	<b>Experiments general setup</b>	<b>51</b>
3.1	Introduction . . . . .	51
3.2	Datasets used in the experiments . . . . .	52
3.2.1	Dataset 1 . . . . .	52
3.2.2	Dataset 2 . . . . .	53
3.2.3	Dataset 3 . . . . .	54
3.3	Platforms used in the study . . . . .	55
3.4	The selection of crowdsourcing task type . . . . .	55
3.5	Evaluation metrics . . . . .	56
3.6	Chapter Summary . . . . .	59
<b>4</b>	<b>Batch ordering and balance for relevance judgement task</b>	<b>61</b>
4.1	Introduction . . . . .	61
4.2	Research Hypotheses . . . . .	63
4.2.1	Class Balance in a HIT Batch . . . . .	63
4.2.2	Ordering HITs in a HIT Batch . . . . .	64
4.3	Design of the experiments . . . . .	64
4.3.1	Dataset . . . . .	64
4.3.2	Participants . . . . .	64
4.4	Experiment 1: Class Imbalance . . . . .	65
4.4.1	Results and Discussion . . . . .	65
4.5	Experiment 2: Class Imbalance and Order . . . . .	66
4.5.1	Results and Discussion . . . . .	67
4.6	Experiment 3: Batch Size . . . . .	69
4.6.1	Results and Discussion . . . . .	70
4.7	Analysis of The Worker Experience . . . . .	70
4.7.1	Perceived Workload . . . . .	70

4.7.2	The Effect of Document Position on judgement Quality and Time . . . . .	71
4.7.3	Completion Time . . . . .	72
4.7.4	Effect on Agreement . . . . .	73
4.8	Discussion . . . . .	74
4.9	Chapter summary . . . . .	74
<b>5</b>	<b>Repeatability and Reproducibility of Crowdsourcing Classification Tasks</b>	<b>77</b>
5.1	Introduction . . . . .	77
5.2	Research Questions and Novelty . . . . .	79
5.3	Methodology . . . . .	80
5.3.1	Dataset . . . . .	80
5.3.2	Task Design . . . . .	80
5.3.3	Pilot Experiment and Sample Size . . . . .	81
5.4	Experiment 1 - Achieving Repeatability . . . . .	82
5.4.1	Experiment 1 - Discussion . . . . .	84
5.5	Experiment 2 - Achieving Reproducibility . . . . .	85
5.5.1	Experiment 2 - Discussion . . . . .	87
5.6	Chapter summary . . . . .	88
<b>6</b>	<b>Payment concerns in crowdsourcing platforms:</b>	
	<b>A case study on Figure Eight channels</b>	<b>91</b>
6.1	Introduction . . . . .	91
6.2	Ethical issues and right regulation around payment . . . . .	92
6.3	Motivation and payment in crowdsourcing tasks . . . . .	93
6.4	Analysis of workers on Figure Eight . . . . .	95
1-	ClixSense channel . . . . .	95
2-	Elite Figure Eight channel . . . . .	96
3-	NeoBux channel . . . . .	96
4-	InstaGC channel . . . . .	96
5-	Swagbucks via Prodege channel . . . . .	97
6.4.1	Data collection and analysis of channels . . . . .	97
6.4.2	Results of survey workers . . . . .	98
*	General information . . . . .	99
*	Crowd-workers experience . . . . .	100
*	Payment per channel . . . . .	102
*	Workers feedback . . . . .	105
6.5	Chapter summary . . . . .	106
<b>7</b>	<b>Conclusion</b>	<b>109</b>
7.1	Summary of the thesis . . . . .	109
7.2	Research challenges and limitations . . . . .	111
7.3	Guidelines for the best design of crowdsourcing task . . . . .	112
7.4	Future work . . . . .	113

<b>A Ethical Consent Forms</b>	<b>117</b>
<b>B NASA Task Load Index questionnaire (NASA-TLX)</b>	<b>121</b>
<b>C GUI of the Relevance Judgment Task</b>	<b>125</b>
<b>D GUI of the classification task (Dataset 1)</b>	<b>129</b>
<b>E GUI of the classification task (Dataset 2)</b>	<b>133</b>
<b>F GUI of the classification task (Dataset 3)</b>	<b>137</b>
<b>G GUI of the Survey workers task</b>	<b>141</b>
<b>Bibliography</b>	<b>147</b>

## List of Figures

1.1	Mind map of the research . . . . .	9
2.1	Adding Semantic web with Human computation as a means of solving computational problems, adapted from (Quinn and Bederson, 2011). . . . .	15
2.2	The key ingredients of crowdsourcing systems as described in (Brabham, 2013). . . . .	18
2.3	The mechanism of crowdsourcing. . . . .	22
2.4	The crowdsourcing process from the worker perspective. . . . .	23
2.5	The crowdsourcing process from the requester perspective. . . . .	23
2.6	The crowdsourcing task process. . . . .	24
2.7	A taxonomy of crowdsourcing tasks, adapted from (Gadiraju, Kawase, and Dietze, 2014). . . . .	26
2.8	Requesters and workers interface in MTurk platform. . . . .	38
2.9	The list of tasks available for the workers in CloudCrowd platform. . . . .	39
2.10	Requesters and workers interface in Figure Eight platform. . . . .	40
2.11	The mechanism of creating the task in Microworkers platform. . . . .	41
2.12	The mechanism of creating the task in Prolific Academia platform. . . . .	42
3.1	Dataset 1 examples include crisis name, type, country, size of collection, tweet text, and category . . . . .	53
3.2	Dataset 2 examples include title, review contents, and images of the fashion item . . . . .	54
3.3	Example of Dataset 3 Topic 421 including the topic title, description, narrative and documents . . . . .	55
4.1	Judgement Accuracy, Positive Predictive Value (PPV), and Negative Predictive Value (NPV) box-plot for E1. The horizontal line in the middle of each box represents the median value. The x-axis label indicates the ratio of relevant documents in the batch. . . . .	66
4.2	Order and balance of document classes for experiments 2 and 3 (blue for relevant and red for non-relevant). . . . .	66
4.3	Judgement Accuracy, Positive Predictive Value (PPV), and Negative Predictive Value (NPV) box-plot for E2. The horizontal line in the middle of each box represents the median value. . . . .	67

4.4	Judgement Accuracy, Positive Predictive Value (PPV), and Negative Predictive Value (NPV) box-plot for Topic 442 (light blue, top plots), Topic 421 (gray, middle plots), Topic 428 (dark green, bottom plots). The horizontal line in the middle of each box represents the median value. . . . .	69
4.5	Comparing Length of the task, Mean judgement PPV for each setting in Experiment 2 and 3. . . . .	70
4.6	Perceived workload using the NASA-TLX assessment tool for each setting in Experiment 1 and Experiment 2. . . . .	71
4.7	Mean Accuracy, PPV, and NPV of the documents in the first, second and last position in a batch over all the experiments. . . . .	72
4.8	Median PPV vs. completion time for each batch in Experiment 1 (a) and Experiment 2 (b). . . . .	73
4.9	(a) Krippendorff's alpha for all batches in Experiment 1 (horizontal line for median value). (b) Krippendorff's alpha for batches in Experiment 1 (blue) with different balance classes, and Experiment 2 (red) with different ordering of documents. . . . .	73
5.1	Average time per assignment for Experiment 1 and Experiment 2 for all 3 datasets.	82
5.2	Accuracy distribution over time for Experiment 1 and Experiment 2 for all 3 dataset. Dataset 1 shows a statistically significant difference in accuracy between the two platforms when using the default payment scheme (Experiment1) . . . .	83
5.3	Average completion time for all batches in Experiment 1 and 2. . . . .	86
6.1	Number of workers per channel in the top 10 countries . . . . .	97
6.2	Number of workers in the top 5 channels . . . . .	98
6.3	Workers' gender in each channel . . . . .	99
6.4	The distribution of workers countries per channel . . . . .	99
6.5	Range of workers' experience in each channel . . . . .	100
6.6	The distribution of completion time for workers in each channel . . . . .	100
6.7	The distribution of earnings from the crowdsourcing jobs in dollars per months .	101
6.8	The criteria of choosing the task . . . . .	101
6.9	Other platforms used by workers . . . . .	102
6.10	The distribution of actual payment received by workers in each channel (The red lines show the median) . . . . .	102
6.11	The relationship between completion time and the amount of payment workers received . . . . .	103
6.12	F8 requester's interface showing wrong amount under 'IN CHANNEL CURRENCY' (0.0035 for 0.35) . . . . .	103
6.13	The variation of the amount of payment in the top 4 channels used in F8 in the last four years . . . . .	104

## List of Tables

2.1	Representative literature references for the studies in the design of crowdsourcing tasks. . . . .	29
2.2	Taxonomy of crowdsourcing systems. . . . .	34
2.3	Crowdsourcing platforms. . . . .	37
2.4	Crowdsourcing approaches. . . . .	44
2.5	Comparison of 32 aggregation methods for crowdsourcing task results. . . . .	46
3.1	TREC8 topic with coherent judgments with Sormunen . . . . .	55
3.2	Confusion matrix for positive and negative labels . . . . .	57
4.1	Kendall's $\tau$ correlation for pairs of order and balance settings over different topics (with FDR-corrected p-values). . . . .	69
5.1	Results of five runs in MTurk and F8 for Experiment 1. . . . .	83
5.2	Two-way ANCOVA for Dataset 1 in Experiment 1. . . . .	84
5.3	Two-way ANCOVA for Dataset 2 in Experiment 1. . . . .	84
5.4	Two-way ANCOVA for Dataset 3 in Experiment 1. . . . .	84
5.5	Results of five runs in MTurk and F8 for Experiment 2. . . . .	86
5.6	Two-way ANCOVA for Dataset 1 in Experiment 2. . . . .	87
5.7	Two-way ANCOVA for Dataset 2 Experiment 2. . . . .	87
5.8	Two-way ANCOVA for Dataset 3 Experiment 2. . . . .	87
6.1	One-way ANOVA for time to complete the survey task per channel. . . . .	101
6.2	The estimate of the money lost for each channel at the workers level and F8 population. . . . .	105
G.1	The results of the Survey workers task for top 5 channels . . . . .	144





## List of Abbreviations

<b>AI</b>	<b>Artificial Intelligent</b>
<b>F8</b>	<b>Figure Eight 8</b>
<b>GUI</b>	<b>Graphical User Interface</b>
<b>HC</b>	<b>Human Computation</b>
<b>HCI</b>	<b>Human Computer Interaction</b>
<b>HITs</b>	<b>Human Intelligence Tasks</b>
<b>MTurk</b>	<b>Amazon Mechanical Turk</b>
<b>NASA-TLX</b>	<b>NASA Task Load Index</b>
<b>NLP</b>	<b>Natural Language Processing</b>
<b>NPV</b>	<b>Negative Predictive Value</b>
<b>PPV</b>	<b>Positive Predictive Value</b>
<b>PTC</b>	<b>Paid to Click</b>
<b>SB</b>	<b>SwagBucks</b>
<b>SW</b>	<b>Semantic Web</b>
<b>TREC8</b>	<b>The Eighth 8 Text Retrieval Conference</b>



To my Father

**Kamal**

&

my Mother

**Salma**

May their soul rest in peace. . .







# 1

## Introduction

The question is: ‘Why do we have to use human minds when we have smart machines?’ Alternatively, ‘When do we need humans to engage in the loop with the machine process?’ These questions and many others led researchers to identify the need to develop a new domain in Artificial Intelligence (AI) and combine individual workers’ skills to lay the foundations for the development of the machine. Several studies over multiple fields showed that there are some approaches where human minds need to collaborate with machines to solve problems (Tsvetkova et al., 2015; Kim and Monroy-Hernandez, 2016).

While Human Computation (HC) methods could theoretically involve only small numbers of contributors, Crowdsourcing approaches leverage the “wisdom of the crowd” by engaging a high number of online contributors to accomplish tasks that cannot yet be automated, often replacing a traditional workforce such as employees or domain experts (Howe, 2006). As such, crowdsourcing methods not only support the creation of research-relevant data, but more importantly they can also help to solve the bottleneck of knowledge experts and annotators needed for the large-scale deployment of Semantic Web and Linked Data technologies.

Organisations and companies rely on a large number of datasets that require manual analysis which will be subsequently utilised. For example, an annotated Twitter dataset will be used to train and evaluate Natural Language Processing (NLP) algorithms and Information retrieval (IR) systems that will process social media information (Imran, Mitra, and Castillo, 2016; Sabou et al., 2014). Building these datasets require to hire the right crowd workers to perform the work and having the datasets ready in a short time.

Several crowdsourcing systems exist. One of the earliest crowdsourcing systems is Wikipedia,

a website where people (who represent workers, or the crowd) participate and add information about different topics and concepts. Another is OpenStreetMap<sup>1</sup> onto which people can add pictures or information about a route or shop locations. Both of these are examples of a volunteer crowd (discussed in Section 2.5).

In a business situation, designing a logo for a company is an example of a crowdsourced job (using platforms like 99designs). A company (the requester) provides a description of a design, asks people outside the firm to submit potential designs, and pays a reward for the selected logo. Online crowdsourcing systems appear to solve a variety of similar problems, giving people outside the organisation the opportunity to participate in activities and to support the business side of operations for the company.

Certain platform services, such as Amazon Mechanical Turk (MTurk)<sup>2</sup> and Figure Eight (F8)<sup>3</sup> (previously known as Crowdflower) depend on humans (online users/workers) and their contributions for the completion of Human Intelligence Tasks (HITs). Online users are influenced by different factors (e.g., incentives, motivation, training, boredom), and it is still unclear to what extent variations in these factors might affect workers' performance on a crowdsourcing platform.

## 1.1 Problem definition

Since the tasks are the only connection between the people commissioning to the work (requesters), and the online users who enter the platforms looking for those crowd jobs (workers), these tasks need to have clear instructions and to present the job requirements in the most effective way. This is even more important in situations in which tasks are paid as workers aim at minimising the time spent on understanding what they are required to do. Thus, designing appropriate crowdsourcing tasks could have a significant impact on the outcome quality.

Several studies in the past have demonstrated that varying some factors in the design of the task (such as presenting examples or giving feedback) affects workers' performance (Finnerty et al., 2013; Alonso, 2013; Oleson et al., 2011; Mitra, Hutto, and Gilbert, 2015). The next two chapters focus on presenting the state of the art including all work on enhancing the design of crowdsourcing tasks in order to improve the overall outcome quality and leading to this research question and contributions in this field.

## 1.2 Aim and objectives

Recent studies have focused on analysing microtask crowdsourcing platforms and evaluating the contribution of the platform actors: workers and requesters (Lasecki et al., 2015; Cheng et al., 2015; Cai, Iqbal, and Teevan, 2016; Owens, 2013). This thesis aims to deliver an in-depth understanding of the development of tasks within crowdsourcing platforms over the past years and to identify the key features of designing tasks by focusing on workers' performance.

---

<sup>1</sup><https://www.openstreetmap.org/>

<sup>2</sup><https://www.mturk.com/>

<sup>3</sup><https://www.figure-eight.com/>



The research focuses on several core factors to measure the workers' performance:

1. The level of accuracy of the results that is achieved in each complete task.
2. The process time: that is, the amount of time the workers spend on a specific task.
3. The time to finish all the tasks that the requester asked to be crowdsourced on a given platform.

Workers' processes when performing a particular task are observed and the way in which the quality of the performance could be affected by changing some aspects in the design of that task are explored.

The starting point is to improve the design of the tasks to be run on a crowdsourcing platform and test the consistency of such design of the task over multiple platforms to optimise the overall performance of human computation systems. The subsequent point is to present the findings of this research as guidelines to help requesters develop the most efficient design of the task in order to ensure highly accurate work and thus save time for the business.

### 1.3 Research questions

The goal of this research is to address the following main question:

*How does the design of a crowdsourcing task affect workers' performance?*

To find an answer to this question a number of sub-questions will be addressed:

- **RQ1:** Do class imbalance and order in a batch of Human Intelligence Tasks (HITs) affect the performance of crowd workers involved in the creation of manually labeled datasets?
- **RQ2:** How should crowdsourcing tasks be split into microtasks between workers? What is an appropriate length of a task?
- **RQ3:** Does providing a training test question or an example improve workers' answers?
- **RQ4:** Is there a significant difference in the quality of the results for the same task repeated on the same crowdsourcing platform at a different point in time?
- **RQ5:** Is there a significant difference in the quality of the results for the same task reproduced on a different platform?
- **RQ6:** Are the results obtained consistent over different classification tasks?
- **RQ7:** Is the crowdsourcing platform transparently communicating the fee payment with the workers? Does the amount of payment affect workers' performance?

First, an in-depth study of previous work in this field was performed in order to tackle the knowledge gap and refine the factors that were to be the focus of this study. Second, practical experiments were performed that addressed each of the sub-question. In Chapter 4, through the first set of experiments, we managed to conduct an analysis focusing on workers' performance to examine **(RQ1)**, **(RQ2)**, and **(RQ3)**. Similarly to the work of Cai, Iqbal, and Teevan

(2016), instead of varying the task type, we focused on sequences of a single task type (relevance judgements) and on the effect of different class distribution settings, including both ordering (e.g., positive cases preceding negative ones) as well as class balance (e.g., one dominant class), on judgment quality and work efficiency (**RQ1**). In contrast to Damessie and Culpepper (2016), we observed that in the situation where most of the documents to be judged are non-relevant and the few relevant ones are presented first, workers perform better. Variation in task length was included in order to assess the difference in the performance of long and short batches of the same type of task (**RQ2**). Moreover, in these experiments, we applied priming by presenting certain data items first, thus showing workers examples of the classes to be labelled in the batch of HITs rather than presenting examples or running training tests before starting the real task (**RQ3**).

To address (**RQ4**), (**RQ5**), and (**RQ6**), in Chapter 5, we set up a second set of experiments for multi-classification tasks with the aim of examining repeatability, reproducibility, and generalisability of the task design in different situations. Answering (**RQ4**) requires conducting a study where the same experiments are repeated on a different time scale. We replicated the experiment using the same part of the dataset for the same assumption as discussed in Blanco et al. (2011) and Tonon, Demartini, and Cudré-Mauroux (2012) for measuring repeatable and reliable evaluation over crowdsourcing systems. These studies show experimental proofs that a crowdsourcing platform produces scalable and reliable results over a repetition time of one month. We also examined a shorter time scale of one week, for the same task design. Furthermore, we investigated the replication of the same task over multiple crowdsourcing platforms (**RQ5**), and over different classification tasks (**RQ6**). For (**RQ7**) we run a survey task to collect information about workers and the actual payment they received from the platform (Chapter 6). This served the aim of developing guidelines and tools to enhance crowdsourcing task design and to improve the overall performance of such tasks.

## 1.4 Research contributions

The contribution of this thesis are as follow:

1. An analysis of the work done in the field of crowdsourcing task design has been performed, which produced a new taxonomy of crowdsourcing systems and a state-of-the-art comparison of 32 aggregation mechanisms that have been used for presenting crowdsourcing task results in the past few years.
2. A presentation of a new classification schema for studies. It is based on six factors relevant to enhancement of task design and three main points of impact that vary with these factors.
3. Through multiple comparisons and practical experiments, we managed to introduce some novel methodological aspects and guidelines as main findings of this research:
  - The first set of factors studied in this research involved priming effects brought on by ordering relevant document first in the batch, rather than using as examples in

the instructions. This work was extended to investigate the impact of balance along with the order of the documents. Furthermore, we tested batches with variations in length to analyse whether with a change in the length of the task the performance was maintained. The findings of this study show that:

- class order and balance within crowdsourcing tasks impact significantly on the quality of the relevance judgements collected across different topics used in these tasks.
  - in terms of evaluating the quality of IR collections, ordering the batch according to the document retrieval rank, rather than using a random order, leads to better quality by allowing workers to identify relevant documents early in the judgement batch.
  - in terms of the length of the task, the results show that increasing the number of documents in unbalanced batches leads to higher performance, while the trend is inverted in balanced batches. Details of all findings are presented in Chapter 4.
- In the study of the second set of factors, the focus shifted from the inter-batch effect to the stability of the results for the same task design at a different time scale. This study looked at a particular subset of data and a fixed task interface. The outcomes of this part of research can be summarised as follows:
    - regarding the evaluation of the reliability of crowdsourcing results, there is a high level of agreement between crowd workers and expert annotators for the datasets we used in our tasks.
    - regarding the evaluation of platform *Consistency*, according to a within-platform analysis, there is significant consistency of results when repeating the same task once every week.
    - regarding the comparison of performance on multiple platforms, there is inconsistency in responses when reproducing the same task at the same time on different platforms. That is, crowdsourcing results are *not reproducible*, more details in Chapter 5.
  - The third factor studied was inspired by the observed motivational effect of paying a bonus for fast and high-quality work. This compelled us to examine some of the ethical issues around payments and review the actual amounts the workers received from each task, more details in Chapter 6.

## 1.5 Research design

We used mind mapping research methods adapted from Crowe and Sheppard (2012) to present a complete visual picture of this research in four linked steps.

Figure 1.1 presents briefly the research and the methodological steps that have been undertaken, described as follows:

- **Step 1** is *initiating* the research problem by presenting state-of-the-art literature review in the fields ranging from human computation to crowdsourcing systems and the factors affecting the quality of the results (Chapter 2), which will lead to the next step.
- **Step 2** is *implementing* the main research question and sub-questions that will address some of the gaps found in the literature to enhance the quality of crowdsourcing tasks in a particular area (Chapter 2).
- **Step 3** is *performing* the practical experiments over the three years of study to identify the ways in which each sub-question can be answered (Chapter 3). The evaluation of the workers' performance in crowdsourcing tasks was carried out using a data-driven approach to design a continuous analysis of crowdsourcing tasks. This approach consists of:
  - *Reflecting*, that is identifying the hypothesis, input data, and output result required.
  - *Coding*, that is the development of the task design, constraints, payment, and quality control settings.
  - *Launching*, that is broadcasting the task on the crowdsourcing platform, monitoring, aggregating, and analysing the results.
  - *Evaluating* the results and improving the current task design, then re-launching again, and going back to the Reflecting step.

Many experiments were performed to study the effects of variation of different settings of the design of the task on workers' performance and how they influence the analysis of the research sub-questions (Chapter 4-6).

- Finally, **Step 4**, is *reporting* findings of each experiment and presenting papers in related academic conferences, as explained in 1.8 and conclusion of this thesis (Chapter 7).

The results of all the experiments performed over the three years of study ascertained the validity and stability of the research hypotheses and addressed the research questions. The rest of the thesis will explain each part of Figure 1.1 and present the research steps that were performed.

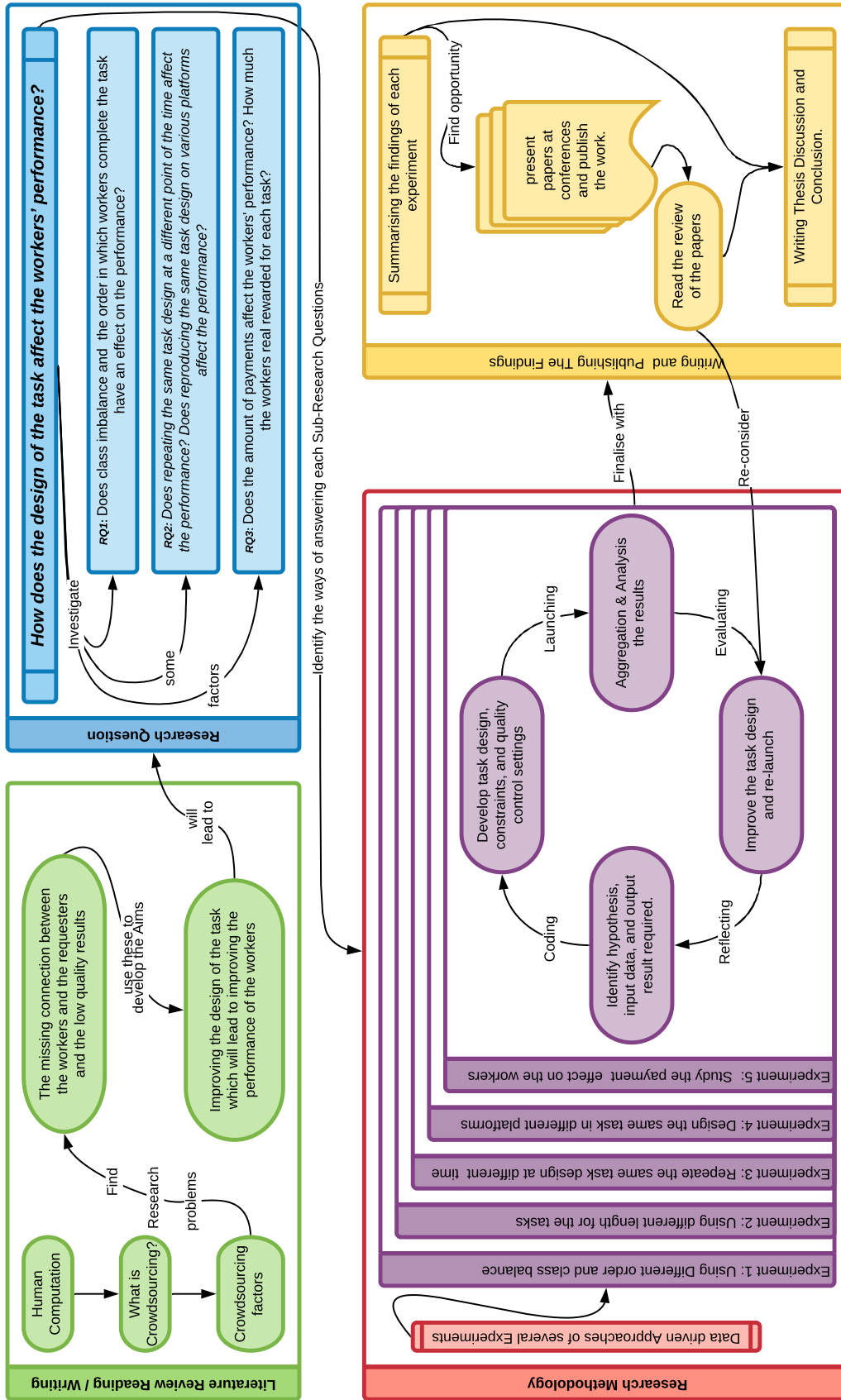


FIGURE 1.1: Mind map of the research

## 1.6 Selection criteria for literature review

To provide a broad background view of crowdsourcing and the interaction between it on the one hand, and Human Computation on the other, a bibliographic analysis of research published in the last decade (2006–2019) was performed.

The literature search was carried out on major digital libraries such as ACM Digital Library (ACM), Scopus, Science Direct (SciDir) and ISI Web of Science (WebScie). Moreover, a number of related conferences that publish research on crowdsourcing and related topics were considered: AAI, CHI, CI, CIKM, CSCW, ECML, ECSCW, HCOMP, ICML, KDD, NIPS, SIGIR, UBICOMP, UIST, VLDB, WSDM, and WWW. Related journals also considered: ACM CSUR, ACM TIIS, ACM TOCHI, ACM TOIS, ACM TOIT, ACM TOSEM, ACM TWEB, Communications of the ACM, CSCW, Information Systems, IEEE Computer, IEEE Internet Computing, IEEE TKDE, IEEE TSC, IEEE TSE, VLDB, and WWW.

In order to keep the selection of references manageable and up to date, articles were retrieved through the advanced search feature of the digital libraries and saved preferences with keywords: Crowd, Crowdsourcing, Human Computation, Task Design, Micro-tasking, Crowd Labour; this advanced search was scheduled to send a monthly email with new published papers. Papers published in conferences were retrieved manually.

## 1.7 Structure of the thesis

This thesis is structured as follows:

Chapter 2 presents a review of relevant literature starting with the introduction of Human Computation as a branch of Artificial Intelligence, and the birth of Crowdsourcing from a human in the loop principle. This is followed by an analysis of the state of the art of crowdsourcing factors, processes and platforms giving a comprehensive taxonomy and survey of what was found in the literature. Follow by introduces a new classification schema for papers and previous research in the task design field and addresses the essential factors that affect the design of crowdsourcing tasks. Furthermore, this chapter presents an overview of the most popular crowdsourcing platforms and the main features in each platform. The previous studies that review the performance of some platforms are also summarised in this chapter.

Chapter 3 introduces the methodology and describes the preliminary setup for the research experiments performed in order to address the research question aiming to enhance the designing of different classification tasks.

Chapter 4 explores the effect of ordering and balance classes in the batch on relevance judgment tasks and examines the effect of different task lengths on the workers' performance.

Chapter 5 investigates the effect of reproducing the same task design on different crowdsourcing platforms and also the effect of repeating the same task over time. By presenting a fixed task design interface for this study, several factors have been varied and examined, such as

a bonus motivation for high-quality performance and using different datasets with multiple classification tasks to measure that effect.

Chapter 6 presents a wide demographical distribution of the workers who have been using one particular crowdsourcing platform for the last four years. Moreover, the effect of the reward in some task designs and the ethical issues around payments were investigated by running a survey task and asking workers about their real motivations for working on the crowdsourcing task.

Chapter 7 summaries the main contribution and discusses the extent to which findings can be generalised. Moreover, it outlines the research limitations, gives recommendations, and presents our future directions.

## 1.8 Previously published material

Some parts of this thesis have been published in the following peer reviewed journals and conferences:

- Part of the work presented in Chapter 2 is partially published in (Sabou et al., 2018). The rest of the work presented in this Chapter is currently under review in ACM Computing Survey Journal (submitted in 30-Apr-2018).
- The preliminary experiment of Chapter 4 has been presented and published in The Fourth AAI Conference on Human Computation and Crowdsourcing (HCOMP 2016) (Qarout, Checco, and Demartini, 2016), and the rest of the chapter are currently under review in ACM Transaction on Information Systems (submitted on 28-May-2019).
- The preliminary experiment of Chapter 5 has been presented and published in The sixth AAI Conference on Human Computation and Crowdsourcing (HCOMP 2018) (Qarout, Checco, and Bontcheva, 2018), and the rest of the chapter are currently under review in The Seventh AAI Conference on Human Computation and Crowdsourcing (HCOMP 2019) (submitted on 5-June-2019).
- The remainder of the thesis is not previously published.





# 2

## Literature review

### 2.1 Introduction

This chapter will present the conceptual overview of the historical background of *Crowdsourcing*, bring together the efforts of researchers to define it. A brief description of the elements of the system is provided, along with the relationships between them. The state of the art of research in the taxonomy and classification of crowdsourcing systems will be summarised, with an in-depth analysis of what has been done in the area of designing the task to enhance the quality of the outcome.

Designing an appropriate task can lead to high-efficiency outcomes and a reduction in disagreements in the results (Garcia-Molina et al., 2016). Catallo and Martinenghi (2017) define a taxonomy of designing crowdsourcing tasks based on four design dimensions inspired by the explicit control aspects of human computation mentioned in Law and Ahn (2011). These dimensions are defined as *What* kind of task needs to be solved, *Who* is going to solve it, *Why* the workers need to work on it, and *How* to process these tasks. This classification, along with low level components, presents the main factors that are involved in the process of designing crowdsourcing tasks. Moreover, Allahbakhsh et al. (2013) considered the task design as one of the main dimensions that control the quality of the crowdsourcing system. Their proposed quality-control approaches are the *design-time* approach, where the requester could use various techniques to control the quality of the task in the design stage; and *run-time* approach, where requesters include some monitoring during the task execution to prevent any mistakes or low-quality performance. These two approaches can help to control the quality of the results and can be applied separately or simultaneously to one task.

A number of studies have been conducted in the crowdsourcing field (Pan and Blevis, 2011; Yuen, King, and Leung, 2011; Xintong et al., 2014; Mao et al., 2017; Chittilappilly, Chen, and Amer-Yahia, 2016). Xintong et al. (2014) presented the state of the art of using crowdsourcing in data mining. Mao et al. (2017) conducted a survey on the use of crowdsourcing in the field of software engineering. Another study by Yuen, King, and Leung (2011) presented a different classification of crowdsourcing systems based on their applications, algorithms, performances and datasets.

A short survey by Pan and Blevis (2011) presented a literature review of crowdsourcing and interaction design in academic, business, and social domains. This study was the first step, providing some insights and recommendations for designing crowdsourcing tasks and highlighting some challenges in task creation within Human Computer Interaction (HCI). The main difference between this work and the current research is that we present full historical background, since the act of crowdsourcing had been carried out in different projects before the concept of crowdsourcing was published by Jeff Howe in 2006 (Howe, 2006). Moreover, this chapter introduces the historical hierarchy of Human Computation and the overlap with other fields such as Semantic web and Collective Intelligence is presented.

In this chapter, we present a new classification schema for papers on different dimensions of task design. Following that, we shed light on the features and services provided by different crowdsourcing platforms. We are presenting an overview of the most popular platforms and assessing the consistency and reliability of the results of crowdsourcing platforms.

## 2.2 Human Computation and Crowdsourcing

Since the rapid growth of Artificial Intelligence (AI), researchers have put considerable effort into creating systems that can simulate human behaviour. This has led to the development of different branches such as pattern recognition, expert systems, machine learning, and natural language processing.

Back in 1950, Alan Turing posed the question “Can a machine think?” and to answer it he proposed the concept of digital and human computers and he wrote:

*“The idea behind digital computers may be explained by saying that these machines are intended to carry out any operations which could be done by a human computer.”*(Turing, 1950)

In his article, Turing presents both human and digital computers as entities that can work together interchangeably in solving problems. After that, another view by Licklider (1960) connects the functionality of the computer to the human. He reports that humans and computers are two sides of one equation and only by working together this equation can achieve balance. Moreover, some computational problems still benefit from a human to solve them, such as image recognition, sentiment analysis, and planning and reasoning (Law, Ahn, and Ahn, 2005).

The term “Human Computation”(HC) in the context of computer science was first coined in the monograph by Law, Ahn, and Ahn (2005) which had the term as its title. In this work, the

authors presented algorithm games that use human skills to complete some specific task, and they defined the term as:

“... a paradigm for utilizing human processing power to solve problems that computers cannot yet solve.”

This definition, along with many others (Yang et al., 2008; Chan, King, and Yuen, 2009; Law and Von Ahn, 2009; Law et al., 2009; Yuen, Chen, and King, 2009; Quinn and Bederson, 2009; Schall, Truong, and Dustdar, 2011), presents the emergence of the problem in computation system and the process of solving this by a human.

HC methods leverage human processing power to solve problems that are still difficult to solve by using solely computers (Quinn and Bederson, 2011), and therefore are well-suited to support some of the computation areas such as Semantic Web research especially in those areas that still require human contributions. For example, HC methods could be used to create training data for advanced algorithms or as means to evaluate the output of such algorithms. However, in order to increase the accuracy and efficiency of data interpretation at scale, increasingly algorithms (machines) and human contributions are brought together in a natural symbiosis (Demartini et al., 2017). Such synergy is often performed as iterative interactions, also known as the Human-in-the-Loop paradigm. In this paradigm the user has the ability to influence the outcome of the machine process by providing feedback on different opinions, perspectives and points of views. Additionally, this paradigm contributes to increasing the explainability and transparency of AI results.

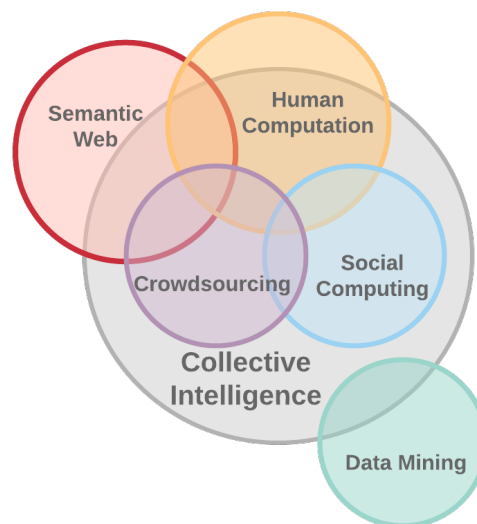


FIGURE 2.1: Adding Semantic web with Human computation as a means of solving computational problems, adapted from (Quinn and Bederson, 2011).

Quinn and Bederson (2011) define HC and other related concepts. They describe the relationship and overlap between *human computation* on one side and *Collective Intelligence*, *Social Computing*, and *Crowdsourcing* concepts on the other side. These relationships depend on the human input and the process of each. Collective intelligence refers to individuals working in

groups to reach a consensus in decision making and social computing refers to online communities such as blogs where people have social interactions. All previous concepts, along with human computation, were considered as part of collective intelligence which relies on a large group of workers to produce some tasks requiring intelligence except in the circumstance when one user could work alone; in this case, it was considered to be part of human computation only. By contrast, crowdsourcing is considered part of human computation because it replaces the traditional worker with online workers from the pool of internet users. Moreover, this research added Semantic web aspect to Quinn and Bederson taxonomy as shown in Figure 2.1, there exist several synergies between the fields of Semantic Web, Human Computation, and Crowdsourcing that open up a number of avenues for research (Sarasua et al., 2015).

Stemming from its original motivation of extending the Web with a layer of semantic representation (Berners-Lee, Hendler, and Lassila, 2001; Glimm and Stuckenschmidt, 2016), the Semantic Web (SW) aims to solve a set of complex problems that computers cannot yet fully master. Examples include creation of conceptual models (e.g., ontologies), semantic annotation of various media types, or entity linking across Linked Open Datasets and Knowledge Graphs. As a result, the largescale deployment of Semantic Web technologies often depends on the availability of significant human contribution that could be provided by crowdsourcing systems. Such contributions are traditionally provided by experts – e.g. ontology engineers to build ontologies, or annotators to create the semantic data or to link between the instances of various datasets.

## 2.3 The rise of Crowdsourcing: definitions and the origins of the concept

### 2.3.1 Definitions

The concept of crowdsourcing incorporates several approaches. This term combines the two key elements of the process, which depends on a large number of workers or online users (crowd) who contribute to out-(sourcing) by performing some tasks or providing potential ideas or solutions.

The concept appeared first in an article written by Howe (2006) where he states that crowdsourcing is:

*“an umbrella term for a highly varied group of approaches that share one obvious attribute in common: they all depend on some contribution from the crowd. But the nature of those contributions can differ tremendously”.*

The primary definition of crowdsourcing quoted by Doan, Ramakrishnan, and Halevy (2011) is:

*“.. enlists a crowd of humans to help solve a problem defined by the owners of the system”*

In other words, crowdsourcing is defined as a system where the requesters (i.e., those who need data-related tasks to be completed) use outsourcing by posting their tasks online (like an

open call) via a website or platform to a crowd of individuals who will perform the tasks for them. Brabham (2008) defines crowdsourcing in his article as:

*“an online, distributed problem-solving and production model that leverages the collective intelligence of online communities to serve specific organisational goals.”*

This definition focuses on the type of the problem that needs to be solved. Estellés-Arolas and González-Ladrón-De-Guevara (2012) came up with a general definition of crowdsourcing by aggregating 40 definitions, which had appeared between 2006 and 2011, by performing accurate analysis of eight factors derived from the three main elements: the Crowd, the Initiator, and the Process. The definition is as follows:

*“Crowdsourcing is a type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit. The user will receive the satisfaction of a given type of need, be it economic, social recognition, self-esteem, or the development of individual skills, while the crowdsourcer will obtain and utilize to their advantage that what the user has brought to the venture, whose form will depend on the type of activity undertaken.”*

Brabham (2013) describes this definition as “wordy but complete” and it covers most aspects of what crowdsourcing represents. He argues that since the rise of the crowdsourcing perspective in 2006, several cases were considered as crowdsourcing in the literature which are technically not. The next section describes the origins of crowdsourcing and the historical background of the concept.

### 2.3.2 The Origin of Crowdsourcing

The theme of Surowiecki’s book was the inspiration for Howe (2006) to outline the concept of crowdsourcing and name some examples that represent the main idea behind it. He used *Threadless.com*, *InnoCentive.com*, *Amazon’s Mechanical Turk*, and *iStockphoto.com* as examples of crowdsourcing models (Surowiecki, 2004). There are many successful examples on the web, where people bring together their knowledge and opinions as resources to support non-profit organisations. Moreover, in other cases, people use their creativity and skills to design a product or serve business goals in solving a particular problem. However, there are some studies that have identified some conditions regarding the definition a crowdsourcing system.

In his book, Brabham (2013), setting out what he considers crowdsourcing is and is not, describes the three factors that a crowdsourcing system requires (see Figure 2.2). These factors are: (1) The traditional top-down management from the organisation or the requester to the crowd. (2) The bottom-up, open creativity process from the crowd. (3) The position of the locus of control of the innovation in an openly exchangeable platform between the organisation and the crowd.

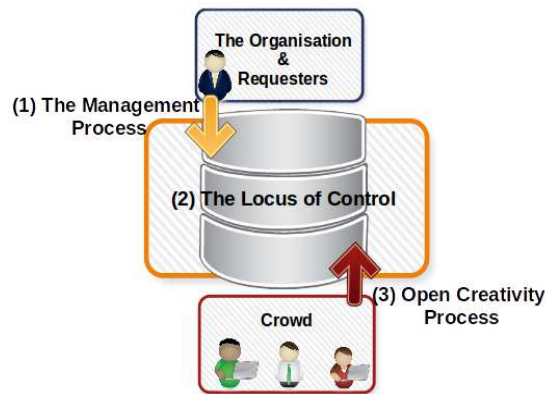


FIGURE 2.2: The key ingredients of crowdsourcing systems as described in (Brabham, 2013).

These key ingredients as distinguished by Brabham are the concepts that are an essential part of crowdsourcing. Both the business - or the requester - and the workers should share the control of creating the solution to the crowdsourced task, and once the locus of control moves down to the crowd or goes up to the business management, the system violates one of the key conditions of crowdsourcing. For example, *Wikipedia*, the world's largest knowledge base on the web, was established in 2001 and has used as a base for contributions for more than 200 projects from different companies (Halder, 2014). The locus of control is based with the users and it is only due to their contributions that the scope of the work increases. However, the company still has the control on removing or editing the content that appears on the *Wikipedia* website by using *WikiProject* which consists of groups of contributors, each group examining articles pertaining to a specific topic and assessing the quality of these articles.

A similar situation applies to any open source project, such as the *Mozilla Firefox Web* browser, where the supervision of management is absent and the collaboration flows horizontally between the users. Although there are users who act as management assessors in these examples, these projects are not considered crowdsourcing systems according to Brabham's definition. The constraints of Brabham's definition conflict with the first definition given by Howe (2006) and, moreover, it is in disagreement with several papers and studies such as Yuen, King, and Leung (2011) and Dawson and Bynghall (2011) that present a survey and taxonomy of crowdsourcing systems and applications as will be described in Sections 2.5 and 2.6.4.

### 2.3.3 The value of using crowdsourcing

AI systems aim to develop an adaptive learning system that can involve, work in, and assess many different situations based on training datasets that contain similar situations stored in knowledge banks. However, in some cases, such as medical diagnoses, the dataset could be limited or sometimes not exist (Holzinger, 2016). In these cases, using human expertise is the best way to find the right diagnosis and adds new information to the database; this has been defined in the literature as a "*Human in the Loop*" approach (Holzinger, 2016; Dautenhahn, 1998).

Using *Humans in the Loop* at every stage will improve machine learning performance by helping to create training data as well as helping to solve unknown cases to gain more accurate results and to make the algorithm smarter. Over the past few years, the term “Crowdsourcing” appears to have replaced “Human Computation” in several cases where employees could be replaced with people outside the business by means of an open call as compared to outsourcing where specific professionals are assigned the job (Holzinger, 2016). Crowdsourcing appears to be an extension of the HC aspect.

The benefit of using crowdsourcing platforms (i.e., where work requests and offers from the crowd come together) is the possibility of carrying out a job incredibly fast, with reasonable quality, and at a low cost in comparison with the traditional way of completing the same job (Alonso, 2013). In this context, a study by Crump, McDonnell, and Gureckis (2013) attempts to validate Amazon Mechanical Turk as a tool for collecting data in behavioural cognitive research. They designed several types of experiments and compared the results with traditional laboratory ways of collecting data. The findings of this study proved that the quality of the data collected under the experimental conditions in Amazon Mechanical Turk is highly similar to the quality of the data collected the traditional laboratory way. Despite some concerns related to the limitations of technical and visual design of the task and unexpected behaviour such as dropping out of a task before finishing it, collecting data with crowdsourcing saves time and money and could reach a wide range of users in a few seconds.

## 2.4 Crowdsourcing factors

A *crowdsourcing platform* depends on specific crowdsourcing processes to achieve its goal. Human factors play a significant role in crowdsourcing platforms with regards to their speed and quality (Difallah et al., 2015).

The atomic units of work in crowdsourcing platforms have been called *Human Intelligence Tasks* (HITs); these tasks are split into (*simple*) *tasks* which are small tasks for humans to complete but still difficult tasks for algorithms to achieve (e.g., natural language understanding, image processing). From this point on in this thesis, the terms *task* and *job* will be used interchangeably, as well as *microtask*, to describe the same concept.

This section will describe the main components/factors in designing crowdsourcing systems: human factor, the type of task factor, the process factor, and other factors influencing the design of a crowdsourcing task.

### 2.4.1 Human Factor

The Human factors in crowdsourcing platforms depend on the different crowdsourcing platform actors: the *requesters*, and the *workers*.

- **Requesters** are entities that represent the organisation or the business side. They are the customers who use the platform to post a description of the microtask which they require to be done by the workers, and details of the payments for that task once it is completed. The

requesters represent the funding source for the crowdsourcing platform, and they will only pay workers who produce valid, or accepted, high-performance results.

- **Workers** (Or *users*) are the key value of the crowdsourcing platform (Howe, 2008). They are the workforce present on the web that is available to complete millions of microtasks presented on a particular platform, usually in exchange for a small monetary reward for each task. The reliance on human capability raises many questions that have been the focus of researchers in the last few years.

Several studies highlight the effect of education level and demographic background on the workers' performance. Ipeirotis (2010b) found that 60% of the workers came from the USA and India and more than 35% of them had a bachelor degree . Another study by Jain et al. (2017) found that a total of almost 50% of the workers on the Figure Eight platform came from five countries: USA, Venezuela, UK, India, and Canada.

One of the new platforms, *Prolific Academia*, provides full information for the requesters about workers' demographic details: over 30% of the workers on this platform came from the USA, followed by 28% from the UK, and 30% of all workers had gained at least an undergraduate education level. More details on different platforms will be discussed in Section 2.6.

Other researchers turned their attention to the psychological behaviour, satisfaction and motivations that influence the workers' efficiency and their ability to produce high-quality results Jain et al., 2017; Brawley and Pury, 2016; McInnis et al., 2016; Harrison et al., 2013; Gadiraju et al., 2015. Targeting specific crowds to do particular tasks was the object of the study by Kazai, Kamps, and Milic-Frayling (2011). They focus on the workers' personality traits and compare them with some workers' behaviours such as the number of completed tasks, average completion time, and percentage of useful labels done by the workers.

According to these behavioural patterns, authors identified five types of workers: (1) *Spammer*, (2) *Sloppy*, (3) *Incompetent*, (4) *Competent*, (5) *Diligent*. High level of Openness correlated with workers' completion times and the number of useful labels. The findings of these studies indicate a strong relationship between workers' characteristics and the overall outcome for different task design. Understanding human factors is necessary for the successful design of a crowdsourcing job.

In crowdsourcing, the task will be performed by a human, not a machine, which is why the psychological aspects of designing the task should be analysed (Alonso, 2013). For that reason, the human factor is one of the main aspects that influences performance. Deng, Joshi, and Galliers (2016) enumerate guidelines for workers, requesters, and platform developers to enhance the services in the crowdsourcing field. They conducted a survey of workers' experiences interacting with a crowdsourcing system. The aim of this study was to enhance the workers' position by providing governance mechanisms to ensure transparency and fairness in the work environment.

Many researchers studied the influence, be it positive or negative, of personality traits on task accuracy, depending on how the side task was designed (Harrison et al., 2013; Kazai, Kamps,



and Milic-Frayling, 2011). For example, using different visual designs for a task could trigger different emotions in the workers leading to a variation in the results.

Kazai, Kamps, and Milic-Frayling (2011) classify workers' behaviour into five different types: *Diligent*, *Competent*, *Sloppy*, *Incompetent*, and *Spammer*. The authors tried to correlate the workers' characteristics and their personality traits with the accuracy and the average task completion time. This study showed that workers' behaviour has a significant effect on the accuracy in labelling tasks. On the other hand, the average time the workers spent on solving the task did not correlate with the accuracy for *Sloppy*, *Incompetent*, and *Spammer* workers. Connecting workers' behaviour with personality traits, *Conscientious* workers achieved higher accuracy in labelling tasks. However, this classification may not be applicable to other kinds of tasks. For this reason, this study could be extended to develop a model that can generate different worker typologies based on their reaction to a particular type of task and vary that classification according to the task type and task design.

Morris, Dontcheva, and Gerber (2012) looked at priming effects in micro-task crowdsourcing environments. They showed that priming workers can increase performance in creative tasks rather than in relevance judgement tasks as we do in this work (see Chapter 4). While they show that priming has positive effects, they also note that it should be unconsciously provided to workers and that it does not substitute training done by means of instructions and examples. Similarly, Harrison et al. (2013) used emotion priming in visual judgement tasks. They pointed out that while positive emotions have significant effects on performance, negative emotions could also prime workers positively in some situations. Moreover, there are environmental priming factors that could affect workers differently and these are beyond the requesters' control. Furthermore, Scholer, Turpin, and Sanderson (2011) and Scholer et al. (2013) studied the effect of priming relevance assessors by showing relevant documents early in the study and measuring the agreement between assessors. Participants from two universities examined documents with different levels of relevance and balance. The findings indicated that priming the assessors by presenting relevant documents early has a significant impact on the relevance label assigned to the rest of the dataset.

In André, Kraut, and Kittur, 2014, authors looked at how a group of workers perform and showed that asking workers to contribute sequentially works better than simultaneous collaboration. This finding proved the importance of crowdsourcing microtasks rather than a group of people working together in parallel to reach a solution. This demonstrates that workers feel more secure when working independently. Several factors need to be investigated, such as the identity of the workers, the time the tasks are released online, and the nature of the tasks; these factors could motivate teamwork among crowdsourcing systems. As compared to them, in this research we instead focus on individual worker performance (as commonly done for microtask crowdsourcing) and do not take into account group dynamics. As presented in the following chapters, when running different experiments, we measured the completion time for the batch in total, which is the time from releasing the task on the platform until the required amount of work has been received. Moreover, the completion time for each worker took into account the differences in setup on different platforms.

Another study looked at inter-task effects for image labelling, that is, how workers are influenced by the type of task they have previously completed when working on a new task (Newell and Ruths, 2016). Similarly to them, in Chapter 4, we look at the effect of tasks completed in a sequence but, rather, from a class distribution point of view.

Moreover, any previous experience that the workers had - including rejection of a completed job - has a significant impact on the workers' expectation for the upcoming task. McInnis et al. (2016) studied the impact of unfair job rejection on workers and the subsequent risk management. As a result of unfair rejections, workers tend to be more risk averse and accept the same type of tasks or select a task from a limited number of requesters who have a good reputation or have previously rewarded them for their work. This risk aversion could keep the workers safe from rejection, but it also prevents them from expanding their experience to any new type of tasks. Furthermore, new requesters face the risk of lack of turnout or have only malicious workers apply for their tasks.

#### 2.4.2 Crowdsourcing process

Despite the fact that different types of tasks exist on a crowdsourcing platform, the process of implementing each of them consists of the same stages. The mechanism of crowdsourcing works according to the following steps (see also Figure 2.3): (1) Define the problem, (2) Collect data requirements, (3) Design the task, (4) Launch the task online via a crowdsourcing platform, (5) Analyse the result, and, if the job has been completed successfully, (6) Send rewards to the workers.

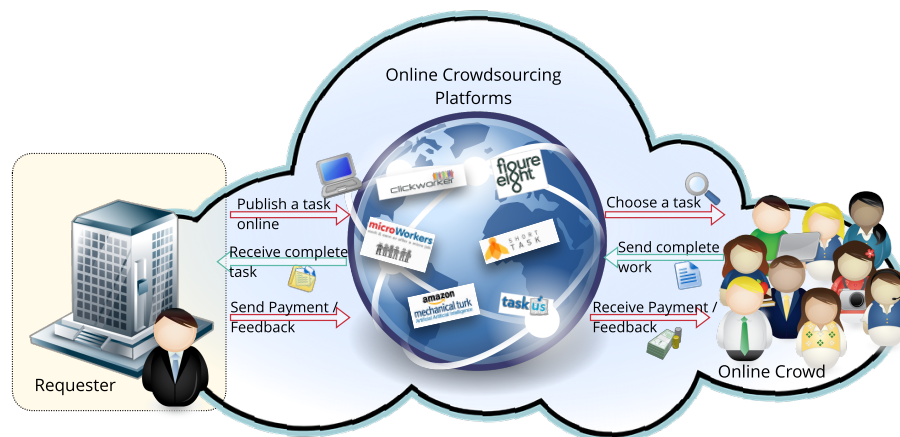


FIGURE 2.3: The mechanism of crowdsourcing.

The process of crowdsourcing could be analysed from three different perspectives:

- **The worker** who registers on a platform and performs some unpaid tasks in order to become qualified in certain skills that might be required. On the platform, the workers will find a list of jobs available along with the specified reward for completing it accurately. The online crowd is invited to an open call for everyone who is interested in providing solutions or performing the tasks on behalf of the company, which will name a price for each task. In a particular

situation, the crowd could be limited by the imposition of some constraints, such as needing certain experience in a given area (Brabham, 2008).

A number of recommendation systems appear to favour some workers for a specific task based on some criteria such as workers' history and their overall performance in a specific type of task. For example, when some workers produce better results than others in some task, the recommendation system will present to them a similar task to the one they did correctly (Schnitzer, Rensing, and Schmidt, 2015; Yuen, King, and Leung, 2015; Geiger and Schader, 2014). Researchers in the field present a number of assignment/recommendation mechanisms that will be discussed in Section 2.7. The worker will choose one of the listed tasks and try to solve it, and he/she could decide at any point to leave the task or submit an answer if they succeed in completing it. The last stage of the worker process is receiving a response to the submitted job, either rejection or the pre-agreed reward (Figure 2.4).

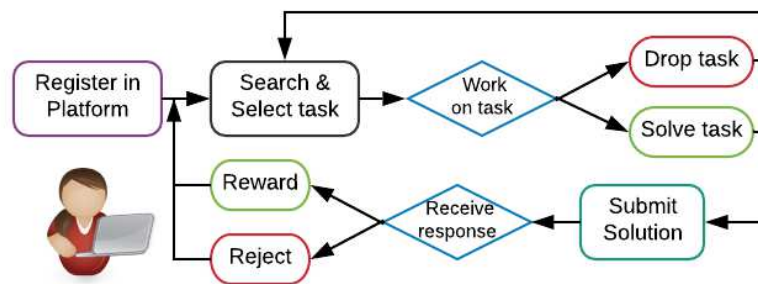


FIGURE 2.4: The crowdsourcing process from the worker perspective.

- **The requester** who represents the company or an academic organisation identifies some tasks or problems that need to be solved. The requester will gather the data and define the requirements, constraints, and output of the job and, for a long or complex task, they have to divide this task into smaller tasks (microtasks) which are released to the crowd online via one of the crowdsourcing platforms. For each task, the requester will determine a specific amount of time for the user to complete this task and submit it to the requester. When this time is up the requesters will analyse the quality of the received work and decide if the problem has been solved by the completed work or whether it should be rejected; based on this decision the worker will receive a response. This mechanism could vary from one requester to another (Figure 2.5).  
Receive

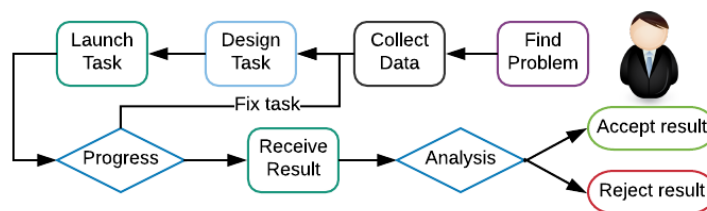


FIGURE 2.5: The crowdsourcing process from the requester perspective.

- **The Task** is developed and executed in three stages as presented in Luz, Silva, and Novais (2015). We simplify these stages in the breakdown shown in Figure 2.6. The first is the “*off-line*” design process, where the task will be outlined using one of the predefined templates provided by the platform or designed from scratch. In this stage, the data or the input will be fed in, and the parameters of the task will be set. The measures for quality control will be implemented at this stage to guarantee efficient results and detect spammers or malicious workers.

Moreover, a long/complex task will be decomposed into micro/simple tasks as it will be discussed in Section 2.4.3. For example, identifying the face of a specific person from a picture of a crowd in a football stadium is a long task to perform by one worker. Such a task can be divided into micro-tasks by cropping the picture into small pieces and crowdsourcing each piece as a simple independent task to workers. The topics of task design, pre-execution quality control, and factors affecting the process will be discussed in detail in Chapter 3.

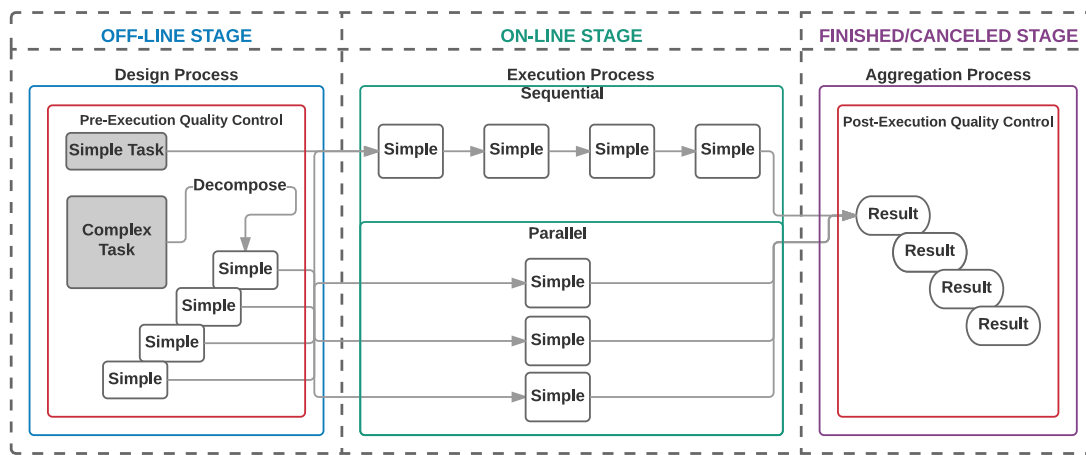


FIGURE 2.6: The crowdsourcing task process.

The second stage is the “*on-line*” execution, where the microtasks go online and become available. The implementation of this process could be in *parallel* when a microtask does not depend on the result of another one. Another way is to implement the microtasks *sequentially* one after another, and the result of each task becomes the input for the next task. In this on-line stage, the task could be paused, if it requires any modifications, and then resumed.

In some situations, the task could be done by a number of different users, and each could be paid if they complete the task successfully. In other cases, such as logo design, the task could be completed in different ways depending on the workers’ understanding and creativity, and the payment given to the best solution, as decided by the requester (Whitla, 2009).

The last stage is reached when the requester receives the completed job and it is assigned either a “*finished*” or “*cancelled*” state. In case of reaching the finished state, the microtasks will go through the aggregation process where these small tasks will be merged together to form the final result of the job. Post-execution quality control methods will be used to identify malicious workers based on their performance and their failure to meet the quality criteria that have been setup in the pre-execution quality control method. The aggregation mechanisms that

have been used in crowdsourcing systems and the post-execution quality control methods will be discussed in Section 2.7.

### 2.4.3 Types of crowdsourcing tasks

The tasks vary in length and complexity (Cheng et al., 2015), and to achieve a high-quality performance, the requester may split the job into smaller tasks (microtasks) which can be crowdsourced separately and their results then merged to reach the overall outcome (Chittilappilly, Chen, and Amer-Yahia, 2016). Some tasks are simple or impossible to be divided into smaller microtasks.

According to Nakatsu, Grossman, and Iacovou (2014), task complexity can be described by eight categories enumerated in their article. These categories are developed from three dimensions that have been found as the most concise representation of task complexity. The first dimension, *Task structure*, defines whether the task is **Well-structured**, for example a task with a specific result such as annotating part of a text according to specific labels, or **Unstructured task** in which workers' creativity is required such as developing an algorithm that provides the best solution to a problem. The second dimension, *Task interdependence*, defines whether the task falls into the category of **Independent**, where the workers can solve the task separately, or **Interdependent**, where the workers will find a solution to a particular task in the virtual community or after solving a series of sub-tasks separately and aggregating their results to form an overall solution. The third dimension, *Task commitment*, is based on the effort needed from the workers to complete the task. The author found that most researchers are aware of this dimension but no one added it as part of the definition of task complexity. This dimension defines as **Low-commitment** tasks which require minimum effort and workers can solve them straightforwardly and **High-commitment** tasks on which workers will spend more time, effort and resources, for example a task of designing software .

Gadiraju, Kawase, and Dietze (2014) present a taxonomy of the task types based on survey data collected from the crowd. They used two levels of categorisation: the first classified tasks based on the *goal* of their design; the second was based on the ways of performing these tasks or, as they called it, the *workflow* of the job. Figure 2.7 interpret the categorisation as outlined in this study.

Some of the categories within these two levels have been previously mentioned in Dawson and Bynghall (2011) along with the application domains where they have been implemented. Difallah et al. (2015) indicate that **Content Creation** was the most popular type of task over the last few years and the trend of using **Surveys** and **Interpretation and Analysis** has increased rapidly since 2009.

These classifications will help the requesters in designing the task in the most appropriate way. Some of the task types become more popular than others. Moreover, one task could be classified under more than one scheme. For example, asking the crowd to extract some words from an image and write them in a text box is a *Media Transcription* task. It could also be correctly classified as **Content Creation** because the worker will create new materials as their answer,

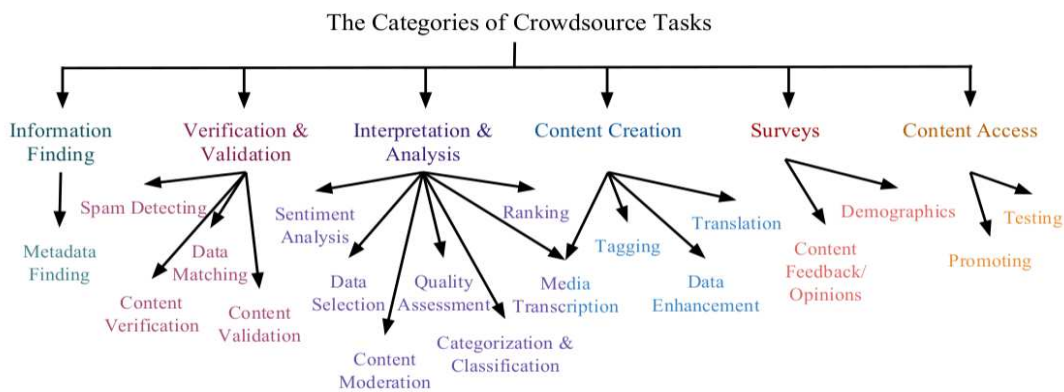


FIGURE 2.7: A taxonomy of crowdsourcing tasks, adapted from (Gadiraju, Kawase, and Dietze, 2014).

and as **Interpretation and Analysis** because the worker will use their interpretation skills to complete the task.

Luz, Silva, and Novais (2015) proposed different types of tasks based on the nature of the task: a *Qualification Task* needs to be completed by the worker to gain a particular skill; an *Aggregation Task* uses an aggregated result from a previous task as the core input data in the design of this task; a *Partition Task* represents a set of microtasks that together form one complex task; a *Grading Task* is designed for expert or high-ranking workers to evaluate the results of a qualification task.

Moreover, Yuen, King, and Leung (2011) describe different types of tasks based on the same schema: *Geometric Reasoning*, tasks that use the crowd in shape and visual analysis; *Named Entity Annotation*, used for recognising and categorising an object by its textual references, such as the name of a person or place; *Opinions and Commonsense*, tasks that collect feedback from the crowd in a specific area; *Relevance Evaluation*, of part of a document or dataset; *Natural Language Annotation*, which is one of the most challenging tasks for a machine and easy for a human to perform; *Spam Identification*, which assesses the validity of, and removes spam from, some contents. All of these types were categorised under the *Voting System* as part of the crowdsourcing application, which will be discussed in Section 2.5.

An optimal task design for one crowdsourcing task might not translate well to a task of different nature. Researchers in the field still investigate the effect of task design on the performance. Each study handles a maximum of two kinds of tasks and often for only one type of task. An analysis of the effect of different variables (e.g., interface, length, the number of items) on task performance has been shown not to generalise when the nature of the task is variable. For example, Marcus et al. (2012) compared counting approaches for image-based tasks. The labelling approach displays a sample and asks the workers to select one of the given labels, for example, if a photograph is that of a man or a woman, whereas in the counting approach the authors show a collection of samples (photos) and ask the workers to estimate the number of people in the photos with specific features, for example, hair colour or gender. Each approach was executed using different task designs (i.e., interface, length, the number of items) and the

results showed that counting approaches outperform labelling ones, while also resulting in lower completion times. However, the results from the image task were different from those of the text processing task. Moreover, the format of the question - closed, such as multiple choice, or open, such as when a text box is provided for the answer - also influences the workers' performance. Some studies found that using predefined answers can save time and result in more accurate answers (Jain et al., 2017), whereas other studies found that this type of tasks can increase the number of malicious workers who complete all answers in the task quickly, just to gain rewards (Eickhoff and Vries, 2013; Gadiraju et al., 2015).

In addition, giving the workers some level of freedom in the way they perform and respond to the task leads to high motivation of the workers (Moussawi and Koufaris, 2013). Similarly, Eickhoff and Vries (2013) state that the use of open questions can lead to obtaining more creative answers and to less cheating. In Alonso (2013), it appears that using questions that require a text answer to get feedback is very helpful. The author conducted a number of repetitive tasks that handle large datasets incorporating some factors that could enhance the overall result. The first factor is forming the question correctly. He recommended that the question should be asked in a simple and straightforward way so that it could be consistently understood by all workers. For answers to labelling questions, he suggests replacing them with a numerical scale to prevent misunderstanding. Moreover, it is preferable to use a broad range of labels (with no more than 6-7 categories) to give the workers some degree of flexibility in giving the right answer. Overall, he favoured this kind of task, that is labelling, for measuring the efficiency and the quality of the task design.

#### 2.4.4 Other factors influencing the design of a crowdsourcing task

Several studies analysed the effect of the other aspects: the length of the task (how long can a task be and still produce a good result), the nature of the required work (for example writing, classifying, or designing), the use of training questions and examples, the graphical user interface which often varies according to the complexity of the task, and also ordering of the data in the batch and how that could prime the workers. From analysing the researchers' efforts in the area of improving the design of crowdsourcing tasks over the last decade we found 4 other factors that effect the design of crowdsourcing task:

1. Task Graphical User Interface (GUI).
2. Training questions.
3. Length of task.
4. Ordering of data in the task.

The following sub-sections will present some of these studies in details.

Furthermore, in this research, the results of the analysis shed light on three main points of impact that vary with the design:

- *The task outcome.*

- *The completion time of the task.*
- *The workers' experience.*

The impact on the task outcome includes the accuracy and precision of the result and any other improvements that can be measured when testing one or more of the design aspects. Several studies considered the fact that task design has a significant effect on the task outcomes. McDonnell et al. (2016) showed that designing the task in a way that reduces the cognitive load on workers significantly increases performance. They used a successful approach to increase crowd reliability by implicitly making workers think about the task by asking them for an explanation of the assigned label. Related to this, Yang et al. (2016) showed how task design properties are highly correlated with perceived task complexity. The time spent completing a particular task, that is, the time from the moment of picking the task up until its submission, can also vary depending on the design aspects of said task. Moreover, multiple studies found that using some variations in the design aspects has a significant impact on the workers' experience and their ability to learn skills necessary to complete the task in question as well as other tasks that they could choose in the future. Table 2.1 presents a summary of the six aspects and the three impact elements covered in this work.

#### **- Graphical User Interface (GUI) of the task**

Since the graphical interface of the task is the main way for the workers to understand the job, it is fundamental to design adequate graphical interfaces that can help the workers to understand the task requirements, the process they need to follow, and the results that are expected from them.

Allahbakhsh et al. (2013) consider graphical design of the task to be one of the factors that affect the quality of the outcome: they found that implementing a simple interface could help the workers to complete the task in a short time and increase the accuracy of the completed job. Furthermore, the study by Jain et al. (2017) showed that writing long instructions that provide a detailed description of the task, and using examples, will have a positive effect on the quality of the result, particularly for complex tasks. Additionally, Alagarai Sampath, Rajeshuni, and Indurkhya (2014) demonstrated that using different background colours in the design of the task to highlight the areas that need to be filled in by the workers, improved the results and reduced the completion time of the tasks.

A study by Kim et al. (2015) used a crowdsourced task to match the appearance of the colour of some products on a website with the real colours of the same products. The lighting and quality of the image used in the task had a strong impact on the accuracy of the result. Other studies, such as Finnerty et al. (2013) compared the outcomes of two tasks with simple and complex interfaces. In the simple interface condition, they used a white background and clear instructions layout, whereas in the complex interface condition they used a patterned background and unstructured layout. The experiment proved that using a simple, clear interface gives better results than using the same task content but with a complex interface.



TABLE 2.1: Representative literature references for the studies in the design of crowdsourcing tasks.

		The Impact elements		
		Task outcome	Completion time	Workers' experience
Task Design Aspects	Psychological aspect	(Harrison et al., 2013), (Kazai, Kamps, and Milic-Frayling, 2011), (Morris, Dontcheva, and Gerber, 2012), (André, Kraut, and Kittur, 2014), (Organisciak, 2014)	(Kazai, Kamps, and Milic-Frayling, 2011)	(Alonso, 2013), (McInnis et al., 2016), (André, Kraut, and Kittur, 2014), (Newell and Ruths, 2016)
	Type of task	(Jain et al., 2017), (Marcus et al., 2012), (Eickhoff and Vries, 2013), (Gadiraju et al., 2015), (Towne, Rosé, and Herbsleb, 2017), (Bozzon et al., 2013), (Feyisetan et al., 2015), (Gadiraju, Kawase, and Dietze, 2014), (Brambilla et al., 2015), (Yang et al., 2016)	(Marcus et al., 2012), (Jain et al., 2017)	(Moussawi and Koufaris, 2013), (Eickhoff and Vries, 2013), (Kim and Monroy-Hernandez, 2016)
	Task GUI	(Allahbakhsh et al., 2013), (Yang et al., 2016), (McDonnell et al., 2016), (Alonso, 2013), (Alagarai Sampath, Rajeshuni, and Indurkha, 2014), (Jain et al., 2017), (Wu and Quinn, 2017), (Gadiraju, Jie, and Bozzon, 2017), (Willett, Heer, and Agrawala, 2012), (Nakatsu and Grossman, 2013), (Crump, McDonnell, and Gureckis, 2013), (Finnerty et al., 2013), (Zhao and Hoek, 2015), (Komarov, Reinecke, and Gajos, 2013), (Alonso, 2015), (Law et al., 2016), (Weidema et al., 2016), (Koyama, Sakamoto, and Igarashi, 2014)	(Alagarai Sampath, Rajeshuni, and Indurkha, 2014), (Allahbakhsh et al., 2013), (Wu and Quinn, 2017), (Koyama, Sakamoto, and Igarashi, 2014)	(Deng, Joshi, and Galliers, 2016), (McDonnell et al., 2016), (Alonso, 2013), (McInnis et al., 2016), (Walsh et al., 2014), (Dontcheva et al., 2014), (Law et al., 2016)
	Training Questions	(Oleson et al., 2011), (Jain et al., 2017), (Wu and Quinn, 2017), (Mitra, Hutto, and Gilbert, 2015), (Doroudi et al., 2016), (Willett, Heer, and Agrawala, 2012), (Gaikwad et al., 2017), (Towne, Rosé, and Herbsleb, 2017), (Kazai and Zitouni, 2016), (Feyisetan et al., 2015)		(Mitra, Hutto, and Gilbert, 2015), (Zhu et al., 2014), (Doroudi et al., 2016), (Drapeau, Chilton, and Weld, 2016), (Feyisetan et al., 2015)
	Length of task	(Cheng et al., 2015), (Kittur, Smus, and Kraut, 2011), (Dai et al., 2015), (Zhao and Hoek, 2015)	(Cheng et al., 2015), (Kittur, Smus, and Kraut, 2011)	(Brambilla et al., 2015), (Walsh et al., 2014)
	Ordering data in task	(Cai, Iqbal, and Teevan, 2016), (Damessie and Culpepper, 2016), (Yang et al., 2016), (Newell and Ruths, 2016), (Williams et al., 2017), (Zhuang et al., 2015), (Alonso and Baeza-Yates, 2011)	(Lasecki et al., 2015), (Krishna et al., 2016)	(Lasecki et al., 2015), (Damessie and Culpepper, 2016), (Zhuang et al., 2015), (Alonso and Baeza-Yates, 2011)

Further to this, a study by Alonso (2013) presented an interface design by following the guidelines of Nielsen (1993) to point out the basics of task design: write clear instructions, show examples, highlight and colour what is important and required for the job. These factors can reduce the effort to complete the task. Also, using a relevant, clear, and attractive title for the job will make it easier for workers to find it quickly when they are searching the platform for

possible tasks to accept and complete.

Moreover, in McInnis et al. (2016) a number of factors that lead to unfair rejections were presented, such as insufficient task design, misleading instructions, technical errors, and requesters with poor knowledge. They concluded their study with a number of suggestions that could reduce the risk and enhance the connection between workers and requesters to achieve a better final outcome. One of these suggestions was to provide, in the design of the task, *an alarm* for a broken task, which notifies the requester of any errors in the task design during the work process.

Recently, Wu and Quinn (2017) outlined best practice guidelines for writing task instructions that could optimise the outcome quality. This study found that regardless of the fact that long and clear instructions will improve the results, workers tend to favour tasks with short guidelines and few lines of instructions. Therefore, the requesters should strike a balance between presenting full instructions and defining attractive short steps which will be easy to read and deliver the full format of the task specification at the same time.

In a similar study, Gadiraju, Jie, and Bozzon (2017) investigate the effect of task clarity on workers' performance. They surveyed workers' opinions on the clarity of tasks they had completed. The feedback they received suggests that the lack of clarity of the task was mostly due to weaknesses in the presentation of the instructions and in the writing style. Also, they reported that an absence of relevant examples made the understanding of job requirements less clear to the workers. The findings of this study show that task clarity can be predicted and supervised via the proposed model and can guide the requester in the task design. Further investigation could draw on this work to examine the relationship between task clarity and complexity on workers' dropout rates.

### - Training questions

Training questions can be formed in a variety of ways some of which may be helpful for certain specific tasks but not others. Several studies have looked at the training of workers before or whilst performing a particular task and a number of training techniques or methods have been used which can be summarised as follows:

- (1) *Control method*: does not have any training questions and the workers read the instructions and start solving the task directly.
- (2) *Solution method*: adding a number of training tests before the real task questions without saying explicitly that the first tasks are for training purposes.
- (3) *Gold Standard*: the same setup as in the solution method but after solving the first training tasks workers are shown the correct answers for the tasks and informed that they had been used for training purposes. Oleson et al. (2011) used this method in tasks as a quality control mechanism rather than using it as a training method.
- (4) *Example method*: design task instructions to explain that workers will be shown some examples completed by an expert and that they are not allowed to start the task until the 30 second

demonstration has been completed; this forces workers to read the examples and understand how they were solved.

Recent studies by Jain et al. (2017) and Wu and Quinn (2017) proved that using examples is crucial and plays a key role in increasing the accuracy of the results and the total agreements. Similarly Mitra, Hutto, and Gilbert (2015), presented some examples for the workers followed by test questions to measure the improvement in their performance and to determine if they had learned from the examples.

(5) *Validation method*: in this method workers were shown two answers by other previous workers and asked to validate these answers by filling out some specific questions about them. Zhu et al. (2014) found that using the validation method in subjective tasks, which required some creativity in devising the solution, was more effective than making the workers do more training tests.

Another study by Doroudi et al. (2016) presents different techniques of using training questions to improve the overall result of what they define as a complex task. They used all five methods to find the most beneficial training method. The findings of this study reported that showing the workers expert examples increased the overall accuracy of the answers compared with using other methods. Moreover, using the validation method was the most effective way of training (Zhu et al., 2014). Rather than using training questions, in this research, we will focus on prime workers by using specific order for documents as a kind of training and learning mechanism.

#### **- Length of the task**

Crowdsourced tasks can be designed with variations in length. To maintain a balance between the length of the task and the desired quality of the outcomes, several solutions have been proposed in different studies. One of these solutions is to decompose a long task into shorter ones (microtasks), which fits best with the crowdsourcing platform paradigm of keeping the tasks simple.

The main purpose for using crowdsourcing platforms is to break down a task into smaller tasks, as we have mentioned previously, which can be solved by the crowd, achieving high-quality performance as well as saving time and money (Cheng et al., 2015; Kittur, Smus, and Kraut, 2011). These microtasks should have a low level of complexity to achieve their purpose.

Doroudi et al. (2016), defined the level of complexity for tasks as a task which cannot be decomposed into micro-tasks and workers can use different mechanisms to perform such tasks. For complex tasks, a high level of accuracy is not achievable with low expertise workers.

Other related work in the area of microtask crowdsourcing has looked at the effect on crowd performance of task granularity (Cheng et al., 2015). Authors showed that shorter tasks lead to increased overall completion time but also to better quality contributions. Similarly, Allahbakhsh et al. (2013) discussed the granularity of long tasks, which affects the quality of the

outcomes. The final result of such a task is a combination of the results of a number of smaller or shorter tasks.

Another solution is to break the long task up with some activities to keep the worker interested in completing the task. Dai et al. (2015) proposed including some entertainment micro-tasks as a short break in performing a long task. They used the MTurk platform to design three different long tasks: (1) Classifying images, (2) Rating Wikipedia articles, and (3) Merging Freebase entities. For each type of task, they inserted three different “micro-diversion” scenarios: no diversion, a narrative webcomic story, and a dice game to keep workers on track and motivate them to continue working on the task. The findings of this study proved that using micro-diversions can significantly maintain workers’ motivation to continue working on a long task as well as enhancing the speed of the answers. There are some variations in the findings depending on the task type and the micro-diversions combination. A complex cognitive task such as rating a Wikipedia article was performed more effectively using a diversions task. Moreover, the story acts better than a game diversion in speeding up workers’ performance especially with the technique of displaying one page of interactive visual design between one task and another. Workers were shown to be more motivated to finish the task in order to see the next page of the story.

Moreover, Brambilla et al. (2015) propose prototyping methods for task design that are implemented first in small datasets in order to gain better results for designing the same task for large datasets. This approach reports significantly better results for image relevance judgment tasks; further work could use the same strategies in other types of tasks.

### - Ordering of the data in the task

In the process of implementing the task, variations in the sequential ordering of microtasks could lead to variations in the overall results. The requester has the ability to organise the data in the batch and present it in the order of increasing difficulty that gradually primes the workers and improves their performance and the level of agreement. Earlier IR evaluation experiments investigated the effect of document order on the relevance assessments.

In this context, it has been shown that the order in which documents are displayed affects the users’ judgements (Eisenberg and Barry, 1988; Park, 1993). In this study, we investigate the effect of order on the relevance judgements using more than one set of classes ordered differently per batch and compare the results of various batches (Chapter 4).

Recent work looked at how sequences of writing tasks impact crowd worker efficiency (Cai, Iqbal, and Teevan, 2016). They observed that by varying the order of task complexity and task type, workers’ performance would vary thus showing potential for optimising worker efficiency by sorting tasks in a batch appropriately.

Another work looked at the effect of interruptions and of changing tasks type (i.e., context switch) on sequences of crowdsourcing tasks showing how worker speed would significantly decrease in such situation (Lasecki et al., 2015). In Shao et al. (2019), authors evaluate the annotation of image search engines, and they found that annotating relevance of images in a

sequence is more efficient than single item-based annotation for the same images. In Chapter 4, we focus on sequences of a single task type (i.e., relevance judgements) and on the effect of ordering tasks on work quality.

Looking at the agreement, in Damessie and Culpepper (2016) authors investigate the impact on the *inter-rater agreement* of presenting documents in two different ways: (1) descending order of relevance, and (2) ordered by document identifier (i.e., similar to random ordering). They designed a judgement task for 30 documents across *easy* and *hard* topics extracted from TREC collections and with a four-level relevance scale. Their results show that ordering by document identifier leads to a higher agreement in both easy and hard topics and a better result in term of identifying the relevant documents. Moreover, in Palotti et al. (2016) authors examined agreement in relevance assessments of medical images and the effect of workers' level of expertise, payment level, and query variation on the obtained result. They used pairwise comparisons to evaluate the agreement among paid expert workers (i.e., medical students). The results show that there is inter-disagreement both between paid experts producing judgements as well as between unpaid workers with no medical background. Such disagreement also leads to changes in the ranking of IR systems being evaluated. While the low agreement between relevance assessors is known to be common Webber, Chandar, and Carterette (2012), Chandar, Webber, and Carterette (2013), and Demeester et al. (2014), in our work we use agreement as a measure to experimentally compare different data distribution setups.

A recent study by Yang et al. (2016) proposed a high-dimensional regression model to measure the impact of task structural features on the complexity of the task and, conversely, used these features to predict the complexity and the tasks outcomes, showing that the semantic description and the visual appearance of the task are the most useful features to predict the complexity of the task and improve the quality of the output.

## 2.5 Classification of crowdsourcing systems

Crowdsourcing systems have been classified in the literature according to different criteria. Some classifications are based on the services that are provided to the online communities and the field to which the workers will make a contribution via these systems. This section summarises the taxonomy of crowdsourcing systems and presents four of the broadest classifications as shown in Table 2.2.

### 2.5.1 Based on the nature of the job

One way of classifying crowdsourcing systems is to focus on the kind of job that the crowd needs to solve. Brabham (2013) proposes a classification with the following divisions: (1) *Knowledge discovery and management*, which uses the crowd to collect data and create a resource of information with a particular format; (2) *Broadcast search*, which uses the crowd to solve scientific problems; (3) *Peer-vetted creative production*, which uses the crowd to design and create ideas as a solution for the task; and (4) *Distributed Human Intelligence tasking*, which uses the crowd to solve problems with large-scale data that needs a human to process it.

TABLE 2.2: Taxonomy of crowdsourcing systems.

	Classification	Types of Systems
Nature of the Job	(Brabham, 2013)	Knowledge discovery and management, Broadcast search, Peer-vetted creative production, and Distributed Human Intelligence tasking.
	(Corney et al., 2009)	Creating, Evaluating, Organising job
	(Yuen, King, and Leung, 2011)	Voting, Information Sharing, Gaming, Creating
	(Doan, Ramakrishnan, and Halevy, 2011)	Rating, Adding media as contributing topic, Editing some existing data, Presenting solutions.
	(Rouse, 2010)	Simple, Moderate, and Sophisticated job.
	(Schenk and Guittard, 2011)	Simple short, Complex, and Creative tasks.
Motivation for the Crowd	(Corney et al., 2009) and (Andrásfalvy et al., 2003)	No reward, Fixed, and Success based reward systems.
	(Rouse, 2010) and (Brabham, 2013)	Self-marketing, Social status, Instrumental, Altruism, Token compensation, Market compensation, and Personal achievement.
	Doan, Ramakrishnan, and Halevy (2011)	Contribute by authority, Pay money, Use Volunteers, Sequence job, and Piggyback on established systems.
Identity of the Crowd	(Rouse, 2010) and (Doan, Ramakrishnan, and Halevy, 2011)	Low-ranking and High-ranking users.
	(Corney et al., 2009)	General, Moderates, and Experts users.
Nature of the Platforms	(Howe, 2008)	Crowd wisdom, Crowd creation, Crowd voting, and Crowd-funding
	(Dawson and ByngHall, 2011)	Distributed innovation, Idea, Innovation prizes, content markets, Prediction markets, and Competition platforms.
	(Geiger, Rosemann, and Felt, 2011)	Processing, Solving, Creating, and Rating.
	(Carr, 2010)	Social-production crowds, Averaging crowds, Data-mine crowds, and Networking crowds

Meanwhile, other systems proposed within this schema have focused on the complexity of the task, for example Corney et al. (2009) differentiate between *creating* (e.g. designing an advertisement), *evaluating* (e.g. giving feedback or completing a survey), and *organising* jobs (e.g. image tagging or website rating). Also Rouse (2010) sets out three different levels of job complexity: *Simple* (e.g. transcribing text), *Moderate* (e.g. testing product improvements), and *Sophisticated* (e.g. entering into a high-status design contest). Moreover, Schenk and Guittard (2011) categorise the jobs according to their level of complexity as *Simple*, *Complex*, and *Creative* tasks.

Another way of classifying crowdsourcing is by focusing on the way in which the crowd can

solve the problem. According to Doan, Ramakrishnan, and Halevy (2011) the crowd can solve the problem in four different ways. First, they can provide a perspective, such as a rating or an opinion, on a situation. Second, they can add some pictures, text, or video to contribute to a specific topic. Third, they can edit some existing data and link to other users on the web. Fourth, they can present solutions to problems that need to be solved, such as designing a logo for a company. These four categories range from *easy* tasks to *complex* ones. More details on crowdsourcing task types were presented in Section 2.4.3.

The types of applications that the system provides have also been considered in classification. Yuen, King, and Leung (2011) provide four such categories: (1) *Voting Systems*, which use the agreement of the crowd's answers to determine the right answer for the task; (2) *Information Sharing Systems*, which are those websites that share different types of information, such as *Yahoo! Answers* and *Wikipedia*; (3) *Gaming systems*, which refer to the online games that depend on players using the network to communicate (ESP Game); (4) *Creative Systems*, which ask the workers to invent work such as drawing or coding.

### 2.5.2 Based on the motivation of the crowd

This categorisation is based on the idea that crowdsourcing depends on the nature of the rewards that the crowd will gain in exchange for their contribution. Corney et al. (2009) and Andrásfalvy et al. (2003) classified crowdsourcing systems as follows: *No reward* (e.g., in social studies where requesters asked people to volunteer for some experiment or give some information); *Fixed reward* (e.g. all the workers are given a pre-defined fee as payment for a completed task); and *Success-based reward systems* (e.g. where workers are rewarded based on achieving some particular goals in the task, not just for completion of the task) (Andrásfalvy et al., 2003; Corney et al., 2009).

Another study by Rouse (2010) reports seven potential motivations and Brabham (2013) arrived at similar conclusions after conducting a survey which asked people about their reasons for working on a crowdsourcing platform. Both studies summarised crowd motivation as: *self-marketing*, where people are trying to find a full time job; *social status*, where people act socially and try to make friends; *instrumental*, which is working on tasks to develop some skills and experience; *altruism*, to share information and volunteer to help others; *token compensation*, to earn some bonus rewards; *market compensation*, those who have real difficulty finding a job and use crowdsourcing to cover their living costs; and finally, *personal achievement*, where people find the task very attractive as a challenge.

A slightly different way of classifying people's motivations is to use the reward systems that might be offered. Doan, Ramakrishnan, and Halevy (2011) describe five recruitment mechanisms in crowdsourcing systems: (1) Allow the crowd to contribute by authority; (2) Pay a specific amount of money for doing the job; (3) Ask the crowd to volunteer in the system; (4) Use solving the task as a required step to use other services on the web; or (5) Piggyback on established systems. The experiments in this research will be classified as a *Fixed reward* as workers were paid a small amount of money to complete the task.

Overall, an even balance should be maintained between the effort spent in processing the task and the incentive reward when designing the task. Even given the low rewards of crowd work, there is an increasing number of workers who use it as their main source of income, particularly in developing countries (Ross et al., 2010; Ipeirotis, 2010b). Different studies draw connections between workers' level of satisfaction and the reward with the type of the task and the effort spent to complete the job, as in Gadiraju, Kawase, and Dietze (2014) where they connect workers' reward satisfaction with proposed categorisation of the task. Although the financial reward is not the primary motivation for the crowd to work on the platform, they found that using standard rewards for different classes of tasks could improve the quality of the results. Mason and Watts (2009) proved that assigning high financial rewards for a task increases the level of interest in doing the job, which leads to results being achieved in a shorter time, but it does not necessarily increase the quality of the output; this will be discussed further in Section 2.4.4.

### 2.5.3 Based on the identity of the crowd

People of different ages, educational backgrounds, employment status, and cultures can act as workers on the platform. Who is qualified to perform the job is the fundamental question in this classification; many studies show a strong relationship between the nature of the job and the level of expertise that the crowd should have. Generalist workers most commonly participate in marketplace platforms such as *Amazon Mechanical Turk* or *Figure Eight* and more specialised workers could be found on a content-based platform such as *Threadless* or *TopCoder* (Brabham, 2008; Ipeirotis, 2010b).

Rouse (2010), states that a *simple* task can be performed by a crowd with an average level of training, while tasks classified as *Sophisticated* need users with a high level of knowledge or that are specialised in solving that particular type of job. Doan, Ramakrishnan, and Halevy (2011) also refer to the importance of designing an appropriate task for an appropriate crowd, as *Low-ranking users* will only be able to perform *Easy* tasks, whilst *High-ranking users* will be able to perform *Hard* tasks. Other studies divided tasks according to who could solve them, such as in Corney et al. (2009) where the authors classify tasks as: ones that all users could solve, ones suitable for most users, and ones only for experts users. This classification of workers could be used in future studies of workers' ability and its importance for task design.

### 2.5.4 Based on the nature of the platforms

This approach is to classify crowdsourcing based on how the platforms function. Howe (2008) classifies crowdsourcing into four categories: *Crowd wisdom*, *Crowd creation*, *Crowd voting* and *Crowd funding*. Another categorisation of crowdsourcing which uses the same approach is provided by Dawson and Bynghall (2011) who identified the following: *Distributed innovation platforms*, *Idea platforms*, *Innovation prizes*, *Content markets*, *Prediction markets*, and *Competition platforms*.

Geiger, Rosemann, and Felt propose four different types of crowdsourcing systems depending on the services provided by each system. These are classified as: *Processing*, *Solving*, *Creating*,



and *Rating* systems (Geiger, Rosemann, and Fielt, 2011). A further classification is found in Carr (2010) where he names the types of crowdsourcing according to how crowds either collaborate and communicate or work individually to perform the task: *Social-production crowds*, *Averaging crowds*, *Data-mining crowds*, and *Networking crowds*.

## 2.6 Crowdsourcing platforms

### 2.6.1 Overview of current crowdsourcing platforms

To understand the contribution of requesters and workers in crowdsourcing, we need to understand the platforms that provide the services and the extensive range of possible features that each one could present. The main structure could be similar across all platforms, although there are some variations in the process of task design and the mechanism of the services provided. However, to clarify some points, Table 2.3 sets out the different uses of terminologies across all six platforms presented in this section (Luz, Silva, and Novais, 2015).

TABLE 2.3: Crowdsourcing platforms.

Platform Features	MTurk	Figure Eight	ShortTask	CloudCrowd	Microworkers	Prolific Academia
Released Date	2005	2007	2009	2009	2009	2014
Task Design Method	Templates	Templates	Templates	-	Templates	-
Evaluation Method	Qualification test	Gold unit	Manual	Credential test	Manual	Manual
Aggregation	Manual	Yes	Manual	-	-	Manual
Population size	Over 500 K	Over 10 K	About 125 K	About 25 K	-	About 60 K
System terminologies						
- Job	Project	Job	Task template	Project	Job	Study/Task
- Unit	HIT	Unit	Task	Task	Task	Study
- Answer	Answer	Judgment	-	Answer	-	Submission
- Worker	Worker	Worker	Solver	Worker	Worker	Participant
- Requester	Requester	Requester	Seeker	-	Employer	Researcher
- Qualification	Qualification	N/A	N/A	Credentials	N/A	Eligibility
- Reference unit	N/A	Gold unit	N/A	Check task	N/A	Pre-screening test

This section presents some of the most common platforms for crowdsourcing services. Overall, the features and services provided for the requesters vary from one platform to another, and no single platform meets all the possible requirements that the requesters may have. For this and other reasons, researchers in the crowdsourcing field are aware of the need for improving (1) the design of the microtasks (Section 2.6.4), (2) the assignment of the right tasks to the most competent worker (Section 2.7), and (3) the aggregation mechanisms of the results of the crowdsourced tasks (Section 2.7). Furthermore, as an extension to MTurk, a new platform has just been released, **TurkPrime**<sup>1</sup>, to provide more features for the requesters and improve the functionality of the crowdsourcing services (Litman, Robinson, and Abberbock, 2017).

#### - Amazon Mechanical Turk

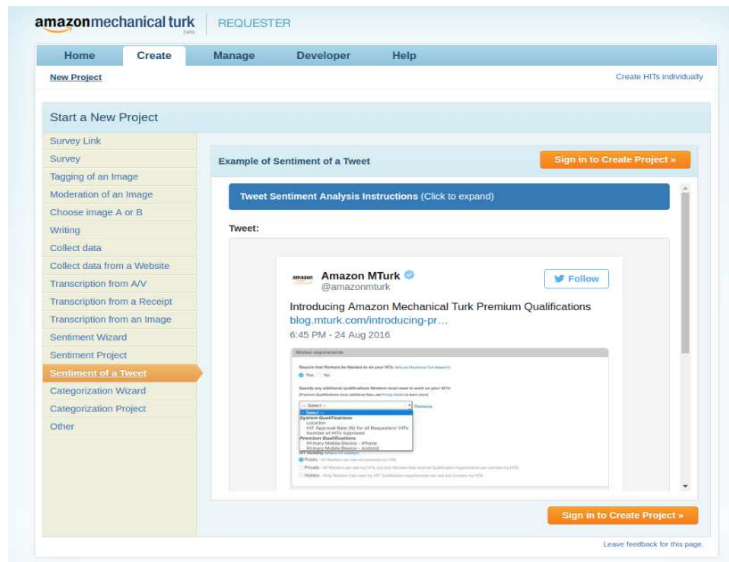
In late 2005, Amazon released its marketplace platform **Mechanical Turk (MTurk)**<sup>2</sup>. As it became the most popular online labour market for crowdsourced tasks, it had been primarily

<sup>1</sup><https://www.turkprime.com/>

<sup>2</sup><https://www.mturk.com/>

featured as the main platform for crowdsource studies such as information retrieval, human computer interaction, economic data, and data mining research (Chen et al., 2011).

The requester creates a *project* which contains a set of tasks called *Human Intelligence Tasks* (HITs). This project can be established from scratch using a HTML layout or by using pre-defined templates as shown in Figure (2.8-a).



(A) Requesters

Requester	Title	HITs	Reward	Created	Actions
Sonny Bonds	Write a 100-200 word summary of your college experience as a student in a Parale...	1	\$2.00	2/17/2017	Preview Accept & Work
Sonny Bonds	Write a 100-200 word summary of your experience as a student in a CRIMINAL J...	1	\$2.00	13d ago	Preview Accept & Work
Alexey Kuznetsov	Write a 520 words article about Calls in Australia	1	\$1.99	4d ago	Preview Accept & Work
FDU Healthy Aging Study	Health Behaviors & Wellness 2017	1	\$1.50	2/15/2017	Preview Accept & Work
20bn	[LIMITED HIGHER PAY] Record 10 short videos of actions with everyday objects	22	\$1.40	6m ago	Preview Accept & Work
20bn	[NEW TASK, LIMITED HIGHER PAY] Record 10 short videos of human actions	20	\$1.40	6m ago	Preview Accept & Work
Wei Xiang	Propose a design that help change bad sitting posture	3	\$1.25	1d ago	Preview Accept & Work
evelyn	Listening test of synthetic speech	30	\$1.00	2h ago	Preview Accept & Work
Gus Gardellini	Transcribe an Audio File approx. 5-minutes long (.mp3)	15	\$1.00	19h ago	Preview Accept & Work
RD Decision Research	Very short simple surveys (Total duration: about 3 minutes)	1	\$1.00	7h ago	Preview Accept & Work
Description: Very short simple surveys about lifestyles, attitudes, preferences, and/or decision-making. Age: must be 18 or more. Time Allowed: 15 Min Expires: in 6d		Qualifications Required: ✓ Total approved HITs is less than 50 ✓ HIT approval rate (%) is not less than 95			
David Chang	Verbal Picture Description 2	2	\$1.00	2d ago	Preview Accept & Work
SocExp	Participate in decision experiment	1	\$1.00	2h ago	Preview Accept & Work
huangxin wang	Categorize tweets. < 10 mins. \$1.0 reward	1	\$1.00	15h ago	Preview Accept & Work
Maastricht University - Social Psych	humbehav fix	1	\$1.00	12h ago	Preview Accept & Work

(B) Workers

FIGURE 2.8: Requesters and workers interface in MTurk platform.

MTurk allows requesters to specify the number of assignments per worker (who can perform single or multiple tasks) and also the number of qualifications the worker should fulfil. Some of these qualifications are already present in the worker's profile, such as demographic location, age range, number of accepted HITs, and first language; and others require some action from the workers, such as completing a sample/practice test to evaluate their eligibility for the real task.

The workers can see a list of HITs which are available for them to select. Each task has the name of the requester, title, description, the time allowed for completing the task, and the qualifications required. They can display the task list according to the number of HITs, the date of creation, and the amount of the reward as shown in Figure (2.8-b).

After receiving the complete HITs from the workers, the requester can accept their solution and approve a reward to be granted or reject their solution. MTurk does not provide an aggregation method for the results, so the requesters will have multiple answers for each task and the analysis needs to be performed manually outside the platform (Luz, Silva, and Novais, 2015).

#### - CloudCrowd

The CloudCrowd platform<sup>3</sup> was launched in October 2009 as an application on Facebook. The main difference between this platform and the others presented in this section is that online registration is only available for workers. To become eligible to work on a task, workers need to get credentials by completing test tasks with different levels of difficulty. Their answers are evaluated by comparing them with a reference unit which is similar to the gold unit in Figure Eight (Luz, Silva, and Novais, 2015).



WRITING AND EDITING	REQUIREMENTS	BONUS?	TODAY'S RATE
<b>Write Business Article</b>	Credibility: 40 Credentials: Writer...	Yes	\$33.00
<b>Write Press Release</b>	Credentials: Market...	Yes	\$14.75
<b>REVIEW: Credential - Write Product Descriptions</b>	Credibility: 85 Credentials: Market...		\$0.10
<b>Write Financial Content</b>	Credibility: 50 Credentials: Writer...	Yes	\$9.50
<b>Grade Written English Text</b>	Credibility: 85 Credentials: Englis...		\$0.10
<b>Write Health Article</b>	Credibility: 40 Credentials: Writer...	Yes	\$6.75

FIGURE 2.9: The list of tasks available for the workers in CloudCrowd platform.

Figure 2.9 shows the workers' interface where they can review the list of available jobs along with the required credentials and the reward. Moreover, based on the complexity level of some tasks, a bonus can be granted for completing a task that requires particular specifications.

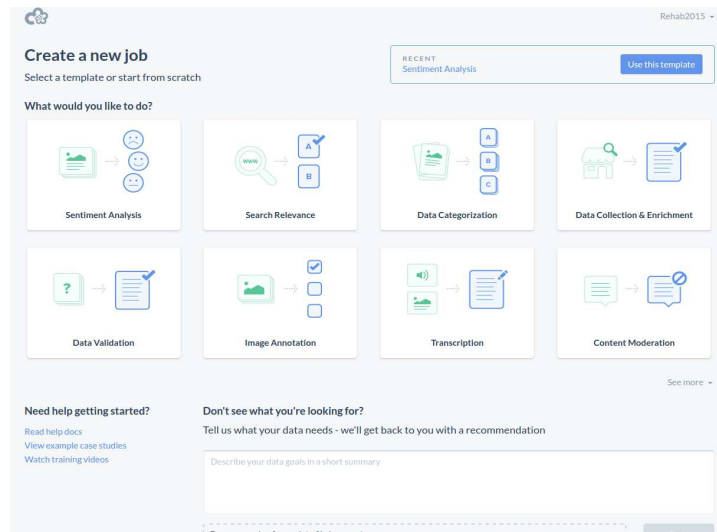
#### - Figure Eight (previously known as CrowdFlower)

The CrowdFlower platform was established in 2007 as a connection platform that distributes a task over more than fifty crowdsourcing channels. In April 2018, the company expanded the platform and changed its name to Figure Eight (F8)<sup>4</sup>. The same services continue to be provided

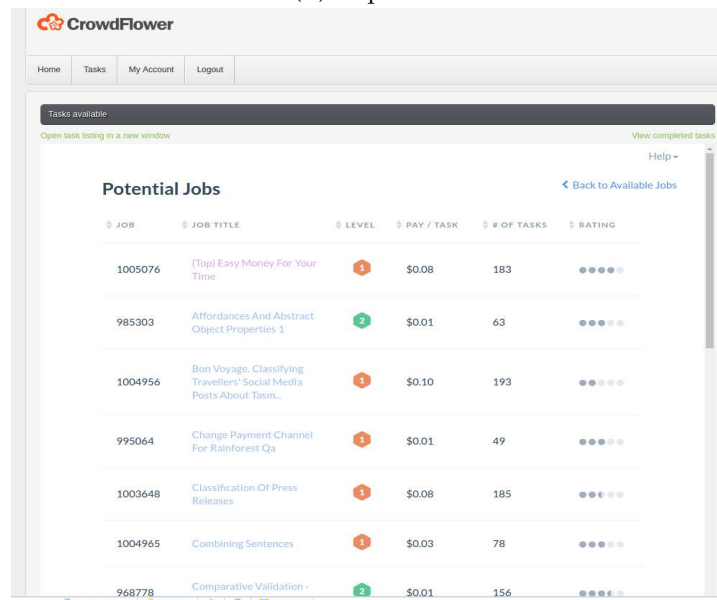
<sup>3</sup><http://www.cloudcrowd.uk.com/>

<sup>4</sup><https://www.figure-eight.com/>

to the requesters and workers, but the policy of the company is shifting toward human-in-the-loop and more integrated solutions.



(A) Requesters



(B) Workers

FIGURE 2.10: Requesters and workers interface in Figure Eight platform.

The mechanism of this platform is similar to that of MTurk: requesters can identify which workers can perform the task and F8 also provides a classification of three levels of workers according to their historical record of completed, accepted tasks. This platform has the ability to use *Gold units* which can be uploaded in the first stage of the design of the task. These units will help control the quality of the task by tracking the workers' performance and remove spammers during the launch time. Moreover, there is an aggregation method for the results which allows the requester to see them based on weighted majority voting, where the weights are related to past worker accuracy (Luz, Silva, and Novais, 2015). Figure 2.10 shows the templates of the job that appears in the requester interface and the workers' list of jobs.

### - Microworkers

The Microworkers platform<sup>5</sup> was launched in 2009 by the Weblabcenter company. A task on this platform can be designed using one of the built-in templates. The nature of the tasks on this platform is less complex than tasks provided by other platforms (Luz, Silva, and Novais, 2015). Unlike other platforms, there is no possibility to modify the instructions with HTML which is considered one of the limitations of this platform. Also, it is not possible to allow multiple units per job. With these restrictions and the simplicity of the tasks, there are no particular requirements for the workers who can perform the job Figure 2.11.



FIGURE 2.11: The mechanism of creating the task in Microworkers platform.

### - ShortTask

The ShortTask platform<sup>6</sup> is an online labour market which was released in July 2009 by "Career Mission", a company based in California. The mechanism of this platform is similar to MTurk in that a requester can create a job and assign multiple units of that job to the workers.

### - Prolific Academia

Prolific Academia<sup>7</sup> is one of the most recent crowdsourcing platforms, released in 2014 by research students from the University of Oxford and Sheffield (Peer et al., 2016). This platform specialises in providing tasks for academic research studies. The aim of Prolific Academia is to offer the ability for researchers to reach the required number of participants with the exact the demographic details.

<sup>5</sup><https://ttv.microworkers.com>

<sup>6</sup><http://www.shorttask.com/>

<sup>7</sup><https://www.prolific.ac>

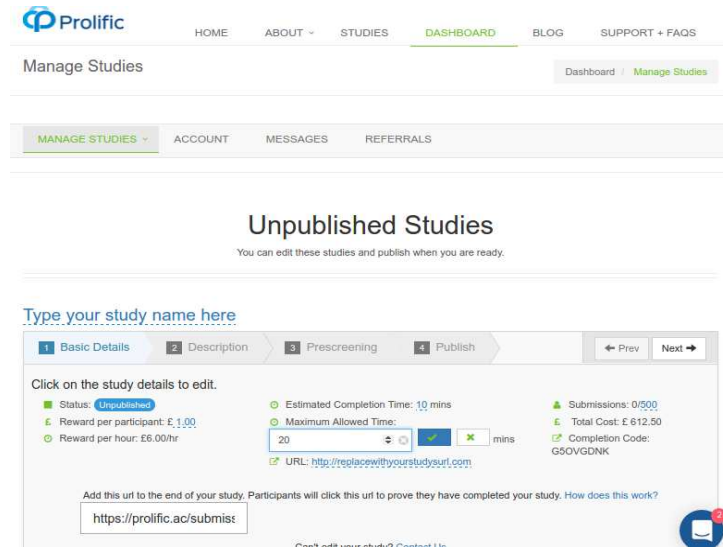


FIGURE 2.12: The mechanism of creating the task in Prolific Academia platform.

The design of the task (study) can be done using four steps as shown in Figure 2.12: (1) *Basic Details*, where the requester manages the task setup such as reward, the time allowed for completing each task, and a unique URL for proving that participants have completed the study. (2) *Description*, for writing and editing the task. (3) *Prescreening*, where the requester can choose the demographic criteria of the people who are eligible to complete the task. (4) *Publish*, which presents a preview of the task and enables the compatibility feature for displaying the task on different device interfaces. Similarly to MTurk, there is no aggregation method for the results on the Prolific Academia platform.

### 2.6.2 The evaluation of crowdsourcing platforms

Few papers in the past have comparatively evaluated the performance of different crowdsourcing platforms and highlighted the differences between them. A study by Crump, McDonnell, and Gureckis (2013) validates MTurk as a tool for collecting data in cognitive behavioural research. Having designed several types of experiments, the researchers performed them online and in a traditional lab setting. After receiving data from both conditions, their findings confirmed that the quality of the data collected under the experimental conditions in MTurk is extremely comparable to the quality of the data collected in the traditional lab-based way.

Bentley, Daskalova, and White (2017) presented a similar case study using three different methodologies to collect data (one traditional and two online surveys) for a study of user behaviours. They compared the quality of the results obtained with MTurk and SurveyMonkey to those obtained using a traditional paper-based survey. This study showed that the results obtained with MTurk are highly similar to, and are obtained much faster when compared to the traditional way of collecting survey data. Although there are some limitations in the technical and visual design of the crowdsourced task and some unexpected behaviours in the crowd (such as dropping out of a task before finishing it), collecting data with crowdsourcing is considered a fast and economic method that reaches a wide range of users within seconds (Crump,

McDonnell, and Gureckis, 2013).

In terms of comparing crowdsourcing platforms, Peer et al. (2016) provided a comparison study between the new platform Prolific Academic (ProA), F8, and MTurk. The results of this study recorded the highest response rate for participants in F8 and the highest data quality for participants in ProA and comparable to those from MTurk. In the same context, Mourelatos, Tzagarakis, and Dimara (2016) presented a ranking model for crowdsourcing platforms to collect data and compared the platforms over two periods of time. They also compared platforms according to: *type of service provided, quality and reliability, region, online imprint*. They discuss the impact of the platform characteristics on their traffic data and popularity (Mourelatos, Tzagarakis, and Dimara, 2016; Mourelatos, Frarakis, and Tzagarakis, 2017). This study showed a theoretical comparison between crowdsourcing platforms based on the Alexa<sup>8</sup> Ranking system, while in this research we present a data-driven comparative analysis based on running experiments on two platforms.

### 2.6.3 The consistency and reliability of platform results

Studying the consistency and reliability of crowdsourcing has previously been done differently than the approach proposed in this research employs. Williams et al. (2017) studied the consistency of results when crowd workers repeat the same task twice. They used a method where they duplicated a task in a queue of tasks presented to the same worker. This method examines the reliability and consistency of workers when completing duplicated tasks.

Blanco et al. (2011) presented an evaluation of repeatable and reliable data generated using crowdsourcing platforms. They investigated the creation of an evaluation dataset for a semantic search task using crowdsourcing. They used a sample of entity-bearing queries from the Yahoo! and Bing search engine logs to create a keyword query set to benchmark. This study experimentally proved that a crowdsourcing platform can produce scalable and reliable results over a single repetition after one month. Moreover, the quality of the results was comparable to that of expert-generated judgements even when repeating the same task over time. Following this work, Tonon, Demartini, and Cudré-Mauroux (2015) proposed a continuous Information Retrieval evaluation methodology using crowdsourcing to extend an existing benchmark dataset by using additional crowdsourcing tasks over time, assuming unvaried reliability of the collected data. Compared to this body of work, in this research we perform a longer-term analysis by means of data collected during a longitudinal study over different crowdsourcing platforms in Chapter 5.

### 2.6.4 Crowdsourcing approaches

The existing crowdsourcing platforms provide many different features and task design mechanisms. However, the procedures for implementing long or complex tasks are not easy for the requesters, and without proper support tools, they may create weak tasks that can lead to poor outcomes. With this in mind, several researchers tried to develop some approaches and tools to enhance the process of designing and maintaining the task.

---

<sup>8</sup><https://www.alexa.com/>

Each of these approaches presents a different perspective and provides various features. Some of these approaches have been used in the design process, while others have been developed to support the evaluation and aggregation of the results. Many of these approaches have been applied only to one platform or are useful in a specific area. Table 2.4 compiles the state of the art of these approaches and tools.

TABLE 2.4: Crowdsourcing approaches.

Approaches	Features Description
TurKit (Little et al., 2010)	Java-Script tool that allows requesters to program a task and pass it to MTurk platform.
CrowdFlow (Quinn et al., 2010)	Tool for tuning speed, cost and quality of the task.
CrowdDB (Franklin et al., 2011)	SQL extension that auto-generates task interface for unsolved queries.
CrowdForge (Kittur, Smus, and Kraut, 2011)	Framework used for systematic and dynamic break-down mechanism of complex tasks and controls the flow and dependencies between tasks.
Turkomatic (Kulkarni, Can, and Hartmann, 2011)	Tool for using the crowd in planning and designing complex tasks along with requesters.
Jabberwocky (Ahmad et al., 2011)	Software consisting of three levels of programming environment: (1) Dormouse, enable programming models for complex tasks along different platforms, (2) ManReduce, framework for parallel data flow for human and machine computation. (3) Dog, high-level procedure focuses on expressive and reuse of task.
AutoMan (Barowy, Berger, and Mcgregor, 2012)	Crowd-programming system used for automatic handling of scheduling, quality control and paying for the task.
CrowdLang (Minder and Bernstein, 2012)	Framework for designing complex tasks that involve large number of actors and data.
CrowdWeaver (Kittur et al., 2012)	Tool for managing complex tasks with the ability of real-time modification and support of reuse and data flow between tasks.
CrowdMAP (Sarasua, Simperl, and Noy, 2012)	Prototype for ontology alignment that uses crowdsource tasks to enhance the quality of the task solutions.
Turkopticon (Irani and Silberman, 2013)	A toolbar extension used for MTurk to help workers find information about the requesters whose giving a rate from other workers.
GATE (Bontcheva et al., 2014)	An open-source plugin the offer mapping documents and generate crowdsourcing task interface for NLP classification and selection tasks.
AskSheet (Quinn and Bederson, 2014)	Extension tool by Google used to implement spreadsheets for a task.
CrowdSearcher (Bozzon et al., 2014)	Framework for reusing and monitoring the data flow of complex tasks in crowd-based systems including crowdsource platforms and social media.
CrowdTruth (Inel et al., 2014)	Framework for creating ground truth data.
CrowdComputer (Tranquillini et al., 2015)	Tool for implementing flexible tasks .
ReTool (Chen et al., 2017)	Web-based tool for designing interactive interface for text and image tasks.



## 2.7 The Assignment and aggregation mechanisms of crowdsourcing tasks

When microtasks are launched on the crowdsourcing platform, workers can choose any available task to complete. The availability of the tasks to any particular worker may vary depending on some pre-execution quality control mechanisms that the requester set up before the tasks are released on the platform. From the requester's point of view it is crucial to assign the job to workers who have the required amount of skills to produce a high-quality result.

The crowdsourcing community gathers workers with different skills and a wide range of knowledge. For this reasons, the assignment mechanisms have been the focus of the attention of researchers looking for an optimal match between the workers and the microtasks.

Ambati, Vogel, and Carbonell (2011) used the workers' profiles in MTurk to identify their preferred tasks in order to recommend new tasks for them. Additionally, in Yuen, King, and Leung (2012) a recommendation framework has been proposed, based on previous successfully completed jobs along with the worker's search history for a particular type of task. These and many more methods found in the literature (Assadi, Hsu, and Jabbari, 2015; Karger, Oh, and Shah, 2011) have been called *Off-line* Assignment mechanisms in Chittilappilly, Chen, and Amer-Yahia (2016). The *Off-line* methods depend on the availability of the workers' records which could be missing from some crowdsourcing platforms, and will not be available for new workers.

On the other hand, the *On-line* mechanisms depend only on defining the requirements of the task in the design stage such as the number of workers, time to solve the problem, and the amount of money to pay, and workers will be filtered by post-execution quality control methods (Yuen, King, and Leung, 2015; Ho, Jabbari, and Vaughan, 2013; Zheng et al., 2015; Boim et al., 2012; Mao et al., 2015).

The last stage is the aggregation mechanism, where the microtasks' outcomes will be collected and aggregated to produce the overall result. Post-execution quality control methods will be applied at this stage to filter spammers and unproductive workers, which will lead to the elimination of low-quality results.

There are several studies of aggregation methods that have been used to represent the outcome of the crowdsourced tasks. Multiple studies surveyed the evaluation and the comparison of existing aggregation methods (Quoc Viet Hung et al., 2013; Venanzi et al., 2016; Chittilappilly, Chen, and Amer-Yahia, 2016). In addition, we evaluate the most recent published methods and present over 32 methods in this survey.

Table 2.5 summarises all these methods and the main features that are provided by each one. All of these methods classified according to the following main features: Type of iterative, type of task used for each method, and learning features (worker accuracy, worker confusion matrix, task difficulty, task duration, worker's type, and quiz question). (Quoc Viet Hung et al., 2013; Venanzi et al., 2016).

TABLE 2.5: Comparison of 32 aggregation methods for crowdsourcing task results.

	Type of Iter.		Type of task				Learning Features					
	non-Iterative	Iterative	Binary-labelling	Multi-labelling	Sentiment analysis	Image tagging	Worker accuracy	Worker confusion matrix	Task difficulty	Task duration	Worker type	Quiz question
<b>Offline Method</b>	DS -(Dawid and Skene, 1979)	●	●	●			●	●				
	GLAD -(Whitehill et al., 2009)		●	●			●		●		●	
	SLIM -(Raykar et al., 2009)		●	●								
	RY -(Raykar et al., 2010)		●	●			●					
	EM -(Ipeirotis, Provost, and Wang, 2010)		●		●			●			●	
	ELICE -(Khattak and Salleb-Aouissi, 2011)	●		●					●		●	●
	ITER -(Karger, Oh, and Shah, 2011)		●						●		●	●
	LDA -(Wang, Faridani, and Ipeirotis, 2011)		●							●		
	KJ -(Kajino and Kashima, 2011)		●	●	●			●			●	
	BCC -(Kim and Ghahramani, 2012)		●	●	●			●	●			
	MACE -(Hovy et al., 2013)		●	●			●			●		
	BLC -(Sheng, 2017)		●	●	●					●		
<b>Online Method</b>	Majority voting	●		●	●							
	CUBAM -(Welinder et al., 2010)	●		●			●				●	
	HP -(Lee, Caverlee, and Webb, 2010)	●										●
	YU -(Yan et al., 2010)		●	●	●			●				
	DARE -(Bachrach et al., 2012)		●	●	●			●		●		
	ZenCrowd -(Demartini, Difallah, and Cudré-Mauroux, 2012)		●	●				●				
	MinMaxEntropy -(Zhou et al., 2012)		●	●	●			●	●			●
	CDAS -(Liu et al., 2012)		●			●	●	●			●	
	MLNB -(Bragg et al., 2013)		●	●	●			●				
	BM -(Bi et al., 2014)		●	●				●		●		
	GP -(Rodrigues, Pereira, and Ribeiro, 2014)		●	●				●				
	LU -(Liu, Peng, and Ihler, 2012)		●	●				●				
	WM -(Li, Zhao, and Fuxman, 2014)		●	●	●			●				
	CBCC -(Venanzi et al., 2014)		●	●	●			●	●		●	
	APM -(Nushi et al., 2015)		●	●	●			●		●		
	BCCTime -(Venanzi et al., 2016)		●	●	●			●	●	●	●	
	RBAM -(Parde and Nielsen, 2017)		●		●			●	●		●	
	cBCMC -(Tam et al., 2017)		●		●						●	
	BMMB -(Wei, Zeng, and Yin, 2017)		●		●			●	●		●	
IDBLA -(Hong, 2017)		●		●				●	●			

Venanzi et al. (2016) presented the Time-Sensitive Bayesian Information Aggregation for Crowdsourcing Systems (BCCTime) method that predicts the reasonable duration for each task and identifies the spammers and cheaters who were too fast or too slow in completing the task.

Parde and Nielsen (2017) have developed one of the most recent methods that uses a learning regression-based model to aggregate labels with a wide range of quality and distribution. This model, which has been used in annotating NLP tasks, automatically discovers any bias and spammers by comparing the predicted labels from the model with non-expert annotators' labels. The presented approach worked efficiently with simple labelling tasks, while it requires further improvements to obtain a better result in more complex tasks.

This method and others presented in Table 2.5 were implemented to aggregate the outcomes of one or two types of crowdsourcing tasks and have been extensively tested on well-known datasets; However, generalising these methods to different types of tasks or different levels of complexity does not guarantee the same level of accuracy, nor does it guarantee unambiguous results. In Checco et al. (2017), the authors show that commons agreement measures used to assess the confidence of aggregation methods suffer from important problems and abnormalities, and they propose a novel measure based on probabilistic parameter estimation to mitigate such problems.

## 2.8 Chapter Summary

This chapter presented the state of the art in the development of the crowdsourcing concept. It also provided an overview of the classification of crowdsourcing systems and listed the previous approaches used along with those platforms. Moreover, it presented the importance of task design and the significant effect of task features on the overall outcomes and the researchers' effort to improve the workers' performance.

Still, there are some limitations in the studies carried out on task design. There is a large number of factors that could affect the quality of the task outcomes such as characteristics of the task, crowd motivations and requesters' needs. That is, existing design approaches could produce a high-quality result in some tasks and low-quality in another. A solution proposed by Allahbakhsh et al. (2013) is a recommender system which could help the requesters in defining the right task design with taking into consideration each requester's profile, past activities, and the task requirements.

Furthermore, this chapter looked at six factors of the task design that affect three main points of impact on the crowdworkers. These factors affect the quality of the crowdsourcing outcome. In summary, while several works looked at task types and task design strategies to improve the efficiency and effectiveness of crowdsourced data collection, this is the first work experimentally comparing different HIT ordering and balancing strategies to collect relevance judgements from crowd workers (Chapter 4).

Finally, we presented in this chapter an analysis of crowdsourcing platforms and show that these platforms are not flexible, due to their dependency on the features provided which vary

from one to another and the requesters cannot modify them based on any specific constraints they may want. We will study the consistency of crowdsourcing platforms by running a longitudinal study where we compare the reliability of results collected with repeated experiments over time and across crowdsourcing platforms (Chapter 5).

The design of some task types will be studied, with some focus on different kinds of tasks (classification tasks for images and documents) than the ones found in previous work and different levels of classification (binary, three classes and eight classes) as a starting point.

Next chapter will discuss the research methodology and represent the datasets and the selected platforms. Furthermore, it explains the basic formulas for the statistic measurements used in the analysis of the results display in the later chapters.





# 3

## Experiments general setup

### 3.1 Introduction

Chapter 2 presented an in-depth analysis of the previous studies in the field of task design and the researcher's contribution to the evaluation of crowdsourcing platforms. Throughout this thesis, the investigation is aimed at accuracy improvements due to interior technical features such as ordering and balance of HITs in a batch rather than improving individual GUI design, which can be applied orthogonally to our techniques.

Over three years of study, a data-driven approach was used - as described in Section 1.5 - to perform long-term analysis of a modern crowdsourcing task and evaluate the effects of different dimensions of the task design on workers' performance.

With relation to the main research question, we performed several crowdsourcing experiments on specific platforms and measured the variation in the results while changing some technical factors in the design of a specific task.

The experiments in this research were open to all workers with no constraints regarding possession of specific qualification or level of experience required from the workers. No personally identifying information was recorded. Participation was entirely voluntary, and workers were free to discontinue at any point. The task had been approved by the Ethics Committee of the The University of Sheffield, Appendix A presents the consent forms that were presented to the workers before they started the task.

## 3.2 Datasets used in the experiments

The datasets considered in this research had already been annotated by experts (acting as the gold standard) and that allowed us to compare our results with this gold standard to measure variation in performance.

As mentioned previously, the research focused on the type of classification task (images and documents) and different levels of classification (binary, three classes, and eight classes). We used three datasets in our crowdsourcing experiments, each with a different classification task. The reasons behind using these particular dataset were: (1) the accessibility to the full dataset, (2) the validity of gold standard data to compare the results with, and (3) the availability of clear documentation on how labels had been collected before.

### 3.2.1 Dataset 1

The first dataset (Dataset 1) was a collection of tweets gathered during a crisis/emergency situation (Imran, Mitra, and Castillo, 2016). This dataset was collected from 2013 to 2015 from 19 different crises and consists of around 52 million disaster-related messages.

Figure 3.1 shows tweet examples from crisis events along with details about the name, type, and country of the crisis as well as the total size of collected tweets. The annotation schema used was to categorise each tweet content into one of the following nine possible categories:

1. *Injured or dead people*: Reports of casualties and/or injured people due to the crisis.
2. *Missing, trapped, or found people*: Reports and/or questions about missing or found people.
3. *Displaced people and evacuations*: People who have relocated due to the crisis, even for a short time (includes evacuations).
4. *Infrastructure and utilities damage*: Reports of damaged buildings, roads, bridges, or utilities/services interrupted or restored.
5. *Donation needs or offers or volunteering services*: Reports of urgent needs or donations of shelter and/or supplies such as food, water, clothing, money, medical supplies or blood; and volunteering services.
6. *Caution and advice*: Reports of warnings issued or lifted, guidance and tips.
7. *Sympathy and emotional support*: Prayers, thoughts, and emotional support.
8. *Other useful information*: Other useful information that helps understand the situation.
9. *Not related or irrelevant*: Unrelated to the situation or irrelevant.





FIGURE 3.1: Dataset 1 examples include crisis name, type, country, size of collection, tweet text, and category

### 3.2.2 Dataset 2

The second dataset (Dataset 2) was a collection of product reviews related to fashion items accompanied by item images (Chernushenko et al., 2018). The size of this dataset is around 2.3 million text reviews collected in eleven different languages for several kinds of fashion items (e.g. shoes, tops, trousers etc.). The review of each item is displayed with a caption and a number of images that show the item from different angles as shown in Figure 3.2.

Crowd workers in this task were asked to identify the issue described in each product review and classify it into one of three aspects (size, fit, or 'other issue'). For the "Size" issues, the comment is expressing feedback about the item's size. When the sentiment is negative, the item's size is either too large or too small compared to the regular one. Description of size can be labels M, L, XL and numbers for apparel; numbers for shoes (43, 44,..); children sizes usually relate to age.

For the "Fit" issues, the comment is expressing feedback about the item's fit, but it does not specify a size issue. The problem could be related to comfort with regards to the fit.

For "No issue with size or fit", the comments are not related to sizing or fit. Moreover, the comments about delivery of a wrong size or missing a size in the inventory are **not** considered a size issue.



FIGURE 3.2: Dataset 2 examples include title, review contents, and images of the fashion item

### 3.2.3 Dataset 3

The third dataset (Dataset 3) was a collection used in the Eighth Text Retrieval Conference<sup>1</sup> (TREC8) (Hawking et al., 2000) which contains documents, queries, and editorial relevance judgements<sup>2</sup> from a general web search.

From this dataset, We chose documents that have a similar length and reading difficulty level to avoid effects due to those dimensions. From this dataset, we sampled documents which had been classified as relevant or non-relevant to the topics by trained human assessors and also were judged in the same way by Sormunen (Sormunen, 2002). There are 2511 documents (85.64%) with the same judgment in TREC8 and Sormunen (0 → 0 and 1, 2, 3 → 1). Table 3.1 shows topics for which only documents with a coherent judgment were considered. We look at three topics (442, 421, and 428) from the collection as they are characterised by a high number of documents in both the relevant and the non-relevant class, giving us more flexibility in the design of different balance settings in the experiments.

Crowd workers were asked to read the search topic description and narrative before they classified documents as relevant or non-relevant to the given topic, (Figure 3.3).

<sup>1</sup><http://trec.nist.gov>

<sup>2</sup>Assessors are human judges hired and trained by NIST.

TABLE 3.1: TREC8 topic with coherent judgments with Sormunen

Topic no.	No. of Relevant	No. of Non-relevant	Total
Topic 418	115	75	140
Topic 421	79	59	138
Topic 428	77	73	150
Topic 431	63	65	128
Topic 440	54	85	139
Topic 442	83	108	191

**Topic 421**

**Title:** Industrial waste disposal

**Description:** How is the disposal of industrial waste being accomplished by industrial management throughout the world?

**Narrative:** Documents that discuss the disposal, storage, or management of industrial waste both standard and hazardous are relevant. However, documents that discuss disposal or storage of nuclear or radioactive waste, or the illegal shipment or dumping of waste to avoid legal disposal methods are not relevant.

Document 1

Britain, under pressure from neighboring countries to improve its pollution practices, agreed to join other North Sea nations in steps to clean up the sea. Britain, the only nation bordering the North Sea that still dumps sewage sludge and industrial wastes into the sea, agreed with other nations at a conference in The Hague, Netherlands, to reduce by 70% from 1985 levels the amount of cadmium, mercury, dioxin and lead it dumps into the sea by 1995.

Document 2

A Kern County waste disposal firm was fined \$150,000 by state Health Director Ken Kizer and ordered to correct alleged toxic pollution violations at its Buttonwillow facilities. Department of Health Services inspectors said Petroleum Waste Inc. failed to implement a required emergency contingency plan following the exposure of three employees to hazardous wastes, failed to submit a written report on the incident to the state and inspectors noted what they called poor laboratory management practices. Kizer said most violations have been corrected.

**Classifications:**

● Relevant
 ● Non-relevant

FIGURE 3.3: Example of Dataset 3 Topic 421 including the topic title, description, narrative and documents

### 3.3 Platforms used in the study

After studying the most common crowdsourcing platforms as presented in Section 2.6, we chose two popular commercial crowdsourcing platforms which have been used for data evaluation and acquisition in industry and academic research: Amazon Mechanical Turk (MTurk) and FigureEight (F8) — previously known as CrowdFlower —, thus we hope that the contribution of this research will be applicable and useful for other academic and researchers in the same field. These were used to perform the experiments pertaining to this research.

### 3.4 The selection of crowdsourcing task type

Manual labelling and human annotation were used to create a corpus and datasets as this kind of task was one of the most common crowdsourcing tasks. As discussed in Section 2.4.3, researchers recommended using a clear and straightforward type of task for measuring the effectiveness of task design. Observing this recommendation makes it possible to compare the variations in the results every time certain factors in the design are changed. The type of task that was used in this part of research is a classification task. This task was chosen not only for being a straightforward one, but also based on the possibility of reproducing the same task GUI.

For this research, the first group of experiments evaluated relevance judgements for documents

from one of the information retrieval systems, that had already been labelled by human experts (Dataset 3). These were evaluated with different settings of order and balance (RQ1 - RQ3) of binary classes in the batch; more details in Chapter 4. After running several batches of these experiments, we observed for the same setup of the batch a variation in worker population and stability on the results, such that when using the same configuration and different topics, the accuracy of the results was not consistent. This led us to RQ4 and RQ5 to examine to what extent the design of the crowdsourcing task can be repeatable and reproducible.

The second group of experiments expanded on our experiments to include the examination of multiple classifications of images (Dataset 2) and tweets (Dataset 1) to examine RQ6 and perform a longitudinal comparison for the same task design by repeating it over time and on two platforms; more details in Chapter 5. During the repetition of the task design, we examined the effect of a motivation bonus, and we observed some significant changes in the workers' performance. Moreover, we received complaints from some workers about low payment, and we found that they were getting less than the set-up payment. With an in-depth analysis, we found that in using F8, the beta channels that workers used were cutting a commission at a rate which was high and varied from one channel to another. This led us to (RQ7), that is the third group of experiments, which included collecting demographic data from past experiments performed on a crowdsourcing platform for the last four years. Additionally, we performed a quantitative analysis by surveying workers about their experience working for the platform and the amount of payment they receive from the used channels.

### 3.5 Evaluation metrics

A variety of evaluation methods in human computation have been developed over the past decade, each of which serves a different purpose. Nielsen (1993) devised four different usability inspection methods to evaluate user interface design: *automatically, empirically, formally, and informally*. Combinations of these methods and others were used to assess our work. The usability of the design of a crowdsourcing task is based on the workers' performance so *empirical* or *User-based Evaluation* methods were used to assess the design by implementing controlled experiments with variations in the technical factors and compare the results of each different design version.

#### **Performance measures: 1- Accuracy**

To assess workers performance, we used statistical measures that have been used with binary as well as with multi-classification crowdsourcing tasks. One such measure is *Accuracy*. This measure is used to calculate the score of workers prediction labels when matched to the gold standard ones. Accuracy is also reported as the rate of correct predictions to the *error rate*, such that  $accuracy = 1.0 - error\ rate$  (Kakas et al., 2011). In a crowdsourcing classification task, accuracy is calculated according to (3.1), using Confusion matrix where we mapped the Actual class (gold standard) with the predicted class (crowd workers labels) as shown in Table 3.2.

TABLE 3.2: Confusion matrix for positive and negative labels

		Worker prediction	
		Positive	Negative
Actual Class (Gold Standard)	Positive	TP	FN
	Negative	FP	TN

Where *True Positive* (TP) is the count of items correctly labelled by workers as positive classes and, *True Negative* (TN) is the count of items correctly labelled by workers as negative classes. On the other hand, *False Positive* (FP) is the count of items incorrectly labelled by workers as positive and *False Negative* (FN) is the count of items incorrectly labelled by workers as negative classes. With accordance to the Confusion matrix, the Accuracy score was calculated as the proportion of true annotated items (both true positives and true negatives) among the total number of all items examined (Metz, 1978; Kohavi and Provost, 1998).

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (3.1)$$

## 2- Positive Predictive Value (PPV) and Negative Predictive Value (NPV)

In addition, for a more concrete evaluation, we used Positive Predictive Value (PPV), that is the ratio of positive results that are true positives (also known as Precision)( 3.2), and Negative Predictive Value (NPV), that is the ratio of negative results that are true negatives, as compared to gold-standard data( 3.3) (Kohavi and Provost, 1998).

$$PositivePredictiveValue(PPV) = \frac{TP}{(TP + FP)} \quad (3.2)$$

$$NegativePredictiveValue(NPV) = \frac{TN}{(TN + FN)} \quad (3.3)$$

## Krippendorff's alpha ( $\alpha$ ) coefficient

The inter-annotator agreement between workers and completion time rate for each worker and for all the batches were compared with different experiment setups for different tasks. We performed several statistical analyses - such as Krippendorff's alpha ( $\alpha$ ), Kendall's  $\tau$  correlation, and Two-way ANCOVA - on these measures to evaluate and formulate the findings for each experiment. Following Krippendorff (2011), to measure the inter-rater reliability coefficient of the workers' agreement, we used Krippendorff's alpha ( $\alpha$ ) general form:

$$\alpha = 1 - \frac{D_o}{D_c} \quad (3.4)$$

Where  $D_o$  is the recorded disagreement among workers, calculated as follows:

$$D_o = \frac{1}{n} \sum_c \sum_k o_{ck\text{metric}} \delta_{ck}^2 \quad (3.5)$$

and  $D_e$  is the expected disagreement, calculated as follows:

$$D_e = \frac{1}{n(n-1)} \sum_c \sum_k n_c \cdot n_{k\text{metric}} \delta_{ck}^2 \quad (3.6)$$

The arguments in the two disagreement measures,  $O_{ck}$ ,  $n_c$ ,  $n_k$  and  $n$  refer to the frequencies of values in coincidence matrices. The value of  $\alpha$  will be one of the following:

- $\alpha = 1$  refers to perfect reliability when workers agree perfectly  $D_o = 0$ .
- $\alpha = 0$  refers to lack of reliability when workers agree exactly as the expected  $D_o = D_e$ .
- $1 > \alpha > 0$  refers to the level of reliability.

### Kendall's tau $\tau$ correlation coefficient

Kendall's  $\tau$  was used to measure the correlation between different setting batches (Kendall, 1990) as follows:

$$\tau = \frac{n_c - n_d}{n(n-1)/2} \quad (3.7)$$

Where  $n$  is the number of items to compare, and  $c, d$  refer to different batches. The coefficient  $\tau$  must be in range  $-1 \leq \tau \leq 1$ , where the correlation between the two ranking groups is near perfect if the coefficient nears 1.

### Analysis of covariance Two-way ANCOVA

An Analysis of covariance (ANCOVA) tests the interaction effect between two or more independent variables based on a continuous response variable (dependent variable). We used two-way ANCOVA to test the significance of differences among group means of two levels of independent variables on a dependent variable (i.e. accuracy) (Keppel, 1973), which was calculated as:

$$Y_{ab} = \mu_a + \beta (X_{ab} - \bar{X}_{..}) + \epsilon_{ij} \quad (3.8)$$

Where  $Y_{ab}$  is the  $b^{\text{th}}$  observation in the  $a^{\text{th}}$  group,  $\mu_a$  represents the true mean of the  $a^{\text{th}}$  group effect, the  $\bar{X}_{..}$  represents the overall means of  $X$ , and  $\epsilon_{ij}$  are the Residuals or errors.

### Relative percentage change

We used relative change ratio to compare the changes between two conditions and by multiplying this ratio by 100 we got the percentage. Hence, the relative percentage change was calculated as:

$$RelativeChange(x, y) = \frac{\Delta}{y} \quad (3.9)$$

$$Relativepercentagechange = RelativeChange * 100 \quad (3.10)$$

Where  $\Delta = x - y$  is the difference between two conditions. By using the Relative percentage change the results can be easily observed, as an increase from one condition to the other will result in a positive value while a decrease - in a negative value. (Törnqvist, Vartia, and Vartia, 1985).

### NASA Task Load Index (NASA-TLX) assessment

Additionally, to measure total task workload, at the end of the task workers were asked to answer a one-page questionnaire about their perceived performance, using the NASA-TLX assessment tool (Hart and Staveland, 1988). This allowed us to measure batch design effects on the task complexity. This questionnaire consisted of six subjective questions described as : *Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration*. Appendix B shows a screenshot of the NASA-TLX questionnaire that was given to the workers.

## 3.6 Chapter Summary

This chapter presented the general setup for all the experiments that have been carried out in this research. To achieve the goal of this research, we performed a set of controlled experiments on three different datasets. The datasets were described and illustrated with examples, as were the criteria for selecting the task types. Furthermore, all the mathematical and statistical measurements used to analyse the results of all the experiments in the study were detailed in this chapter.

The following chapters will address the sub-research questions and describe the setup of each set of experiments in detail, along with an analysis of the results and the main findings that will shape the answer for the main research question.





# 4

## Batch ordering and balance for relevance judgement task

### 4.1 Introduction

After studying the state of the crowdsourcing concept and presented an overview of the classification of crowdsourcing systems, in chapter 3, we displayed the methodological methods that will be used starting from this chapter. In this research, we aim to enhance the design of the task to improve workers' performance. The first set of experiments will conduct an analysis focusing on workers' performance to examine **(RQ1)**, **(RQ2)**, and **(RQ3)** which consider the evaluation of existing dataset by design the order and balance of the labels in the task in a particular way.

The evaluation of Information Retrieval (IR) systems is based on the degree of relevance search results have with respect to a search query. To measure the performance of such systems, relevance judgments are created by human assessors. Crowdsourcing relevance judgements have become a popular approach to scale the creation of IR evaluation collection. Studies have shown how crowdsourced judgements lead to similar results to standard Text Retrieval Conference (TREC) assessments (Alonso and Mizzaro, 2012; Maddalena et al., 2017) and that evaluation initiatives built on top of crowdsourced collections are reliable and repeatable (Blanco et al., 2011).

Crowdsourcing is often used to scale-out the collection of manual labels that are then used, for example, to train machine learning models or to create large IR evaluation collections. In such cases, it often happens that the frequency of the classes to be labelled in the data is unbalanced:

For example, there are typically few relevant documents as compared to the number of non-relevant ones in a judgment pool or there are few fMRI images which are positive for a certain disease as compared to the negative cases.

Several factors that influence the accuracy of human judgments have been studied in the past (Eisenberg and Barry, 1988; Park, 1993; Clemmensen and Borlund, 2016). It is also well known that class imbalance has negative effects on training supervised machine learning models. This creates problems such as biasing the model towards the class which is most frequent in the training data (Ali, Shamsuddin, and Ralescu, 2015). In this chapter, we apply priming effects by presenting certain data items first thus introducing workers with examples of the classes to be labelled in the batch of HITs.

In our work, compared to Cai, Iqbal, and Teevan (2016) work discussed in Section 2.4.4, we instead focus on the impact of ordering tasks in a batch on worker effectiveness. We also focus on a different task type, that is, relevance judgements rather than creative tasks. We also investigate the effect of class imbalance on user completion time. Related work in this area has looked at how limiting available task time influences relevance judgements quality in crowdsourcing setting (Maddalena, Basaldella, and Innocenti, 2016).

Our hypothesis is that similar effects may be present in crowdsourcing where class imbalance situations may bias workers in the way they assign labels to data items. We also hypothesise that presenting workers with instances of a certain class first can help them label data more accurately later in the batch of tasks.

As presented in Section 1.3, in this chapter, we report results towards the investigation of the following questions:

- **RQ1:** Do class imbalance and order in a batch of Human Intelligence Tasks (HITs) affect the performance of crowd workers involved in the creation of manually labelled datasets?
- **RQ2:** How should crowdsourcing tasks be split into microtasks between workers? What is an appropriate length of a task?
- **RQ3:** Does providing a training test question or an example improve workers' answers?

To answer these questions, we run comparative experiments on a popular commercial crowdsourcing platform where we measure judgment quality and work efficiency for different class distribution settings both including class balance (e.g., one dominant class) as well as ordering (e.g., positive cases preceding negative ones).

Our results show that being able to *train* workers with the positive class (i.e., showing them the items they are looking for) yields to significantly better quality judgments and reach similar conclusions to those performed by experts in Scholer, Turpin, and Sanderson (2011), (see Section 2.4.1). Opposite to Damessie and Culpepper (2016) results' (see Section 2.4.4), our results show that there is a positive effect in presenting relevant documents first to train assessors with positive examples. We additionally expand on this by examining different task lengths and different ordering and balance settings.

We additionally show that workers tend to agree more and be faster when many not-relevant documents are present in the batch and when relevant documents are presented early in the batch.

The main contributions of this chapter are the identification of order and class balance factors that affect crowdsourcing relevance judgment quality and a set of recommendations for crowdsourcing relevance judgment design best practices. Our results show significant effects across different topics of task order and class balance on the quality of the relevance judgments collected by means of crowdsourcing. Ordering HITs based on document retrieval rank rather than at random may allow crowd workers to encounter relevant documents early in the judgment batch thus leading to better quality IR evaluation collections.

The rest of the chapter is structured as follows. First, we present our hypotheses on the effect of ordering HITs in a batch and of different class balance situations on worker behaviours. Next, we present our experimental results measuring such effect in different conditions both looking at order and balance effects. Finally, we conclude by summarising our main findings and lessons learned that can inform the design of future crowdsourcing relevance judgement experiments.

## 4.2 Research Hypotheses

In this section we present our hypotheses on 1) how datasets with different class balance situations can generate biases in crowd answers and on 2) how the order in which HITs are presented to crowd workers may impact the results collected back from a crowdsourcing platform.

We discuss this in the context of a binary classification problem (e.g., positive/negative sentiment classification or relevant/non-relevant judgments), but it can be easily generalised to multi-class classification problems.

### 4.2.1 Class Balance in a HIT Batch

First, we claim that (similarly to what happens when training machine learning models) there is a bias effect on crowd workers in an imbalanced class situation where the batch of HITs they complete contains significantly more data points of one class as compared to the other class. For example, out of 50 relevance judgment HITs, it is common to encounter 40-45 non-relevant documents and just a few relevant ones. For this case, we define the following hypothesis:

**H1** A class imbalance situation (many HITs of the same class in a batch) will bias worker labeling behavior tending to favor the **dominant** class because of a *developed habit* of labeling instances from such class.

On the other hand, it would also be valid to assume that **H1** does not hold in the case of an *implicit expectation* of crowd workers to find a similar number of instances from the two classes.

### 4.2.2 Ordering HITs in a HIT Batch

We also claim that the order in which HITs of a homogeneous batch (e.g., of relevance judgment tasks) are presented to workers has an effect on their performance.

Our hypotheses in this case are the following:

**H2** The order in which workers complete HITs has an effect on their training and performance (i.e., both efficiency and effectiveness).

**H3** We can *train* workers on distinguishing between classes (e.g., relevant or not-relevant) by presenting them first with certain instances from the dataset, improving the overall judgment quality.

While crowd performance may depend on many factors including worker background, experience, and intrinsic motivation, in this chapter we run several controlled experiments to test our hypotheses and draw conclusions that can help address the research questions (RQ1, RQ2, and RQ3) presented in Section 1.3, and design better crowdsourced relevance judgment experiments and create higher quality IR evaluation collections. In this chapter we verify which of these hypotheses hold in a crowdsourcing relevance judgment setting and what this means in terms of designing such HITs.

## 4.3 Design of the experiments

### 4.3.1 Dataset

In this study, we used Dataset 3 which was described in Section 3.2.

Crowd workers participating in our experiments are asked to read the topic description and narrative before they can start the task, similarly to expert assessors having previously judged the same documents. Appendix C shows GUI for the task design instructions and examples of the questions.

All experiments were performed on the Figure Eight platform (F8). As a quality control mechanism, we discarded answers from crowd workers who completed a 10-judgments batch in less than 3 minutes (this threshold has been selected based on a pilot study where participants took an average of 15 to 20 minutes to perform the 10-judgments task).

### 4.3.2 Participants

We collected 550 (Experiment 1), 750 (Experiment 2), and 80 (Experiment 3) judgments on Figure Eight to complete a 10-30 documents task that included a sequence of documents to be judged as relevant or not. Compensation was computed based on the expected HIT time duration at around \$8.00 per hour.

We provided crowd workers with a brief description of the topic they had to judge documents for, and some guidance to help them recognize whether a document is relevant to the topic or not, similarly to the procedure for TREC assessors.

In Experiment 1, we asked 50 workers to judge 10 documents in a batch (as Relevant or not Relevant to specific topic), where each batch had a different ratio (10% - 100%) of relevant documents. In Experiment 2, for each batch represented in Figure 4.2, we asked 50 workers to judge 10 documents in one of the pre-defined orders. In Experiment 3 we changed the batch length (i.e., more judgment tasks to be completed). In this experiment, we tested our hypothesis with an extended batch to see if the results hold when changing one of the experimental setups. For this longer batch we adapted the reward to keep the hourly rate equivalent to Experiment 1 and 2.

To prevent memory bias, each worker was allowed to participate in only one of the experiments. For all three experiments, all batches with the same setup were run at the same time. To avoid bias caused by judging documents from different topics in one task interface as mentioned in Eickhoff (2018), in each batch we asked workers to judge documents on the same topic.

#### 4.4 Experiment 1: Class Imbalance

Experiment 1 (E1) is an assessment of the sole effect of class imbalance on relevance judgment performance, when the effect of the HIT ordering in the batch is removed. Each participant is asked to perform a sequential judgment of 10 documents. We vary the class ratio from 10% to 100% relevant, in steps of 10%, thus obtaining 10 batches of HITs with different ratios of relevant documents.

To remove the effect of the ordering in which relevant and non relevant documents appear to workers, from the set of all possible orderings we select the ten that differ most in terms of adjacent swaps between a relevant and a non-relevant document (as crowdsourcing all possible orderings for each balance situation would be infeasible). We then let 5 workers judge documents for each batch, thus performing a total of 550 judging experiments, each of which consists of 10 sequential judgment tasks.

##### 4.4.1 Results and Discussion

We start by looking at judgment accuracy (measured against TREC judgments) to test H1 and to study the differences between the two relevance classes.

As shown in Figure 4.1a, we observe an increase in accuracy as the ratio of relevant documents in the HIT batch increases. Moreover, the two classes are clearly asymmetrical: a high ratio of relevant documents leads to higher accuracy as compared to the case with same ratio of relevant and non-relevant documents (Figure 4.1).

We can thus conclude that crowd workers are more prone to error in unbalanced (low number of relevant documents) HIT batches showing similar challenges to what Machine Learning algorithms face when trained over unbalanced datasets (Figure 4.1). However, when looking at inter-assessor agreement among crowd workers judging the same documents, in such situations workers tend to agree more (Figure 4.9a). To confirm these intuitions, we perform a linear

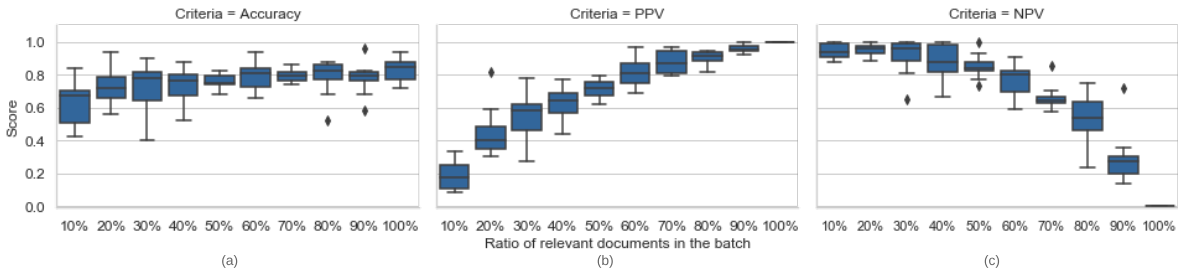


FIGURE 4.1: Judgement Accuracy, Positive Predictive Value (PPV), and Negative Predictive Value (NPV) box-plot for E1. The horizontal line in the middle of each box represents the median value. The x-axis label indicates the ratio of relevant documents in the batch.

regression, that confirms the effect of the balance on the accuracy ( $b = 0.011, t(99) = 21.19, p < 10^{-4}, R^2 = 0.819$ ).

With respect to H1, we can conclude that a balanced batch does not necessarily lead to higher judgement quality.

### 4.5 Experiment 2: Class Imbalance and Order

Experiment 2 (E2) focuses on the effect of class imbalance and document ordering on more realistic relevance judgment scenarios.

Regarding class imbalance, we used two different relevant/non-relevant ratios in a batch of judging tasks. The first is composed of 10% relevant and 90% non-relevant documents (i.e., batch 1 and 2). The second is composed of 50% relevant and 50% non-relevant documents (i.e., batch 3-5). Regarding class ordering, we use two different configurations, each one characterized by a different order of the two classes in the batch (i.e., relevant documents first, as in batch 1 and 3, or non-relevant documents first, as in batch 2 and 4), as shown in Figure 4.2.

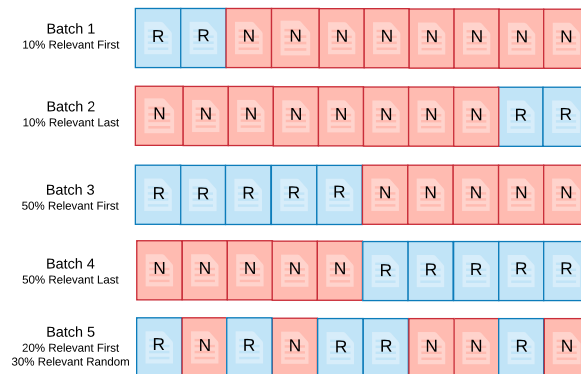


FIGURE 4.2: Order and balance of document classes for experiments 2 and 3 (blue for relevant and red for non-relevant).

We additionally tested a fourth ordering composed of 20% of the relevant documents at the beginning followed by a random ordering of the remaining 30% of relevant and 50% non-relevant ones (i.e., batch 5). This ordering mimics a real setting in which few editorial relevance judgments may be available (e.g., to be used as quality check to validate workers' answers) and thus allows the experimenter to put a few relevant documents first followed by all the other unjudged documents to be manually assessed (i.e., the latter 80% randomized documents). We test whether this setting can be used as a surrogate (in terms of quality of assessment) as compared to the 'relevant first' setting.

#### 4.5.1 Results and Discussion

For the second experiment, we analyze the judgment quality for each of the settings, as shown in Figure 4.3. The trend we observe is that showing relevant results first improves accuracy and PPV (especially for balanced batches) and, in general, a more balanced ratio between relevant and non relevant documents improves PPV and accuracy.

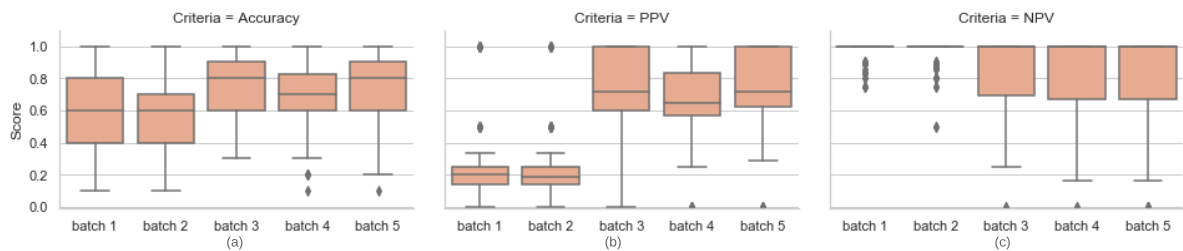


FIGURE 4.3: Judgement Accuracy, Positive Predictive Value (PPV), and Negative Predictive Value (NPV) box-plot for E2. The horizontal line in the middle of each box represents the median value.

To validate these observations we first perform a two-tailed ANOVA test on the PPV with the two factors being the order ('relevant first', 'relevant last') and the balance (10%-90% and 50%-50%). The results show that both balance and order affect PPV scores ( $F(1,499)=510.8, p < 10^{-78}$  and  $F(1,499)=5.65, p = 0.01$ ). This test does not include batch 5 as we applied this order only to the balanced setting (because the imbalanced one does not contain enough relevant documents) and we will thus include this in a post-hoc analysis below.

##### **Class imbalance.**

As shown by the statistical test above, highly balanced batches obtain higher PPV as compared to unbalanced ones ( $F(1,499)=510.8, p < 10^{-78}$ ). This effect is consistent with results from E1 showing that having more relevant documents in a batch leads to higher accuracy.

##### **Class ordering.**

As a post-hoc analysis (with FDR correction for multiple tests) we investigate whether 'relevant first' does, indeed, lead to a better PPV. Thus, we perform two one-tailed t-tests, one for each balance setting (i.e., 10%-90% and 50%-50%). For the unbalanced setting the results are

not statistically significant, while for the balanced case, indeed, crowd judgement accuracy is higher when relevant results are presented first in the batch, with an effect size of  $d = 0.04$  and  $p = 0.002 < 0.05$ .

### Relevant first (Batch 5).

We did not include the batch 5 setting in the aforementioned two-way ANOVA test as this ordering approach cannot be applied to a very imbalanced setting (because of not having enough relevant documents). We thus perform, for the balanced case, a one-way ANOVA on the PPV and order ('relevant first', 'relevant last', and 'batch 5'), that confirms that PPV is affected by the order ( $F(2, 381) = 5.45$ ,  $p = 0.004 < 0.05$ ). We then perform a post-hoc Tukey HSD test (included in the FDR correction used above), that confirms that 'relevant first' and 'batch 5' are significantly different than 'relevant last' ( $p = 0.03 < 0.05$ ,  $p = 0.008 < 0.05$ ), while 'relevant first' and 'batch 5' are not significantly different. This result corroborates the intuition that batch 5 can be used as surrogate for batch 4, when it is not possible to put all relevant documents at the beginning of the batch as the relevant labels are still unknown.

In conclusion, showing a small portion the relevant documents first (20% of the whole dataset) is enough to obtain an increase in performance that is statistically indistinguishable from the case in which all relevant documents are shown first.

### The effect in different topics.

Figure 4.4 shows a breakdown over three topics of the results presented above and depicted in Figure 4.3.

The results indicate only minor variations across topics, especially looking at judgment accuracy and PPV. For example, across topics we observed similar PPV (e.g., for batch 1 and 2) and similar accuracy values.

While for some topics scores are lower (e.g., PPV values for topic 421 tend to be lower than those obtained for the other two topics) we observe consistent performance in terms of batch ordering and class balance settings.

### The evaluation of different batches

Since multiple workers judged the same batch of documents, we build *document rankings* using, as a score, the sum of relevant judgments minus the sum of non-relevant judgments given by workers who judged that document. This technique generates a ranking score for each document based on the number of distinct relevance judgments for that document.

Table 4.1 shows Kendall's  $\tau$  correlation values between the document rankings generated by different batch settings in E2 (p-values obtained after FDR correction).

Kendall's  $\tau$  correlation for similar balance and different order of the classes in the batches (batches 1&2, batches 3&4, batches 3&5, and batches 4&5) indicates that, as could be expected, the batches with lower Kendall's  $\tau$  correlation are the ones in which the order is reversed



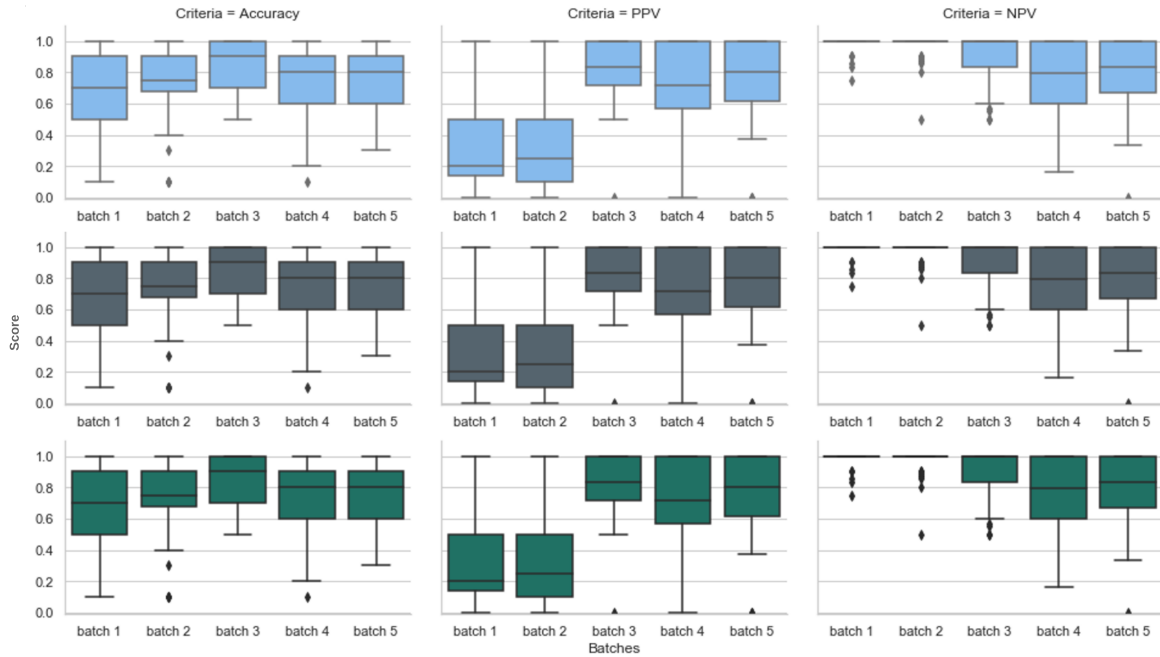


FIGURE 4.4: Judgement Accuracy, Positive Predictive Value (PPV), and Negative Predictive Value (NPV) box-plot for Topic 442 (light blue, top plots), Topic 421 (gray, middle plots), Topic 428 (dark green, bottom plots). The horizontal line in the middle of each box represents the median value.

(batches 1&2 and batches 3&4), with an average Kendall’s  $\tau$  below 0.67, while the average  $\tau$  for the rest of the batches is above 0.73 ( $p < 0.05$ ).

Moreover, the direct comparison of the correlation coefficients of batches 1&2 and batches 3&4 indicates that priming might have a slightly stronger effect on the ranking when the number of relevant documents is low, as batches 1&2 have the lowest average  $\tau$ .

TABLE 4.1: Kendall’s  $\tau$  correlation for pairs of order and balance settings over different topics (with FDR-corrected p-values).

		batch 1 & 2	batch 3 & 4	batch 3 & 5	batch 4 & 5
Topic 442	$\tau$	0.5	0.7	0.7	0.8
	$p$	0.0436	0.0057	0.0057	0.0028
Topic 421	$\tau$	0.9	0.6	0.9	0.8
	$p$	0.00001	0.0023	0.0003	0.0006
Topic 428	$\tau$	0.5	0.7	0.6	0.7
	$p$	0.0103	0.0039	0.0046	0.0039

## 4.6 Experiment 3: Batch Size

As a follow-up study to assess the effect of the number of documents judged by a worker (RQ2), in Experiment 3 (E3) we repeated E2 increasing the number of documents from 10 to 30 documents per batch. We asked 20 workers to judge documents for batches where order differs most, that is, batches 1, 2, 3 and 4. This allows us to compare with E2 where only 10 documents

were used. We are then able to observe whether a longer batch (30 vs 10 HITs) displays similar effects of imbalance bias and priming, or not.

#### 4.6.1 Results and Discussion

From the results (shown in Figure 4.5) we can see two opposing trends: for unbalanced batches (1 and 2), increasing the number of documents improves the performance, whereas for balanced batches (3 and 4) the trend is inverted.

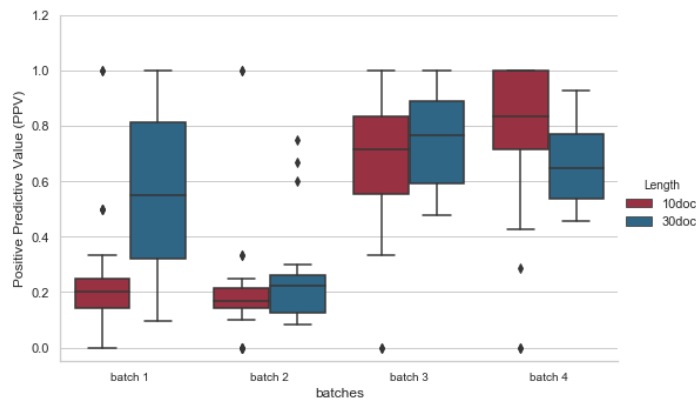


FIGURE 4.5: Comparing Length of the task, Mean judgement PPV for each setting in Experiment 2 and 3.

From Experiment 1 and 3, we can conclude that the more relevant documents workers encounter, the more accurate work they provide. Again, this is in line with the proposed approach of priming workers towards the positive class.

In unbalanced situations, the longer the batch, the better the workers' performance (Figure 4.5 batch 1 and 2). This is in line with related work looking at how breaking down complex crowdsourcing tasks helps improve worker accuracy (Cheng et al., 2015). On the one hand, the increased temporal demand and fatigue can lead to a reduction in judgment quality, but, on the other hand, the possibility of seeing more *rare class* cases can counteract this negative effect in the case of heavily unbalanced batches.

## 4.7 Analysis of The Worker Experience

### 4.7.1 Perceived Workload

For all the experiments we run, no significant differences have been observed in the workload perceived by workers across the different settings. In Figure 4.6 (top) we show the result of the NASA-TLX questionnaire for E1. We can observe that, as more relevant documents are present in the batch, frustration tends to decrease together with an increased perceived performance.

In Figure 4.6 (bottom) we show the result of the NASA-TLX questionnaire for E2. While the differences between the scores are not statistically significant, we can observe that the maximum

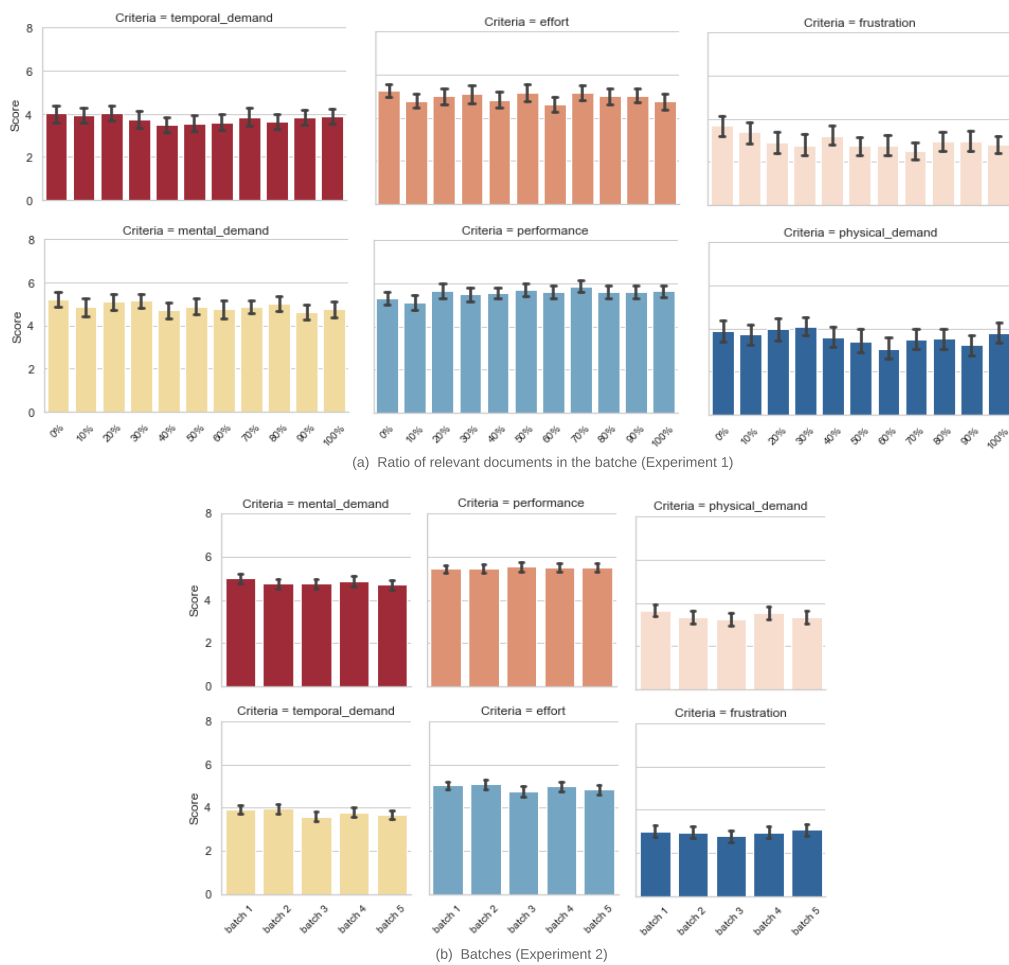


FIGURE 4.6: Perceived workload using the NASA-TLX assessment tool for each setting in Experiment 1 and Experiment 2.

perceived performance and minimum effort were observed for batch 3 (50% of relevant documents shown first, followed by 50% of non-relevant). The corresponding more realistic version of it (batch 5) shows the lowest level of frustration and, together with the results of Figure 4.3, corroborates the idea that it is a suitable candidate for a re-balancing technique to maximize performance without affecting the assessor’s perceived workload. Similar results are observed for E3 where batch length has no significant impact on perceived judgment complexity.

The effort required to complete the HIT batch is not affected by the class balance or by the order of items presented to workers (Figure 4.6). This is a positive result that allows us to re-order HITs in a batch without impacting on the crowd worker experience.

#### 4.7.2 The Effect of Document Position on judgement Quality and Time

Since workers completed HITs in sequence, we also analysed the effect of the HIT position on their performance, regardless of the class balance setting. In this way we can answer questions like, for example: is the judgment accuracy of the first document appearing in a batch different than the judgment accuracy of the document in the last position? Even if the differences in

judgment accuracy were not statistically significant, we noticed that documents presented first in a batch have the lowest accuracy showing a possible learning effect of workers getting into a new batch (Figure 4.7). This finding is consistent with previous work (Maddalena, Basaldella, and Innocenti, 2016).

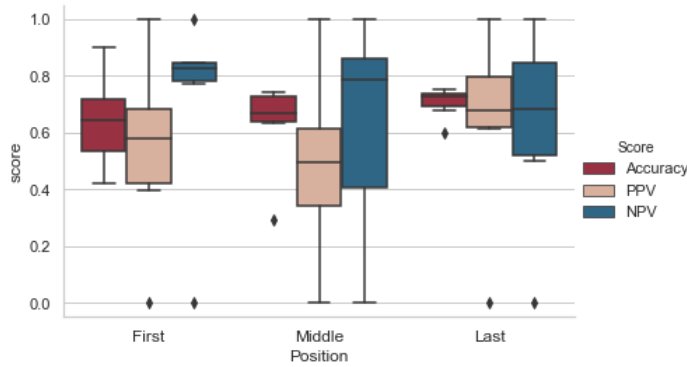


FIGURE 4.7: Mean Accuracy, PPV, and NPV of the documents in the first, second and last position in a batch over all the experiments.

On average, the first document being judged in a batch shows lower accuracy levels. More interestingly, documents in the first position of the batch show high precision and low NPV values: When the first document is relevant, workers tend to be very accurate while when it is non-relevant, workers make more mistakes. This supports even further the ‘batch 5’ alternative in E2, that is, to include in the first positions documents known to be relevant from editorial judgments: This will both train workers on relevance as well as allow for training. We also observed that the position of the document to be judged does not affect the completion time in a significant way for any of the batches

### 4.7.3 Completion Time

We analysed the relationship between judgment quality and HITs completion time for Experiment 1, 2 and 3. For Experiment 1, we found that workers that spent between 3 - 5 minutes working on the experiment had a low accuracy. Similarly, for Experiment 2, the majority of the workers who spent between 500 and 1800 seconds on the experiment had an accuracy between 0.6 and 1.

Figure 4.8 shows the average completion time for all batches considered in E1 and E2 compared to PPV values. We can observe that in E1 completion time shows no clear pattern as compared to balance and order settings. In E2, fastest completion time was achieved in balanced batches (30-50%). Comparing time with judgment effectiveness, we can see no strong correlation of PPV with the average completion time. We conclude that while introducing lower bounds in task completion time allows to filter out workers who randomly judge relevance, in general,

completion time is not a sufficient indicator of judgment quality: a result in agreement with previous work (e.g., (Cai, Iqbal, and Teevan, 2016)).

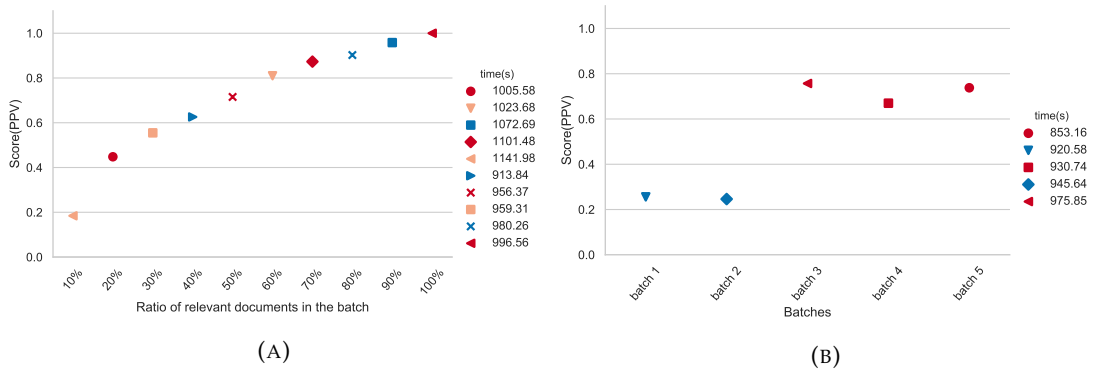


FIGURE 4.8: Median PPV vs. completion time for each batch in Experiment 1 (a) and Experiment 2 (b).

#### 4.7.4 Effect on Agreement

Since different workers have been judging the same documents in the same order and balance conditions, we are also able to measure the effect of document order and class balance on assessor agreement across experimental settings.

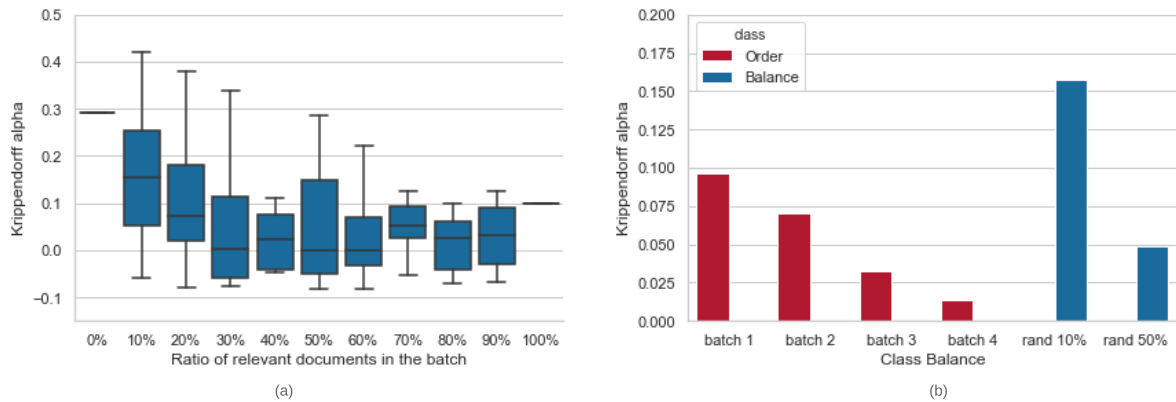


FIGURE 4.9: (a) Krippendorff's alpha for all batches in Experiment 1 (horizontal line for median value). (b) Krippendorff's alpha for batches in Experiment 1 (blue) with different balance classes, and Experiment 2 (red) with different ordering of documents.

Figure 4.9-a shows Krippendorff's alpha scores computed in different class balance situations (E1). We can observe that inter-annotator agreement scores tend to be higher when fewer relevant documents are present in the batch of tasks (which is the most realistic setting). Lowest agreement levels are observed around 50% balance levels.

Figure 4.9-b shows average assessor agreement levels computed on documents appearing at the beginning or at the end of a batch. We can observe that higher worker agreement levels are

observed when relevant documents are presented first and when few relevant documents are present in the batch (10% vs 50%) consistently with Figure 4.9-a.

## 4.8 Discussion

Regarding class balance we formulated hypothesis **H1**. Given our results, we have observed a more convoluted situation: We can conclude that an unbalanced situation with **many** instances of the positive (i.e., relevant) class leads to good quality results, especially when such positive instances are presented early in the batch as a form of *priming*. Based on our results, we can conclude that **H2** holds with respect to worker *effectiveness*; on the other hand, order has no significant effect on worker *efficiency*. With respect to **H3**, we can conclude that showing relevant documents first in a batch leads to a priming effect that leads to more accurate judgments across the entire batch.

While our experimental results do not directly generalise to non-binary classification problems, we expect that a similar priming effect toward the minority class would be present similarly in non-binary situations. That is, by showing instances from all different classes at the beginning of a batch or by starting with one of the minority class (like we did with relevant documents) should lead to more accurate labels.

Thus, when running a relevance judgment HIT batch we can leverage *gold* questions (to be used, for example, as a quality check) early in the batch. In the absence of gold labels, we can decide to order HITs by decreasing likelihood of relevance (e.g., based on the document ranking generated by the IR systems contributing to the judgment pool).

## 4.9 Chapter summary

In this chapter, we looked at the bias effect of class imbalance in a batch of crowdsourced relevance judgment tasks. For our experimental design, we focused on binary classification (i.e., binary relevance judgments) tasks and on how imbalance and order affect worker performance.

We observed that in the cases in which the number of relevant and non-relevant documents is approximately the same (i.e., balanced classes), crowd workers perform better when the relevant ones are presented first. An analysis of document rank correlation corroborates this observation: priming workers has a significant effect on the resulting ranking, and this effect is consistent within TREC topics.

Similarly, inter-annotator agreement is higher when relevant documents are shown at the beginning of the batch. This is a positive result which can be applied to real IR evaluation settings, e.g., based on pooling documents retrieved by different IR systems. While in a real setting it is not possible to put relevant documents before non-relevant ones as their relevance label is unknown, it would still be possible to order documents by attributes indicating their relevance (e.g., retrieval rank, number of IR systems retrieving the document, etc.) thus presenting first to the workers the documents with higher probability of being relevant, similarly to Damessie et al. (2018) but keeping our configuration for the document ordering.

Moreover, we found that showing a small portion of the relevant documents first (20% of the whole dataset) is enough to obtain an increase in performance that is statistically indistinguishable from the case in which all relevant documents are shown first. In the cases where few gold relevance judgments are available, it is possible to use this technique to train workers on what a relevant document looks like thus enabling them to make comparisons when looking at the subsequent documents in the batch.

We state that ordering tasks in an appropriate manner can be useful to increase crowd worker performance in unbalanced label situations. This confirms previous results showing that inter-task effects can be leveraged to increase outcome quality for image labeling tasks (Newell and Ruths, 2016).

A different way to deal with the class imbalance problem in crowdsourced labeling tasks would be to perform an activity similar to over-sampling for supervised learning training: When a dataset to be labeled is known to be unbalanced, it is possible to introduce additional (possibly artificial) data points to re-balance the dataset. To make such additional tasks useful, they can be exploited as gold questions with known answers to check for worker reliability.

Furthermore, in this chapter we compared our results with human expert assessors who labelled for these documents more than ten years ago. Researchers pointed that replicating tasks over time is needed to measure the consistency in human-based experiments (Thimbleby et al., 2011). We assume that the level of expertise and knowledge of the crowd workers has changed over time compared to the experts and for this reason and others we are asking the question: are the results coming from TREC8 still reliable? Is it possible to get the same results if we repeat the task on different workers? And what if we reproduce the same task on a different platform? These questions will be answered in Chapter 5.





# 5

## Repeatability and Reproducibility of Crowdsourcing Classification Tasks

### 5.1 Introduction

The rise of several crowdsourcing platforms has enabled the collection of human labels at scale. Researchers using such platforms (as requesters) aim to obtain reliable, repeatable, and reproducible results from the crowd, as required by scientific best practice.

In a crowdsourcing setting, we adapt these standard definitions in scientific experimentation as follows:

- *Reliable* results are obtained when the crowdsourced data shows a high level of accuracy compared to gold-standard data or according to other quality measures like, for example, inter-annotator agreement. Using quality control mechanisms to obtain reliable results is identified as one of the main challenges in crowdsourcing (Kittur, Nickerson, and Bernstein, 2013; Assis Neto and Santos, 2018).
- *Repeatable* results are obtained when holding consistency after repeating the same experiment multiple times. In Wilson et al. (2013) authors refer to it as “*Conceptual Replication*”, a common form of replication in Human Computer Interaction (HCI) where a study is to be replicated with alternative methods to confirm its finding. Prior work in human assessment research showed inconsistency when repeating the same experiment over time, revealing the need for new approaches when assessing repetitive tasks (Harter, 1996). In Paritosh (2012) he use of thresholds on Krippendorff’s alpha values is suggested as a

form of consistency measurement for human computation tasks. However, it has been argued in later studies that this measure may be not appropriate for crowdsourcing (Checco et al., 2017). While previous work has addressed this issue by providing guidelines for requesters (Paritosh, 2012), these guidelines are not sufficient to assess workers performance (Waterhouse, 2013).

- *Reproducible* results are obtained when consistent observations can be made across different crowdsourcing platforms. Previous studies (Campo et al., 2018; Blohm et al., 2018; Mourelatos, Frarakis, and Tzagarakis, 2017; Kohler, 2018; Peer et al., 2017) have discussed output variability across crowdsourcing platforms by studying external and internal factors affecting it, as shown in Section 2.6.2; Nevertheless, reproducing the results for identical tasks over multiple platforms has not been previously explored.

Previous studies in machine learning (Rosten, Porter, and Drummond, 2010) and human-computer interaction (Thimbleby et al., 2011; Hornbæk et al., 2014) used reliability as a measure of consistency. In the crowdsourcing field, a limited number of studies have examined result consistency (Blanco et al., 2011; Sun and Stolee, 2016; Bentley, Daskalova, and White, 2017; Cheng et al., 2015). Thus, many questions still need to be addressed:

1. Does an experiment on the same platform give different result quality levels when repeating the same task over the same dataset?
2. Is it possible to obtain the same result quality level when the same task is launched on different platforms (and thus with potentially different crowds)?

In this chapter, we present the first experimental study showing how crowdsourcing results are more or less consistent with such requirements of scientific research. We execute a longitudinal experiment over time and across different crowdsourcing platforms (i.e., MTurk and FigureEight) showing how the result reliability significantly changes across platforms (thus not resulting in reproducible experiments), while repeating experiments on the same platform produces consistent results.

This work is the first to address the reproducibility of a crowdsourcing task on different platforms in a rigorous and controlled manner (by ensuring identical user experience on different platforms). Moreover, the time scale used in this work (weeks) is novel as compared to previous work, and allows obtaining useful insights on using crowdsourcing for tasks that require a continuous, regular polling of the crowd over time. Another important novel contribution of this work is the uncovering of the fundamental effect of the payment scheme on the reproducibility of the results. The aim of this study is to reach an understanding of what the best strategies are in designing a crowdsourcing task and to advise crowdsourcing experimenters on the best way to achieve reliable results from the platforms they use.

The rest of the chapter is organised as follows. Section 5.2 presents our research questions and summarises the contributions of our work. Section 5.3 introduces our methodology, the dataset used in the experiment, task design, and the pilot experiments that validate our design and determine the sample size for the main experiments. Section 5.4 presents our experimental

results and findings on obtaining repeatable results. Section 5.5 presents our experimental results and findings on achieving reproducible results. We conclude with a discussion on the implications of our findings and directions for future work in Section 5.6.

## 5.2 Research Questions and Novelty

In this chapter, we examine the following research sub-questions:

- **RQ4 - Repeatability:** Is there a significant difference in the quality of the results for the same task repeated on the same crowdsourcing platform at a different point in time?
- **RQ5 - Reproducibility:** Is there a significant difference in the quality of the results for the same task reproduced on a different platform?
- **RQ6 - Generalisability:** Are the results obtained consistent over different classification tasks?

To address **RQ4**, we repeated the same experiment over multiple weeks to measure the reliability and consistency of the results over time (i.e., repeatability). When addressing **RQ5**, to compare the quality of data obtained through different crowdsourcing platforms (i.e., reproducibility) we chose two popular commercial crowdsourcing platforms which have been used for data evaluation and acquisition in industry and academic research studies: Amazon Mechanical Turk (MTurk) and FigureEight (F8).

To generalise our findings (**RQ6**) we used the same task design as in Experiment 1 and 2 over three different classification tasks described in Section 5.3. We reproduced the experiment on both platforms and over five weeks.

Overall, we collected data from over 4500 unique workers over the timespan of a week for each run.

Our results have implications for AI researchers using crowdsourcing platforms to perform experiments and to collect datasets over time or across multiple platforms. We have observed:

- A high level of agreement between crowd workers and expert annotators for the dataset we used in our tasks. In other words, crowdsourced results are *reliable*;
- *Consistency* of results when repeating the same task once every week according to a within-platform analysis;
- Inconsistency in responses when reproducing the same task at the same time on different platforms. That is, crowdsourcing results are *not reproducible*.
- We notice consistent performance for each dataset and on each platform over multiple weeks.

### 5.3 Methodology

We performed the first experiment to address **RQ4** and **RQ5**. After analysing the results of Experiment 1, we observed a statistically significant difference in accuracy between the results collected from the two platforms. Thus, we constructed a hypothesis to explain this difference and designed a follow-up experiment (Experiment 2) to test it, as explained in Section 5.5.

Furthermore, to answer **RQ6** we repeated Experiments 1 and 2 on two additional datasets to assess the generalisability of our findings.

The crowdsourcing tasks have been launched on the two platforms, MTurk and F8, at the same time and day of the week and repeated five times (once a week). We strived to create the same setup on both platforms to produce results that are statistically comparable. For this reason, we avoided using any qualifications such as Master workers in MTurk which would not have a comparable qualification in F8.

#### 5.3.1 Dataset

For both Experiment 1 and 2, we used three different classification tasks with three different kinds of labelling: documents, tweets, and images. We used all the three datasets that were mentioned in Section 3.2, each with a different classification task and difficulty level.

#### 5.3.2 Task Design

The task consisted of one batch of 10 documents from Dataset 1 and 20 documents from Dataset 2 and Dataset 3, obtained by sampling uniformly at random from the datasets. Counter to the design of the task in chapter 4 where we used short and long batch, the number of documents in this experiments was selected to ensure each task could be finished in approximately 5-6 minutes.

The interface was designed to appear identical in both platforms, thus, we used an external server to host the task interface and visualised it into each platform using iframes. The only difference between the worker experience on the two platforms was the way the task preview was visualised and the way the workers could reach the task (e.g., with platform search functionalities). These variables might have an effect on both completion time and population selection bias. Appendices D, E, and F show an example of the GUI for the task design for Experiment 1 and 2 as it appeared to the workers on both platforms.

Crowd workers were rewarded according to US minimum wage rates (\$8 per hour) after internal tests to estimate the average task execution time. Since our focus was on the differences between platforms, we run a unique Human Intelligent Task (HIT) consisting of 20 individual judgements, that was functionally equivalent to 20 HITs, each with a single judgement. This design choice removes the confounding effects caused by the order of HITs being decided by the platform, by the fact that workers will typically complete a different number of HITs, and by other learning effects.

To ensure unbiased results, crowd workers in each platform were allowed to perform the task only once: after that, worker identifiers were not allowed to participate in future batches of the same task. It is important to notice that the goal here is to assess the variability of the workers' behaviour over time and across different populations, to achieve bounds on the reproducibility of tasks. Based on a recent study by Difallah, Filatova, and Ipeirotis (2018), the likelihood of having the same workers participating in future tasks is very low. However, this approach allowed us to assess all workers equally as they all had the same level of experience when completing the task. With regard to quality control, we also checked task completion time and removed workers who took less than 3 minutes (i.e., the 20<sup>th</sup> percentile over the entire experiment) to complete the task.

To reduce the effect of external information gathering on the classification task, we asked workers to base their judgement only on the content presented in the task, and we advised them not to access any of the URLs present in the data item; to encourage this behaviour, we made the URLs appear without hyperlinks.

### 5.3.3 Pilot Experiment and Sample Size

We ran a pilot experiment on both platforms to test the validity of the task design and to calculate the ideal sample size for the main experiment. The settings and the interface used in this experiment were the same as the ones that were later used in Experiment 1. Isaac and Michael (1995) and Hill (1998) suggested 10–30 participants for studies where the population size is unknown and influenced by many factors. For that reason, we used 30 participants per platform for the pilot experiment.

To calculate the sample size needed for our main experiments, we used equation 5.1, that allows to estimate the sample size when comparing the means of a continuous outcome variable in two independent populations Thompson, 2012.

$$n_i = 2 \left( \frac{Z_{1-\alpha/2} + Z_{1-\beta}}{ES} \right)^2 \quad (5.1)$$

where  $n_i$  is the sample size required on each platform  $i$ ,  $\alpha$  is the selected level of significance and  $Z_{1-\alpha/2}$  is the value from the standard normal distribution holding  $1 - \alpha/2$  below it,  $1 - \beta$  is the selected power and  $Z_{1-\beta}$  is the value from the standard normal distribution holding  $1 - \beta$  below it.  $ES$  is the effect size:

$$ES = \frac{|\mu_1 - \mu_2|}{\sigma} \quad (5.2)$$

According to the results of the pilot,  $ES = 0.29$ , and, to have 75% statistical power, the sample size needs to be  $n_i = 150$  workers on each platform  $i$  for each weekly run. Using this number of workers guarantees that we can have a statistically significant sample size to make an observation on repeatability and reproducibility, but it does not require requesters to use this sample size. Should this experiment observe similar results across time or platforms, the requester will then be able use a small number of workers confidently, knowing that the variability that will be obtained is statistically bounded over time. In other words, should the results from this

experiment indicate that crowdsourced classification tasks are repeatable and reproducible, a requester might confidently run longitudinal tasks over multiple platforms using a small number of workers.

## 5.4 Experiment 1 - Achieving Repeatability

For Experiment 1, we used the same task design as presented in Section 5.3.3. We launched the task on the same day of the week and at the same time of the day on each of the two platforms and repeated the same experiment five times (once every week). Each week, we had 150 different workers completing the tasks on each of the platforms.

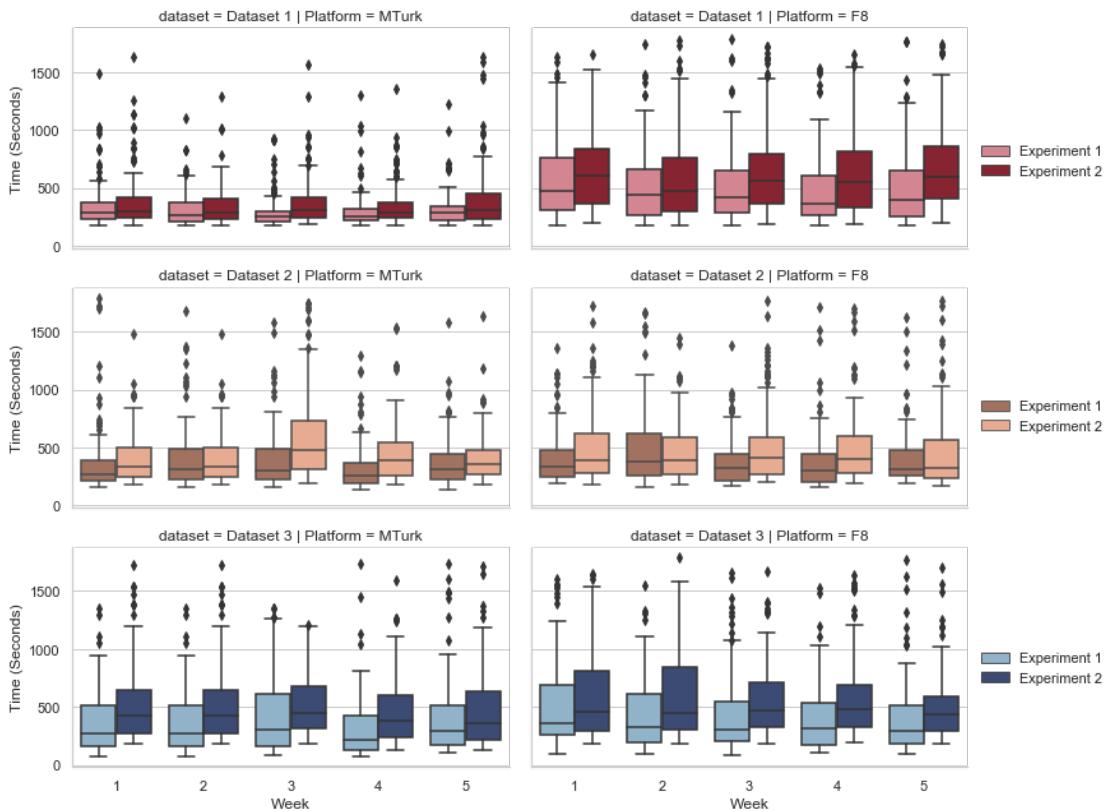


FIGURE 5.1: Average time per assignment for Experiment 1 and Experiment 2 for all 3 datasets.

For the three datasets we used in the experiment, the results show a high level of label quality consistency over the five repetitions. For Dataset 1, crowd workers in MTurk were individually faster than those in F8. MTurk workers took an average of 4 minutes to complete the task while it took approximately 6 minutes for workers in F8. For Dataset 2, each worker took an average of 5 minutes in MTurk and 4 minutes in F8 and similar results were observed for Dataset 3, as shown in Figure 5.1 and Table 5.1 (Average time per assignment).

Moreover, Figure 5.2 and Table 5.1 (Avg. accuracy) show the same consistency level in the distribution of the result accuracy over time on each platform and for each dataset. Overall, the accuracy of each run on MTurk was over 75% whereas on F8 it was in the 70% range for

TABLE 5.1: Results of five runs in MTurk and F8 for Experiment 1.

		Data 1		Data 2		Data 3	
		Mturk	F8	MTurk	F8	MTurk	F8
Average Time per Assignment	Week 1	4 m, 16 s	6 m, 09 s	5 m, 17 s	4 m, 50 s	6 m, 36 s	5 m, 10 s
	Week 2	4 m, 49 s	6 m, 33 s	5 m, 55 s	5 m, 16 s	5 m, 06 s	4 m, 24 s
	Week 3	4 m, 24 s	6 m, 18 s	5 m, 47 s	4 m, 29 s	5 m, 53 s	4 m, 15 s
	Week 4	4 m, 25 s	5 m, 30 s	4 m, 40 s	4 m, 20 s	4 m, 15 s	4 m, 17 s
	Week 5	4 m, 37 s	5 m, 49 s	5 m, 19 s	4 m, 46 s	5 m, 31 s	3 m, 59 s
Avg. Accuracy & Standard deviation	Week 1	0.73 ± 0.20	0.63 ± 0.28	0.71 ± 0.20	0.65 ± 0.20	0.72 ± 0.17	0.70 ± 0.17
	Week 2	0.76 ± 0.17	0.66 ± 0.25	0.67 ± 0.22	0.64 ± 0.18	0.69 ± 0.18	0.68 ± 0.19
	Week 3	0.76 ± 0.14	0.67 ± 0.25	0.64 ± 0.23	0.61 ± 0.21	0.71 ± 0.19	0.65 ± 0.17
	Week 4	0.74 ± 0.19	0.66 ± 0.27	0.58 ± 0.27	0.63 ± 0.20	0.70 ± 0.18	0.68 ± 0.19
	Week	0.76 ± 0.14	0.64 ± 0.28	0.68 ± 0.22	0.69 ± 0.16	0.73 ± 0.17	0.70 ± 0.17
Completion Time for the Batch	Week 1	72 h, 14 m	05 h, 11 m	14 h, 20 m	13 h, 22 m	151h, 01 m	54 h, 54 m
	Week 2	73 h, 29 m	04 h, 45 m	49 h, 37 m	13 h, 29 m	168 h, 02 m	64 h, 06 m
	Week 3	56 h, 36 m	07 h, 10 m	18 h, 16 m	15 h, 42 m	143 h, 57 m	60 h, 58 m
	Week 4	85 h, 54 m	04 h, 43 m	24 h, 30 m	28 h, 41 m	168 h, 00 m	25 h, 31 m
	Week 5	75 h, 28 m	04 h, 04 m	42 h, 20 m	50 h, 19 m	167 h, 55 m	66 h, 18 m

Dataset 1, over 60% on MTurk and 68% on F8 for Dataset 2, while for Dataset 3 the average accuracy was over 70% on MTurk and 65 % on F8.

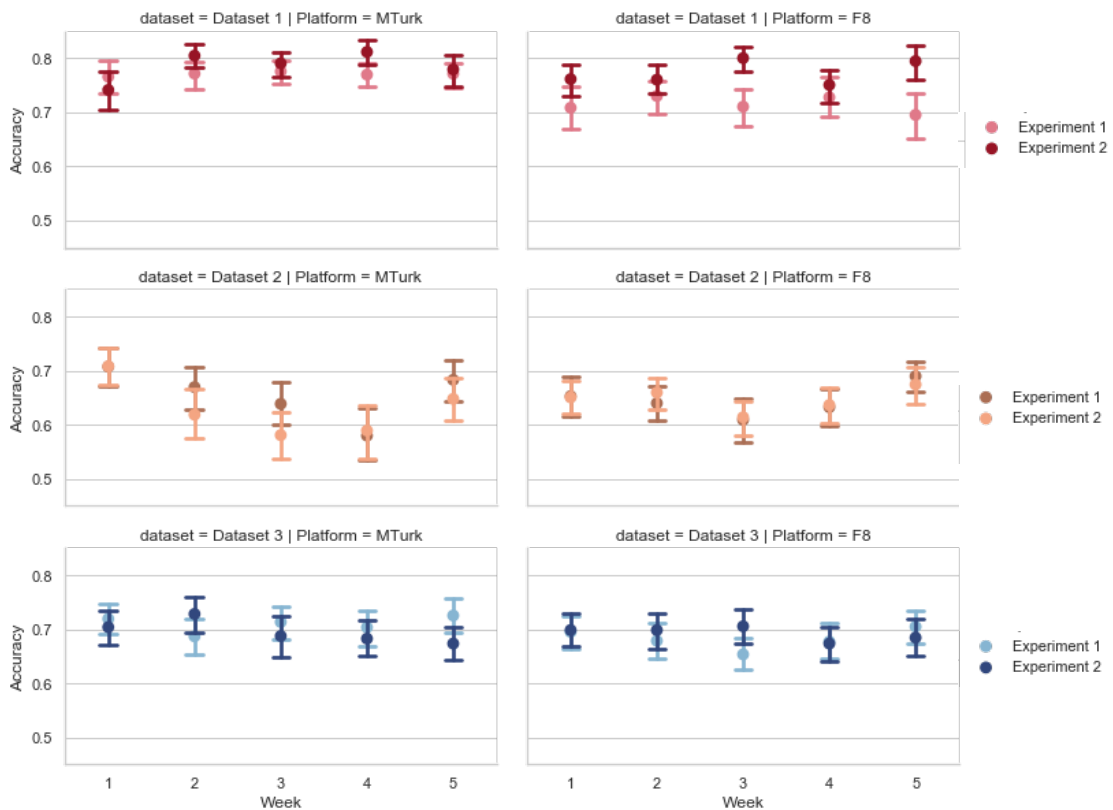


FIGURE 5.2: Accuracy distribution over time for Experiment 1 and Experiment 2 for all 3 dataset.

Dataset 1 shows a statistically significant difference in accuracy between the two platforms when using the default payment scheme (Experiment1)

We carried out a statistical analysis on accuracy as the dependent variable and studied two factors: *Week* and *Platform*. Consecutive repetitions of the same experiment are called Week 1-5. *Platform* refers to the two crowdsourcing platforms used to reproduce the experiment: MTurk and F8 (Tables 5.2-5.8). The effect of the platform on accuracy is statistically significant ( $p < 0.05$ ), while repetition effect and joint repetition-platform effects are not significant. This indicates the consistency of the outcome of each platform: we have successfully obtained the *repeatability* of the experiment, but we observe a problem of *reproducibility* over different platforms. These results are still statistically significant after Bonferroni-Holm (BH) correction over the whole set of experiments.

The total completion time (to obtain 150 results) for the entire batch was, on average, 3 days in MTurk and 4 to 7 hours in F8 for Dataset 1, 30 hours in MTurk and 23 hours in F8 for Dataset 2, and for Dataset 3 it took 6 days in MTurk and 2 days in F8, as shown in Figure 5.3 and Table 5.1 (Completion time for the batch).

We further investigate the reasons behind such differences in accuracy between the two platforms and in the long completion time in 5.5.

TABLE 5.2: Two-way ANCOVA for Dataset 1 in Experiment 1.

	sum_sq	df	F	PR(>F)
Platform	1.96	1.0	47.27	$9.6 \times 10^{-12}$
Week	0.00002	1.0	0.0006	$9.8 \times 10^{-1}$
Platform: Week	0.02	1.0	0.54	$4.6 \times 10^{-1}$
Residual	53.16	1283.0	NaN	NaN

After Bonferroni-Holm (BH) correction, only the effect of factor Platform is statistically significant ( $p^* = 1.15 \times 10^{-10}$ ).

TABLE 5.3: Two-way ANCOVA for Dataset 2 in Experiment 1.

	sum_sq	df	F	PR(>F)
Platform	0.36	1.0	0.78	0.37
Week	0.03	1.0	0.68	0.40
Platform: Week	0.25	1.0	5.62	0.02
Residual	54.00	1195	NaN	NaN

After BH correction, no factor has a statistically significant effect.

TABLE 5.4: Two-way ANCOVA for Dataset 3 in Experiment 1.

	sum_sq	df	F	PR(>F)
Platform	0.23	1.0	7.58	0.006
Week	0.008	1.0	0.26	0.60
Platform: Week	0.0008	1.0	0.02	0.87
Residual	36.64	1161.0	NaN	NaN

After BH correction, no factor has a statistically significant effect.

#### 5.4.1 Experiment 1 - Discussion

In Experiment 1, we observed a consistent superiority of MTurk over F8 in terms of accuracy. One potential explanation for this result is that the user interface of F8 explicitly shows whether a quality control system based on gold questions is being used or not. Moreover, workers in F8 get paid as soon as the task is completed (even if the quality is not satisfactory), while in



MTurk the requester has the option to reject and not pay for a task. Since we did not use any of the embedded quality control schemes provided by F8 (for better comparability across platforms), workers in F8 had access to that information, whereas the workers in MTurk did not. Additionally to that, F8 workers knew that completing a task would guarantee them the payment even if the quality of the provided labels were unsatisfactory. Based on these results, we can construct the following hypotheses:

**H1** Knowledge of the absence of a quality control scheme reduces crowd worker performance.

**H2** The potential for work rejection increases crowd worker performance.

To test these hypotheses, we designed a second experiment to equalise the conditions related to these two hypotheses on the two platforms, as explained in the next section.

## 5.5 Experiment 2 - Achieving Reproducibility

To equalise the conditions between platforms as described in Section 5.4.1, we adapted the task instructions by promising crowd workers that their submissions would not be rejected, and by offering a bonus to workers able to achieve at least 80% accuracy. This has two effects: 1) it motivates F8 crowd workers with the potential bonus (H1); 2) it reassures MTurk workers that no rejection would be performed (H2).

Workers on MTurk still recorded faster results than F8 workers (as in Experiment 1), completing tasks with an average time per assignment of 5–6 minutes, where the average in F8 was 7–9 minutes for Dataset 1, while for Dataset 2 and Dataset 3, there was no difference in the completion time observed for each task, as opposed to what was observed in Experiment 1 and 2 for Dataset 1. The same completion time of approximately 6 minutes was recorded for both platforms, as shown in Figure 5.1. This can be related to the level of content complexity as we discuss later in this Section.

The reasons why significant differences between platforms in completion time per single task for Dataset 1 were observed (as shown in Figure 5.1 and Table 5.5 (Average time per assignment)) could be related to language and demographics distribution of crowd workers on these platforms. The majority of workers on MTurk are based in the US Difallah, Filatova, and Ipeiro-tis, 2018 and as such they could be native English speakers and also more familiar with the data items present in the tasks, as the tweets are all in English and describe incidents that mostly happened or were discussed in the US. This may have led them to finish the task faster than workers in F8 who constitute a more demographically diverse group and may be from other countries around the world.

The modification that we introduced in the task instructions had a significant effect on the number of workers attracted to our task in MTurk: the completion time for the whole batch (which is related to how often workers would choose this task) is remarkably lower than the completion time for Experiment 1 for all 3 datasets on both platforms, as shown in Figure 5.3 and Table 5.5 (Completion time for the batch). This can be explained by the fact that the workers were reassured that they would receive a guaranteed payment for the time spent on

the task, reducing the uncertainty in payment. Even more importantly, the rejection uncertainty was also reduced with this payment scheme.

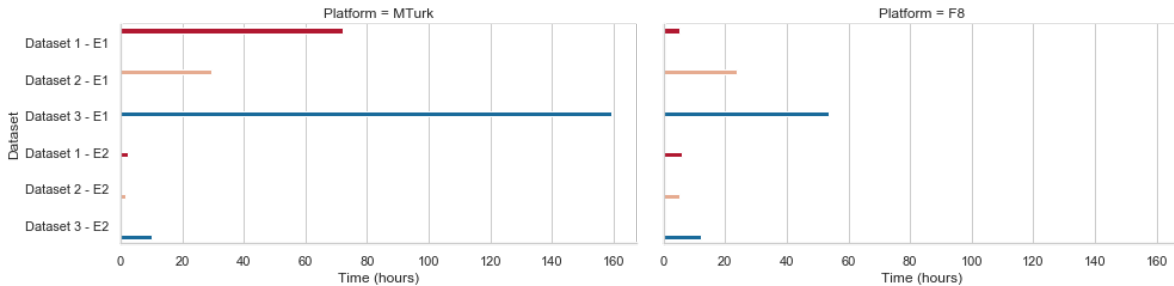


FIGURE 5.3: Average completion time for all batches in Experiment 1 and 2.

Despite the guaranteed payment, workers did not reduce their effort in completing the task: on the contrary, workers performed significantly better on difficult classification tasks (Dataset 1). The results from Experiment 2 show significant improvements in the performance on the F8 platform, Figure 5.2 and Table 5.5 (Avg. accuracy) show the distribution of the accuracy of the results over time on each platform and for each dataset. The average accuracy of each run on MTurk was over 80% and over 70% in F8 for Dataset 1, which shows some improvement compared to the results of Experiment 1.

The results for Dataset 2 and Dataset 3 recorded the same consistency in performance with repeating the task over multiple weeks as we had presented previously in Experiment 1 over various platforms with an overall accuracy of 65% for Dataset 2 and 70% for Dataset 3 on both platforms. After Bonferroni-Holm correction, we do not observe a statistically significant effect of the factors on accuracy.

TABLE 5.5: Results of five runs in MTurk and F8 for Experiment 2.

		Data 1		Data 2		Data 3	
		Mturk	F8	Mturk	F8	Mturk	F8
Average Time per Assignment	Week 1	5 m, 37 s	9 m, 00 s	5 m, 21 s	5 m, 43 s	5 m, 57 s	6 m, 24 s
	Week 2	5 m, 09 s	7 m, 46 s	5 m, 30 s	5 m, 56 s	6 m, 31 s	6 m, 13 s
	Week 3	5 m, 37 s	8 m, 54 s	8 m, 27 s	5 m, 44 s	6 m, 26 s	6 m, 12 s
	Week 4	5 m, 20 s	8 m, 27 s	6 m, 20 s	6 m, 16 s	6 m, 27 s	6 m, 43 s
	Week 5	6 m, 03 s	9 m, 13 s	6 m, 01 s	4 m, 38 s	6 m, 34 s	6 m, 08 s
Avg. Accuracy & Standard deviation	Week 1	0.71 ± 0.23	0.71 ± 0.25	0.71 ± 0.19	0.65 ± 0.17	0.70 ± 0.17	0.70 ± 0.17
	Week 2	0.77 ± 0.18	0.73 ± 0.21	0.62 ± 0.24	0.66 ± 0.18	0.73 ± 0.17	0.70 ± 0.18
	Week 3	0.78 ± 0.15	0.77 ± 0.21	0.58 ± 0.28	0.61 ± 0.20	0.69 ± 0.20	0.71 ± 0.18
	Week 4	0.80 ± 0.16	0.70 ± 0.25	0.59 ± 0.27	0.64 ± 0.20	0.68 ± 0.18	0.67 ± 0.18
	Week 5	0.76 ± 0.20	0.76 ± 0.24	0.65 ± 0.22	0.67 ± 0.20	0.67 ± 0.17	0.68 ± 0.19
Completion Time for the Batch	Week 1	01 h, 38 m	04 h, 45 m	03 h, 09 m	02 h, 29 m	12 h, 11 m	07 h, 08 m
	Week 2	03 h, 01 m	04 h, 33 m	01 h, 31 m	03 h, 26 m	05 h, 12 m	08 h, 32 m
	Week 3	02 h, 39 m	04 h, 46 m	01 h, 54 m	08 h, 11 m	15 h, 31 m	07 h, 14 m
	Week 4	02 h, 59 m	08 h, 54 m	01 h, 45 m	08 h, 48 m	10 h, 54 m	23 h, 02 m
	Week 5	03 h, 58 m	06 h, 45 m	02 h, 16 m	03 h, 02 m	08 h, 45 m	13 h, 55 m

Similarly to Experiment 1, a two-way ANCOVA was performed to analyse the effect of repeating the same task every week and reproducing it over two different platforms. Table 5.6 shows that none of the factors have a significant effect on accuracy, corroborating the idea that by taking into account the difference in payment schemes (being guided by H1 and H2) it is possible to achieve both repeatability and reproducibility (see Table 5.6, 5.7 and 5.8). It is important to notice that, differently than for Dataset 1, we did not observe a difference in accuracy between the two platforms for Datasets 2 and 3. This can be explained by the fact that Dataset 1 was obtained from a more difficult task, where the elements to be classified have 9 potential classes, and can also explain why this effect has not been observed in the past in the literature: Dataset 1 has an extreme correction by chance factor, by having 9 potential classes, and requires a higher cognitive effort than the other two datasets, where a quick glance at the text could be sufficient to allow an average quality classification level.

TABLE 5.6: Two-way ANCOVA for Dataset 1 in Experiment 2.

	sum_sq	df	F	PR(>F)
Platform	0.05	1.0	1.8	0.18
Week	0.12	1.0	4.7	0.03
Platform: Week	0.004	1.0	0.2	0.7
Residual	34.3	1298.0	NaN	NaN

After BH correction, no factor has a statistically significant effect.

TABLE 5.7: Two-way ANCOVA for Dataset 2 Experiment 2.

	sum_sq	df	F	PR(>F)
Platform	0.10	1.0	2.11	0.14
Week	0.09	1.0	1.99	0.15
Platform: Week	0.18	1.0	3.73	0.053
Residual	60.56	1260.0	NaN	NaN

After BH correction, no factor has a statistically significant effect.

TABLE 5.8: Two-way ANCOVA for Dataset 3 Experiment 2.

	sum_sq	df	F	PR(>F)
Platform	0.002	1.0	0.07	0.78
Week	0.143	1.0	4.52	0.03
Platform: Week	0.016	1.0	0.51	0.47
Residual	36.36	1143.0	NaN	NaN

After BH correction, no factor has a statistically significant effect.

### 5.5.1 Experiment 2 - Discussion

While the inability to reject the null hypothesis can be indicative of repeatability and reproducibility, it is important to consider that equivalence tests should be carried out to corroborate these findings (Parkhurst, 2001).

Despite H1 and H2 being potentially confounded by additional factors (like the motivation induced by the presence of a payment scheme), the findings suggest that H1 should be confirmed, while H2 should be rejected; reducing the uncertainty of being paid did not reduce quality: instead, it significantly increased the attractiveness of the task and, in turn, decreased

the batch completion time (these changes affected MTurk). On the other hand, letting the workers know that the quality is monitored, while guaranteeing a bonus for high quality results, has statistically increased the quality of the results for difficult tasks (these changes affected F8).

## 5.6 Chapter summary

In this chapter, we have looked at how crowdsourcing experiments can be repeatable and reproducible. Our findings show that:

1. **(RQ4)** it is possible to obtain *repeatable* experiments in each of the studied crowdsourcing platforms, but we have observed a problem of *reproducibility* over different platforms when the task is extremely difficult (Dataset 1);
2. **(RQ5)** by setting the same expectations on payment and rejection rate across different platforms, it is possible to achieve both *repeatability* and *reproducibility* of crowdsourcing results;
3. using standard crowdsourcing platform settings, the same data collection experiment may finish orders of magnitude faster on F8 as compared to MTurk, but with lower accuracy;
4. by using different datasets, we observed consistent results over time and over different platforms. After the equalisation of the payment scheme, our findings generalise over different classification tasks **(RQ6)**.

While the absence of quality control does reduce the labelling quality for difficult tasks, we can observe that the threat of unpaid rejection of a task does not increase its labelling quality, but reduces the attractiveness of the task and thus its overall completion time. On the other hand, introducing a bonus for high-quality labelling has a positive effect on the labelling quality.

Our future work will consider equivalence testing (Parkhurst, 2001) to corroborate our findings, investigate other realistic settings in terms of rejection and quality control, examine task *repeatability* and *reproducibility* on datasets with varying complexity levels, and consider other crowdsourcing tasks, and also consider additional crowdsourcing platforms in our comparative analysis.

After receiving a complaint from a worker about a low payment that was not the amount we set up for the task, we noticed a variation in the actual payment amounts the workers received on F8, especially for those who are using beta channels to access crowdsourcing tasks instead of using the F8 platform directly (Elite channel). This issue led us to investigate the payment scheme on the F8 platform and brought to light some issues related to the money transactions between different channels before the payment reaches the workers. In Chapter 6 we will present this study along with a demographic analysis of the workers who have been using F8 in the last four years and a survey crowdsourcing task on the platform to ask the workers using a specific channel about their experience and motivation for choosing this channel and task.





# 6

## Payment concerns in crowdsourcing platforms: A case study on Figure Eight channels

### 6.1 Introduction

In chapter 6, we compared the performance of MTurk and F8 platforms and observed that a motivational bonus allowed to achieve reproducibility and repeatability. The comparison of the performance between the two platforms raised payment concerns noticed when using different channels on F8 platform, which led us to develop (RQ7) presented in Section 1.3.

**RQ7:** Is the crowdsourcing platform transparently communicating the fee payment with the workers? Does the amount of payment affect workers' performance?

Workers on MTurk get their payment directly after the requester has approved the results. During the task design, the requesters are trying to adjust payment to be ethically acceptable and pay around the current minimum wage. However, Hara et al. (2018) point out that workers are getting around 4-6\$ per hour on the MTurk platform, which uses a direct payment process, but still the payment is far lower than the US minimum wage (8\$ per hour). There are many factors that cause this gap between the intended payment and the actual reward received by a worker. One of these is the time workers spend to find the right task (Hara et al., 2018), workers leaving the task before submitting because they don't understand something or fear rejection (Han et al., 2019; McInnis et al., 2016), requesters estimating shorter time to complete the task which leads them to set up lower payments (Silberman et al., 2018), and many other reasons, such as the difference in the economic growth between countries which makes the payment for a task more valuable in some countries more than in others.

During the literature review of the field, it was discovered that several studies discussed the fair reward for crowdsourcing tasks in MTurk (Horton and Chilton, 2010; Silberman et al., 2018; Hara et al., 2018; Ipeirotis, 2010a; Ipeirotis, 2010b; Ross et al., 2010; Marshall and Shipman, 2013; Williamson, 2016; Tate, Johnstone, and Felt, 2017) and a few were comparing the performance of other platforms (Borromeo and Toyama, 2016; Peer et al., 2016; Difallah, Filatova, and Ipeirotis, 2018).

Furthermore, other platforms, such as F8 use intermediary channels. These channels connect workers and give them access to the F8 platform. In exchange for their services, these channels take a commission that will be subtracted from the workers and the requesters. Based on the F8 platform, we noticed that the commission amount varies from one channel to another. However, when a requester is designing a task, they only take into consideration the general ethical regulations that they should follow, not the commission rate at which a particular channel will reduce the payment. A large number of requesters do not pay attention to the issue of reduced rate which will affect the actual payments the workers receive after they finish the task. As far as we know, no study has been conducted in the area of analysing channel payments in crowdsourcing platforms. This chapter investigates the variation of the payment schema in F8 channels. We ran a survey task asking the workers in F8 how much they had been rewarded, and we compared the payment amount for each channel. Furthermore, we collected results for over 150 tasks running between 2016 - 2018. We provide a comparative analysis of the demographics and channel distribution for over 50k workers over time.

In this chapter we will present all of these efforts to view payments as motivation and measure the effectiveness of incentive rewards. Before that, we will present some of the researchers' efforts that address the ethical and moral issues around payment in crowdsourcing tasks. Researchers address some of the rules for payment in human interaction studies. We present these rules and how one could apply them to crowdsourcing tasks. These rules should be consistent for all requesters and need to be adjusted for all kinds of crowdsourcing tasks.

## 6.2 Ethical issues and right regulation around payment

The globalisation of crowdsourcing work makes it difficult to apply any regulations or ethical conduct guidelines that are already used in business and academic field. For example, the requester could be from Europe following the ethical code of payments like the German ethical code (Martin et al., 2016) and workers could be from the US or Asia and they could not agree with the payment amount for a given task. Another issue is the variation of time spent on a task which would be different from one worker to another and each should be rewarded according to the time they spent on the task. These and many other issues around payment in crowdsourcing tasks are not easy to deal with.

Using crowdsourcing to collect data in social science studies brings together the benefit of having a varied demographic distribution and fast responses. However, many researchers raised the issues of improper payments and low rewards, especially in the kind of studies where researchers are not experts in the domain of crowdsourcing task design. Andersen and Lau



(2018) discussed the pay rate in social science experiments that have been carried out using crowdsourcing. Two experiments of different length were conducted to measure the effect of payment rate on the quality of the performance of the workers. The findings of this study confirmed that high or low pay rates did not have a significant effect on workers' performance, but they did have a different kind of effect, such as the speed of finishing the task on the platform. A similar study by Haug (2018) explored the moral and ethical issues in collecting data for survey research using crowdsourcing. Several studies considered using crowdsourcing as a fast, cheap and high-quality tool for managing data for social science. However, others (Borromeo et al., 2017; Williamson, 2016; Fort, Adda, and Cohen, 2011) have risen concerns, claiming that low pay rate could ruin the ethical essence of the collected data for such studies. In Haug (2018), the author discusses both scenarios (high and low payments), and points out that raising the payment did raise the risk of having workers who are used to doing the same kind of task which would increase the level of bias and that could harm the results of their survey study that require workers who are not familiar with this kind of task.

Paul and Lars (2018) developed a model to test the fairness of the payment during the execution of the task and after the submission. Another work by Goel and Faltings (2018) discussed the fairness and the workers' trust in crowdsourcing platforms. They proposed a mechanism that used peers' answers to verify workers and reduced the number of gold questions needed in the task. A deeper study by Archambault, Purchase, and Hoßfeld (2017), pp 27-69, discusses the ethical issues around the use of crowdwork in academic research. The authors recommended following the guidelines provided by the Dynamo project and the Crowdworking Code of Conduct as a moral guide for the researchers planning to use crowd tasks in their work.

In most of the studies around payment issues, researchers strive to pay attention to the fair payments when using crowdsourcing tasks (Silberman et al., 2018; Brawley and Pury, 2016; Ipeirotis, 2010a). Silberman et al. (2018) notes the ethical responsibility of paying workers fair wages and discusses the importance of money as a motivation factor for most of the workers as it had been considered in previous studies (Ross et al., 2010; Ipeirotis, 2010b; Ho, Jabbari, and Vaughan, 2013; Ye, You, and Robert, 2017; Finnerty et al., 2013). Moreover, they point out that fair payment leads to high-quality performance from the crowd. Researchers tried to develop models or implement criteria as a base for giving the right payment for each kind of task. However, even if the requesters are paying an acceptable rate with accordance to the minimum wage, workers still complain that they are not getting a fair payment and that could be due to multiple reasons. In our study, we examine these issues and focus on the one related to intermediary channels and the gap between the actual payment made by the requester and the payment received by the workers on the F8 platform.

### 6.3 Motivation and payment in crowdsourcing tasks

One of the earliest studies that examined the effectiveness of financial incentives on the crowdsourcing task outcomes was by Mason and Watts (2009). The authors discussed the impact of increasing the task rewards on the workers' expectations of the task, and they found that high rewards make the tasks more attractive to the workers but did not increase the quality of the

outcome as we stated in the previous chapter. A similar study by Borromeo and Toyama (2016) compared the performance in an unpaid crowdsourcing task with a paid one. They used Py-Bossa for the unpaid crowdsourcing task and Figure Eight for the paid one. The findings of the study show that the results of the task used (sentiment analysis and data extraction) were highly similar in the paid and the unpaid condition, but it took longer to finish the unpaid tasks. In contrast, Kost, Fieseler, and Wong (2018) define incentive rewards as one of the four sources of experience meaningfulness for the workers. During their experiments, they found that the level to which the payment affects the workers depends on their employment status in real life and how much they rely on the crowdsourcing work.

These studies and others show that the impact of the payment cannot be ignored even if it may have only a slight effect on the workers' performance. Ye, You, and Robert (2017) investigated the impact of the payment amount on the workers' performance in two types of crowdsourcing tasks. They introduced the concept of Perceived Fairness in Pay (PFP) and measured it in their experiments. This study aimed to clarify the relationship between fair payment and the quality of the results.

More studies investigated extensively the effect of fair payments and the loss of time in crowdsourcing tasks. Researchers found that there is a massive gap between the earnings and the time and the effort spent to complete a task. They warn the academics and all requesters in general that discarding these details could threaten the popularity of crowdsourcing jobs in the future. Hara et al. (2018) discussed workers' earnings on MTurk and considered the non-payment time, e.g., time spent finding a task and working on tasks that are rejected. The authors expressed their concerns about some waste time that affects the hourly wage, such as time spent searching for the right task and time spent on work that will be rejected or withdrawing before submitting.

Another study by Borromeo et al. (2017) discusses the implementation and evaluation of transparency and fairness principles on a crowdsourcing platform. On the one hand, the authors discussed the fairness in task assignment, completion time and payment. On the other hand, they recommended having a special framework to encourage a more transparent process for requesters and platform developers. Moreover, in Ho et al. (2015), the authors suggested a different payment scheme such as payment per unit as well as a bonus for achieving a specific target.

Furthermore, other researchers show that workers could be motivated and work on a task with low or unfair payment or even work as volunteers if the task has deep meaning to them. Some researchers claim that workers will respond to the good humanitarian causes such as tasks for World Health Organisation (WHO) or disaster responses. For example, in Spatharioti et al. (2017) the authors point out that workers tend to do more work in - as the authors refer to it - a "meaningful task" such as a disaster response task. They designed a task with a fixed amount of minimum units to be completed by the workers and gave the workers choice to complete as many units as they wanted. With this kind of a flexible number of units and an interesting task, workers did more work and showed more commitment to finish it.

Most of the studies that discuss the payments and intensive rewards used MTurk to analysis the quality of the results vs payments. However, most of the work in MTurk - as mentioned in Ho et al. (2015) - is "performance-based" which means workers tend to submit a high-quality piece of work because they are afraid of rejection if their work does not meet the task criteria or the requesters' level of expectation. On the other hand, on the F8 platform low payment could affect the workers' performance as we saw in Chapter 5, since the workers know that they are getting paid regardless of the requester feedback (accept or reject) for their work, thus low payment will not motivate them to expand effort to submit high performance results. In this work, we focused on the F8 platform and the variation of payment due to different commission rates taken by the channels. Based on an analysis of over 150 tasks in the last four years, we present the most common channels and show how they work.

## 6.4 Analysis of workers on Figure Eight

Following the appearance of crowdsourcing platforms, researchers made effort to publish a statistical analysis for these platforms and provide users in academic and business fields with fruitful data about these platforms in terms of workers population and diversity. Each platform claims that they have the largest crowd population and the most diverse expertise of online users. MTurk provides the number of workers available to work on the task and claims that they hire over 500,000 workers on the platform. Other platforms -such as F8, CROWD, Swagbucks, and Clixsence- combine and use network channels to provide a large crowd across these platforms (Schmidt and Jettinghoff, 2016). We chose to focus on the F8 platform as it considered one of the most commonly used platforms that work with these channels.

Many studies used MTurk as their crowdsourcing platform as it allows for the analysis the workers giving a wide range of information about the demographic, gender and academic qualifications. This platform has also been used to study the ethical and moral issues around payments as we discussed in Section 6.3 and Section 6.2. Archambault, Purchase, and Hoßfeld (2017), compared the performance of the workers in the USA and India using qualitative studies and close observations. In this study, the authors tried to make sense of how workers behave when trying to understand the task and how that impacts their work. In our study, we analysed the workers in F8 marketplace and compared the payment through different channels. We also used a survey task to collect information from workers and to understand their perspective as to what motivates them to use and work on this crowdsourcing platform.

The following sections present an overview of the top 5 channels and looks into the reasons that motivate workers to choose to use these channels rather than access the crowdsourcing task directly from the website (Elite channel).

### 1- ClixSense channel

Established in 2007, Clixsense<sup>1</sup> is one of the most popular PTC websites. On this platform, a weekly contest is run and the top 10 workers (with the highest number of tasks completed) win

---

<sup>1</sup><https://www.clixsense.com/>

\$ 100 in total prizes. The tasks include completing surveys, offers, F8 tasks, watching a video, and others. There is a standard and a premium membership option; the difference between the two levels is the percentage amount the worker gets from doing the daily checklist and how much they receive from their referrals. Members are given referrals links, and a worker can get a 20% commission on what their referrals earn at Clixsense. Payments are issued every Monday if the worker has earned more than \$8 for Standard Members and \$6 for Premium members. As a motivation, this channel offers a \$5 bonus if the worker earns \$50. The minimum reward for a task is 1 cent and if the worker completes a task worth less than 1 cent they will not get paid in Clixsense unless they complete another task for the same job.

## 2- Elite Figure Eight channel

Most of the other channels offer crowdsourcing tasks as one of many services and ways to earn money. The Elite<sup>2</sup> channel is the most straightforward way of accessing a task in F8 and provides a slightly higher payment as there is no commission like in other channels, but workers may prefer other channels because they offer bonuses and other ways to gain extra points by completing surveys, games, and ads. Workers have to complete at least 100 test questions and pass them to be officially working on the real tasks. There is no payment for these test questions, but rewards follow when further levels of accuracy are maintained. To reach level 1, workers must achieve 70% accuracy, 80% to reach level 2, and 85% accuracy to reach level 3. Workers were moved to a new "Contributor Portal"<sup>3</sup> after the change to the company administration and the changes to the platform policy are still unknown.

## 3- NeoBux channel

NeoBux<sup>4</sup> is an online Paid to Click (PTC) website established in 2008 and is still operating under the same name and policy. It offers free registration (Standard membership) and pays members for every click on ads (the clicks, however, are limited and workers need to be active daily to avoid suspension or cancellation of their membership). Workers can earn more when they upgrade to Golden membership for \$90 per year, for which they will get up to 2000 clicks per month at 0.01, and rent referrals. Once the workers earn money, they can withdraw it to their Paypal and Payza accounts with the minimum withdrawal amount of \$2 for the first time, and then when they reach a fixed minimum amount of \$10. The crowdsourcing tasks come as **mini jobs** to earn extra money.

## 4- InstaGC channel

Set up in 2011, InstaGC<sup>5</sup> channel offers free registration and is similar to Clixsense and Neobux, in the kind of services and referral system it provides; the only advantages of using it rather

---

<sup>2</sup><https://elite.figure-eight.com/>

<sup>3</sup><https://contributors.figure-eight.work/>

<sup>4</sup><https://www.neobux.com/>

<sup>5</sup><https://www.instagc.com/>

than the previous two is that the payout threshold is only \$1 for 100 collection points. The payment is in the form of a gift card or a cash payment made through bitcoins or other electronic money transaction with a small fee due to the cash exchange process.

### 5- Swagbucks via Prodege channel

This channel supports online shopping to earn Swagbucks (SB) points. It is connected to the Swagbucks<sup>6</sup> website where workers can enter F8 and complete a task to get SB points that will be exchanged for real money. The primary ways to redeem Swagbucks are Paypal, Visa gift card, and Merchant gift cards. For each 100 SB, a worker gets \$ 1 at the end of the month. There are several ways to earn SB points such as using the SB search engine, playing games, watching videos, shopping online, answering surveys, and completing F8 tasks.

#### 6.4.1 Data collection and analysis of channels

The data has been collected from the Figure Eight (F8) platform. We aggregated data from the last four years (2016 - 2018) from multiple requesters' accounts. This gave us access to over 130 tasks with 54,000 rows of data. From this data, we found that workers in F8 come from over 110 countries. We looked carefully at the channel distributions in each country. Channel popularity varies from one country to another. Figure 6.1 shows the top 10 countries for the workers in F8 platforms in the last four years and the frequency with which the channels are used in each country. Workers in the USA tend to use many different channels compared to workers in other countries. Over 5000 workers used InstaGC channel, which was the most popular channel in USA, while Elite channel came in the second place with over 4000 workers in the last four years. Venezuela came in the second place with more than 90% of workers using NeoBux channel. India was third with over 4000 workers using Clixsense in the first place and Elite the second most used channel.

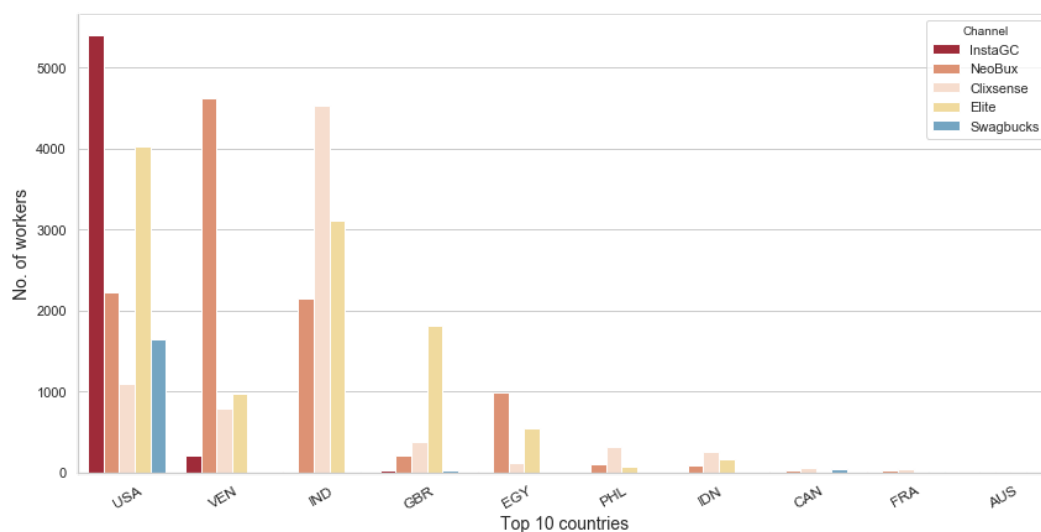


FIGURE 6.1: Number of workers per channel in the top 10 countries

<sup>6</sup><https://www.swagbucks.com/>

In addition, Figure 6.1 shows that more than 70% of workers in the GBR rely on using Elite channel rather than using the other channels. The variation of the uses and popularity of these channels depends mostly on the payment amounts and the advertising campaigns by the channels to gain workers' loyalty. For example, in Anand (2018), the author presents some guidelines and recommendations for earning from crowdsourcing websites, especially from the Figure Eight platform. He shows why Clixsense is one of the most popular channels as it provides some benefits to the workers and maximises their rewards.

Most of these results were collected from tasks that did not have any restrictions and workers from all over the world could access and perform the task. We analysed the demographic distribution of the workers over the world. With many channels providing the same services, some channels were more attractive to the workers. Figure 6.2 shows the number of workers in the top 5 channels used in the last four years. NeoBux, Elite, and ClixSense were the most popular channels for F8 workers from all over the world, followed by InstaGC and Swagbucks channels.

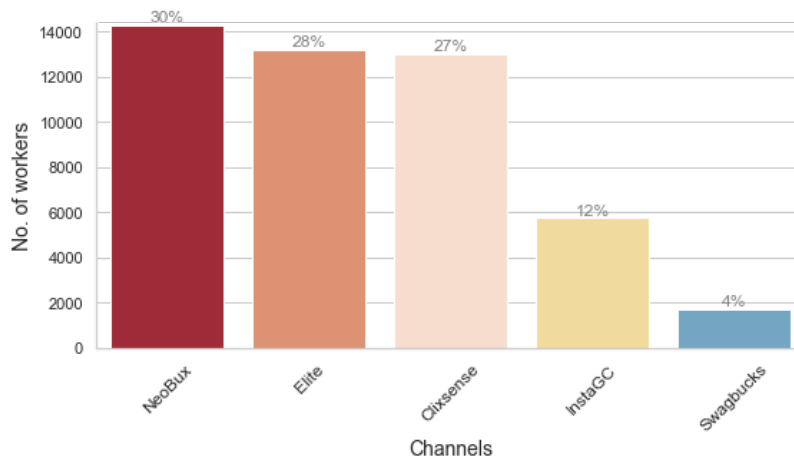


FIGURE 6.2: Number of workers in the top 5 channels

#### 6.4.2 Results of survey workers

We ran a survey task, as shown in Appendix G, on F8, in which we focused on the top five channels shown in Figure 6.2. We collected information from 60 workers from each channel. Based on the results, the channels were divided into two different groups:

- **Group 1:** Clixsense, Elite, and NeoBux;
- **Group 2:** InstaGC and Swagbucks.

All the results are reported in Table G.1, and this section summarises the most important conclusions from this survey.

### \* General information

For Group 1 channels, more than 70% of the responders were male and 30% female, while nearly the opposite was true for Group 2 with over 60% female and less than 40% male as shown in Figure 6.3.

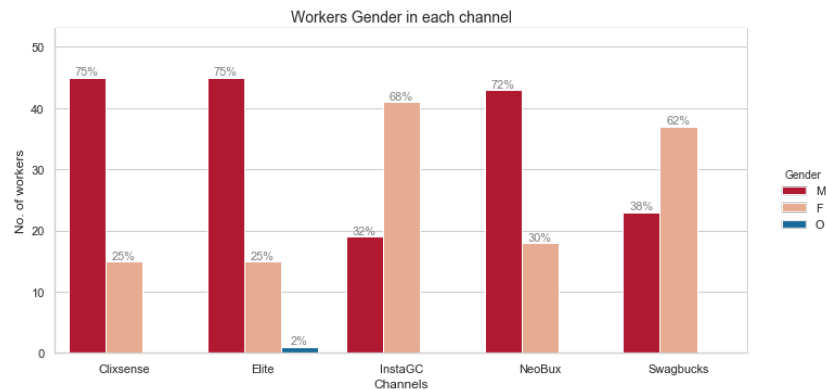


FIGURE 6.3: Workers' gender in each channel

The majority of workers in Group 1 (over 60 %) were in the 18-34 age range, while in Group 2 the majority of workers were in the 25-44 age range. Most of the workers in all channels hold a good level of education - a bachelor or a higher degree -.

Over 35 % of workers in Elite and NeoBux are self-employed whereas on the other channels, more than 45% already have a full time job. We asked the workers about the device they use when performing the task and the majority of them -over 95%- said they use a personal computer or laptop; thus we know that the workers see the task GUI as intended. Moreover, when workers are using a desktop computer or laptop to perform the task there is high possibility that they are paying full attention when they are working on the tasks.

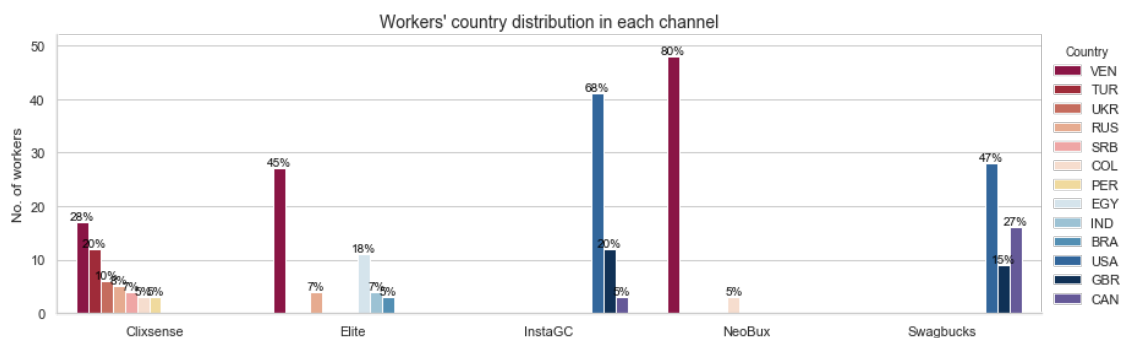


FIGURE 6.4: The distribution of workers countries per channel

As presented in Figure 6.1 NeoBux channel is very popular with workers in Venezuela. In our survey task, 80% of the responders from this country said they were using it. Elite and

Clixsense are widely used around the world with a high variation between countries. Furthermore, platforms from Group 2 are most popular with workers in USA, Canada, and the UK, see Figure 6.4.

**\* Crowd-workers experience**

When it comes to the level of experience in performing crowdsourcing tasks, over 40% of workers in Group 1 have around two years of experience in performing crowdsourcing tasks, while over 60% of workers in Group 2 reported that they have more than three years of experience in crowdsourcing tasks, as shown in Figure 6.5.

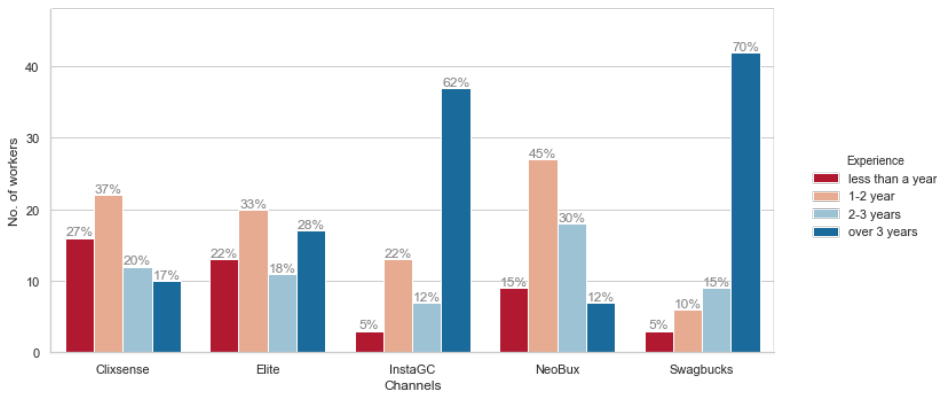


FIGURE 6.5: Range of workers' experience in each channel

The level of workers' experience is consistent with the completion time recorded for each worker in F8 and shows that workers in Group 2 finished our survey in less than 4 minutes, while workers in Group 1 took an average of 8 minutes to do the same task, see Figure 6.6.

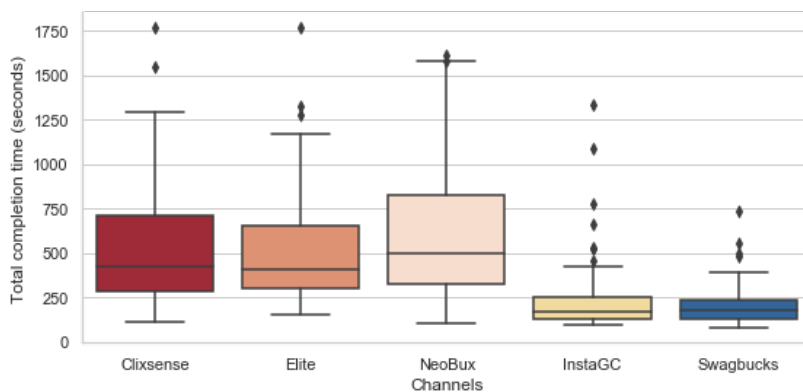


FIGURE 6.6: The distribution of completion time for workers in each channel

We carried out a statistical analysis on time required to complete the task as the dependent variable and studied *Channels* as independent variables. The effect of the channel on task completion time is statistically significant ( $p < 0.05$ ) as shown in Table 6.1.



TABLE 6.1: One-way ANOVA for time to complete the survey task per channel.

	sum_sq	df	F	PR(>F)
Channel	$7.96 \times 10^6$	4	21.7	$9.05 \times 10^{-16}$
Residual	$2.68 \times 10^7$	297.0	NaN	NaN

70% of workers who are using Elite and NeoBux were found to be relying more on the crowdsourcing tasks as the main source of income compared to workers who are using other channels in this study. However, 35% of workers from these two channels specified that crowdsourcing jobs provide them with over 80% contribution to their total income. Furthermore, over 90% of workers in Group 2 stated that they do not rely on crowdsourcing jobs, and specified that the contribution of the crowdsourcing jobs to their income is less than 20%, or less than \$50/month.

Figure 6.7 shows the average earnings(\$)/month for workers in each channel. Although workers on Clixsense are not considering crowdsourcing job as the main source of income, they recorded the highest earnings compared to workers on other channels, with a median average of 100\$ per month.

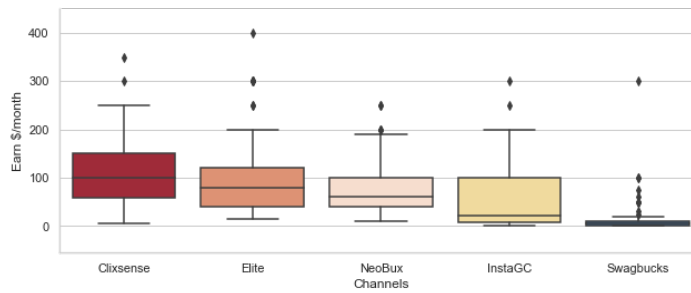


FIGURE 6.7: The distribution of earnings from the crowdsourcing jobs in dollars per months

When asked about their preference and the reasons behind choosing a particular task from the list of tasks on the platform, over 80% of workers from all channels chose 'Reward' as the main reason. 'Time required for completing', 'Difficulty', and 'The most interesting' were chosen as reasons 50% of the time, as shown in Figure 6.8.

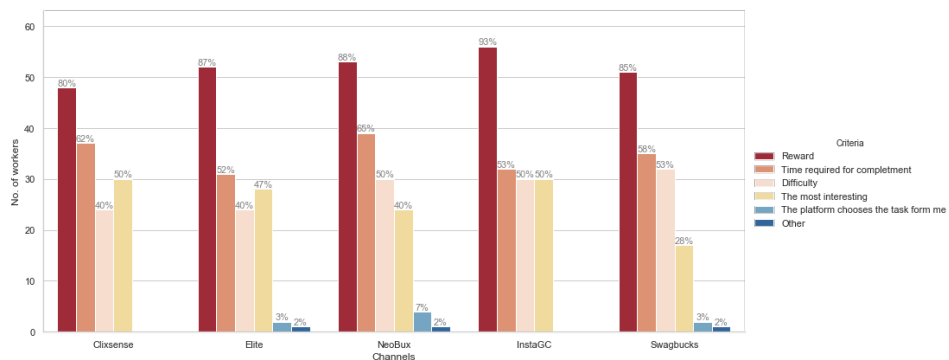


FIGURE 6.8: The criteria of choosing the task

Workers seem to be loyal to some platforms, since we asked them if they were using any other channels and less than 20% said 'Yes'. Figure 6.9 displays the other platforms that workers are using to access the F8 platform. Workers in Swagbucks channel are the most versatile, using more than one channel at the same time.

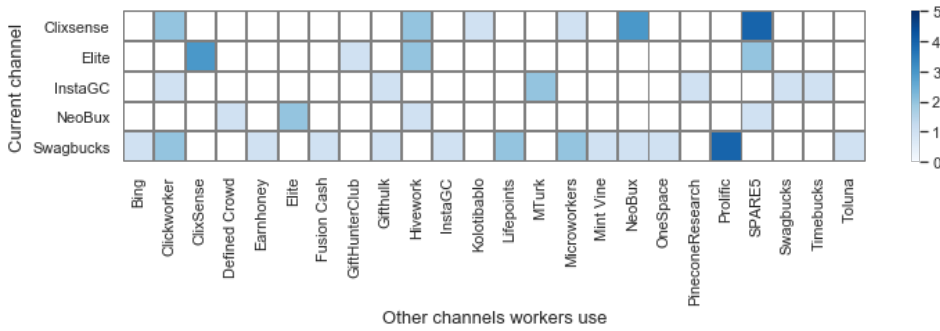


FIGURE 6.9: Other platforms used by workers

**\* Payment per channel**

To investigate the payments in each channel, we asked the workers if they get paid for crowdsourcing tasks. Over 90 % of workers in all channels said that they choose to work for tasks that have a monetary reward, not as volunteers. This was followed by more detailed questions if the workers had answered that they are rewarded. In that case we asked about the nature of rewards, process time, and how they are receiving it. Most of the workers get paid for every task or group of tasks they complete. Concerning the nature of the reward workers get from the crowdsourcing task when they use channels, around 60% of workers in Group 2 said that they received money as an electronic payment that they can use later, and over 35% of workers in Group 1 said that they received money directly to their bank accounts. When it comes to the time between submitting the task and receiving the payment, 60 % of workers in InstaGC and NeoBux receive the payments within a few minutes, while in Clixsense and Elite, 60 % of workers said that the payment appears in their accounts within a few days.

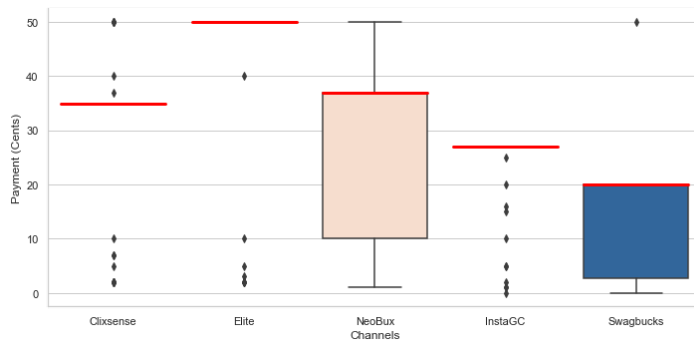


FIGURE 6.10: The distribution of actual payment received by workers in each channel (The red lines show the median)

For this survey task, we paid the workers 50 cents for completing the task, and we asked them to specify the final amount they will receive from it. Figure 6.10 shows the distribution of the payments the workers said they got. Workers in Elite recorded almost the same amount of 50 cents. However, workers in Group 2 recorded less than 50% of the actual payment, as shown in Figure 6.10. This variation of payments results from the commission rate that the channels cut from the money that should go straight to the workers.

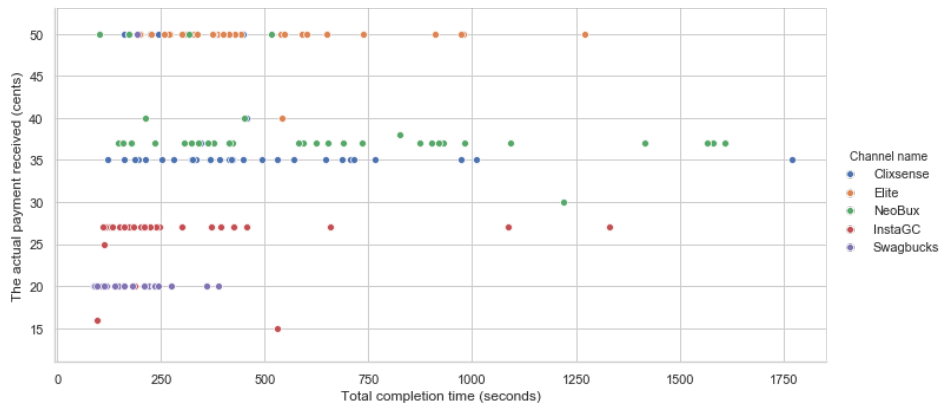


FIGURE 6.11: The relationship between completion time and the amount of payment workers received

We analysed the amount of payment the workers thought they are getting from the task with the time they spend completing the task. We found that the amount of payment is consistent over channels but not consistent with the time spent to complete the task, see Figure 6.11.

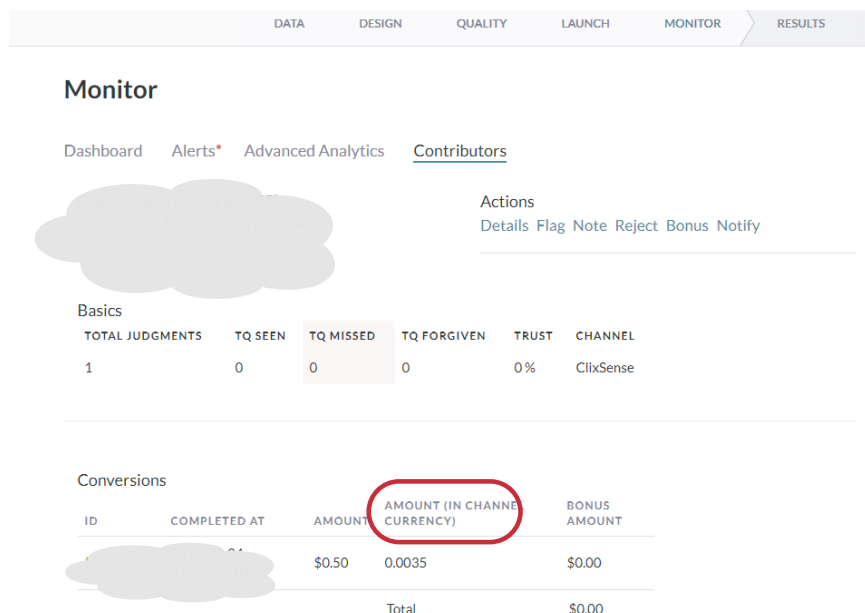


FIGURE 6.12: F8 requester’s interface showing wrong amount under ‘IN CHANNEL CURRENCY’ (0.0035 for 0.35)

To compare payments by channel we launched another group of tasks and we asked 500 workers how much they got paid for each task. Furthermore, we verified the answers provided in these experiments with the data collected for the last four years (discussed in Section 6.4.1) and we compared the differences between the payment specified by the requester and the one actually received by the worker. From our analysis of the payment channels, we found that in F8 the requester’s interface wrongly presents the amount of payment since the channel commission rate can only be inferred from the field labelled "AMOUNT" (IN CHANNEL CURRENCY) as shown in Figure 6.12.

Moreover, we found that only Elite awards 100% of the original payment (given by the requester) to the workers. Most of the beta channels, on the other hand, take a commission from the payments and even though some of these channels take a high amount, still a high number of workers use them. Figure 6.13 shows the original payment provided by the requesters and the actual payment received by the workers in the rest four channels, that is: Clixsense, InstaGC, NeoBux, and Swagbucks, which are the most used channels in the last four years.

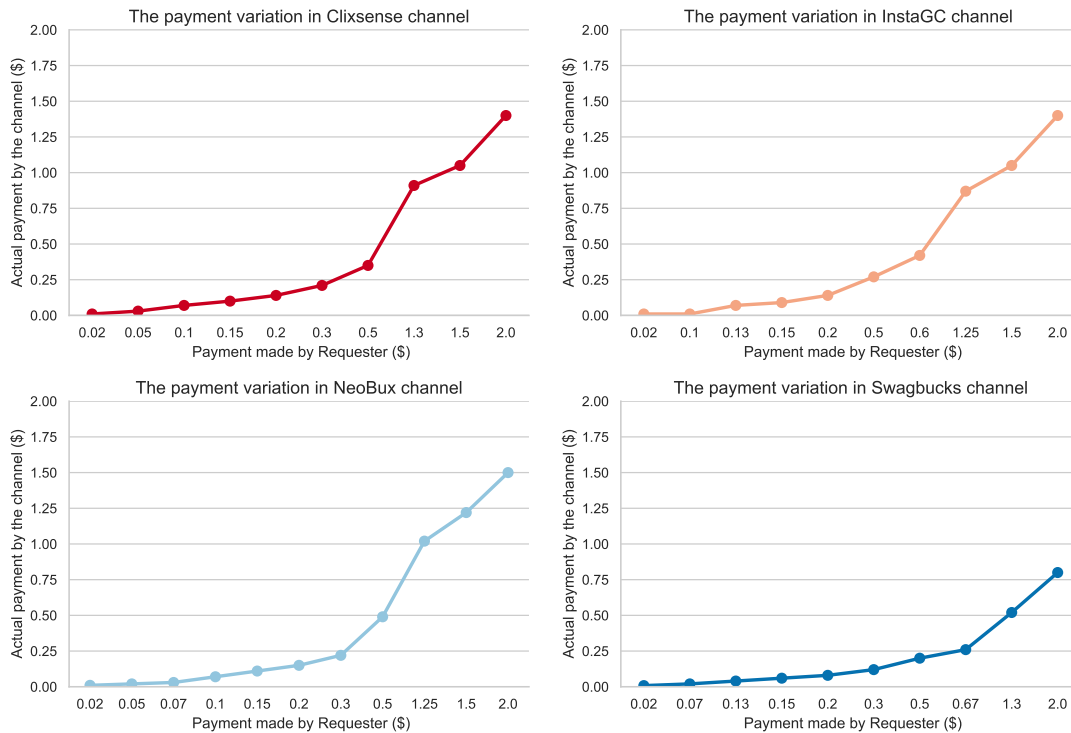


FIGURE 6.13: The variation of the amount of payment in the top 4 channels used in F8 in the last four years

A use-case scenario can simulate how much money the workers will lose based on which channel they choose to use for the crowdsourcing tasks. Using the values shown in Figure 6.13 we calculated the *Percentage of Relative Change* between the payment received by the worker and that made by the requester. After that, we estimated the amount of money lost per year for the workers based on which channel they choose to access crowdsourcing tasks. For example, if the workers use Elite they will be paid 100 % of the set payment and they will not lose any money, but if they choose NeoBux, they will lose 20% of the amount and 30% if they choose

Clixsense or InstaGC channel. However, for workers who choose to work in SwagBucks, they will lose 60 % of the total amount of the set payment per year.

From the data we have for the last four years presented in Section 6.4.1, and as shown in Figure 6.2, 30% of workers use the NeoBux channel, followed by 28% & 27% of workers who use Elite and Clixsense respectively, and less than 12% use InstaGC channel and only 4% use SwagBucks. We used this information to estimate the total loss of money for workers on the F8 platform by using the percentage of relative change calculation. Table 6.2 shows the estimate of the money lost for each worker — based on which channel they choose to work on —, and for F8 workers population per year in general -based on the estimate of the number of workers using these channels in the last four years-.

TABLE 6.2: The estimate of the money lost for each channel at the workers level and F8 population.

Channels	Channels	
	% of loss for worker	% of loss for F8 population
Elite	0	0
Clixsense	30 %	8.1 %
NeoBux	20 %	6.2 %
InstaGC	30 %	3.6 %
Swagbucks	60 %	2.4 %

#### \* Workers feedback

More than 60% of the feedback was positive; workers were impressed to see this kind of survey where we investigate the workers perspective and they were interested in answering the survey. They expressed that our task was one of the easiest and quickest survey tasks they have ever completed as they see many surveys with long and complicated questions. Some of the positive comments regarding the survey and the topic we are discussing in general are quoted below:

*“I am glad that you are carrying out this type of survey since, I am sure, you could obtain results that demonstrate the injustices that are committed against workers who do not have any type of defence in relation to work and rights that this entails.”*

*“I am pleased that you are asking these questions with the motivation you have explained, the weight of consideration for task authors/clients vs. workers feels quite unbalanced at times so it’s heartening that there is here clear consideration for workers. Thanks! I have found this type of work has been a lifeline having been on debt for years it has enabled me to add to my regular income and now I am nearly able to afford a house deposit. It had changed my opportunity to improve my life considerably.”*

*“Liked this survey - short, concise and keeps interest. Hope this has been useful.”*

*“Very interesting survey, apart I would like that Figure Eight could enable more jobs and leave them for a long time to complete it completely and thus be able to obtain better earnings for the work done, because sometimes a job comes out and does not last long and only does 4 or 6 tasks Thank you for letting me be part of this prestigious company. Regards”*

There were repeated requests to examine the fairness of the payments and poor task quality. Moreover, several workers pointed out that task instructions sometimes are ambiguous and not clear, which leads them to abandon the task or submit wrong answers. Workers point out that the number of tasks on the platform has decreased and the platform has been suffering from many technical issues.

## 6.5 Chapter summary

In this chapter, we studied the crowdsourcing task from the workers perspective and focused on some of the ethical issues around payments. We found in the literature that researchers spend a lot of effort to create a safe and fair environment for the workers and requesters.

Despite that, the gap between workers and requesters still persists and so in order to present the best task design methods for requesters we took steps to understand the workers' point of view. We performed a qualitative study focused on the top 5 channels used in the last four years. We surveyed workers on four aspects: general demographic information, crowd-working experience, payment schema, and their feedback and ability to participate in further studies.

The findings of this study show an unbalance in the treatment of workers due to the differences in channels' policies. Furthermore, the distribution of workers demographics and the level of experience vary from one channel to another. In general, in addition to other reasons, for all workers the main motivation for completing a crowdsourcing task is the reward.

The workers complain about unfair payments, while they do not know that there is a high possibility that the intended payment is much higher than the one they are actually getting. Besides crowdsourcing tasks, we found that channels offer many services to the workers and doing a crowdsourcing task is only one of the extra jobs that they can do to get extra rewards or points.

We believe that more knowledge about the channels' policies and procedures will help workers get the best service and save their time. In case workers focus only on crowdsourcing tasks, they should choose the right channels, which might lead them to get higher benefits and get the right payment for the crowdsourcing task. From another point of view, requesters expect high-quality performance for crowdsourcing tasks as they are small, short, and paid jobs. They should pay attention to the amount of payment they set and the target channel they choose for their task.

The country distribution of workers differs between channels. If the requesters want to achieve results fast while offering an easy task with low payment, they may consider to target workers from Group 1. For a task with high level of difficulty or one that requires some advanced

efforts, requesters may want to consider workers from Group 2. Moreover, if the task requires a specific gender, this may influence the requesters' choice of channel as females are more likely using channels from Group 2 more than other channels.

In future work we will continue to further investigate the economic changes over the last 4 years that have had an influence on workers' readiness to work on crowdsourcing platforms in specific countries. Moreover, we will interview some of the workers who did this survey task. We are planning to design the questions for the interview based on the results of this survey. Furthermore, the study will be extended to survey workers from other channels on F8 platform.





# 7

## Conclusion

In this thesis, we presented a variety of methods for improving the design of crowdsourcing task, taking into account different aspects involved. The goal is to enhance the design of crowdsourcing tasks to achieve better performance. This work strives to deliver to the requester the best methods in designing the task and make them achieve a better understanding of the workers' standpoint when they perform the task. This information will serve both sides (workers and requesters) to enhance the overall performance by saving time and money, and obtain better crowdsourcing services. This aim was achieved by performing experiments exploring different design aspects and investigating some of the design factors that we discussed in Chapter ??.

This chapter revisits the research sub-questions (presented in Section 1.3), and summarises the findings and contributions that addressed these questions throughout the previous chapters and indicates possible directions for future work.

### 7.1 Summary of the thesis

Chapter 4 looked at designing crowdsourcing relevance judgment experiments. We investigated the effect of class order and balance in the batch of binary classification tasks on the quality of workers' results. To answer **(RQ1)**, we found that appropriately ordering tasks can be useful to increase crowd workers' performance in unbalanced label situations. We observed that in the cases in which the number of relevant and non-relevant documents is approximately the same (i.e., balanced classes), crowd workers perform better when the relevant ones are presented first. To answer **(RQ2)**, which takes into account order and balance of classes in the batch, the results indicate an opposite trend: for unbalanced batches, increasing the number of

documents improves the performance, whereas for balanced batches the result is opposite. Requesters should note that longer unbalanced batches might lead to classifying a rare class more accurately, although it might lead to increased temporal demand and fatigue, which can reduce the overall performance. For the balanced classes situation, we can achieve the same results as for a shorter batch, that is - the more relevant documents workers encounter, the more accurate work they provide. To answer **(RQ3)**, in the absence of gold labels or training questions, an analysis of document rank correlation corroborates the observation that priming workers has a significant effect on the resulting ranking, and this effect is consistent within TREC topics. Similarly, inter-annotator agreement is higher when relevant documents are shown at the beginning of batch. Observing this level of agreement motivated us to measure the level of agreement between the crowd workers themselves and study the stability of the task design and the likelihood of achieving the same results when running the task several times and on multiple platforms.

Chapter 5 focused on the inter-batch effect and investigated the consistency of crowdsourcing a task with the same design over time and multiple platforms. We ran continuous experiments on MTurk and F8 every week. The task interface and datasets used for this study were fixed. To answer **(RQ4)**, we measured the repeatability of a crowdsourcing task over time. The findings of this study show a significant consistency of results when repeating the same task once every week and a high level of agreement between crowd workers and expert annotators for the datasets we used in the tasks, especially when the level of the task is challenging and not easy to define, such as in binary classification. Moreover, to measure the reproducibility of a crowdsourcing task over multiple platforms **(RQ5)**, we recorded inconsistency in responses when reproducing the same task at the same time on different platforms. Using different datasets to address **(RQ6)**, we observed consistent results over time and different platforms even when equalising the payment scheme in both; thus, we can confirm that crowdsourcing tasks can be generalised over different classification tasks. This findings allow for another level of comparison between platforms while taking into account the different payment amounts due to different channels.

Chapter 6 discussed some of the ethical issues around payment setup for crowdsourcing tasks. Several previous studies discussed the low payment for crowdsourcing tasks and presented suggestions to save requester's money and prevent workers from wasting their time. Most of these studies used MTurk for their experiments, and several others used the F8 platform. In our study, we focused on the commission rate cuts by the intermediary channels which the workers used to access the F8 platform. To address **(RQ7)** and **(RQ8)**, we performed a series of qualitative experiments and surveyed workers as they were performing crowdsource work about their experience, motivations, and the received payments. To answer **(RQ7)**, we found that the workers' primary motivation are the rewards, followed by their interests, then time required to complete the task and its level of difficulty. The findings of this study show a broad variation in the demographic distribution and level of experience between different channels. Furthermore, the investigation of **(RQ8)** showed another level of variation which lies in the amount of payment received for the same task when workers use different channels.

This study showed that in addition to the low payment for the crowdsourcing task there is a gap between the original payment made by the requester and the one worker receives caused by the intermediary channels on the F8 platform.

## 7.2 Research challenges and limitations

The wisdom of the crowd is shallow without the right arrangement of the crowd work. One of the most recognised challenges of crowdsourcing tasks is avoiding creating a poorly defined task that will lead to low-quality results (Whitla, 2009). As researchers stated in the past, crowdsourcing tasks is powerful and fruitful if requesters take time to design clear instructions and intelligible task interface (Ikediego et al., 2018; Schmidt and Jettinghoff, 2016).

In this thesis, we showed that enhancing the design of the task leads to better results, even though the the accuracy of the crowdsourcing results was not higher than 85% compared to the gold standard, which could be explained as follows:

- we used three different datasets and the gold standards collected for these datasets were generated in different situations: Dataset 1 was annotated in 2016, Dataset 2 was annotated in 2018, and Dataset 3 was annotated in 2000. Due to the time differences between creating the gold standard data and running the crowdsourcing task, the experience and the background of the participants (workers and the experts) could have been dissimilar.
- we noticed consistently low accuracy for the results from Dataset 2 (around 65% accuracy) and as Whitla (2009) states, the results might need to be re-evaluated because a vast amount of noise could not be noise. Filtering the results will need more work as well as running extra tasks for this dataset should be considered.
- all the experiments in this research were launched on a working day and due to time zone differences, we may have targeted the same demographic distribution every time and unintentionally excluded some others.

Moreover, the availability of resources used in previous work was one of the challenges that we faced in this study. In order to expand the work of Chapter 5, we needed to have a clear and well documented dataset that contained full information about the task GUI, the batch formation setup (e.g. number of classes and number of judgments per batch), and gold standard data for each setup. Furthermore, the work in Chapter 6 faced the regular known challenges for any qualitative research, that is the limitations of the pre-set goals for the questions. We aimed to be open-minded in predicting the workers answers and setup clear questions that pertain to our aim, however, some of the workers expressed that they found some of the questions to be ambiguous. Others expressed a wish for more detailed questions.

### 7.3 Guidelines for the best design of crowdsourcing task

To keep a balance between high-quality performance for the requester and fair treatment for the workers, requesters should pay attention to some key factors while they design and set up a crowdsourcing task on the F8 platform:

- **For the design of the task:**
  - When evaluating an existing dataset, requesters can use relevant labels at the beginning of the batch as priming for the workers instead of using examples or train tests before they start the original task; this will save time and will improve the learning of the workers and eliminate the bias of seeing examples.
  - Requesters could use the relevance attributes of documents to order them when dealing with a new dataset with unknown labels.
  - Performing a batch with a balanced number of classes is important to achieve better performance and rise the accuracy of identifying each class.
  - The requester should avoid using a long batch containing more than 10 elements or one that requires more than 3 minutes to complete, as we found during the experiments in this research that while longer batches increase workers' learning level, they also increase frustration and the level at which workers abandon the task, which ultimately leads to lower performance.
- **For the execution of the task:**
  - Requesters may recognise that a task with a specific setup can produce consistent results even when targeting new workers every time if the task is repeated on the same platform.
  - Motivating the workers with high payment will increase the probability of finishing the entire job in a shorter amount of time (as this is related to the number of workers needed) but it may not necessarily lead to higher performance.
- **For the selection of the platforms, channels, and workers for the task:**
  - If requesters wish to use different platforms for the same task, they will have to spend some time to make sure that the task is presented in the same way on different platforms and pay all workers equally.
  - Requesters may consider channel popularity and that some channels are more popular than others and platform preference varies from country to country. For example, channels in Group 1 (Clixsense, Elite, and NeoBux) are more prevalent in countries with lower employment rate and lower cost of living such as Venezuela, Turkey, and Ukraine; thus, workers in these channels accept working even on tasks that offer minimum reward as it still contributes to their income, and requesters will receive a complete batch for the task faster than using all workers on the F8 without specifying particular channels.

- If a task that requires workers with a specific gender, it should be noted that females tend to prefer channels in Group 2 (InstaGC, Swagbucks), and they recorded a high level of trust and experience; however, we noticed two disadvantages in using these two channels. First, the total completion time for a task in these channels is low as workers come from the developed countries such as the US, UK, and Canada. Thus, they do not expect a task that does not have an attractive objective or provides a high payment. Second, a high commission rate taken by the channels should be considered by the requester if they want to attract the attention of workers in these channels.
- For a task that requires unique workers - working on the task one time only -, the requester may avoid using Group 2 channels where workers are more likely to use more than one channel, and the possibility of having dual accounts from multiple channels will provide them with access to the same task on F8.
- Using workers from level 1 will open the opportunity for newcomers to perform the task and reduce the knowledge bias in answers which appears with workers at a higher level who have been doing similar tasks for a long time.

## 7.4 Future work

The work in this thesis opens many directions for possible future work.

In Chapter 4, we examined the effects of class ordering and balance in a relevance judgment crowdsourcing task. The findings of this study show promising results for priming workers by presenting relevant documents first in the batch. This is a positive and useful result that could be used in real IR evaluation settings, e.g., based on pooling documents retrieved by different IR systems. However, as in real settings the workers are labelling new data (where the relevant class is unknown), these results cannot be directly applied. However, it still might be possible to order documents by attributes indicating their relevance (e.g. retrieval rank, number of IR systems retrieving the document, etc.). Therefore, a good practice is to present to the workers the documents with a higher probability of being relevant early in the batch, similarly to Damessie et al. (2018), but keeping our configuration for the document order. It is possible to use around 20% of the relevant documents of the whole dataset as gold data and show it at the beginning of the batch as a priming technique rather than using examples in the instructions, thus the learning level will be enhanced by practice and it will enable workers to make comparisons when looking at the subsequent documents in the batch. We look forward to examining the stability of reproducing the current results on a different platform, using the same order and balance of the batch and comparing the results. Moreover, a possible new direction for this study is to examine the effect of order and balance on new datasets, such as labelling images and sentiment analysis of tweets.

In Chapter 5, we studied the possibility of achieving the same results when repeating the same crowdsourcing task over a different time scale and by reproducing the same task on different platforms and targeting new crowd workers every time. The findings of this study show a

high level of consistency for repeating and reproducing a task over time and over different platforms but one should bear in mind that a more difficult task could produce inconsistent results when reproduced over different platforms. Furthermore, when we managed to equalise the payment and add a bonus for high-quality results as a motivation, we achieved the same effects on both platforms and thus, reproducible crowdsourcing tasks. Our future work will examine task *repeatability* and *reproducibility* in advanced realistic settings in terms of rejection and quality control. Furthermore, we will expand the work by considering different crowdsourcing tasks, such as Entity tagging and sentiment analysis - subject to the availability of the dataset documentation-. Additionally, while we presented several crowdsourcing platforms in Section 2.6, we will include more in the comparison, such as *Microworkers* and *Prolific Academia*.

In Chapter 6, we presented up-to-date information about workers using the F8 platform. Using a survey task that consisted of subjective and some open questions, we studied the variation of the amount of payment in the top 5 channels used in F8. It would be insightful to study the economic changes that have occurred in the past four years and the rise of the unemployment rate in some countries, which affected the direction of the labour force to obtain additional work, such as crowdsourcing jobs, regardless of the high commission rate that some channels are taking. We aim to extend this study by surveying workers from other channels in F8. Furthermore, we will interview workers who have already participated in this survey and expressed their acceptance to proceed. Through the planned interview, we will design questions that will look deeply into some of the variations that we found in the results presented in this chapter: (1) the reasons behind choosing to work for a specific channel, (2) the gender variation in using some channels more than others, (3) the effect of payments on workers life, and (4) the experience of using multiple channels to access crowdsourcing tasks on the F8 platform.

*We hope this work can contribute to the field of crowdsourcing and enrich the methods of designing the optimal tasks for workers that will deliver a satisfying experience to the requester.*









## Ethical Consent Forms

conferences and journals. We will make the anonymised collection of data publicly available to enable further research such as training and evaluating information systems. This anonymized data may also be used by others outside of the project for the purposes of evaluating the performance of systems.

The data recorded will be securely stored on password protected computers at Sheffield University. A copy will be stored on the researcher's university laptop for analysis purposes and it will be backed up on an external drive kept in a locked drawer in the Information Retrieval Lab at Sheffield.

#### Will my participation be confidential?

All the information that we collect about you during the course of the research will be kept strictly confidential, and will be stored without any personal identifying information. Each participant will be anonymised and identified by a randomly chosen code, e.g. P01, P25. You will not be identifiable in any reports, publications, or presentations. All data you provide through the online experiment will be stored securely as described above.

#### What will happen to the results of the research project?

The results of the research will be included in academic papers, presentations and reports which will be publicly available. If you wish to be given a copy of any reports or publications based on the research, please email us to add you to our circulation list. We will make the anonymised collection of data publicly available for further research.

- I confirm that I have read and understand the description of the research project, and that I have had an opportunity to ask questions about the project.
- I understand that my participation is voluntary and that I am free to withdraw at any time before three months after the collection of the data without any negative consequences.
- I understand that if I withdraw I can request for the data I have already provided to be deleted, however this might not be possible if the data has already been a nonymised or findings published.
- I understand that I may decline to answer any particular question or questions, or to do any of the activities.
- I understand that my responses will be kept strictly confidential, that my name or identity will not be linked to any research materials, and that I will not be identified or identifiable in any report or reports that result from the research, unless I have agreed otherwise.
- I give permission for all the research team members to have access to my responses.
- I give permission for the research team to re-use my data for future research as specified above.
- I agree to take part in the research project as described above.

Date:

Wed Jun 12 2019 22:06:13 GMT+0100 (British Summer Time)

**Note: If you have any difficulties with, or wish to voice concern about, any aspect of your participation in this study, please contact Dr Paul Reilly, Research Ethics Coordinator, Information School, The University of Sheffield (jschool\_ethics@sheffield.ac.uk), or the University Registrar and Secretary.**

### The University of Sheffield Information School

FashionBrain: Understanding Europe's Fashion Data Universe

#### Researchers

Dr Alessandro Checco  
Information School, The University of Sheffield, Regent Court, 211 Portobello, Sheffield, S1 4DP, UK Telephone: +44 (0) 114 222 2637  
Email: a.checco@sheffield.ac.uk

#### Purpose of the research

The purposes of this research are to improve the quality of Human Computation applications by developing better quality and less expensive crowdsourcing techniques; to develop novel crowdsourcing quality assurance techniques; to understand dimensions like bias and subjectivity and to deal with it by developing priming and training approaches for crowdsourcing.

#### Who will be participating?

We are inviting all adults (people aged 18 and over) registered on a crowdsourcing platform such as Amazon Mechanical Turk and Crowdflower to participate in our study.

#### What will you be asked to do?

We will ask you to complete the online experiment, which is described in the task instructions. You will be presented with a data item (for example an image, a tweet, or a document) and some questions. You will then need to answer the questions as described in the instructions.

Please note that participation is entirely voluntary and that you can withdraw from the study at any time. You can request the deletion of your data within 3 months from the data collection. After this point your data will have been anonymised, so that it will not be possible to identify your data in order to delete it. To withdraw from participation or to request data deletion please contact the Research Ethics Coordinator via the email provided at the end of this document.

#### What are the potential risks of participating?

The risks of participating are the same as those experienced in everyday life.

#### What data will we collect?

We will collect some demographic information about you to enable a picture of our participant group as a whole. We will track various browser events related to your activity on our study's web page, including the answers you provide, how long you spend on each task, the mouse clicks you make and the quantity of scrolling you do on each page. We will record the answers you provide to the questions after providing each answer.

#### What will we do with the data?

We will analyse the data to understand the process people go through when they complete crowdsourcing tasks, the factors that can influence this process and whether crowdsourcing is a viable means for big data processing. The data will be used for the purposes of academic research by the project team, with results being published in reputable

anonymized data may also be used by others outside of the project for the purposes of evaluating the performance of systems.

The data recorded will be securely stored on password protected computers at Sheffield University. A copy will be stored on the researcher's university laptop for analysis purposes and it will be backed up on an external drive kept in a locked drawer in the Information Retrieval Lab at Sheffield.

#### **Will my participation be confidential?**

All the information that we collect about you during the course of the research will be kept strictly confidential, and will be stored without any personal identifying information. Each participant will be anonymised and identified by a randomly chosen code, e.g. P01, P25. You will not be identifiable in any reports, publications, or presentations. All data you provide through the online experiment will be stored securely as described above.

#### **What will happen to the results of the research project?**

The results of the research will be included in academic papers, presentations and reports which will be publicly available. If you wish to be given a copy of any reports or publications based on the research, please email us to add you to our circulation list. We will make the anonymised collection of data publicly available for further research.

- I confirm that I have read and understand the description of the research project, and that I have had an opportunity to ask questions about the project.
- I understand that my participation is voluntary and that I am free to withdraw at any time without any negative consequences.
- I understand that if I withdraw I can request for the data I have already provided to be deleted, however this might not be possible if the data has already been anonymised or findings published.
- I understand that I may decline to answer any particular question or questions, or to do any of the activities.
- I understand that my responses will be kept strictly confidential, that my name or identity will not be linked to any research materials, and that I will not be identified or identifiable in any report or reports that result from the research, unless I have agreed otherwise.
- I give permission for all the research team members to have access to my responses.
- I give permission for the research team to re-use my data for future research as specified above.
- I agree to take part in the research project as described above.

Date: Wed Jun 12 2019 22:19:21 GMT+0100 (British Summer Time)

**Note: If you have any difficulties with, or wish to voice concern about, any aspect of your participation in this study, please contact Dr Jo Bates, Research Ethics Coordinator, Information School, The University of Sheffield (jschool\_ethics@sheffield.ac.uk), or the University Registrar and Secretary.**

## **The University of Sheffield Information School**

BetterCrowd: Human Computation for Big Data

### **Researchers**

Dr Gianluca Demartini, Dr Alessandro Checco  
Information School, The University of Sheffield, Regent Court, 211 Portobello, Sheffield, S1  
4BP, UK Telephone: +44 (0) 114 222 2637  
Email: g.demartini@sheffield.ac.uk

### **Purpose of the research**

The purpose of this research is the improvement of Human Computation (HC) quality and scalability for Big Data processing. The project deals with the fundamental scientific challenge of understanding how humans and machines can better interact and collaborate in computation problems. The findings will make possible to deal with more complex Big Data analytics problems. While most current Big Data solutions focus on volume, this project aims to improve HC to make it applicable to the analysis of heterogeneous data with variable quality.

### **Who will be participating?**

We are inviting all adults (people aged 18 and over) and registered on a crowdsourcing platform such as CrowdFlower to participate in our study.

### **What will you be asked to do?**

We will ask you to complete the online experiment, which is described in the task instructions. You will be presented with a data item (for example an image, a tweet, or a document) and some questions. You will then need to answer the questions as described in the instructions.

Please note that participation is entirely voluntary and that you can withdraw from the study at any time.

### **What are the potential risks of participating?**

The risks of participating are the same as those experienced in everyday life.

### **What data will we collect?**

We will collect some demographic information about you to enable a picture of our participant group as a whole. We will track various browser events related to your activity on our study's web page, including the answers you provide, how long you spend on each task, the mouse clicks you make and the quantity of scrolling you do on each page. We will record the answers you provide to the questions after providing each answer.

### **What will we do with the data?**

We will analyse the data to understand the process people go through when they complete crowdsourcing tasks, the factors that can influence this process and whether crowdsourcing is a viable means for big data processing. The data will be used for the purposes of academic research by the project team, with results being published in reputable conferences and journals. We will make the anonymised collection of data publicly available to enable further research such as training and evaluating information systems. This



# B

## NASA Task Load Index questionnaire (NASA-TLX)











## GUI of the Relevance Judgment Task



## GUI of the relevance judgments task

← → ↻ ↑

### Classify The Relevance Of 10 Documents

Instructions ▾

#### Overview

We ask your help to categorize some documents according to their relevance to a given topic.

#### Steps

1. Read and Understand the topic you should judge documents against.
2. Review the 10 documents given below, each document consists of one or two pages maximum.
3. Select the level of relevance for each document.
4. At the end find 5 questions to give feedback about your experience with the job.

#### Rules & Tips

- Make sure you review and understand the information we've provided about the topic before making your decision.
- You must answer all 10 documents to complete the task.
- Note that the maximum time to complete the job is 30 minutes and it should take 15 - 20 minutes normally to be complete.

#### Relevance Definitions

- You should choose **Relevant** if:
  - The content of the document is about the topic.
  - The information in the document is highly or moderately relevant to the topic.
- Choose **Not Relevant** if:
  - The content of the document clearly doesn't match the topic.
  - The information in the document is irrelevant.

#### The Topic

##### Heroic Acts

- Find accounts of selfless heroic acts by individuals or small groups for the benefit of others or a cause.
- Relevant documents will contain a description of specific acts. General statements concerning heroic acts are not relevant.

← → ↻ ↑

Please read the following communication about our data collection policy.  
(required)

I agree and wish to work on this task

- Please make your judgement for all 10 documents in this page

Document 1: [Click Here to See the Document](#)

Is the document Relevant to the topic? (required)

Relevant

Not Relevant

Document 2: [Click Here to See the Document](#)

Is the document Relevant to the topic? (required)

Relevant

Not Relevant

Document 3: [Click Here to See the Document](#)

Is the document Relevant to the topic? (required)

Relevant

Not Relevant

Document 4: [Click Here to See the Document](#)

Is the document Relevant to the topic? (required)

Relevant

Not Relevant

Document 5: [Click Here to See the Document](#)

Is the document Relevant to the topic? (required)

Relevant

Not Relevant

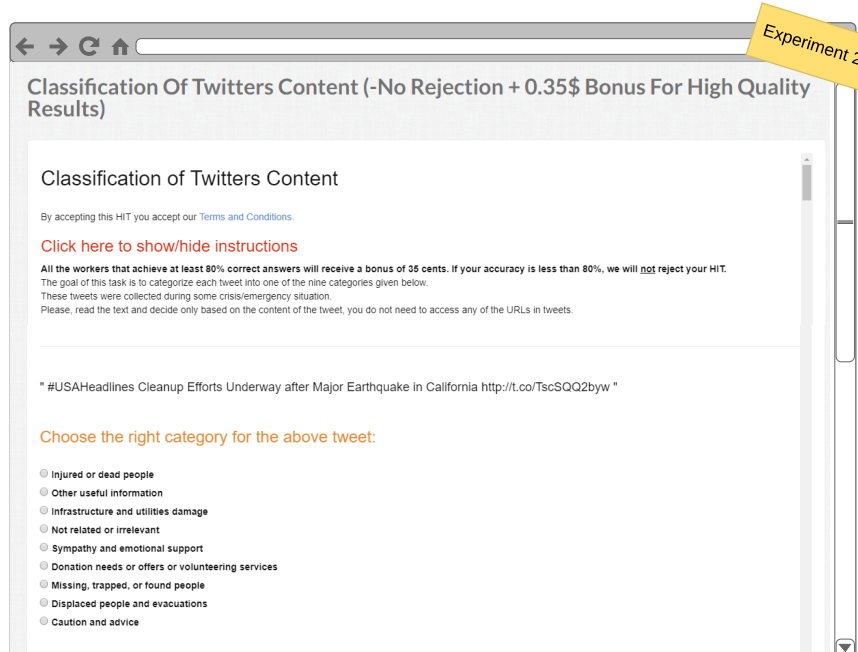
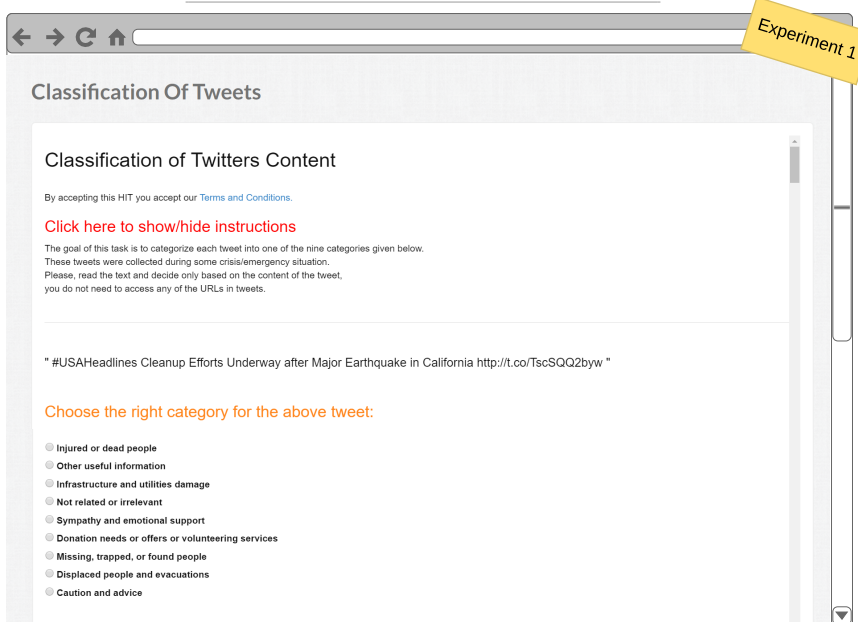


# D

GUI of the classification task (Dataset 1)



GUI of the classification task (Dataset 1)







# E

GUI of the classification task (Dataset 2)

← → ↻ ⬆

## Classify The Relevance Of Reviews Related To Fashion Items 1

### Classify Sizing/Fitting Issue In Item Reviews

By accepting this HIT you accept our [Terms and Conditions](#).

[Click here to show/hide instructions](#)

In this job, you will be presented with reviews related to fashion items.

Each review can be classified to one of the following three classes:

**Size aspect**

In these examples, the comment is expressing a feedback about the item's size. When the sentiment is negative, the item's size is either too large or too small compared to the regular one.

Description of size can be labels M, L, XL and numbers for apparel; numbers for shoes (43, 44, ...); children size usually talk about age.

- "Fantastic! Shipping was fast. Can't wait to shop here again! These shoes are around one size larger than normal."
- "Large fit."
- "bigger than expected"
- "this is not XL"
- "way too HUGE"
- "Wrong sizes"
- "lovely top but way too big"
- "the size is OK"
- "I recommend to buy the this shoes one size bigger"

**Fit aspect**


In these examples, the comment is expressing a feedback about the item's fit, but it does not specify a size issue. The problem could be related to a comfortability regarding fitting aspect.

- "doesn't fit"
- "The shoes really rub on your ankle"
- "Perfect fit"

**No issue with size or fit**

In these examples, the comments are not related to the sizing or fitting aspect. Note that delivering the wrong size or missing a size in the inventory is **not a size aspect**

- "boots arrived 6 days after ordering which was fine - but they had sent me the wrong style and the wrong size !!!" - (Related to delivery)
- "great shoes" - (General remark)
- "Nice Perfume" - (Data error)
- "the sleeves are too long" - (Data error)
- "Very comfortable shoes" - (Comfortability remark)
- "loved the jacket but a shame no larger sizes were available" - (Data error)
- "There is no size available on the web site" - (Shopping issue)



**Lovely shoes by Ricosta**  
 "As always Ricosta has come up trumps! Trying to find well-made shoes with an excellent fit has become quite a mission; however, my daughter loves these, they are comfy and even have in sole that can be taken out and washed. Highly recommended! "

Choose the right category for the above review:

- Size aspect
- Fit aspect
- No issue with size or fit

## Classification Of Fashion Reviewers 5 (No Rejection, 0.35\$ Bonus For High Quality Results)

### Classify Sizing/Fitting Issue In Item Reviews

By accepting this HIT you accept our [Terms and Conditions](#).

[Click here to show/hide instructions](#)

All the workers that achieve at least 80% correct answers will receive a bonus of 35 cents. If your accuracy is less than 80%, we will not reject your HIT.

In this job, you will be presented with reviews related to fashion items.

Each review can be classified to one of the following three classes:

#### Size aspect

In these examples, the comment is expressing a feedback about the item's size. When the sentiment is negative, the item's size is either too large or too small compared to the regular one.

Description of size can be labels M, L, XL, and numbers for apparel; numbers for shoes (43, 44, ...); children size usually talk about age.

- o "Fantastic! Shipping was fast. Can't wait to shop here again! These shoes are around one size larger than normal."
- o "Large fit"
- o "bigger than expected"
- o "this is not XL"
- o "way too HUGE"
- o "Wrong sizes"
- o "lovely top but way too big"
- o "the size is OK"
- o "I recommend to buy the this shoes one size bigger"

#### Fit aspect

In these examples, the comment is expressing a feedback about the item's fit, but it does not specify a size issue. The problem could be related to a comfortability regarding fitting aspect.

- o "doesn't fit"
- o "The shoes really rub on your ankle"
- o "Perfect fit"

#### No issue with size or fit

In these examples, the comments are not related to the sizing or fitting aspect. Note that delivering the wrong size or missing a size in the inventory is **not a size aspect**

- o "boots arrived 6 days after ordering which was fine - but they had sent me the wrong style and the wrong size III" - (Related to delivery)
- o "great shoes" - (General remark)
- o "Nice Perfume" - (Data error)
- o "the sleeves are too long" - (Data error)
- o "Very comfortable shoes" - (Comfortability remark)
- o "loved the jacket but a shame no larger sizes were available" - (Data error)
- o "There is no size available on the web site" - (Shopping issue)



**Lovely shoes by Ricosta**  
"As always Ricosta has come up trumps! Trying to find well-made shoes with an excellent fit has become quite a mission; however, my daughter loves these, they are comfy and even have in insole that can be taken out and washed. Highly recommended!"

Choose the right category for the above review:

- Size aspect
- Fit aspect
- No issue with size or fit



# F

GUI of the classification task (Dataset 3)



GUI of the classification task (Dataset 3)

Experiment 1

### Classify The Relevance Of Documents 1

#### Classify relevance of web documents

By accepting this HIT you accept our [Terms and Conditions](#).

[Click here to show/hide instructions](#)

In this job, you will be presented with documents with some content and you should read it and decide if it relevant to the giving topic or not.

**Topic: Industrial Waste Disposal**

**Description:**  
How is the disposal of industrial waste being accomplished by industrial management throughout the world?

**Narrative:**  
Documents that discuss the disposal, storage, or management of industrial waste—both standard and hazardous—are relevant. However, documents that discuss disposal or storage of nuclear or radioactive waste, or the illegal shipment or dumping of waste to avoid legal disposal methods are not relevant.

10- Please have a look at the following document:

A cement plant near Gorman, which is under review by state health authorities because it burns hazardous waste, has agreed to pay a \$5,000 fine for burning illegal fuel last fall, Kern County air quality officials said Tuesday.

The penalty settlement between National Cement Co.'s Los Robles plant and the Kern County Air Pollution Control District was half the amount originally asked by air pollution officials, and far less than the maximum \$43,000 that could have been assessed.

The company was accused of burning a petroleum-based material known as carbon black during a 42-day period last September and October. The company has permits to burn hazardous wastes as fuel in its cement-making process, but carbon black was not one of its permitted fuels.

A spokesman for the state Department of Health Services said tests showed the material apparently did not pose a health hazard to neighbors. But the spokesman added that the department is still evaluating broader health concerns about the plant's emissions from hazardous waste incineration.

Is this document relevant to the topic?

Relevant

Not Relevant

reset submit

Experiment 2

### Classification Of Web Documents 1 (No Rejection , 0.35\$ Bonus For High Quality Results)

#### Classify relevance of web documents

By accepting this HIT you accept our [Terms and Conditions](#).

[Click here to show/hide instructions](#)

**All the workers that achieve at least 80% correct answers will receive a bonus of 35 cents. If your accuracy is less than 80%, we will not reject your HIT.**

In this job, you will be presented with documents with some content and you should read it and decide if it relevant to the giving topic or not.

**Topic: Industrial Waste Disposal**

**Description:**  
How is the disposal of industrial waste being accomplished by industrial management throughout the world?

**Narrative:**  
Documents that discuss the disposal, storage, or management of industrial waste—both standard and hazardous—are relevant. However, documents that discuss disposal or storage of nuclear or radioactive waste, or the illegal shipment or dumping of waste to avoid legal disposal methods are not relevant.

10- Please have a look at the following document:

A cement plant near Gorman, which is under review by state health authorities because it burns hazardous waste, has agreed to pay a \$5,000 fine for burning illegal fuel last fall, Kern County air quality officials said Tuesday.

The penalty settlement between National Cement Co.'s Los Robles plant and the Kern County Air Pollution Control District was half the amount originally asked by air pollution officials, and far less than the maximum \$43,000 that could have been assessed.

The company was accused of burning a petroleum-based material known as carbon black during a 42-day period last September and October. The company has permits to burn hazardous wastes as fuel in its cement-making process, but carbon black was not one of its permitted fuels.

A spokesman for the state Department of Health Services said tests showed the material apparently did not pose a health hazard to neighbors. But the spokesman added that the department is still evaluating broader health concerns about the plant's emissions from hazardous waste incineration.

Is this document relevant to the topic?

Relevant

Not Relevant

reset submit





G

GUI of the Survey workers task

- Hispanic or Latino
- Black or African American
- Native American or American Indian
- Asian / Pacific Islander
- Other

**4- Education: What is the highest degree or level of school you have completed? (required)**

- No schooling completed
- Nursery school to 8th grade
- Some high school, no diploma
- High school graduate, diploma or the equivalent
- Some college credit, no degree
- Trade/technical/vocational training
- Bachelor's degree
- Master's degree
- Professional degree
- Doctorate degree

If currently enrolled, highest degree received.

**5- What is your marital status? (required)**

- Single, never married
- Married or domestic partnership
- Widowed
- Divorced
- Separated

**6- What is your employment status? (required)**

- Employed for wages
- Self-employed
- Out of work and looking for work
- Out of work but not currently looking for work
- A homemaker
- A student
- Military
- Retired
- Unable to work

**As you are working in Figure Eight, please answer the following questions.**

**7- How long have you been working on crowdsourcing tasks? (required)**

- less than a year
- 1-2 year
- 2-3 years
- over 3 years

# Survey Workers Per Channel (Instagc.3.2)

Instructions ▾

**Who we are:** A group of researchers who want to make online job platforms more fair, ethical, and transparent environments. We aim at supporting virtuous online job services which behave honestly and professionally with their workers, spotting out those services that make huge profits by exploiting human labour take advantage of the weak negotiating power of crowdworkers. Participating in this job you can give us a vital contribution to pursuing our (and probably also yours) objective.

**Instructions:** This task consists of a survey that takes you almost ten minutes and asks for a few simple questions about your experience performing tasks. It is going to be easy, and just by being genuine you will complete it smoothly.

**Disclaimer:** Please, read these instructions carefully before deciding whether to complete the task. Note that there are some checks throughout the task, and if you do not perform these correctly you will not be able to complete the task and get paid. The data from this task is being gathered for research purposes. No personally identifying information is recorded. Participation is entirely voluntary, and you are free to discontinue at any point. The task has been approved by the Ethics Committee ([https://www.alessandrochecco.online/static/Ethics\\_info\\_consent.html](https://www.alessandrochecco.online/static/Ethics_info_consent.html)) of The University of Sheffield (<https://www.sheffield.ac.uk/>).

Thank you in advance for your collaborations.

Let us begin.

## General information about you

**1- Gender: Are you... (required)**

- Male
- Female
- Prefer not to say

**2- Age: What is your age? (required)**

- Under 12 years old
- 12-17 years old
- 18-24 years old
- 25-34 years old
- 35-44 years old
- 45-54 years old
- 55-64 years old
- Over 64 years old

**3- Ethnic origin: Please specify your ethnicity. (required)**

- White

## Keep in touch

15- Would you like to be possibly contacted for a paid Skype interview in the future? (required)

- Yes  
 No

Is there any feedback, opinion, or question about this survey you want to share with us? (required)

Thank You.

Test Validators

8- Which type of device do you use for performing crowdsourcing tasks? (required)

- Mobile (smartphone or tablet)  
 Desktop (personal computer or laptop)  
 Other

9- Is crowdsourcing your main source of income? (required)

- Yes  
 No

10- How much it contributes in your life income? (required)

- 0 % - 20 %  
 21 % - 40 %  
 41 % - 60 %  
 61 % - 80 %  
 81 % - 100 %

11- How much do you earn from the crowdsourcing jobs (dollars per months)? (required)

12- Are you getting rewarded for completing this job (e.g., monetary payment, voucher, game credits, or bonuses)? (required)

- Yes, I get rewarded  
 No, I am a volunteer  
 Other, I am not sure

Can you describe where you can find this kind of jobs in InstaGC ? (required)

Do you use other websites to access similar jobs? (required)

- Yes  
 No

Which are the criteria you care the most when choosing a job to perform? (required)

- Reward  
 The most interesting  
 Time required for completion  
 Difficulty  
 Other  
 The platform chooses the task form me

TABLE G.1: The results of the Survey workers task for top 5 channels

General information	Channels				
	Clixsense	Elite	NeoBux	InstaGC	Swagbucks
1 - Gender:					
- Male	75 %	73 %	70 %	32 %	38 %
- Female	25 %	15 %	30 %	68 %	62 %
- Prefer not to say		2 %			
2 - Age:					
- 12-17 years old					
- 18-24 years old	17 %	27 %	30 %	10 %	
- 25-34 years old	43 %	38 %	43 %	27 %	28 %
- 35-44 years old	27 %	25 %	15 %	32 %	33 %
- 45-54 years old	8 %	7 %	10 %	15 %	22 %
- 55-64 years old	5 %	2 %	3 %	13 %	15 %
- Over 64 years old		3 %		3%	2 %
3 - Ethnicity:					
- White	55 %	42 %	28 %	73 %	80 %
- Hispanic or latino	37%	43 %	67 %	5 %	2 %
- African			3 %		
- Asian pacific islander	7 %	12 %	2 %	12 %	12 %
- Other	2 %	5 %	2 %	7 %	3 %
4 - Education level:					
- bachelors degree	25 %	45 %	40 %	37 %	32 %
- masters degree	25 %	12 %	7 %	8 %	12 %
- high school graduate diploma	17 %	18 %	15 %	18 %	12 %
- professional degree	15 %	3 %	7 %	2 %	5 %
- some college credit no degree	10 %	13 %	17 %	20 %	20 %
- trade technical vocational training	5 %	7 %	10 %	10 %	15 %
- doctorate degree	2 %			2 %	2 %
- some high school no diploma	2 %	3 %	2 %	3 %	2 %
- no schooling completed			2 %		
5- Material status:					
- Single	22 %	52 %	60 %	43 %	43 %
- Married or domestic relationship	75 %	42 %	37 %	52 %	57 %
- Widowed	2 %	3 %	2 %		
- Divorced	2 %	5 %		5 %	
- Separated			3 %		
6 - Employment status:					
- Employee for wage	45 %	18 %	25 %	58 %	63 %
- Self-employed	28 %	37 %	38 %	8 %	10 %
- Out of work and looking for one	12 %	15 %	8 %	8 %	5 %
- Out of work and not currently looking for one	2 %	7 %	8 %	3 %	3 %
- A home worker	8 %	8 %	3 %	15 %	8 %
- A student	5 %	15 %	13 %	2 %	3 %
- Retired		2 %	8 %	3 %	5 %
- Unable to work			%	2 %	2 %

<b>As working in Figure Eight:</b>					
7- Working experience :					
- less than year	27 %	22 %	15 %	5 %	5 %
- 1-2 years	37 %	33 %	45 %	22 %	10 %
- 2-3 years	20 %	18 %	30 %	12 %	15 %
- Over 3 years	17 %	28 %	12 %	62 %	70 %
8- Type of device used for performing the task:					
- Mobile (smartphone or tablet)	5 %	3 %	2 %	5 %	7 %
- Desktop (personal computer or laptop)	93 %	95 %	98 %	93 %	90 %
- Other					
9- Is crowdsourcing the main source of income?					
- Yes	47 %	72 %	70 %	3 %	
- No	53 %	30 %	32 %	97 %	100 %
10- How much it contributes in your life income?					
- 0%-20%	33 %	13 %	10 %	83 %	95 %
- 21%-40%	17 %	22 %	17 %	12 %	5 %
- 41%-60%	15 %	18 %	13 %	2 %	
- 61%-80%	8 %	15 %	23 %		
- 81%-100%	27 %	33 %	38 %	3 %	
11- How much earn from the crowdsourcing jobs (\$/months)?	Avg. 116.6	Avg. 104.6	Avg. 77.5	Avg. 61.5	Avg. 18.6
12- Are you getting rewarded for completing this job (e.g., monetary payment, voucher, game credits, or bonuses)?					
- Yes, I get rewarded	87 %	79 %	90 %	92 %	97 %
- No, I am a volunteer	5 %	18 %	8 %	7 %	2 %
- Other, I am not sure	7 %	3 %	2 %		
12.1- When do you get rewarded?					
- Every single job	70 %	69 %	79 %	87 %	89 %
- Every multiple jobs completed	16 %	10 %	10 %		7 %
- Other, I am not sure			2 %	5 %	2 %
12.2- How do you get rewarded for this job?					
- Monetary payment (direct in you bank account or similar)	34 %	33 %	43 %	10 %	
- Monetary payment (you can somehow redeem later)	48 %	43 %	46 %	70 %	56 %
- Cryptocurrency (e.g., Bitcoins, Litecoins, Ethereum, others)					
- Vouchers				2 %	5 %
- Other, credits for games or points for collecting voucher later on					

12.2.1- Can you specify the amount of your payment? (for example, write 2 if you are getting 0.02\$ and write 50 if you are getting 0.50\$)?	Avg. 35	Avg. 50	Avg. 37	Avg. 27	Avg. 20
12.2.1- How long do you have to wait to be able to withdraw you payment?					
- In a few minutes	10 %	18 %	56 %	61 %	18 %
- Less than one hour	3 %		2 %	5 %	3 %
- Less than one day	2 %	2 %	2 %	2 %	
- In a few days	64 %	54 %	28 %	5 %	23 %
- I do not know	3 %	2 %	2 %	8 %	11 %
<b>As crowd-worker:</b>					
Do you use other websites to access similar jobs?					
- Yes	8 %	23 %	13 %	13 %	17 %
- No	92 %	78 %	88 %	87 %	83 %
Which are the criteria you care the most when choosing a job to perform?					
- Reward	80 %	87 %	88 %	93 %	85 %
- The most interesting	50 %	47 %	40 %	50 %	28 %
- Time required for complement	62 %	52 %	65 %	53 %	58 %
- Difficulty	40 %	40 %	50 %	50 %	53 %
- The platform chooses the task form me		3 %	7 %		3 %
- Other		2 %	2 %		2 %
<b>Keep in touch:</b>					
Would you like to be possibly contacted for a paid Skype interview in the future?					
- Yes	67 %	55 %	78 %	50 %	27 %
- No	33 %	47 %	23 %	50 %	73 %

## Bibliography

- Ahmad, Salman, Alexis Battle, Zahan Malkani, and Sepander Kamvar (2011). "The Jabberwocky Programming Environment for Structured Social Computing". In: *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11)*, p. 53.
- Alagarai Sampath, Harini, Rajeev Rajeshuni, and Bipin Indurkha (2014). "Cognitively Inspired Task Design to Improve User Performance on Crowdsourcing Platforms". In: *Proceedings of the 32nd annual ACM conference on Human factors in computing systems (CHI' 14)*. Toronto, pp. 3665–3674.
- Ali, Aida, Siti Mariyam Shamsuddin, and Anca L. Ralescu (2015). "Classification with class imbalance problem: A Review". In: *International journal of advances in soft computing and its applications* 7.3, pp. 176–204.
- Allahbakhsh, Mohammad, Boualem Benatallah, Aleksandar Ignjatovic, Hamid Reza Motahari-Nezhad, Elisa Bertino, and Schahram Dustdar (Mar. 2013). "Quality Control in Crowdsourcing Systems: Issues and Directions". In: *IEEE Internet Computing* 17.2, pp. 76–81.
- Alonso, Omar (2013). "Implementing crowdsourcing-based relevance experimentation: An industrial perspective". In: *Information Retrieval* 16.2, pp. 101–120.
- (2015). "Practical Lessons for Gathering Quality Labels at Scale". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*, pp. 1089–1092.
- Alonso, Omar and Ricardo Baeza-Yates (2011). "Design and implementation of relevance assessments using crowdsourcing". In: *Advances in information retrieval*, pp. 153–164.
- Alonso, Omar and Stefano Mizzaro (2012). "Using crowdsourcing for TREC relevance assessment". In: *Information Processing and Management* 48.6, pp. 1053–1066.
- Ambati, Vamshi, Stephan Vogel, and Jg Carbonell (2011). "Towards Task Recommendation in Micro-Task Markets." In: *Human Computation*, pp. 1–4.
- Anand, John (2018). *Figure Eight (CrowdFlower) Tasks - The Definitive Guide*.
- Andersen, David J and Richard R Lau (2018). "Pay Rates and Subject Performance in Social Science Experiments Using Crowdsourced Online Samples". In: *Journal of Experimental Political Science*, pp. 1–13.
- Andrásfalvy, Bertalan K., Mark A. Smith, Thilo Borchardt, Rolf Sprengel, and Jeffrey C. Magee (Oct. 2003). "Impaired regulation of synaptic strength in hippocampal neurons from GluR1-deficient mice." In: *The Journal of physiology* 552.Pt 1, pp. 35–45.

- André, Paul, Robert E Kraut, and Aniket Kittur (2014). "Effects of Simultaneous and Sequential Work Structures on Distributed Collaborative Interdependent Tasks". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 139–148.
- Archambault, Daniel, Helen Purchase, and Tobias Hoßfeld (2017). *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments: Dagstuhl Seminar 1[1] D. Archambault, H. Purchase, and T. Hoßfeld, Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments: Dagstuhl Seminar 15481, Dagstuhl Castle, Germa.* Vol. 10264, pp. 154–190.
- Assadi, Sepehr, Justin Hsu, and Shahin Jabbari (2015). "Online Assignment of Heterogeneous Tasks in Crowdsourcing Markets". In: *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing (HCOMP '15)*, pp. 12–21.
- Assis Neto, Fábio R. and Celso A.S. Santos (July 2018). "Understanding crowdsourcing projects: A systematic review of tendencies, workflow, and quality management". In: *Information Processing and Management* 54.4, pp. 490–506.
- Bachrach, Yoram, Thore Graepel, Tom Minka, and John Guiver (2012). "How To Grade a Test Without Knowing the Answers — A Bayesian Graphical Model for Adaptive Crowdsourcing and Aptitude Testing". In: *Proceedings of the 29th International Conference on Machine Learning (ICML '12)*, pp. 1183–1190.
- Barowy, Daniel W, Emery D Berger, and Andrew Mcgregor (2012). "AUTOMAN: A Platform for Integrating Human-Based and Digital Computation". In: *ACM SIGPLAN Notices* 47.10, pp. 639–654.
- Bentley, Frank R, Nediya Daskalova, and Brooke White (2017). "Comparing the Reliability of Amazon Mechanical Turk and Survey Monkey to Traditional Market Research Surveys". In: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*, pp. 1092–1099.
- Berners-Lee, Tim, James Hendler, and Ora Lassila (2001). "THE SEMANTIC WEB". In: *Scientific American* 284.5, pp. 34–43.
- Bi, Wei, Liwei Wang, James T Kwok, Zhuowen Tu, Hong Kong, United States, and United States (2014). "Learning to Predict from Crowdsourced Data". In: *Proceedings of the 30th Conference Uncertainty in Artificial Intelligence (UAI 2014)*, pp. 82–91.
- Blanco, Roi, Harry Halpin, Daniel M Herzig, Peter Mika, Jeffrey Pound, and Henry S Thompson (2011). "Repeatable and Reliable Search System Evaluation using Crowdsourcing". In: *Journal of Web Semantics* 21, pp. 923–932.
- Blohm, Ivo, Shkodran Zogaj, Ulrich Bretschneider, and Jan Marco Leimeister (2018). "How to manage crowdsourcing platforms effectively?" In: *California Management Review* 60.2, pp. 122–149.
- Boim, Rubi, Ohad Greenshpan, Tova Milo, Slava Novgorodov, Neoklis Polyzotis, and Wang Chiew Tan (2012). "Asking the right questions in crowd data sourcing". In: *Proceedings - International Conference on Data Engineering*, pp. 1261–1264.
- Bontcheva, Kalina, Ian Roberts, Leon Derczynski, and Dominic Rout (2014). "The GATE Crowdsourcing Plugin: Crowdsourcing Annotated Corpora Made Easy". In: *Proc. Demonstr. 14th Conf. Eur. Chapter Assoc. Comput. Linguist.* Association for Computational Linguistics, pp. 97–100.



- Borromeo, Ria Mae and Motomichi Toyama (2016). *An investigation of unpaid crowdsourcing*.
- Borromeo, Ria Mae, Thomas Laurent, Motomichi Toyama, and Sihem Amer-Yahia (2017). "Fairness and Transparency in Crowdsourcing". In: *Proceedings of the 20th International Conference on Extending Database Technology*, pp. 466–469.
- Bozzon, Alessandro, Marco Brambilla, Stefano Ceri, and Andrea Mauri (2013). "Reactive crowdsourcing". In: *Proceedings of the International Conference on World Wide Web*. Rio de Janeiro, Brazil, pp. 1–11.
- Bozzon, Alessandro, Marco Brambilla, Stefano Ceri, Andrea Mauri, and Riccardo Volonterio (2014). "Pattern-based specification of crowdsourcing applications". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 8541, pp. 218–235.
- Brabham, Daren C. (2008). "Crowdsourcing as a Model for Problem Solving: An Introduction and Cases". In: *Convergence: The International Journal of Research into New Media Technologies* 14.1, pp. 75–90.
- (2013). *Crowdsourcing*. Mit Press, p. 176.
- Bragg, Jonathan, Mausam, Daniel S Weld, Jonathan Bragg Mausam, and Daniel S Weld (2013). "Crowdsourcing Multi-Label Classification for Taxonomy Creation". In: *Proceedings of the First Conference on Human Computation and Crowdsourcing*, pp. 25–33.
- Brambilla, Marco, Stefano Ceri, Andrea Mauri, and Riccardo Volonterio (2015). "An Explorative Approach for Crowdsourcing Tasks Design". In: *Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion*, pp. 1125–1130.
- Brawley, Alice M and Cynthia L S Pury (2016). "Work experiences on MTurk: Job satisfaction, turnover, and information sharing". In: *Computers in Human Behavior* 54, pp. 531–546.
- Cai, Carrie J, Shamsi T Iqbal, and Jaime Teevan (2016). "Chain Reactions: The Impact of Order on Microtask Chains". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI' 16)*. CHI '16. New York, NY, USA: ACM, pp. 3143–3154.
- Campo, Simon à, Vassilis Javed Khan, Konstantinos Papangelis, and Panos Markopoulos (2018). *Community heuristics for user interface evaluation of crowdsourcing platforms*.
- Carr, Nicholas G (2010). *The Shallows: What the Internet Is Doing to Our Brains*. W. W. Norton.
- Catallo, Ilio and Davide Martinenghi (2017). "The Dimensions of Crowdsourcing Task Design". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 10360 LNCS, pp. 394–402.
- Chan, Kam Tong, Irwin King, and Man-Ching Yuen (2009). "Mathematical modeling of social games". In: *Proceedings of the International Conference on Computational Science and Engineering (CSE'09)*. Vol. 4. IEEE, pp. 1205–1210.
- Chandar, Praveen, William Webber, and Ben Carterette (2013). "Document features predicting assessor disagreement". In: *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '13)*, p. 745.
- Checco, Alessandro, A Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini (2017). "Let's agree to disagree: Fixing agreement measures for crowdsourcing". In: *Proceedings of the Fifth AAI Conference on Human Computation and Crowdsourcing (HCOMP 17)*, pp. 11–20.

- Chen, Chen, Xiaojun Meng, Shengdong Zhao, and Morten Fjeld (2017). "ReTool". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*, pp. 3551–3556.
- Chen, Jenny J, Natala J Menezes, Adam D Bradley, and T A North (2011). "Opportunities for Crowdsourcing Research on Amazon Mechanical Turk". In: *Human Factors* 5.3, p. 3.
- Cheng, Justin, Jaime Teevan, Shamsi T Iqbal, and Michael S Bernstein (2015). "Break it down: A comparison of macro-and microtasks". In: *CHI '15*. ACM, pp. 4061–4064.
- Chernushenko, Iurii, Felix A Gers, Alexander Löser, and Alessandro Checco (2018). *Crowd-Labeling Fashion Reviews with Quality Control*. Tech. rep.
- Chittilappilly, Anand Inasu, Lei Chen, and Sihem Amer-Yahia (Sept. 2016). "A Survey of General-Purpose Crowdsourcing Techniques". In: *IEEE Transactions on Knowledge and Data Engineering* 28.9, pp. 2246–2266.
- Clemmensen, Melanie Landvad and Pia Borlund (2016). "Order effect in interactive information retrieval evaluation: an empirical study". In: *Journal of Documentation* 72.2, pp. 194–213.
- Corney, J R, C Torres-Sánchez, P Jagadeesan, A Lynn, and W Regli (2009). "Outsourcing labour to the cloud". In: *International Journal of Innovation and Sustainable Development* 4.4, pp. 294–313.
- Crowe, Michael and Lorraine Sheppard (2012). "Mind mapping research methods". In: *Quality and Quantity* 46.5, pp. 1493–1504.
- Crump, Matthew J C, John V. McDonnell, and Todd M. Gureckis (2013). "Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research". In: *PLoS ONE* 8.3.
- Dai, Peng, Jeffrey M Rzeszotarski, Praveen Paritosh, and Ed H Chi (2015). "And Now for Something Completely Different: Improving Crowdsourcing Workflows with Micro-Diversions". In: *Crowd Work and Crowd Process*. Vancouver.
- Damessie, Tadele T and J Shane Culpepper (2016). "The Effect of Document Order and Topic Difficulty on Assessor Agreement". In: *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval (ICTIR'16)*, pp. 2–5.
- Damessie, Tadele T., J. Shane Culpepper, Jaewon Kim, and Falk Scholer (2018). "Presentation Ordering Effects On Assessor Agreement". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. Toronto, pp. 723–732.
- Dautenhahn, Kerstin (1998). "The art of designing socially intelligent agents: science, fiction, and the human in the loop". In: *Applied Artificial Intelligence* 12.7-8, pp. 573–617.
- Dawid, A P and A M Skene (1979). "Maximum likelihood estimation of observer error-rates using the EM algorithm". In: *Journal of the Royal Statistical Society Series C Applied Statistics* 28.1, pp. 20–28.
- Dawson, R and S Bynghall (2011). *Getting Results from Crowds: The Definitive Guide to Using Crowdsourcing to Grow Your Business*. Advanced Human Technologies.
- Demartini, Gianluca, Djellel Eddine Difallah, and Philippe Cudré-Mauroux (2012). "ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-Scale Entity Linking". In: *Proceedings of the 21st international conference on World Wide Web (WWW '12)*, pp. 469–478.

- Demartini, Gianluca, Djellel Eddine Difallah, Ujwal Gadiraju, and Michele Catasta (2017). "An Introduction to Hybrid Human-Machine Information Systems". In: *Foundations and Trends® in Web Science* 7.1, pp. 1–87.
- Demeester, Thomas, Robin Aly, Djoerd Hiemstra, Dong Nguyen, Dolf Trieschnigg, and Chris Develder (2014). "Exploiting user disagreement for web search evaluation". In: *Proceedings of the 7th ACM international conference on Web search and data mining (WSDM '14)*, pp. 33–42.
- Deng, Xuefei, K. D. Joshi, and Robert D. Galliers (2016). "The Duality of Empowerment and Marginalization in Microtask Crowdsourcing: Giving Voice to the Less Powerful Through Value Sensitive Design". In: *Management Information Systems Quarterly (MIS)* 40.X, pp. 1–24.
- Difallah, Djellel, Elena Filatova, and Panos Ipeirotis (2018). "Demographics and Dynamics of Mechanical Turk Workers". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. Vol. 9, pp. 135–143.
- Difallah, Djellel Eddine, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux (2015). "The Dynamics of Micro-Task Crowdsourcing". In: *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. WWW '15. New York, New York, USA: ACM Press, pp. 238–247.
- Doan, Anhai, Raghu Ramakrishnan, and Alon Y. Halevy (Apr. 2011). "Crowdsourcing Systems on the World-Wide Web". In: *Communications of the ACM* 54.4, pp. 86–96.
- Dontcheva, Mira, Robert R. Morris, Joel R. Brandt, and Elizabeth M Gerber (2014). "Combining crowdsourcing and learning to improve engagement and performance". In: *Proceedings of the 32nd annual ACM conference on Human factors in computing systems (CHI' 14)*, pp. 3379–3388.
- Doroudi, Shayan, Ece Kamar, Emma Brunskill, and Eric Horvitz (2016). "Toward a Learning Science for Complex Crowdsourcing Tasks". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. San Jose, CA, USA, pp. 2623–2634.
- Drapeau, Ryan, Lydia B Chilton, and Daniel S Weld (2016). "MicroTalk: Using Argumentation to Improve Crowdsourcing Accuracy". In: *The 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 16)*, pp. 32–41.
- Eickhoff, Carsten (2018). "Cognitive Biases in Crowdsourcing". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. Marina Del Rey: WSDM, pp. 162–170.
- Eickhoff, Carsten and Arjen P. de Vries (2013). "Increasing cheat robustness of crowdsourcing tasks". In: *Information Retrieval* 16.2, pp. 121–137.
- Eisenberg, Michael and Carol Barry (1988). "Order effects: A study of the possible influence of presentation order on user judgments of document relevance". In: *Journal of the American Society for Information Science* 39.5, pp. 293–300.
- Estellés-Arolas, Enrique and Fernando González-Ladrón-De-Guevara (Apr. 2012). "Towards an Integrated Crowdsourcing Definition". In: *Journal of Information Science* 38.2, pp. 189–200.
- Feyisetan, Oluwaseyi, Elena Simperl, Max Van Kleek, and Nigel Shadbolt (2015). "Improving Paid Microtasks through Gamification and Adaptive Furtherance Incentives". In: *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*, pp. 333–343.

- Finnerty, Ailbhe, Pavel Kucherbaev, Stefano Tranquillini, and Gregorio Convertino (2013). "Keep It Simple: Reward and Task Design in Crowdsourcing". In: *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI. CHIItaly '13*. Trento, Italy: ACM, 14:1–14:4.
- Fort, Karèn, Gilles Adda, and K Bretonnel Cohen (2011). "Amazon Mechanical Turk: Gold Mine or Coal Mine?" In: *Computational Linguistics* 37.2, pp. 413–420.
- Franklin, Michael J, Donald Kossmann, Tim Kraska, Sukriti Ramesh, and Reynold Xin (2011). "CrowdDB: answering queries with crowdsourcing". In: *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pp. 61–72.
- Gadiraju, Ujwal, Yang Jie, and Alessandro Bozzon (2017). "Clarity is a Worthwhile Quality - On the Role of Task Clarity in Microtask Crowdsourcing". In: *Proceedings of 28th ACM Conference on Hypertext and Hypermedia (HT' 17)*, pp. 5–14.
- Gadiraju, Ujwal, Ricardo Kawase, and Stefan Dietze (2014). "A Taxonomy of Microtasks on the Web". In: *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, pp. 218–223.
- Gadiraju, Ujwal, Gianluca Demartini, Ricardo Kawase, and Stefan Dietze (2015). "Human Beyond the Machine: Challenges and Opportunities of Microtask Crowdsourcing". In: *IEEE Intelligent Systems* 30.4, pp. 81–85.
- Gaikwad, S et al. (2017). "Prototype Tasks: Improving Crowdsourcing Results through Rapid, Iterative Task Design". In: *Proceedings of the fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP '17)*, pp. 2012–2017.
- Garcia-Molina, Hector, Manas Joglekar, Adam Marcus, Aditya Parameswaran, and Vasilis Verroios (Apr. 2016). "Challenges in Data Crowdsourcing". In: *IEEE Transactions on Knowledge and Data Engineering* 28.4, pp. 901–911.
- Geiger, David, Michael Rosemann, and Erwin Fieft (2011). "Crowdsourcing information systems: a systems theory perspective". In: *Proceedings of the 22nd Australasian Conference on Information Systems (ACIS 2011)*.
- Geiger, David and Martin Schader (2014). "Personalized task recommendation in crowdsourcing information systems - Current state of the art". In: *Decision Support Systems* 65.C, pp. 3–16.
- Glimm, Birte and Heiner Stuckenschmidt (2016). "15 Years of Semantic Web: An Incomplete Survey". In: *Proceedings of the Künstliche Intelligenz (KI 2016)* 30.2, pp. 117–130.
- Goel, Naman and Boi Faltings (2018). "Deep Bayesian Trust : A Dominant Strategy and Fair Reward Mechanism for Crowdsourcing".
- Halder, Buddhadeb (2014). "EVOLUTION OF CROWDSOURCING: Potential Data Protection, Privacy and Security Concerns under the New Media Age ". In: *Revista Democracia Digital e Governo Eletrônico*, pp. 377–393.
- Han, Lei, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Madalena, and Gianluca Demartini (2019). "All Those Wasted Hours: On Task Abandonment in Crowdsourcing". In: *Proceedings of the 12th ACM International Conference on Web Search and Data Mining (WSDM '19)*. Vol. 9. Melbourne, ACM, pp. 321–329.
- Hara, Kotaro, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham (2018). "A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical

- Turk". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pp. 1–14.
- Harrison, Lane, Drew Skau, Steven Franconeri, Aidong Lu, and Remco Chang (2013). "Influencing visual judgment through affective priming". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*, pp. 2949–2958.
- Hart, Sandra G and Lowell E Staveland (1988). "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research". In: *Advances in psychology* 52, pp. 139–183.
- Harter, Stephen P (1996). "Variations in relevance assessments and the measurement of retrieval effectiveness". In: *Journal of the American Society for Information Science* 47.1, pp. 37–49.
- Haug, Matthew C (2018). "Fast, Cheap, and Unethical? The Interplay of Morality and Methodology in Crowdsourced Survey Research". In: *Review of Philosophy and Psychology* 9.2, pp. 363–379.
- Hawking, David, Ellen Voorhees, Nick Craswell, and Peter Bailey (2000). "Overview of the TREC-8 Web Track". In: *Proceedings of Eighth Text Retrieval Conference (TREC8). National Institute of Standards and Technology Special Publication 500-246*, pp. 131–148.
- Hill, Robin (1998). "What sample size is "enough" in internet survey research?" In: *An Electronic Journal for the 21st Century* 6.3-4, pp. 1–10.
- Ho, Chien-Ju, Shahin Jabbari, and Jennifer Wortman Vaughan (2013). "Adaptive Task Assignment for Crowdsourced Classification". In: *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. Vol. 28, pp. 534–542.
- Ho, Chien-Ju, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan (2015). "Incentivizing High Quality Crowdsourcing". In: *The International World Wide Web Conference Committee (IW3C2)*.
- Holzinger, Andreas (2016). "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" In: *Brain Informatics* 3.2, pp. 119–131.
- Hong, Chi (2017). "Generative Models for Learning from Crowds".
- Hornbæk, Kasper, Søren S. Sander, Javier Andrés Bargas-Avila, and Jakob Grue Simonsen (2014). "Is Once Enough?: On the Extent and Content of Replications in Human-computer Interaction". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '14*. Toronto, Ontario, Canada: ACM, pp. 3523–3532.
- Horton, John and Lydia Chilton (2010). "The Labor Economics of Paid Crowdsourcing". In: *Proceedings of the 11th ACM conference on Electronic commerce (EC'10)*.
- Hovy, Dirk, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy (2013). "Learning Whom to Trust with MACE." In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Naacl-Hlt '13)*. Vol. 3. June, pp. 1120–1130.
- Howe, Jeff (2006). "The rise of crowdsourcing". In: *Wired magazine* 14.6, pp. 1–4.
- (2008). *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. 1st ed. New York, NY, USA: Random House Business Books.

- Ikediego, Henry Oluchukwu, Mustafa Ilkan, A. Mohammed Abubakar, and Festus Victor Bekun (2018). "Crowd-sourcing (who, why and what)". In: *International Journal of Crowd Science* 2.1, pp. 27–41.
- Imran, Muhammad, Prasenjit Mitra, and Carlos Castillo (2016). "Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC '16)*, pp. 1638–1643.
- Inel, Oana, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips (2014). "CrowdTruth: Machine-Human Computation Framework for Harnessing Disagreement in Gathering Annotated Data". In: *Proceedings of the 13th International Semantic Web Conference (ISWC2014)*, pp. 486–504.
- Ipeirotis, P G (2010a). "Analyzing the amazon mechanical turk marketplace". In: *XRDS: Crossroads* 17.2, pp. 16–21.
- Ipeirotis, Panagiotis G (2010b). "Demographics of mechanical turk". In:
- Ipeirotis, Panagiotis G, Foster Provost, and Jing Wang (2010). "Quality management on Amazon Mechanical Turk". In: *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10)*, p. 64.
- Irani, Lily C. and M. Six Silberman (2013). "Turkopticon: Interrupting worker invisibility in amazon mechanical turk". In: *Proceedings of the ACM Special Interest Group on Computer-Human Interaction (SIGCHI '17)*. Paris, pp. 611–620.
- Isaac, Stephen and William B Michael (1995). *Handbook in research and evaluation: A collection of principles, methods, and strategies useful in the planning, design, and evaluation of studies in education and the behavioral sciences, 3rd ed.* San Diego, CA, US: EdITS Publishers, pp. viii, 262–viii, 262.
- Jain, Ayush, Akash Das Sarma, Aditya Parameswaran, and Jennifer Widom (2017). "Understanding workers, developing effective tasks, and enhancing marketplace dynamics: a study of a large crowdsourcing marketplace". In: *Proceedings of the 43rd International Conference on Very Large Databases (VLDB '17) Endowment*. Vol. 10. 7, pp. 829–840.
- Kajino, Hiroshi and Hisashi Kashima (2011). "A Convex Formulation of Learning from Crowds". In: *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. Vol. 111. 275, pp. 231–236.
- Kakas, Antonis C. et al. (2011). *Accuracy*. Boston, MA: Springer US, pp. 9–10.
- Karger, David R, Sewoong Oh, and Devavrat Shah (2011). "Iterative Learning for Reliable Crowdsourcing Systems Accessed Iterative Learning for Reliable Crowdsourcing Systems". In: *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, pp. 1–9.
- Kazai, Gabriella, Jaap Kamps, and Natasa Milic-Frayling (2011). "Worker types and personality traits in crowdsourcing relevance labels". In: *Proceedings of the 2011 ACM international conference on Information and knowledge management*, pp. 1941–1944.
- Kazai, Gabriella and Imed Zitouni (2016). "Quality Management in Crowdsourcing using Gold Judges Behavior". In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM '16)*. San Francisco, California, pp. 267–276.
- Kendall, Maurice G. (Maurice George) (1990). *Rank correlation methods*. eng. 5th ed. London: Edward Arnold.

- Keppel, Geoffrey (1973). *Design and analysis : a researcher's handbook*. eng. Prentice-Hall series in experimental psychology. Englewood Cliffs ; (Hemel Hempstead): Prentice-Hall.
- Khattak, Faiza Khan and Ansaf Salleb-Aouissi (2011). "Quality Control of Crowd Labeling through Expert Evaluation". In: *Proceedings of the Advances in Neural Information Processing Systems 24 (NIPS 11)*.
- Kim, Hyun-Chul and Zoubin Ghahramani (2012). "Bayesian Classifier Combination". In: *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 619–627.
- Kim, Jaejeung, Sergey Leksikov, Punyotai Thamjamrassri, Uichin Lee, and Hyeon-Jeong Suk (2015). "CrowdColor: Crowdsourcing Color Perceptions Using Mobile Devices". In: *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '15)*. Copenhagen, Denmark, pp. 478–483.
- Kim, Joy and Andres Monroy-Hernandez (2016). "Storia : Summarizing Social Media Content based on Narrative Theory using Crowdsourcing". In: *Conference on Computer-Supported Cooperative Work and Social Computing (CSCW '16)*, pp. 1018–1027.
- Kittur, Aniket, J Nickerson, and M Bernstein (2013). "The Future of Crowd Work". In: *Proceedings of the 16th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '13)*, pp. 1–17.
- Kittur, Aniket, Boris Smus, and Robert Kraut (2011). "CrowdForge Crowdsourcing Complex Work". In: *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems (CHI EA '11)*, p. 1801.
- Kittur, Aniket, Susheel Khamkar, Paul André, and Robert Kraut (2012). "CrowdWeaver". In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW '12)*, p. 1033.
- Kohavi, Ron and Foster Provost (1998). "Glossary of terms: Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process". In: *Journal of Machine Learning* 30.2/3, pp. 271–274.
- Kohler, Thomas (2018). "How to Scale Crowdsourcing Platforms". In: *California Management Review* 60.2, pp. 98–121.
- Komarov, Steven, Katharina Reinecke, and Krzysztof Z Gajos (2013). "Crowdsourcing Performance Evaluations of User Interfaces". In: *roceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Paris.
- Kost, Dominique, Christian Fieseler, and Sut I Wong (May 2018). "Finding meaning in a hopeless place? The construction of meaningfulness in digital microwork". In: *Computers in Human Behavior* 82, pp. 101–110.
- Koyama, Yuki, Daisuke Sakamoto, and Takeo Igarashi (2014). "Crowd-powered parameter analysis for visual design exploration". In: *Proceedings of the 27th annual ACM symposium on User interface software and technology (UIST '14)*, pp. 65–74.
- Krippendorff, Klaus (2011). *Computing Krippendorff's Alpha-Reliability Part of the Communication Commons*. Tech. rep.

- Krishna, Ranjay, Kenji Hata, Stephanie Chen, Joshua Kravitz, David A. Shamma, Li Fei-Fei, and Michael S. Bernstein (2016). "Embracing Error to Enable Rapid Crowdsourcing". In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI '16)*, p. 10.
- Kulkarni, Anand P., Matthew Can, and Bjoern Hartmann (2011). "Turkomatic". In: *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems (CHI EA '11)*, p. 2053.
- Lasecki, Walter S, Jeffrey M Rzeszotarski, Adam Marcus, and Jeffrey P Bigham (2015). "The Effects of Sequence and Delay on Crowd Work". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Vol. 1. New York, NY, USA: ACM, pp. 1375–1378.
- Law, Edith, Luis von Ahn, and Luis Von Ahn (2005). "Human Computation". PhD thesis, pp. 1–121.
- Law, Edith and Luis von Ahn (2011). *Human Computation(book)*. Vol. 5. 3, pp. 1–121.
- Law, Edith and Luis Von Ahn (2009). "Input-agreement: a new mechanism for collecting data using human computation games". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp. 1197–1206.
- Law, Edith, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie (2009). "Evaluation of Algorithms Using Games: The Case of Music Tagging." In: *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 09)*, pp. 387–392.
- Law, Edith, Ming Yin, Joslin Goh, Kevin Chen, Michael A. Terry, and Krzysztof Z Gajos (2016). "Curiosity Killed the Cat, but Makes Crowdwork Better". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*, pp. 4098–4110.
- Lee, Kyumin, James Caverlee, and Steve Webb (2010). "The social honeypot project: protecting online communities from spammers". In: *Proceedings of the 19th international conference on World wide web*. October 2007, pp. 2008–2009.
- Li, Hongwei, Bo Zhao, and Ariel Fuxman (2014). "The wisdom of minority: discovering and targeting the right group of workers for crowdsourcing". In: *Proceedings of the 23rd International Conference on World Wide Web*, pp. 165–175.
- Licklider, Joseph C. R. (1960). "Man-computer symbiosis". In: *IRE transactions on human factors in electronics HFE-1.1*, pp. 4–11.
- Litman, Leib, Jonathan Robinson, and Tzvi Abberbock (Apr. 2017). "TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences". In: *Behavior Research Methods* 49.2, pp. 433–442.
- Little, Greg, Lydia B Chilton, Max Goldman, and Robert C Miller (2010). "TurKit : Human Computation Algorithms on MTurk". In: *Proceedings of the 23rd annual ACM symposium on User interface software and technology*, pp. 57–66.
- Liu, Qiang, Jian Peng, and Alexander Ihler (2012). "Variational Inference for Crowdsourcing". In: *Nips*, pp. 701–709.
- Liu, Xuan, Meiyu Lu, Beng Chin Ooi, Yanyan Shen, Sai Wu, and Meihui Zhang (2012). "CDAS: A Crowdsourcing Data Analytics System". In: *Proceedings of the VLDB Endowment* 5.10.
- Luz, Nuno, Nuno Silva, and Paulo Novais (2015). "A survey of task-oriented crowdsourcing". In: *Artificial Intelligence Review* 44.2, pp. 187–213.



- Maddalena, Eddy, Marco Basaldella, and Dante Degl Innocenti (2016). "Crowdsourcing Relevance Assessments : The Unexpected Benefits of Limiting the Time to Judge". In: *The Fourth AAAI Conference on Human Computation and Crowdsourcing (HCOMP '16)*, pp. 129–138.
- Maddalena, Eddy, Stefano Mizzaro, Falk Scholer, and Andrew Turpin (Jan. 2017). "On Crowdsourcing Relevance Magnitudes for Information Retrieval Evaluation". In: *ACM Transactions on Information Systems* 35.3, 19:1–19:32.
- Mao, Ke, Ye Yang, Qing Wang, Yue Jia, and Mark Harman (2015). "Developer recommendation for crowdsourced software development tasks". In: *Proceedings of the 9th IEEE International Symposium on Service-Oriented System Engineering (IEEE SOSE '15)*. Vol. 30, pp. 347–356.
- Mao, Ke, Licia Capra, Mark Harman, and Yue Jia (2017). *A Survey of the Use of Crowdsourcing in Software Engineering*. Tech. rep., pp. 57–84.
- Marcus, Adam, David Karger, Samuel Madden, Robert Miller, and Sewoong Oh (2012). "Counting with the crowd". In: *Proceedings of 38th the International Conference on Very Large Data Bases (VLDB)*. Vol. 6. 2, pp. 109–120.
- Marshall, Catherine C and Frank M Shipman (2013). "Experiences surveying the crowd: Reflections on Methods, Participation, and Reliability". In: *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13*, pp. 234–243.
- Martin, David, Jacki O'Neill, Neha Gupta, and Benjamin V Hanrahan (2016). "Turking in a Global Labour Market". In: *Computer Supported Cooperative Work (CSCW)* 25.1, pp. 39–77.
- Mason, Winter and Duncan J Watts (2009). "Financial incentives and the performance of crowds". In: *Proceedings of the ACM SIGKDD Workshop on Human Computation*. Vol. 11. 2, pp. 77–85.
- McDonnell, Tyler, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed (2016). "Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments". In: *Proceeding of the 4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP '16)*, pp. 139–148.
- McInnis, Brian, Dan Cosley, Chaebong Nam, and Gilly Leshed (2016). "Taking a HIT: Designing around Rejection, Mistrust, Risk, and Workers' Experiences in Amazon Mechanical Turk". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*, pp. 2271–2282.
- Metz, Charles E (1978). "Basic principles of ROC analysis". In: *Seminars in Nuclear Medicine* 8.4, pp. 283–298.
- Minder, Patrick and Abraham Bernstein (2012). *CrowdLang: programming human computation systems*. Tech. rep. Zürich, University of Zurich Department of Informatics (IFI), p. 13.
- Mitra, Tanushree, C J Hutto, and Eric Gilbert (2015). "Comparing Person- and Process-centric Strategies for Obtaining Quality Data on Amazon Mechanical Turk". In: *Proceedings of the ACM CHI'15 Conference on Human Factors in Computing Systems*. Vol. 1, pp. 1345–1354.
- Morris, Robert R., Mira Dontcheva, and Elizabeth M. Gerber (2012). "Priming for better performance in microtask crowdsourcing environments". In: *IEEE Internet Computing* 16.5, pp. 13–19.
- Mourelatos, Evangelos, Nikos Frarakis, and Manolis Tzagarakis (2017). "A Study on the Evolution of Crowdsourcing Websites". In: *European Journal of Social Sciences Education and Research* 11.1, pp. 2411–9563.

- Mourelatos, Evangelos, Manolis Tzagarakis, and Efthalia Dimara (2016). "A REVIEW OF ON-LINE CROWDSOURCING PLATFORMS". In: *South-Eastern Europe Journal of Economics* 14.1, pp. 59–74.
- Moussawi, Sara and Marios Koufaris (2013). "The Crowd on the Assembly Line: Designing Tasks for a Better Crowdsourcing Experience". In: *Thirty Fourth International Conference on Information Systems*, pp. 1–17.
- Nakatsu, Robbie and Elissa Grossman (2013). "Designing effective user interfaces for crowdsourcing: An exploratory study". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8016 LNCS.PART 1, pp. 221–229.
- Nakatsu, Robbie T., Elissa B. Grossman, and Charalambos L. Iacovou (2014). "A taxonomy of crowdsourcing based on task complexity". In: *Journal of Information Science* 40.6, pp. 823–834.
- Newell, Edward and Derek Ruths (2016). "How One Microtask Affects Another". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*, pp. 3155–3166.
- Nielsen, Jakob (1993). *Usability Engineering*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.
- Nushi, Besmira, Adish Singla, Anja Gruenheid, Erfan Zamanian, Andreas Krause, and Donald Kossmann (2015). "Crowd Access Path Optimization: Diversity Matters". In: *Proceedings, The Third AAAI Conference on Human Computation and Crowdsourcing (HCOMP 15)*. AUGUST, pp. 130–139.
- Oleson, David, Alexander Sorokin, Greg Laughlin, Vaughn Hester, John Le, and Lukas Biewald (2011). "Programmatic Gold: Targeted and Scalable Quality Assurance in Crowdsourcing". In: *Human Computation: Papers from the 2011 AAAI Workshop*, pp. 43–48.
- Organisciak, Peter (2014). *Reliable Task Design for Descriptive Crowdsourcing*. Tech. rep.
- Owens, Trevor (2013). "Digital Cultural Heritage and the Crowd". In: *Curator: The Museum Journal* 56.1, pp. 121–130.
- Palotti, J., G. Zuccon, J. Bernhardt, A. Hanbury, and L. Goeuriot (2016). "Assessors agreement: A case study across assessor type, payment levels, query variations and relevance dimensions". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9822.
- Pan, Yue and Eli Blevis (May 2011). "A survey of crowdsourcing as a means of collaboration and the implications of crowdsourcing for interaction design". In: *Proceedings of the 2011 International Conference on Collaboration Technologies and Systems, CTS 2011*. IEEE, pp. 397–403.
- Parde, Natalie and Rodney Nielsen (2017). "Finding Patterns in Noisy Crowds: Regression-based Annotation Aggregation for Crowdsourced Data". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1908–1913.
- Paritosh, Praveen (2012). "Human Computation Must Be Reproducible". In: *Proceedings of the 21st International Conference on World Wide Web (WWW '12)*. Lyon.
- Park, Taemin Kim (1993). "The Nature of Relevance in Information Retrieval: An Empirical Study." In: *The Library Quarterly: Information, Community* 63.3, pp. 318–51.

- Parkhurst, David F (2001). "Statistical Significance Tests: Equivalence and Reverse Tests Should Reduce Misinterpretation". In: *Bioscience* 51.12, pp. 1051–1057.
- Paul, Aplar and Osterbrink Lars (2018). "ANTECEDENTS OF PERCEIVED FAIRNESS IN PAY FOR Research in Progress". In: *Twenty-Sixth European Conference on Information Systems (ECIS2018)*.
- Peer, Eyal, Sonam Samat, Laura Brandimarte, and Alessandro Acquisti (2016). "Beyond the Turk : Alternative platforms for crowdsourcing behavioral research". In: *Journal of Experimental Social Psychology* 70.January, pp. 153–163.
- Peer, Eyal, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti (2017). "Beyond the Turk: Alternative platforms for crowdsourcing behavioral research". In: *Journal of Experimental Social Psychology* 70, pp. 153–163.
- Qarout, Rehab K, Alessandro Checco, and Gianluca Demartini (2016). "The Effect of Class Imbalance and Order on Crowdsourced Relevance Judgments". In: *Computing Research Repository (CoRR)*. Vol. abs/1609.0.
- Qarout, Rehab Kamal, Alessandro Checco, and Kalina Bontcheva (2018). "Investigating stability and reliability of crowdsourcing output". In: *CEUR Workshop Proceedings*. Vol. 2276, pp. 83–87.
- Quinn, Alexander J and Benjamin B Bederson (2009). "A taxonomy of distributed human computation". In: *Human-Computer Interaction Lab Tech Report, University of Maryland*.
- (2011). "Human Computation: A Survey and Taxonomy of a Growing Field". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '11. New York, NY, USA: ACM, pp. 1403–1412.
- (2014). "AskSheet : Efficient Human Computation for Decision Making with Spreadsheets". In: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*, pp. 1456–1466.
- Quinn, Alexander J., Benjamin B. Bederson, Tom Yeh, and Jimmy Lin (2010). "CrowdFlow: Integrating Machine Learning with Mechanical Turk for Speed-Cost-Quality Flexibility". In: *Human Computer Interaction Lab, 2010-05*, pp. 1–8.
- Quoc Viet Hung, Nguyen, Nguyen Thanh Tam, Lam Ngoc Tran, and Karl Aberer (2013). "An evaluation of aggregation techniques in crowdsourcing". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8181 LNCS.PART 2, pp. 1–15.
- Raykar, Vikas C, Shipeng Yu, Linda H. Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy (2009). "Supervised learning from multiple experts". In: *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*, pp. 1–8.
- Raykar, Vikas C, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, Linda Moy, and Linda Moy@nyumc Org (2010). "Learning From Crowds". In: *Journal of Machine Learning Research* 11, pp. 1297–1322.
- Rodrigues, Filipe, Francisco C Pereira, and Bernardete Ribeiro (2014). "Gaussian Process Classification and Active Learning with Multiple Annotators". In: *Proceedings of the 31st International Conference on Machine Learning*. Vol. 32, pp. 433–441.

- Ross, Joel, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson (2010). "Who are the crowdworkers?" In: *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems (CHI EA '10)*. March 2017, p. 2863.
- Rosten, Edward, Reid Porter, and Tom Drummond (2010). *Faster and better: A machine learning approach to corner detection*. Tech. rep. 1, pp. 105–119.
- Rouse, Anne C (2010). "A preliminary taxonomy of crowdsourcing". In: *ACIS 2010 Proceedings, 21st Australasian Conference on Information Systems*. Vol. 76.
- Sabou, Marta, Kalina Bontcheva, Leon Derczynski, and Arno Scharl (2014). "Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines." In: *The International Conference on Language Resources and Evaluation (LREC)*, pp. 859–866.
- Sabou, Marta, Lora Aroyo, Kalina Bontcheva, Alessandro Bozzon, and Rehab K Qarout (2018). "Semantic Web and Human Computation: The status of an emerging field". In: *Semantic Web 9.3*, pp. 291–302.
- Sarasua, Cristina, Elena Simperl, and Natalya F Noy (2012). "CROWD MAP: Crowdsourcing Ontology Alignment with Microtasks". In: *The Semantic Web (ISWC 2012 Lecture Notes in Computer Science)*, pp. 525–541.
- Sarasua, Cristina, Elena Simperl, Natasha F. Noy, Abraham Bernstein, and Jan Marco Leimeister (2015). "Crowdsourcing and the Semantic Web: A Research Manifesto". In: *Human Computation 2.1*, pp. 3–17.
- Schall, Daniel, Hong-Linh Truong, and Schahram Dustdar (2011). "The human-provided services framework". In: *Socially Enhanced Services Computing*. Springer, pp. 1–15.
- Schenk, Eric and Claude Guittard (2011). "Towards a characterization of crowdsourcing practices". In: *Journal of Innovation Economics 0.1*, pp. 93–107.
- Schmidt, Gordon B. and William M. Jettinghoff (2016). "Using Amazon Mechanical Turk and other compensated crowdsourcing sites". In: *Business Horizons 59*, pp. 391–400.
- Schnitzer, Steffen, Christoph Rensing, and Sebastian Schmidt (2015). "Demands on task recommendation in crowdsourcing platforms - the worker's perspective". In: *In workshop of Crowdsourcing and human computation for recommender systems (CrowdRec 15), ACM RecSys*. Vol. 2015. September, pp. 1–7.
- Scholer, Falk, Andrew Turpin, and Mark Sanderson (2011). "Quantifying Test Collection Quality Based on the Consistency of Relevance Judgements". In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 1063–1072.
- Scholer, Falk, Diane Kelly, Wan-Ching Wu, Hanseul S Lee, and William Webber (2013). "The effect of threshold priming and need for cognition on relevance calibration and assessment". In: *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '13)*, p. 623.
- Shao, Yunqiu, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma (Mar. 2019). "On Annotation Methodologies for Image Search Evaluation". In: *ACM Transactions on Information Systems 37.3*, 29:1–29:32.

- Sheng, Victor S (2017). "Label Aggregation for Crowdsourcing with Bi-Layer Clustering". In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. Vol. 1, pp. 921–924.
- Silberman, M.S., B. Tomlinson, R. LaPlante, J. Ross, L. Irani, and A. Zaldivar (2018). "Responsible research with crowds: Pay crowdworkers at least minimum wage". In: *Communications of the AMC* 61.3, pp. 39–41.
- Sormunen, Eero (2002). "Liberal relevance criteria of TREC - Counting on negligible documents?" In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '02)*, p. 324.
- Spatharioti, Sofia Eleni, Sofia Eleni Spatharioti, Rebecca Govoni, Jennifer S Carrera, Sara Wylie, and Seth Cooper (2017). "A Required Work Payment Scheme for Crowdsourced Disaster Response: Worker Performance and Motivations". In: *Proceedings of the 14th International Conference on Information Systems for Crisis Response And Management (ISCRAM '17)*. May. Albi, France, pp. 475–488.
- Sun, Peng and Kathryn T. Stolee (2016). "Exploring crowd consistency in a mechanical turk survey". In: *Proceedings of the 3rd International Workshop on CrowdSourcing in Software Engineering (CSI-SE '16)*. New York, New York, USA: ACM Press, pp. 8–14.
- Surowiecki, James (2004). *The Wisdom of Crowds*, p. 296.
- Tam, Nguyen Thanh, Huynh Huu Viet, Nguyen Quoc Viet Hung, MatthiasWeidlich, Hongzhi Yin, and Xiaofang Zhou (2017). "Multi-Label Answer Aggregation for Crowdsourcing". In: *ACM Computing Surveys* 1828, pp. 53–59.
- Tate, Mary, David Johnstone, and Erwin Fieft (2017). "Ethical issues around crowdwork: How can blockchain technology help?" In: *Australasian Conference on Information Systems*. Hobart, pp. 1–11.
- Thimbleby, Harold, Michael Bernstein, Dan Russell, Ed Chi, Wendy Mackay, and Max L. Wilson (2011). "RepliCHI - CHI should be replicating and validating results more". In: *Proceedings of the 2011 CHI Conference on Human Factors in Computing Systems (CHI '11)*, p. 463.
- Thompson, Steven K. (2012). "Sample Size". In: *Sampling*. Wiley-Blackwell. Chap. 4, pp. 53–56.
- Tonon, Alberto, Gianluca Demartini, and Philippe Cudré-Mauroux (2012). "Combining inverted indices and structured search for ad-hoc object retrieval". In: *Proceedings of the 35th International ACM SIGIR conference on research and development in Information Retrieval (SIGIR '12)*, p. 125.
- (2015). "Pooling-based continuous evaluation of information retrieval systems". In: *Information Retrieval* 18.5, pp. 445–472.
- Towne, W Ben, Carolyn P Rosé, and James D Herbsleb (2017). "Conflict in Comments". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI' 17)*, pp. 655–666.
- Tranquillini, Stefano, Florian Daniel, Pavel Kucherbaev, and Fabio Casati (2015). "Modeling, Enacting, and Integrating Custom Crowdsourcing Processes". In: *ACM Transactions on the Web* 9.2, 7:1–7:43.
- Törnqvist, Leo, Pentti Vartia, and Yrjö Vartia (Feb. 1985). "How Should Relative Changes Be Measured?" In: *The American Statistician* 39, pp. 43–46.

- Tsvetkova, Milena, Taha Yasseri, Eric T Meyer, J Brian Pickering, Vegard Engen, Paul Wal-land, Marika Lüders, Asbjørn Følstad, and George Bravos (2015). "Understanding Human-Machine Networks: A Cross-Disciplinary Survey". In: *ACM Computing Surveys*. Vol. 50. 12.
- Turing, Alan M (1950). "Computing machinery and intelligence". In: *Mind* 59.236, pp. 433–460.
- Venanzi, Matteo, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi (2014). "Community-based bayesian aggregation models for crowdsourcing". In: *Proceedings of the 23rd international conference on World wide web (WWW' 14)*, pp. 155–164.
- Venanzi, Matteo, John Guiver, Pushmeet Kohli, and Nicholas R. Nick Jennings (2016). "Time-Sensitive Bayesian information aggregation for crowdsourcing systems". In: *Journal of Artificial Intelligence Research* 56, pp. 517–545.
- Walsh, Brandon, Claire Maiers, Gwen Nally, Jeremy Boggs, and Praxis Program Team (2014). "Crowdsourcing individual interpretations: Between microtasking and macrotasking". In: *Literary and Linguistic Computing* 29.3, pp. 379–386.
- Wang, Jing, Siamak Faridani, and Panagiotis G. Ipeirotis (2011). "Estimating the Completion Time of Crowdsourced Tasks Using Survival Analysis Models". In: *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM '11)*. 1, pp. 31–34.
- Waterhouse, Tamsyn P (2013). "Pay by the bit". In: *Proceedings of the 2013 conference on Computer supported cooperative work (CSCW '13)*, pp. 623–638.
- Webber, William, Praveen Chandar, and Ben Carterette (2012). "Alternative assessor disagreement and retrieval depth". In: *Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM '12)*, p. 125.
- Wei, Xuan, Daniel Dajun Zeng, and Junming Yin (2017). *Multi-Label Annotation Aggregation in Crowdsourcing*. Tech. rep.
- Weidema, Edgar R. Q., Consuelo López, Sahand Nayebaziz, Fernando Spanghero, and André van der Hoek (2016). "Toward microtask crowdsourcing software design work". In: *Proceedings of the 3rd International Workshop on CrowdSourcing in Software Engineering (CSI-SE '16)*, pp. 41–44.
- Welinder, Peter, Steve Branson, Pietro Perona, and Serge J Belongie (2010). "The multidimensional wisdom of crowds". In: *Advances in neural information processing systems* 23, pp. 2424–2432.
- Whitehill, Jacob, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan (2009). "Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise". In: *Advances in Neural Information Processing Systems* 22.1, pp. 1–9.
- Whitla, Paul (Mar. 2009). "Crowdsourcing and Its Application in Marketing Activities". In: *Contemporary Management Research* 5.1, pp. 15–28.
- Willett, Wesley, Jeffrey Heer, and Maneesh Agrawala (2012). "Strategies for crowdsourcing social data analysis". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 227–236.
- Williams, Alex C., Joslin Goh, Charlie G. Willis, Aaron M. Ellison, James H. Brusuelas, Charles C. Davis, and Edith Law (2017). "Deja Vu: Characterizing Worker Reliability Using Task Consistency". In: *The fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP '17)*.

- Williamson, Vanessa (2016). "On the Ethics of Crowdsourced Research". In: *PS - Political Science and Politics* 49.1, pp. 77–81.
- Wilson, Max L., David Coyle, Paul Resnick, and Ed H Chi (2013). *RepliCHI-The Workshop*. Tech. rep.
- Wu, Meng-han and Alexander J Quinn (2017). "Confusing the Crowd : Task Instruction Quality on Amazon Mechanical Turk". In: *The Fifth AAAI Conference on Human Computation and Crowdsourcing*. Hcomp, pp. 206–215.
- Xintong, Guo, Wang Hongzhi, Yangqiu Song, and Gao Hong (2014). "Brief survey of crowdsourcing for data mining". In: *Expert Systems with Applications* 41.17, pp. 7987–7994.
- Yan, Yan, Romer Rosales, Glenn Fung, and Mark Schmidt (2010). "Modeling annotator expertise: Learning when everybody knows a bit of something". In: *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 9, pp. 932–939.
- Yang, Jie, Judith Redi, Gianluca Demartini, and Alessandro Bozzon (2016). "Modeling Task Complexity in Crowdsourcing". In: *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing (HCOMP '16)*. October, pp. 249–258.
- Yang, Yang, Bin B Zhu, Rui Guo, Linjun Yang, Shipeng Li, and Nenghai Yu (2008). "A comprehensive human computation framework: with application to image labeling". In: *Proceedings of the 16th ACM international conference on Multimedia*. ACM, pp. 479–488.
- Ye, Teng, Sangseok You, and Lionel P Robert (2017). "When Does More Money Work? Examining the Role of Perceived Fairness in Pay on the Performance Quality of Crowdworkers". In: *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*. Icwsm, pp. 327–336.
- Yuen, Man-Ching, Ling-Jyh Chen, and Irwin King (2009). "A survey of human computation systems". In: *International Conference on Computational Science and Engineering (CSE'09)*. Vol. 4. IEEE, pp. 723–728.
- Yuen, Man Ching, Irwin King, and Kwong Sak Leung (2011). "A survey of crowdsourcing systems". In: *Proceedings of the IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing*, pp. 766–773.
- (2015). "TaskRec: A Task Recommendation Framework in Crowdsourcing Systems". In: *Neural Processing Letters* 41.2, pp. 223–238.
- Yuen, MC, Irwin King, and KS Leung (2012). "Task recommendation in crowdsourcing systems". In: *Proceedings of the First International Workshop on Crowdsourcing and Data Mining (CrowdKDD'12)*.
- Zhao, Mengyao and Andre Van Der Hoek (2015). "A Brief Perspective on Microtask Crowdsourcing Workflows for Interface Design". In: *Proceedings of the 2nd International Workshop on Crowdsourcing in Software Engineering, CSI-SE 2015*.
- Zheng, Yudian, Jiannan Wang, Guoliang Li, Reynold Cheng, and Jianhua Feng (2015). "QASCA". In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15)*, pp. 1031–1046.
- Zhou, Dengyong, John Platt, Sumit Basu, and Yi Mao (2012). "Learning from the wisdom of crowds by minimax entropy". In: *Advances in Neural Information Processing Systems 25*, pp. 2204–2212.

- Zhu, Haiyi, Steven P Dow, Robert E Kraut, and Aniket Kittur (2014). "Reviewing versus Doing: Learning and Performance in Crowd Assessment". In: *Proceedings Conference on Computer Supported Cooperative Work and Social Computing (CSCW '14)*, pp. 1445–1455.
- Zhuang, Honglei, Aditya Parameswaran, Dan Roth, and Jiawei Han (2015). "Debiasing Crowdsourced Batches". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*, pp. 1593–1602.